Stony Brook University

Partial Correlation Network Analysis for Mixed Data

A Dissertation Presented

by

Shirley Hui Yee Leong

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

May 2012

Stony Brook University

The Graduate School

Shirley Hui Yee Leong

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Wei Zhu – Dissertation Advisor
Professor, Deputy Chair, Department of Applied Mathematics and Statistics

Hongshik Ahn - Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Song Wu – Faculty Committee Member
Assistant Professor, Department of Applied Mathematics and Statistics

Roman Kotov – Outside Committee Member
Assistant Research Professor, Department of Psychiatry

This dissertation is accepted by the Graduate School

Charles Taber
Interim Dean of the Graduate School

Abstract of the Dissertation

Partial Correlation Network Analysis for Mixed Data

by

Shirley Hui Yee Leong
Doctor of Philosophy
in
Applied Mathematics and Statistics
Stony Brook University
2012

The partial correlation is well defined for continuous data and popularly used in network analysis. Its strength is in its interpretation as the relationship between two variables after removing the effects of other variables. We follow up on a recent proposal of such a measure for categorical data, but the properties of which were not well studied. The new partial correlation is defined as the first canonical correlation of Pearson residuals from logistic regressions. This is analogous to the continuous case, where the partial correlation is obtained from correlating residuals from linear regressions. A simulation study is presented to examine the properties of the new partial correlation and compare it to other measures, such as the partial phi coefficient. In the limiting case, the new partial correlation and the partial phi coefficient converge in estimate and inference. However, the partial phi coefficient cannot be applied to multi-categorical data. Furthermore, it is not an efficient measure to control for more than one variable. The new partial correlation is well defined for the multi-categorical case and can readily control for more than one variable. Being derived as the canonical correlation, the new partial correlation can also measure the relationship between continuous and categorical variables as the multiple correlation between the Pearson residuals from the logistic regression and the usual residual from the linear regression when the response variables are categorical and continuous respectively.

Now that we are fully capable of obtaining partial correlation networks for any data types, continuous, categorical or mixed, our next goal is to compare the network structure between different groups and to examine the impact of continuous, in addition to categorical covariates, on the pathway connections. This is accomplished by extending the two-level regression approach for continuous data originally developed by our research group (Pradhan, 2009) to categorical data and mixed data network analysis. By linearly regressing the first canonical variates and replacing the slope coefficient with an expression of the covariates, we can test for the effect of covariates (both categorical and continuous) on the partial correlation and the network structure. This new covariate partial correlation network analysis approach is illustrated through two studies on the links between human genotypes (single-nucleotide polymorphisms) and disease phenotypes.

Dedications

To my family, friends, and mentors who have supported me in my work. This work would not be possible without their constant encouragement.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to acknowledge the following people for their invaluable insight and commentary on this work, without which I would not have been able to clarify my thoughts and ideas.

Dr. Wei Zhu

Dr. Ellen Li

Dr. Hongyan Chen

Dr. Xiao Wu

Jinmiao Fu

1 Introduction

A basic question in scientific research is whether two variables are related and if so, how they are related. However, in the presence of another variable or covariate, the answer is not so clear. A significant association between two variables may be due to the influence of a third variable or covariate confounder on both variables of interest. The problem increases in complexity with a large set of variables: one may be interested in all pair-wise relationships independent of all other variables. Network analysis provides a means to visualize such associations in an organized structure. If additional covariates or confounders affect the network structure, it is important to be able to detect and measure such an effect. These concepts and theories have been well studied for continuous variables, but the literature for categorical and mixed data is sparse. We propose extensions of several existing methods to categorical variables and the general case of mixed variable types.

This work is organized into three parts. The first part discusses the impact of confounding variables on statistical models. Various modeling techniques that address this problem are presented. We discuss residual logistic regression (Baez-Revueltas 2009), a new method analogous residual linear regression. The second part discusses various correlation and association measures. The detection and measurement of pair-wise relationships, while controlling for other variables, is examined in the context of partial correlations and partial associations. In particular, we consider a new partial correlation coefficient for categorical data, obtained by correlating Pearson residuals from logistic regressions (Chen 2011). This is analogous to the continuous case, where the partial correlation is obtained from correlating residuals from linear regressions. The third part deals with network analysis and covariate in network analysis. The use of the partial correlation and partial association measures from part two in network analysis is presented. A novel extension of two-level regression (Pradhan 2009) to categorical and mixed data is proposed to analyze the effect of covariates on categorical and mixed network structure. The performances of the new methods are studied under various simulation settings; the new methods are compared to existing ones based on bias, standard error, power, and specificity.

An overview of existing methods for controlling confounders in statistical models is provided, followed by a discussion on what development is needed this area. Another overview of correlation and association measures and their partial counterparts is given. Subsequently, we present various proposals and novel methods that are in development. For the problem of confounded regression models, we propose the residual logistic regression as an extension of residual linear regression for categorical data. A new partial correlation recently proposed for categorical and mixed variable types, but never adequate

examined, is studied using simulated data and compared to other suitable partial correlation and partial association measures for such data. Finally, two-level regression, originally proposed for continuous data to measure the effect of covariates in network structure, will be presented for categorical and mixed variables types based on the new partial correlation.

The new methods proposed will be compared to existing methods using simulated data; methods will be compared based on bias, standard error, power, and specificity. The new methods were applied to three individual data sets. One data set is from a study conducted at the NYU Alzheimer's Disease Core Center (ADCC). The other two datasets are genetic studies, one from the Collaborative Genetic Study of Nicotine Dependence (COGEND) and the other from Washington University Digestive Diseases Research Core Center Tissue Procurement Facility database. A discussion about the results and work that will be done in the future will conclude this paper.

1.1 Overview: Regression methods to control confounding

Existing regression methods to control for confounding will be introduced within the context of two common statistical models. Consider a single response variable Y and p covariates. When the Y is continuous, the most basic statistical model is the linear model:

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon, \text{ where } \varepsilon \sim N\left(0, \sigma^2\right).$$

When the Y is binary (dichotomous), a logistic model is generally applied:

$$\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \sum_{i=1}^{p} \beta_i X_i, \text{ where } \pi = P\left(Y = 1\right).$$

Traditionally, all confounders and variables of interest are included as individual covariates in the regression model. This model is commonly referred to as the full regression model and is equivalent to the above models. This method, which for the remainder of this paper will be referred to as traditional multiple regression, does not differentiate between confounders and variables of interest. To reduce the complexity of the model, one may also implement a variable selection procedure on the set of variables of interest while forcing the confounders to remain in the model.

Another common strategy, to be referred to in this paper as one-at-a-time regression, focuses on measuring the individual effects of each variable of interest while controlling for the confounders. Individual regression models are developed such that only a single variable of interest is considered at a time. If X_i, $i = 1, \ldots, k$, $k < p$, are the confounders and the remaining $p - k$ covariates are variables of interest, then for each variable of interest, our linear model and logistic model would be, respectively,

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \beta_j X_j + \varepsilon$$

$$\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \beta_j X_j, \text{where } j = k+1, \ldots, p.$$

When the main focus of the study is to compare the response between two different groups, we may still implement the above method for other variables of interest, but an indicator variable for group would also be included in each regression model.

$$Y = \beta_0 + \alpha * Group + \sum_{i=1}^{k} \beta_i X_i + \beta_j X_j + \varepsilon$$

$$\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \alpha * Group + \sum_{i=1}^{k} \beta_i X_i + \beta_j X_j, \text{where } j = k+1, \ldots, p.$$

On the other hand, propensity score methods can be applied to control for confounding when studying group effects. This strategy reduces the k confounders to a single score, $\hat{\pi}(X_1, \ldots, X_k)$, that balances the distribution of the confounders between groups. The method of propensity score adjustment replaces the confounders in the regression model with $\hat{\pi}(X_1, \ldots, X_k)$:

$$Y = \beta_0 + \alpha * Group + \gamma * \hat{\pi}(X_1, \ldots, X_k) + \sum_{i=k+1}^{p} \beta_i X_i + \varepsilon$$

$$\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \alpha * Group + \gamma * \hat{\pi}(X_1, \ldots, X_k) + \sum_{i=k+1}^{p} \beta_i X_i$$

These strategies are also commonly applied to longitudinal models such as the Cox proportional hazards model, the accelerated failure time model, and the mixed effects logistic model.

- Cox PH model: $\log h_i(t) = \alpha(t) + \sum_{j=1}^{p} \beta_j X_j$, where $h_i(t)$ is the hazard function and $\alpha(t)$ is the baseline log hazard
- Accelerated failure time (AFT) model: $\log T = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \sigma \times \varepsilon^*$, where ε^* follows the extreme minimum value distribution
- Mixed effects logistic model: $\log\left[\frac{\pi_{ij}}{1-\pi_{ij}}\right] = \sum_{k=1}^{p} \beta_k X_{ijk} + \sum_{l=1}^{q} \delta_{ijl} U_{il}$, where X_{ijk} are fixed effects, U_{il} are random effects and $\pi_{ij} = P\left(Y_{ij} = 1 \mid X_{ij1}, ..., X_{ijp}, U_{i1}, ..., U_{iq}\right)$

A two stage residual linear regression procedure has been well studied in the case of a continuous outcome. In this case, a stage one linear regression of the outcome on the confounders only provides residuals. These residuals are then regressed on the remaining variables of interest in a stage two linear regression.

Pearson residual analysis is similar to the two stage residual linear regression, but is applied to dichotomous outcomes. The first stage consists of a logistic regression, from which the Pearson residual is calculated and linearly regressed on the variables of interest. A more analogous procedure for dichotomous outcomes, residual logistic regression, has been recently suggested in order to maintain the odds ratio interpretation. However, the method needs further theoretical development and extension to other research settings.

1.2 Overview: Correlation and their partials

Correlation and association are the two most basic concepts that describe how two variables are related. Correlations measure the amount of linear covariance between two variables, assuming the variables have an inherent ordering. Ideally, a good correlation measure should have the following properties:

- A positive correlation would indicate that as one variable increases (decreases), the other variable also increase (decreases).
- A negative correlation would indicate that as one variable increases (decreases), the other variable would trend towards the opposite direction or decrease (increase).

- A zero correlation indicates no relationship.
- A correlation of magnitude 1 would indicate perfect correlation.

Hence due to the requirement of a scale interpretation of the variables, correlation concepts have only been applied to continuous or ordinal categorical data, with binary data being the exception to this rule under certain assumptions about the underlying distributions. The relationship between two nominal categorical variables can be described by association measures, which are often interpreted as the distance from independence; if the association measure is equal to zero, then the two variables are independent.

Suppose we have N observations for continuous variables *X* and *Y*. The Pearson product-moment correlation (Pearson 1895, 1920), also known as the Pearson correlation, between any two variables *X* and *Y* is

$$r_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} = \frac{N\sum xy - \sum x \sum y}{\sqrt{N\sum x^2 - (\sum x)^2}\sqrt{N\sum y^2 - (\sum y)^2}}$$

To measure the correlation between two continuous variables while controlling for a third variable *Z*, one would correlate the residuals obtained from linear regressions of each variable onto the third variable. The result is the partial correlation (Yule 1897).

$$X = \beta_0 + \beta_1 Z + \varepsilon_x \rightarrow e_x = \hat{\beta}_0 + \hat{\beta}_1 Z$$
$$Y = \gamma_0 + \gamma_1 Z + \varepsilon_y \rightarrow e_y = \hat{\gamma}_0 + \hat{\gamma}_1 Z$$
$$r_{xy.z} = cor(e_x, e_y) = \frac{\Sigma \left(e_x - \frac{1}{N}\sum e_x\right)\left(e_y - \frac{1}{N}\sum e_y\right)}{\sqrt{\sum \left(e_x - \frac{1}{N}\sum e_x\right)}\sqrt{\sum \left(e_y - \frac{1}{N}\sum e_y\right)}}$$

It can be shown that this partial correlation can be rewritten in terms of each bivariate Pearson correlation.

$$r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1-r_{xz}^2)}\sqrt{(1-r_{yz}^2)}}$$

Assuming the data come from a multivariate normal distribution, we can test the null hypothesis that the partial correlation is equal to zero against the alternative that it is non-zero via an approximate test using Fisher's Z transformation or *t* or *F* test with *k* being the number of control variables (Fisher 1924).

$$Z = \frac{1}{2}\ln\left(\frac{1+r_{xy(z)}}{1-r_{xy(z)}}\right) \rightarrow Z\sqrt{N-2-k} \sim N(0,1)$$

$$t = \frac{r_{xy(z)}}{\sqrt{1-r_{xy(z)}^2}}\sqrt{N-2-k} \sim \text{Student's } t_{N-2-k}$$

$$F = \frac{r_{xy(z)}^2}{1-r_{xy(z)}^2}(N-2-k) \sim F_{1,N-2-k}$$

Suppose *X* and *Y* are categorical variables. For a random sample taken from the joint distribution, let n_{ij} be the number of observations where *X*=*i* and *Y*=*j*, $n_{i+} = \sum_j n_{ij}$ the number of observations where *X*=*i*, $n_{+j} = \sum_i n_{ij}$ the number of observations where *Y*=*j*, and $N = \sum_i \sum_j n_{ij}$ the total number of observations.

To test whether or not *X* and *Y* are independent, the chi-square test (Agresti 2007) is commonly used.

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{N}\right)^2}{\frac{n_{i+}n_{+j}}{N}}$$

Under the null hypothesis that *X* and *Y* are independent, χ^2 has chi-square distribution with (*I*-1)(*J*-1) degrees of freedom. *X* and *Y* are completely independent if χ^2 obtains the value of zero. The farther away from independency the observed sample of *X* and *Y is*, the larger the value of χ^2.

Many association measures involve some normalized version of the χ^2 statistic such that its values range from -1 to 1 and so that it can be interpreted analogously to the Pearson correlation coefficient. These χ^2 based association measures can be interpreted as the distance to independence, with zero being complete independence. To obtain their partial counterparts, one usually obtains the measure within stratas of the control variable and then calculates the partial association as a weighted average Chen 2011(Ritschard, et al. 1996). Let *S* represent a set of stratas of the variable *Z*. For each *s* in *S*, let θ_{xyls} be the association between *X* and *Y* within that strata and ω_s be the corresponding weight. The weights are typically the proportion of observations in each stratum. The general partial association measure is

$$\theta_{xy(z)} = \sum_{s \in S} \omega_s \theta_{xyls}$$

Suppose X and Y are binary variables. The application of the Pearson correlation formula applied to such data will result in the phi coefficient (Pearson, et al. 1900).

$$\phi_{xy} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{(n_{12}+n_{22})(n_{11}+n_{21})(n_{21}+n_{22})(n_{11}+n_{21})}} = \sqrt{\frac{\chi^2}{N}}$$

The phi coefficient ranges from -1 to 1, attaining zero when there is perfect independence, -1 when there is perfect discordance (diagonal is empty), and 1 when there is perfect concordance (off diagonal is empty). Since it is also a direct application of the Pearson correlation, it can be used in the partial correlation formula, assuming that the control variable Z is also binary. The partial phi correlation is then

$$r_{phi[xy(z)]} = \frac{\phi_{xy} - \phi_{xz}\phi_{yz}}{\sqrt{(1-\phi_{xz}^2)}\sqrt{(1-\phi_{yz}^2)}}$$

If Y is continuous and X is dichotomous then the direct application of the Pearson's product-moment correlation will results in the point-biserial correlation.

$$r_{pbis} = \frac{\bar{y}_{X=1} - \bar{y}_{X=0}}{Ns_y}\sqrt{(\sum x)(1-\sum x)}$$

If X is multi-categorical and ordinal, then the point-polyserial correlation is the result of direct application on Pearson's product moment correlation. Because the phi coefficient, the point-biserial correlation, and point-polyserial correlation is a direct application of Pearson's product-moment correlation, so statistical inferences on partial correlations constructed from such measures is carried out a similar way. If the data are from an underlying multivariate normal distribution whose true partial correlation is equal to zero, then an F-test can be applied. In other words, we assume that X and Y are manifest variables that are from a categorized multivariate normally distributed variables X^* and Y^*.

$$F = \frac{r_{xy(z)}^2}{1-r_{xy(z)}^2}(N-2-k) \sim F_{1,N-2-k}$$

While the phi coefficient, point-biserial correlation, and point-polyserial correlation measure the observed correlation between binary X and Y, the actual correlation of underlying continuous variables X^* and Y^* can be estimated via other correlations. The tetrachoric correlation (Pearson, et al. 1900, Ekstrom

2008) estimates the actual correlation between two variables from bivariate normal distributions which were both dichotomized. The tetrachoric correlation was originally derived as a mathematical formula involving the tetrachoric series. However, recent developments have led to maximum likelihood methods that are more commonly used to estimate this correlation (Olsson 1979).

If the *X* and *Y* are multi-categorical, the extension of the tetrachoric correlation is the polychoric correlation. Maximum likelihood procedures have also been developed to estimate the polychoric correlation (Martinson, et al. 1972, Olsson 1979, Drasgow 1986). These estimates can be used to estimate the partial correlation of the underlying continuous distribution. Suppose *X*, *Y*, and *Z* are categorical manifestations of multivariate normally distributed variables *X**, *Y**, *Z**.

$$r_{tet[xy(z)]} = \frac{r_{tet(xy)} - r_{tet(xz)} r_{tet(yz)}}{\sqrt{1 - r_{tet(xz)}^2}\sqrt{1 - r_{tet(yz)}^2}}$$

$$r_{pch[xy(z)]} = \frac{r_{polychoric(xy)} - r_{polychoric(xz)} r_{polychoric(yz)}}{\sqrt{1 - r_{polychoric(xz)}^2}\sqrt{1 - r_{polychoric(yz)}^2}}$$

If *Y* is actually observed as continuous, then the correlation between *X** and *Y* can be estimated by the biserial correlation (Pearson 1909, Kelley 1923). The extension of this to multi-categorical variables is called the polyserial correlation (Jaspen 1946). These correlations can also be calculated using maximum likelihood methods (Olsson, et al. 1982, Drasgow 1986, Poon, et al. 1987). Suppose that *X* and *Z* are trichotomous manifestations of normal variables *X** and *Z**, and *Y* is a continuous variable. Then the partial correlation between *X** and *Y* while controlling for *Z** can be estimated as

$$r_{mixed[xy(z)]} = \frac{r_{polyserial(xy)} - r_{polychoric(xz)} r_{polyserial(yz)}}{\sqrt{1 - r_{polychoric(xz)}^2}\sqrt{1 - r_{polyserial(yz)}^2}}$$

Despite all these numerous partial correlation options for categorical data, none of these measures can relate back to the idea of a correlation of residuals, since linear regression would not be suitable for these types of variables. In addition, since they rely on bivariate correlations to construct a partial correlation, they cannot control for more than one variable. These measures require that there is an underlying continuous distribution from which the data are observed, which is not suitable for truly categorical variables. In the case of the where tetrachoric, polychoric, biserial, and polyserial correlations are used in the construction of the partial correlation, statistical inference is not available.

1.3 Proposals

1.3.1 Residual logistic regression

The first proposal is to extend residual analysis to categorical outcomes via residual logistic regression (Baez-Revueltas 2009). This new method will be compared with existing regression methods to control for confounding, such as Pearson residual analysis, traditional multiple logistic regression, one-at-a-time logistic regression, and propensity score adjustment. The Pearson residual is the generalized linear model equivalent of a standardized residual from ordinary linear regression. The propensity score is the probability of being in a particular treatment group, given a particular covariate pattern.

In the case when the data is longitudinal in nature, the strategy really depends on whether or not there are (1) censoring, and (2) time-dependent covariates. The first data set to be used in conjunction with this paper is a longitudinal follow-up study examining the key prognostic factors of dementia conducted at NYU. It contains censoring information as well as time dependent variable (follow-up time). Hence, we will focus on the Cox proportional hazards model and the accelerated failure time model, and compare propensity score methods to the traditional multiple regression.

1.3.2 Partial correlation for categorical data and mixed data

Alternatively, the relationship between two variables while controlling for a third variable can be measure using partial correlation. The partial correlation is the correlation between residuals calculated from regressions of each variable onto the third variable; it measures the amount of covariance between two variables after removing the variance due to a third variable. While the partial correlation and its properties are well defined for continuous variables, a corresponding measure for the categorical data and mixed variables are not.

Chen (2011) proposed a new partial correlation measure applicable to multi-categorical variables and mixed data. Its novelty is that it is obtained analogously to the continuous partial correlation; Pearson residuals from logistic regressions are correlated. However, the theoretical properties of this new partial correlation measure have not been studied adequately. The performance of the new partial correlation measure is examined under two contexts: using simulated data from a categorized multivariate normal distribution and simulated data from an exponential model called the Ising model. Although the Ising

model was developed by the physicist Ernst Ising to model ferromagneticism (Ising 1925), it has since been extended to graph theory and network analysis.

1.3.3 Partial correlation network analysis (PCNA)

With a network of nodes, or variables, partial correlation can be applied to infer network edges, or pair-wise relationships while controlling for all other network variables. Partial correlation network analysis has been widely applied for fMRI studies where variables are continuous. Chen (2011) developed his new partial correlation to apply PCNA to SNP data, which are naturally categorical and may be coded to have an ordinal interpretation (number of minor alleles). We discuss the theoretical implication of such an application.

In addition to the network variables, there may be covariates which when included into the network may change the structure of the network by affecting individual edges. Pradhan (2009) developed a two-level regression method which measures the effect of covariates on the partial correlation for continuous variables. In this work, the two-level regression is extended to categorical and mixed data using Chen's novel partial correlation measure.

1.4 Case studies

1.4.1 NYU Alzheimer's Disease Core Center (ADCC) study

Regression methods of part one will be applied to a longitudinal study conducted at the NYU Alzheimer's Disease Core Center (ADCC) (Reisberg, et al. 2010). The goal is to determine if the mental decline rate is the same for subjects with or without subjective complaints of cognitive impairment. Mental decline is defined as decline to mild cognitive impairment (MCI) or dementia. Time to decline was either time to progression to MCI or dementia or, if no progression, time to the last follow-up before 2002.

Subjects were allocated to one of two groups based on the Global Deterioration Scale (GDS) for age-associated cognitive decline and dementia (Reisberg, et al. 1982). The No Cognitive Impairment (NCI) group was constructed from GDS stage 1 subjects: subjects who are normatively functioning and

free of subjective complaints or objective evidence of cognitive impairment. The Subjective Cognitive Impairment (SCI) group was constructed from GDS stage 2 subjects: subjects who are normatively functioning and have subjective complaints in the absence of objectively manifest deficits. Decline was measured as movement to higher stages on the GDS scale.

A list of 69 covariates that were observed is provided below. These variables account for possible demographic differences between the two groups, as well as various psychological measurements that may be associated with initial grouping classification and follow up outcome.

List of covariates observed in the ADCC Study

Age*
Gender*
Years of education*
Length of follow-up*
Mental status assessment
Mini-Mental State Examination (MMSE) score – 30 point scale
Clinical cognitive and cognitively based functioning examination
Brief Cognitive Rating Scale – 7 point scale for each axis; optimally concordant with other axes, GDS stages, MCI, and Alzheimer's Disease
5 axes (BCR01-BCR05)
Total BCR (BCRTOT) score
Affective status
Hamilton Depression Scale
21 axes (HDS01-HDS21)
Total HSD (HDTTOT) score
Comprehensive behavioral changes
Behavioral Pathology in Alzheimer's Disease – 4 point scale
25 axes (BEH01-BEH25)
Total BEH (BEHTOT) score
Neuropsychometric evaluation variables
Memory
paragraph recall, initial and delayed (PARI, PARD)
paired associate recall of pairs of initial words, initial and delayed (PRDI, PRDD)
design recall of abstract shapes (DESN)
Working memory
digit span subtests of Wechsler Intelligence Scale Revised (WAIS-R), forward and backward (WASDIGF, WASDIGB)
Perceptual motor skills
WAIS-R digit symbol substitution subtest (DSST)
Language function
WAIS-R vocabulary subtest (WASV)
Combined psychometric score
Psychometric Deterioration Score (PDS); sum of the nine other scores in neuropychometric evaluation
*confounders

1.4.2 The Collaborative Genetic Study of Nicotine Dependence (COGEND)

The COGEND data is a subset from a GWAS containing 2022 subjects and 215 SNPs located in eight different chromosomes. It was ensured all three possible genotypes exist in samples for every SNP. However, the 15 SNPs with rare minor allele frequencies < 5% (Figure 1) were not excluded since it has been found that less common SNPs could be also associated with nicotine dependence (Saccone, et al. 2009). Nicotine dependence is an additional covariate.

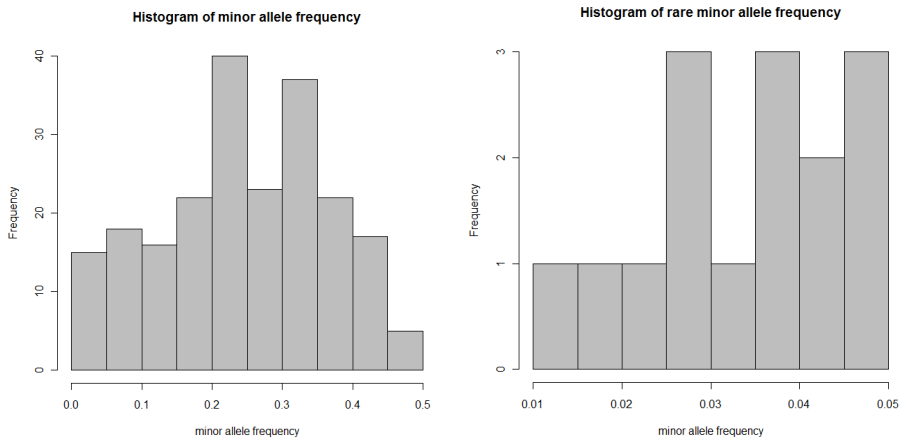


Figure 1 Minor allele frequency distribution of COGEND data. The left panel is the overall distribution of all 215 SNPs. The right histogram is for a subset of SNPs that are rare minor alleles (frequency <5%).

Given the large number of SNPs to work with, it is difficult to present an interpretable network for the sake of method performance. Hence we further combined clustering and network techniques to perform a sequential analysis: hierarchical clustering using linkage disequilibrium as a measure of similarity with all 215 SNPs was conducted first and then SNP representatives were selected within each cluster. The SNP that has overall highest similarities (linkage disequilibrium) to all the other SNPs within each cluster will be chosen. 11 SNPs from 11 clusters and 7 SNPs from an outlier cluster were included in analysis.

1.4.3 Crohn's disease study

The last dataset is from the Washington University Digestive Diseases Research Core Center Tissue Procurement Facility database; the goal is to find pathways between SNPs in Crohn's disease patients that would be of interest to do further research on.

Crohn's disease (CD) is a chronic relapsing inflammatory intestinal disorder that can affect any segment of the intestine often in a discontinuous manner (Goyette, et al. 2007, Abraham, et al. 2009). One great advantage of GWAS on CD is that it has been intensively explored and more than 30 susceptibility loci have now been identified through genome-wide association studies (Barrett, et al. 2008).

CD patients are phenotypically heterogeneous. Efforts have been made to subphenotype the patients in order to facilitate genoytpe-phenotype correlations. Both the Vienna and Montreal classifications have classified the patients on the basis of three major parameters: age of diagnosis, disease location and disease behavior (Louis, et al. 2001, Satsangi, et al. 2006). While disease behavior changes over time (Unkart, et al. 2008), disease location remains fairly stable. Based on both the Vienna and Montreal classifications, there are four major patterns of disease location: L1, ileal disease with or without cecal disease (ileal CD); L2, colonic disease only (Crohn's colitis); L3, ileal disease with colonic disease beyond the cecum; L4, proximal intestinal disease. Most CD patients have ileal and/or colonic disease (L1, L2, and L3). Only a small number of patients have disease restricted to the proximal gut (L4). Two identified CD-related genes (*NOD2* and *ATG16L1*) have been previously associated with the subset of CD patients with ileal disease location compared to control patients without inflammatory bowel diseases. These studies incorporated a relatively limited set of susceptibility loci (Cuthbert, et al. 2002, Lesage, et al. 2002, Prescott, et al. 2007, Fowler, et al. 2008, Van Limbergen, et al. 2008, Márquez, et al. 2009).

The dataset contains 628 CD patients recruited between April 2005 - February 2010 that have complete genotype information on 31 established CD risk alleles (Barrett, et al. 2008) and complete clinical information on disease location (L1-L4), smoking, gender, race and age of diagnosis. From a case-control study point of view, we intend to carry out comparison between (L1 + L3) vs. L2, or ileal (case) vs. non-ileal (control). The only six L4 observations were excluded and the information for three SNPs on *NOD2* was combined. The final dataset includes 622 samples and 29 SNPs in total. Covariate analysis is shown in Table 1.

Table 1 Crohn's disease covariate analysis. Demographic information by disease location groups. Smoking is the only covariate that is significantly associated with disease location ($p<0.01$).

		Disease location				
		Non-ileum		Ileum		
Covariate		%	n	%	n	Pvalue
Gender	Male	49.62	65	44.6	219	0.3059
	Female	50.38	66	55.4	272	
Race	White	83.97	110	90.43	444	0.1072
	Black	13.74	18	8.35	41	
	Other	2.29	3	1.22	6	
Montreal_Age	<17	16.03	21	15.27	75	0.372
	17-40	60.31	79	66.19	325	
	>40	23.66	31	18.53	91	
Smoking	Never	61.83	81	55.6	273	0.0043
	Current	12.98	17	6.52	32	
	Ex	25.19	33	37.88	186	

2 Confounding and confounders

Confounding is an important issue to consider in experimental design and statistical analysis. When confounding is not taken into account, effect estimates can be severely biased. In order to identify when confounding occurs and which covariates are confounders, confounding and confounders must be clearly defined. Unfortunately, many statistical texts fail to pay proper homage to the concept. Perhaps this is so because in a purely statistical approach, there are no distinctions among covariates that will affect the outcome. However, in fields such as epidemiology, there are two "classifications" of covariates: specific variables of interest and variables that are not of interest but have an effect on the outcome, also known as confounders. In this context, confounding and confounders must be properly studied.

2.1 Comparability versus collapsibility

According to Greenland and Robins (1986), there are two different but well known approaches to define confounding: comparability based and collapsibility based. The comparability based definition credits inherent differences in risk of outcome between different treatment groups as the source of confounding. The collapsibility based definition considers confounding as a result of the difference between certain stratified (conditional) statistical measures of association and the corresponding crude (unconditional or "collapsed") measure. The problem with collapsibility based definitions is that collapsibility can occur when there is confounding and, conversely, there may be no confounding with noncollapsibility. Another problem with collapsibility-based definitions is that confounding would depend on the parameter chosen to measure the effect.

Miettinen and Cook (1981) defined confounders separately for follow-up studies and case-control studies. In follow-up studies, confounders must be a predictor of the outcome and have different distributions between treatment populations. In case-control studies, there are two possible ways confounding can occur. A priori confounders correlate with exposure in joint overall population of cases and controls; they are determinants of the outcome or have different selection implications between cases and control. Factors that affect the accuracy of exposure information are also confounders if distributed differently between cases and controls. Although confounding was defined separately for the two different types of studies, the criterions for both types ultimately deal with the covariates' relationships with observing the outcome and observing the treatment given.

2.2 Counterfactual model for effects

The two criterions given by Miettinen and Cook for confounding in follow-up studies (must be a predictor of the outcome and must have different distributions between treatment populations) are not sufficient to define a confounder (Greenland, et al. 1986). The problem of confounding is rooted in failure to identify, or estimate, from the data alone the causal parameters that determine what is observed. By combining the data with an assumption about the exchangeability of the treatment populations, the parameters are partially identifiable (Greenland, et al. 1986). The exchangeability assumption assumes equality of incidence proportions of cohorts when exposure is absent. It is deduced that this exchangeability assumption is equivalent to comparability definitions of confounding. Hence, confounding is consistent with comparability based definitions and contradictory with collapsibility based definitions.

To formalize the definition of confounding, Greenland *et al.* (1999) suggest using the counterfactual model for effects. Suppose the goal is to measure the effect of applying a treatment x_1 on a parameter μ of the distribution of the outcome y in target population A, relative to applying treatment x_0. If x_1 is applied to population A then $\mu = \mu_{A1}$; if x_0 is applied to population A, then $\mu = \mu_{A0}$. Then the causal effect of x_1 relative to x_0, or the association between the treatment and the outcome, can be measured as a difference or a ratio:

$$\mu_{A1} - \mu_{A0} \text{ or } \frac{\mu_{A1}}{\mu_{A0}} \text{ if } \mu \text{ is strictly positive.}$$

If A is observed under x_1 then $\mu = \mu_{A1}$ is observable and μ_{A0} is unobservable. Then, supposing a control population B under treatment x_0 results in $\mu = \mu_{B0}$, we assume that $\mu_{A0} = \mu_{B0}$. Confounding is present if μ_{A0} does not equal μ_{B0}. If confounding is present, the crude association parameter obtained by substituting μ_{B0} for μ_{A0} in effect measure will not equal causal parameter and the association parameter is confounded. If $\mu_{A0} \neq \mu_{B0}$, then populations A and B must differ by factors that affect μ and these factors are confounders.

With regards to the inconsistency of the collapsibility definition of confounding, when the effect measure can be expressed as the average of the effect on population members, conditions for non-collapsibility and confounding will be identical, provided the covariates in question form a sufficient set for control.

2.3 Occasional confounder and uniformly irrelevant factor

Geng *et al*. (2002) gave a formal definition of a confounder based on the criterions produced by Miettenin and Cook (1981), Kleinbaum *et al*. (1982), Greenland and Robins (1986) and Greenland *et al*. (1999). Their definitions do not require a set of sufficient confounder to control for confounding. They proposed two new concepts, occasional confounder and uniformly irrelevant factor, to unify Miettinen and Cook's criterion with the criterion based on collapsibility of differences in risk or relative risks.

Let E be exposure with values e (presence) and $\overline{e}$ (absence), and let D_e and $D_{\overline{e}}$ be the corresponding responses that take on values 1 or 0 (presence or absence of disease, respectively). $P\left(D_e = 1 \mid E = e\right)$ is the proportion of diseased individuals in exposed population. $P\left(D_{\overline{e}} = 1 \mid E = \overline{e}\right)$ is the proportion of diseased individuals in unexposed population, also known as the crude proportion. $P\left(D_{\overline{e}} = 1 \mid E = e\right)$ is the hypothetical proportion of individuals in exposed population who would have developed the disease even if they had not been exposed, also known as the counterfactual model. Confounding bias would then be the difference between the hypothetical proportion and the crude proportion.

Let C be a covariate with possible values 1, …, K that is not an intervening variable in a causal pathway from exposure to disease. The crude proportion of diseased individuals in the unexposed subpopulation of $C = k$ is $P\left(D_{\overline{e}} = 1 \mid E = \overline{e}, C = k\right)$ and the hypothetical proportion of diseased individuals in the exposed subpopulation $C = k$ is $P\left(D_e = 1 \mid E = e, C = k\right)$. Confounding bias in the subpopulation $C = k$ is, correspondingly, the difference between the hypothetical proportion and the crude proportion. Then the standardized proportion $P_{\Delta}\left(D_{\overline{e}} = 1 \mid E = \overline{e}\right)$ is obtained by adjusting the distribution of C in the unexposed population to that in the exposed population is

$$P_{\Delta}\left(D_{\overline{e}} = 1 \mid E = \overline{e}\right) = \sum_{k=1}^{K} P\left(D_{\overline{e}} = 1 \mid E = \overline{e}, C = k\right) P\left(C = k \mid E = e\right).$$

A covariate C is a confounder if the standardized proportion obtained by adjusting for C is closer to the hypothetical proportion than the crude proportion. This definition does not assume subpopulation comparability. If C is a confounder than C is dependent on E and $D_{\overline{e}}$ is conditionally dependent on C given $E = \overline{e}$. This roughly translates to Miettinen and Cook's (1981) two criterions: C has different distributions in the exposed and unexposed populations and is predictive of risk in the unexposed

population. A covariate *C* is an irrelevant factor if the hypothetical proportion is equal to the standardized proportion obtained by adjusting for *C*. These definitions are based on the categorization of the sample space of *C*, and could lead to contradictory conclusions about whether or not *C* is a confounder. These possible contradictions motivated Geng *et al.* (2002) to introduce the following concepts.

An occasional confounder *C* has a partition of its sample space such that confounding can be reduced or eliminated by controlling for *C* with respect to the partition p. If *C* is an occasional confounder then *C* is dependent on *E* and $D_{\bar{e}}$ is conditionally dependent on *C* given $E = \bar{e}$. A factor *C* is a uniformly irrelevant factor if, for any partition *p* of its sample space, the standardized proportion is equal to the hypothetical proportion. *C* is a uniformly irrelevant factor if and only if *C* is independent from *E* or $D_{\bar{e}}$ is conditionally independent from *C* given $E = \bar{e}$. So when there is confounding in subpopulations induced by a potential confounder, non-collapsibility of risk differences is neither equivalent to Miettinen and Cook's criterion nor a necessary condition of a confounder, but Miettinen and Cook's criterion is still necessary for an occasional confounder.

Extension to multiple potential confounders is derived by Geng *et al.* (2002) as follows: let Δ_A be the set of all possible values of the covariate set *A*. The standardized proportion conditional on $S = s$ by adjusting the distribution of *A* in the unexposed population to that in the exposed population is

$$P_A\left(D_{\bar{e}} = 1 \mid E = \bar{e}, S = s\right) = \sum_{a \in \Delta_A} P\left(D = 1 \mid E = \bar{e}, A = a, S = s\right) P\left(A = a \mid E = e, S = s\right).$$

Let *C* be a set of potential confounders, *S* be a subset of *C* and *R* be the remainder subset. *R* is a confounder set conditional on $S = s$ if the difference between the hypothetical proportion, $P\left(D_{\bar{e}} = 1 \mid E = e, S = s\right)$, and the standardized proportion by adjusting for the distribution of *R* is smaller than the confounding bias in the subpopulation $S = s$.

R is an irrelevant set conditional on $S = s$ if the standardized proportion by adjusting for the distribution of R is equal to the hypothetical proportion. *R* contains at least one occasional confounder conditional on $S = s$ if there are both a subset $F = \{F_1, ..., F_m\}$ of R and a partition $p_k = \{\omega_{k1}, ..., \omega_{kI_k}\}$, $k = 1, \ldots, m$ of Δ_{F_k} for each covariate F_k in *F* such that

$$P_{F_p}\left(D_{\bar{e}} = 1 \mid E = \bar{e}, S = s\right) = \sum_{\omega_1 \in p_1} \ldots \sum_{\omega_m \in p_m} P\left(D_{\bar{e}} = 1 \mid E = \bar{e}, F_1 \in \omega_1, ..., F_m \in \omega_m, S = s\right)$$
$$\times P\left(F_1 \in \omega_1, ..., F_m \in \omega_m \mid E = e, S = s\right)$$

R is a uniformly irrelevant set conditional on $S = s$ if the standardized proportion by adjusting for the distribution of any subset F and any partition p_k of each F_k in F is equal to the hypothetical proportion. So, a uniformly irrelevant set must be an irrelevant set and that a uniformly irrelevant set does not contain any occasional confounders.

If R can be decomposed into two disjoint subsets R_1 and R_2 such that either

1) R_1 is conditionally independent from E given $S = s$ and R_2 is conditionally independent from $D_{\bar{e}}$ given $E = \bar{e}$, R_1, and $S = s$, or
2) R_2 is conditionally independent from E given R_1 and $S = s$, and R_1 is conditionally independent from $D_{\bar{e}}$ given $E = \bar{e}$ and $S = s$

then R is an irrelevant set conditional on $S = s$ and R_1 is a uniformly irrelevant set conditional on $S = s$. If R contains at least one occasional confounder conditional on $S = s$, then R is not conditionally independent from E given $S = s$ and R is not conditionally independent from $D_{\bar{e}}$ given $E = \bar{e}$, $S = s$. If R is a confounder set conditional on $S = s$, then both

1) R_1 is not conditionally independent from E given $S = s$ or R_2 is not conditionally independent from $D_{\bar{e}}$ given $E = \bar{e}$, R_1, and $S = s$ and
2) R_2 is not conditionally independent on E given R_1 and $S = s$ or R_1 is not conditionally independent from $D_{\bar{e}}$ given $E = \bar{e}$ and $S = s$ for any possible decomposition R_1 and R_1 of R.

With the introduction of the occasional confounder and uniformly irrelevant factor, the seemingly different approaches to defining confounding, comparability and collapsibility, can coexist. Based on these developed concepts, Geng *et al*. (2002) conclude that Miettinen and Cook's criterion, derived from comparability based definitions of confounding, identify uniformly irrelevant factors and the collapsibility based approach only identifies irrelevant factors, which may also be occasional confounders.

2.4 Common methods to controlling for confounding

There are five general methods to control for confounding: randomization, restriction, matching, stratification, and regression analysis (covariance adjustment) (Greenland, et al. 2001). Each method has its own advantages and disadvantages. Randomization is considered the gold standard to control for

confounding, but often outside factors such as ethical considerations or rarity of events of interest making randomization impossible. Restriction removes confounding by reducing the sample to a specific covariate pattern, but this can make the subject pool too small and generalizability of the results difficult to justify. Matching subjects removes confounding by balancing the distribution of the confounding covariates between the groups, but matching on too many covariates makes finding a match difficult. Stratification on a confounder groups subjects such each stratum is homogenous with respect to the confounder, but stratifying on too many confounders can easily result in empty stratums, also known as the sparse-data problem.

Although still limited by sample size, covariate adjustment can work around the sparse-date problem. In addition, the simplicity in interpreting the results obtained from regression methods makes regression a popular method. For this reason, we will focus on the comparing the aforementioned regression techniques for controlling for confounding.

2.5 Confounding in the ADCC study

We will identify confounders based on criterions derived above through the definition of uniformly irrelevant factors from Geng, et al. (2002). In particular, we will focus on comparability based definitions of confounding and the counterfactual model for effects. We will also identify covariates as confounders based on background knowledge, regardless of the outcome for significance in the data.

In the ADCC study, mental decline rate is to be compared between subjects with subjective complaints of cognitive impairment (SCI) and subjects without subjective complaints of cognitive impairment (NCI). Age, gender, education, and length of follow-up will be considered as confounders. Table 2 compares the distribution of the confounders between the two groups. Age is significantly different between NCI and SCI, while gender and education are marginally significantly different.

Table 3 shows the within group distribution of the confounders between those who declined and those who were stage, where once again, within each group, age is significantly different between decliners and those who were stable. Length of follow-up is also significantly different between decliners and those who were stable.

Table 2 Confounder distribution for followed subjects. Age is significantly different between NCI (no subjective cognitive impairment) and SCI (subjective cognitive impairment) groups.

	NCI (n = 47)		SCI (n=166)			
Baseline Variable	Mean	Std Dev	Mean	Std Dev		P value
Age	64.1	8.9	67.5	8.9		0.020
Gender						
Female, no. (%)	26	55.3	108	65.1		0.226
Male, no. (%)	21	44.7	58	34.9		
Education, yrs	16.1	2.4	15.6	2.6	(n=164)	0.295
Length of follow-up	6.7	3.1	6.8	3.4		0.796

Table 3 Confounder distribution for followed NCI and SCI subjects. Stratified confounder distribution within NCI (no subject cognitive impairment) and SCI (subjective cognitive impairment) subjects; comparison of those who had stable cognitive states against those who experience decline in cognitive state. The independent samples t-test was performed when the normality assumption was met (**) while the Wilcoxon Rank Sum Test (*) was used when the assumption was not attainable.

NCI Subjects	STABLE (n=40)		DECLINE (n=7)		
	Mean	Std Dev	Mean	Std Dev	P value*
Age	62.540	8.4	73.0	5.9	0.004
Gender					
Female, no. (%)	23	57.5	3	42.9	0.684
Male, no. (%)	17	42.5	4	57.1	
Education, yrs	16	2.4	16.6	2.5	0.651
Length of follow-up	6.0	2.4	10.7	3.6	0.002

SCI Subjects	STABLE (n=76)		DECLINE (n=90)			
	Mean	Std Dev	Mean	Std Dev		P value**
Age	65.3	8.9	69.4	8.4		0.002
Gender						
Female, no. (%)	51	67.1	57	63.3		0.612
Male, no. (%)	25	32.9	33	36.7		
Education, yrs	16.2	2.0	15.148	3.0	(n=88)	0.009
Length of follow-up	5.8	3.1	7.703	3.5		0.0003

3 The Pearson residual

3.1 Dichotomous outcome

In linear regression, the true model for each subject i, $i = 1, \ldots, n$, is of the form

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i \text{, where } \mathbf{X}_i = \{1, X_{i1}, \ldots, X_{ip}\} \text{ and } \boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_p\}.$$

It is assumed that the errors ε_i are independent, identically distributed normal with mean zero and constant variance σ. This error variance is also independent from the conditional mean $E[Y_i \mid \mathbf{X}_i]$, which allows one to calculate the residuals, e_i, directly from the regression equations. For linear regression, the residual is difference between the observed and predicted values of Y, e.g. $e_i = Y_i - \hat{Y}_i$.

Residual calculations are not so straightforward in the case of logistic regression, because the error variance is a function of the conditional mean (Hosmer, et al. 2000). A residual for logistic regression that has common properties as the residual for linear regression is the Pearson residual.

Suppose there are J observed covariate patterns for covariate vector $\mathbf{X}$, and the number of subjects with $\mathbf{X} = \mathbf{x}_j$ is m_j, for $j = 1, \ldots, J$. Let y_j be the number of subjects with covariate pattern $\mathbf{X} = \mathbf{x}_j$ that had outcome $Y = 1$. The Pearson residual for a particular covariate pattern $\mathbf{X} = \mathbf{x}_j$ is defined to be

$$r_j = \frac{y_j - m_j\hat{\pi}(\mathbf{x}_j)}{\sqrt{m_j\hat{\pi}(\mathbf{x}_j)\left[1-\hat{\pi}(\mathbf{x}_j)\right]}},$$

where $\hat{\pi}(\mathbf{x}_j) = \dfrac{e^{\mathbf{x}_j'\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_j'\hat{\boldsymbol{\beta}}}}$, the estimated probability that $Y = 1$.

For subject i with covariate vector $\mathbf{X} = \mathbf{x}_i$ and observed outcome $Y = y_i$, the Pearson residual is

$$r_i = \frac{y_i - \hat{\pi}(\mathbf{x}_i)}{\sqrt{\hat{\pi}(\mathbf{x}_i)\left[1-\hat{\pi}(\mathbf{x}_i)\right]}}.$$

The Pearson residual is asymptotically normal with mean zero and variance equal to one (Hosmer, et al. 2000, Agresti 2007).

3.2 Multinomial outcome

When the outcome *Y* is polychotomous with $I > 2$ possible outcomes, the Pearson residual can be extended as follows (Lesaffre, et al. 1989):

For outcome *i*, the Pearson residual for an individual with covariate pattern covariate pattern $\mathbf{X} = \mathbf{x}_j$ is

$$r_{ij} = \frac{y_{ij} - \hat{\pi}_i\left(\mathbf{x}_j\right)}{\sqrt{\hat{\pi}_i\left(\mathbf{x}_j\right)}},$$

where $\hat{\pi}_i\left(\mathbf{x}_j\right)$ is the estimated probability of resulting in outcome *i* given covariate pattern $\mathbf{X} = \mathbf{x}_j$. Thus, each individual will have *I* residuals.

3.3 Pearson residual analysis

The Pearson residual will also be incorporated into a two stage logistic regression technique for comparison purposes to the proposed residual logistic regression. Pearson residual analysis will be carried out as follows:

Let $X_1,...,X_k$ be a set of potential confounding variables, $X_{k+1},...,X_p$ be a set of variables of interest (VOI's), and let *Y* be a dichotomous outcome variable where $\mathrm{E}[Y] = \pi$.

Stage 1: Fit the logistic model for *Y* on the set of potential confounding variables

$\ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ to obtain $\hat{\pi}\left(\mathbf{x}_i\right) = \frac{e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}}$ and calculate the Pearson residuals

$$r_i = \frac{y_i - \hat{\pi}\left(\mathbf{x}_i\right)}{\sqrt{\hat{\pi}\left(\mathbf{x}_i\right)\left[1-\hat{\pi}\left(\mathbf{x}_i\right)\right]}}$$

Stage 2: Fit the a linear regression model for r_i on the set of VOI's

$$r_i = \beta_0^* + \beta_{k+1} X_{k+1} + \cdots + \beta_p X_p .$$

3.4 Application to the ADCC study

For the ADCC study, Pearson residuals, r_i, will be obtained from the following fitted model

$$\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 * AGE + \beta_2 * GENDER + \beta_3 * EDUC + \beta_4 * TIME2END ,$$

where $\pi = P[\text{Decline}]$. Effects of variables of interest will be estimated based on the following linear regression with stepwise variable selection:

$$r_i = \beta_0 + \alpha * Group + \sum_{i=5}^{69} \beta_i X_i + \varepsilon, \text{ where } \varepsilon \sim N\left(0, \sigma^2\right).$$

Stepwise variable selection will then be implemented to reduce the model.

4 The propensity score

Consider the situation in which responses under treatment ($E = 1$) and control ($E = 0$) are to be compared. A balancing score, $b(\mathbf{x})$, is a function of the observed vectors of covariates $\mathbf{x}$ such that the conditional distribution of $\mathbf{x}$ given $b(\mathbf{x})$ is the same for treatment and control units. All functions of $\mathbf{x}$ are balancing scores and the coarsest function of $\mathbf{x}$ that is a balancing score is the propensity score of Rosenbaum and Rubin (1983). The propensity score is the probability of treatment given the observed covariates $\mathbf{x}$; namely, the propensity score is

$$\pi(\mathbf{x}) = P(E = 1 \mid \mathbf{x})$$

The propensity score can be modeled through the logistic regression.

Treatment assignment is strongly ignorable given a vector of covariates $\boldsymbol{\nu}$ if the joint distribution of the responses under each treatment is conditionally independent from the treatment assignment given $\boldsymbol{\nu}$. If treatment assignment is strongly ignorable given observed covariate vector $\mathbf{x}$, then the difference between treatment and control means at each value of the propensity score is an unbiased estimate of the treatment effect at that value (Rosenbaum, et al. 1983). Thus pair matching, stratification and covariance adjustment on the propensity score can produce unbiased estimates of the average treatment effect (Rosenbaum, et al. 1983). Hence, adjusting for the propensity score would control for confounding.

4.1 Matching, stratification and covariance adjustment

Rosenbaum and Rubin (1985) give three matching techniques based on the propensity score. It is suggested that matching be carried out through the logit of the estimated propensity score because it is approximately normally distributed. The first step for each of these techniques is to randomly order the treatment and control subjects.

Nearest available matching on the estimate propensity score selects the first treated subject and finds the control subject with the closest propensity score.

Mahalanobis metric basing including the propensity score selects the first treated subject and finds the closest control subject, where distance is defined by Mahalanobis distance:

$$d(\mathbf{u},\mathbf{v}) = (\mathbf{u}-\mathbf{v})^T \mathbf{C}_0^{-1} (\mathbf{u}-\mathbf{v}),$$

where u and v are observed values of the confounders and the estimated propensity score, and $\mathbf{C}_0$ is the sample covariance matrix of the observed values of confounders and the estimated propensity score in the control group.

Nearest available Mahalanobis metric matching within calipers defined by the propensity score is a mix method of the previous two. For each treated subject, a subset of potential controls that have propensity scores close to that of the treated subject is first formed. Then the control subject that is closest to the treated subject based on Mahalanobis metric matching is selected. The matched pairs are then removed from the set, and the process would repeat for the remaining unmatched subjects.

Based on the empirical findings of Rosenbaum and Rubin (1985), all three methods were able to reduce bias. The first method was less computational. The second method resulted in smaller standardized differences for individual variables, but had a substantial difference along the propensity score. The third method was best with respect to balancing covariates, their squares, and cross products.

Cochran (1968) showed that stratification on a covariate divided into five strata or subclasses can reduce bias by roughly 90%. However, increasing the number of covariates used for stratification will increase the number of subclasses exponentially (Cochran, et al. 1965), and result in increasing the risk of empty strata. Stratification based on the propensity score can help alleviate this problem by condensing the number of stratification variables to one. Dividing the estimated propensity score into five strata will remove more than 90% of bias due to each of the covariates most of the time (Rosenbaum, et al. 1984).

There are several approaches to covariance adjustment based on the propensity score. In regards to how the propensity score is considered as a covariate, it can be added directly or under some linear transformation or as a categorical variable where its categories are decided by its quintiles. The propensity score can replace all the confounding variables in the regression model. When the set of confounders is large, the propensity score can be included along with a subset of the confounding variables in the regression model.

D'Agostino (1998) and Newgard *et al*. (2004) provide rough overviews of using the propensity score with accompanying applications of these three general methods.

4.2 Longitudinal data: time-lagged responses and time-varying treatments

Some effort has been made to extend the use of propensity scores to longitudinal data. A propensity score estimator of the treatment effect with time-lagged responses and possible censoring was developed by Anstrom and Tsiatis (2001). They used the propensity score to obtain an inverse-probability-weighted estimator based on that from Rosenbaum (1987). The estimator is $\hat{\delta} = \hat{\mu}_1 - \hat{\mu}_0$ where $\hat{\mu}_1$ and $\hat{\mu}_0$ are solutions to

$$\sum_{i=1}^{n} \frac{A_i \Delta_i \left(R_i - \hat{\mu}_1\right)}{\pi\left(X_i, \hat{\gamma}\right) \hat{K}_1\left(U_i\right)} = 0 \text{ and } \sum_{i=1}^{n} \frac{\left(1 - A_i\right) \Delta_i \left(R_i - \hat{\mu}_0\right)}{\left\{1 - \pi\left(X_i, \hat{\gamma}\right)\right\} \hat{K}_0\left(U_i\right)} = 0,$$

respectively. A_i is the treatment indicator variable, Δ_i is the complete-case indicator (not censored), R_i is the observed response, $\pi\left(X_i, \hat{\gamma}\right)$ is the estimated propensity score, U_i is the time to response ascertainment or censoring, and $K_j(U_i)$ is the treatment-specific Kaplan-Meyer estimated probability of censoring occurring beyond time U_i given treatment group j. This estimator is asymptotically normal. A simulation study with sample sizes 500 and 1000 found that the estimator performs well. Caution should be taken, as inverse-probability-weighted estimators are unstable when weights are large.

Patients may receive treatment at different times. For this situation, Li *et al*. (2001) proposed a balanced risk set matching design which matches a patient receiving the treatment at time t to another patient with similar history of symptoms up to time t who has not received the treatment yet. This matching method would balance the marginal distribution, but not necessarily produce closest matches on covariate values and is not practical when the covariates are of high dimension.

The time dependent propensity score is the hazard of receiving the treatment at time t given that the treatment has not been given before time t. It can be model by a proportional hazards model with time varying covariates:

$$h_m\left(t\right) = h_0\left(t\right) \exp\left\{\beta^T X_m\left(t\right)\right\}$$

where $h_m(t)$ is the hazard for patient m at time t.

Time-dependent propensity scores balance the distribution of observed covariates in matched treated and control groups, and they do so at every time point (Lu 2005). Two matching algorithms based on the time-dependent propensity have been suggested (Lu 2005). Sequential matching is done within risk

sets, where a risk set for time t contain all patients at risk of treatment at time t. If there is only one treated patient in the risk set, then it is matched with the control with the closest time-dependent propensity score. If there is more than one treated patient in a risk set then they all compete for controls. Matches are made when the total distance within matched sets are minimized. Minimization is obtained through optimal bipartite matching (Bergstralh, et al. 1996, Rosenbaum 2002). Matched subjects are removed and the process continues with the next risk set. Simultaneous matching forces all patients to have only one match. Optimal simultaneous matching is achieved by comparing all possible combinations of matched pairs at once. The set of matched pairs is found by optimal non-bipartite matching (Derigs 1988).

At the time of this paper, properties of time-dependent propensity score through stratification and covariate adjustment approaches and the combination of the time-dependent propensity score in the time-lagged response scenario do not appear to have been explored in the literature.

4.3 Dealing with more than two treatment levels

Suppose treatment levels are ordinal. When there is a single variable, say $b(X)$, that determines not just the expected dose given X but the entire distribution of doses given X, then a single balancing score is available with more than two treatment levels. If the entire distribution of treatments Z depends on covariates X only through $b(x)$, so that $P(Z|X) = P(Z|b(x))$, then $b(X)$ is a balancing score and persons with the same balancing score in different treatment groups have the same distribution of the covariates X (Joffe, et al. 1999). Then adjusting for $b(X)$ would balance X across the treatment levels.

The multiple propensity score of Wang, *et al*. (2001) is an extension of the propensity score such that it could be applied to situations in which there are more than two treatment levels, not necessarily ordinal. It is the conditional probability of a patient receiving a particular treatment given all observed covariates and can be estimated with multinomial logistic regression. Once the multiple propensity score has been estimated, it can be used to control for confounding through the same three general methods: matching, stratification, and covariate adjustment. Although the multiple propensity score has been shown to be viable and general results from the usual propensity score appear to extend to the multiple propensity score, few formal results have been developed (Spreeuwenberg, et al. 2010).

4.4 Previous findings concerning propensity scores

Although the true propensity score is unknown, it is argued that adjusting for the estimated propensity score removes bias better because it accounts for the chance imbalances in *X* (Joffe, et al. 1999).

It should be stressed that adjusting for propensity scores removes bias from observed confounders, but does not account for unmeasured confounding. However, the amount of bias that may occur from omitting a confounder from the propensity score is positively correlated with the degree of association between the confounder and the treatment and between the confounder and the outcome (Weitzen, et al. 2005).

Cepeda, *et al*. (2003) compared logistic regression against propensity score covariate adjustment in the situation when the number of events is low and multiple confounders exist. The logistic regression model included all individual confounders and the exposure variable as independent variables and the clinical outcome as the dependent variable. The propensity score was estimated and divided into five strata based on its quintiles. Another logistic model was then built, with the clinical trial as the dependent variable, and the exposure variable and the five categories of the propensity score as the independent variables. Propensity score estimates were less biased, more robust, more precise, and had more power than logistic regression when there were seven or fewer events per confounder. It was concluded that propensity scores are a good alternative to control for imbalances when there are seven or fewer events per confounder.

Martens, *et al*. (2008) compared the three propensity score methods (matching, stratification, and covariate adjustment) with a multivariable Cox proportional hazards regression to estimate an adjusted effect of drug treatment for hypertension on the incidence of stroke. Stratification was based on quintiles. Matching on the propensity score was based on pair matching through a greedy algorithm. Covariate adjustment for the propensity score consisted for replacing all confounders in the Cox proportional hazards model with the propensity score. Matching and stratification on the propensity score gave larger treatment effect and more precision than the multivariable Cox proportional hazards regression and propensity score covariate adjustment. Propensity score covariate adjustment appeared to perform about the same as the multivariable Cox proportional hazards regression.

Austin (2007) compared the performance of the three propensity score methods in estimating marginal odds ratios through Monte Carlo simulations. They found that matching on the propensity score

resulted in the least biased estimates of the marginal odds ratios. Stratifying on the quintiles of the propensity score resulted in the greatest degree of bias. When there was true non-null treatment effect, then all methods resulted in confidence intervals with sub-optimal coverage. Propensity score matching tended to give estimated of marginal odds ratios with the lowest mean square error.

Senn, *et al*. (2007) compared treatment effects using stratification on the propensity score versus least squares regression. They show that, when both propensity score stratification and least squares regression produce conditionally unbiased estimators, the true variance of the propensity score estimator would be higher. When the estimators are compared marginally, then depending on the ability of a balanced covariate to predict the outcome, the propensity score estimator may outperform the least squares estimator with respect to marginal variance. It is conjectured that if a linear model holds, propensity score stratification based on estimated propensity scores produces an estimator with the same expectation but greater variance than the least squares estimator.

4.5 Application to the ADCC study

Although the propensity score was originally developed for the analysis of treatment effects, propensity score methods can easily be extended to analysis of group effects (Reeve, et al. 2008). We will consider SCI as the group of interest and NCI has the control group. The conditional probability of SCI will be modeled on the confounding variables through logistic regression.

$$\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 * AGE + \beta_2 * GENDER + \beta_3 * EDUC + \beta_4 * TIME2END \quad ,.$$

where π=P$\left(SCI \mid AGE, GENDER, EDUC, TIME2END\right)$

Many research studies have applied propensity score adjustment by using the raw propensity score as a covariate in the model. However, there is another body of research studies that have used the logit of the estimated propensity score because Rosenbaum and Rubin (1985) found that it is more normally distributed than the raw propensity score. We will compare each method by considering two different models:

1) logistic regression with raw propensity

$$\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \alpha * Group + \gamma * \hat{\pi} + \sum_{i=5}^{69} \beta_i X_i \text{ and}$$

2) logistic regression with the logit propensity score

$$l\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \alpha * Group + \gamma * \log\left[\frac{\hat{\pi}}{1-\hat{\pi}}\right] + \sum_{i=5}^{69} \beta_i X_i$$

where π = P[decline]. Stepwise variable selection will also be incorporated in the second regression. Similar implementation is carried out for Cox PH regression and AFT regression, with corresponding models in the second regression.

5 Residual logistic regression

For multiple linear regression analysis where the response variable is continuous, the two-stage residual linear regression analysis strategy has been well developed and adopted (Freund, et al. 1961a, Freund, et al. 1961b, Kabe 1963, Zyskind 1963, Alley 1987). Let $X_1,...,X_k$ be a set of potential confounding variables, $X_{k+1},...,X_p$ be a set of variables of interest (VOI's), and let *Y* be a continuous outcome variable where $\mathrm{E}[Y]=\mu$.

Stage 1: Fit the linear regression model for y on the set of confounding variables

$$Y=\beta_0+\beta_1X_1+\cdots+\beta_kX_k+\varepsilon \text{ to obtain } \hat{Y}=\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_kX_k$$

Stage 2: Fit the residual linear regression model for $Y-\hat{Y}$ on the set of VOI's

$$Y-\hat{Y}=\beta_0^*+\beta_{k+1}X_{k+1}+\cdots+\beta_pX_p+\varepsilon$$

For dichotomous response variables, Baez-Revultas (2009) introduced the analogous residual logistic regression to control for confounding. Now, let *Y* be a dichotomous outcome variable where $\mathrm{E}[Y]=\pi$.

Stage 1: Fit the logistic model for y on the set of potential confounding variables

$$\ln\left[\frac{\pi}{1-\pi}\right]=\beta_0+\beta_1X_1+\cdots+\beta_kX_k$$

to obtain

$$T=\ln\left[\frac{\hat{\pi}}{1-\hat{\pi}}\right]=\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_kX_k$$

Stage 2: Fit the residual logistic model for *Y* on the set of VOI's

$$\ln\left[\frac{\pi}{1-\pi}\right]-T=\beta_0^*+\beta_{k+1}X_{k+1}+\cdots+\beta_pX_p.$$

This is equivalent to fitting

$$\ln\left[\frac{\pi}{1-\pi}\right] = T + \beta_0^* + \beta_{k+1}X_{k+1} + \cdots + \beta_p X_p .$$

The procedure maintains the interpretative odds ratio accounted by the VOI's since both levels are logistic. Furthermore, variable selection from the set of VOI's can easily be incorporated in the procedure within the second stage.

5.1 Simple Proof

Ideally, we would like to show that

$$\ln\left(\frac{\pi(\mathbf{x}_c, g, \mathbf{x}_{VOI})}{1-\pi(\mathbf{x}_c, g, \mathbf{x}_{VOI})}\right) = \ln\left[\frac{\pi(\mathbf{x}_c)}{1-\pi(\mathbf{x}_c)}\right] + \ln\left[\frac{\pi(g, \mathbf{x}'_{VOI})}{1-\pi(g, \mathbf{x}'_{VOI})}\right]$$

We consider a simple scenario of one confounder X and one variable of interest G and attempt to derive

$$\ln\left(\frac{\pi(X, G)}{1-\pi(X, G)}\right) = \ln\left[\frac{\pi(X)}{1-\pi(X)}\right] + \ln\left[\frac{\pi(G)}{1-\pi(G)}\right]$$

Let Y be a binary response variable, G be a discrete variable and X a discrete confounding variable. Consider the odds of Y=1 given G=g and X=x.

$$\frac{\pi(g,x)}{\pi^c(g,x)} = \frac{P(Y=1 \mid G=g, X=x)}{P(Y=0 \mid G=g, X=x)} = \frac{P(Y=1, G=g, X=x) \,/\, \cancel{P(G=g, X=x)}}{P(Y=0, G=g, X=x) \,/\, \cancel{P(G=g, X=x)}}$$

$$= \frac{P(Y=1, G=g, X=x)}{P(Y=0, G=g, X=x)}$$

$$= \frac{P(G=g, X=x \mid Y=1)\,P(Y=1)}{P(G=g, X=x \mid Y=0)\,P(Y=0)}$$

Now suppose X and G are conditionally independent variables.

$$P(G=g, X=x \mid Y=y) = P(G=g \mid Y=y)\,P(X=x \mid Y=y)$$

Then,

$$\frac{P(G=g,X=x\mid Y=1)P(Y=1)}{P(G=g,X=x\mid Y=0)P(Y=0)}=\frac{P(G=g\mid Y=1)P(X=x\mid Y=1)P(Y=1)}{P(G=g\mid Y=0)P(X=x\mid Y=0)P(Y=0)}$$

$$=\frac{\frac{P(Y=1\mid G=g)\,P(G=g)}{P(Y=1)}\,\frac{P(Y=1\mid X=x)\,P(X=x)}{P(Y=1)}\,P(Y=1)}{\frac{P(Y=0\mid G=g)\,P(G=g)}{P(Y=0)}\,\frac{P(Y=0\mid X=x)P(X=x)}{P(Y=0)}\,P(Y=0)}$$

$$=\left[\frac{P(Y=1\mid G=g)}{P(Y=0\mid G=g)}\right]\left[\frac{P(Y=1\mid X=x)}{P(Y=0\mid X=x)}\right]\left[\frac{1/P(Y=1)}{1/P(Y=0)}\right]$$

$$\frac{\pi(g,x)}{\pi^{c}(g,x)}=\left[\frac{P(Y=1\mid G=g)}{P(Y=0\mid G=g)}\right]\left[\frac{P(Y=1\mid X=x)}{P(Y=0\mid X=x)}\right]\left[\frac{1}{P(Y=1)/P(Y=0)}\right]$$

$$=\left[\frac{\pi(g)}{\pi^{c}(g)}\right]\left[\frac{\pi(x)}{\pi^{c}(x)}\right]\left[\frac{1}{\pi_0/(1-\pi_0)}\right]$$

$$\ln\left[\frac{\pi(g,x)}{\pi^{c}(g,x)}\right]=\ln\left[\frac{\pi(g)}{\pi^{c}(g)}\right]+\ln\left[\frac{\pi(x)}{\pi^{c}(x)}\right]-\ln\left[\frac{\pi_0}{1-\pi_0}\right]$$

Thus,

$$\ln\left[\frac{\pi(g,x)}{\pi^{c}(g,x)}\right]=\ln\left[\frac{\pi(g)}{\pi^{c}(g)}\right]+\ln\left[\frac{\pi(x)}{\pi^{c}(x)}\right]-\ln\left[\frac{\pi_0}{1-\pi_0}\right]$$

We will modify Baez-Revultas's method by including the correctional crude estimate of the log odds.

5.2 Simulations

To compare the new residual logistic regression to the other modeling techniques to control for confounding, 1000 datasets were simulated containing 500 observations each with three binary variables: outcome *Y*, group variable *G*, and confounder *X*. The goal of the simulation is to study the performance of the methods to detect and estimate the effect of *G* on *Y* based on the interplay of the strength of the correlation between *G* and *X* and the strength of the true effect of *G* on *Y*.

We take advantage of the odds and odds ratio interpretation of the logistic model to determine reasonable values of the parameters. We also use the conditional probability interpretation to calculate the true correlation between predictor *X* and confounder *G*.

The data were simulated as follows:

- X ~ Bernoulli($P\left(X = x\right)$)
- Given X, generate G ~ Bernoulli($P\left(G = 1 \mid X = x\right) = \frac{\exp(a+bx)}{1+\exp(a+bx)}$)
- Given X and G, generate Y ~ Bernoulli($P\left(Y = 1 \mid G = g, X = g\right) = \frac{\exp(\alpha+\beta_G g+\beta_X x)}{1=\exp(\alpha+\beta_G g+\beta_X x)}$)

To determine appropriate values of the parameters needed in the simulation, we note the following interpretations.

$$
\begin{aligned}
&P\left(X = 1\right) = p_X \\
&a = \log(\text{odds G}) = \log\left(\tfrac{p_0}{1-p_0}\right) \\
&b = \log(OR_X) = \log\left(\tfrac{p_1}{1-p_1} / \tfrac{p_0}{1-p_0}\right) \\
&\log\left(\tfrac{P(G=1|X=x)}{1-P(G=1|X=x)}\right) = a + bx \\
&\alpha = \log(\text{odds}) = \log\left(\tfrac{p_2}{1-p_2}\right) \\
&\beta_G = \log(OR_G) = \log\left(\tfrac{p_3}{1-p_3} / \tfrac{p_2}{1-p_2}\right) \\
&\beta_X = \log(OR_X) = \log\left(\tfrac{p_4}{1-p_4} / \tfrac{p_2}{1-p_2}\right) \\
&\log\left(\tfrac{P(Y=1|G=g,X=x)}{1-P(Y=1|G=g,X=x)}\right) = \alpha + \beta_G g + \beta_X x
\end{aligned}
$$

Each of p_X, p_0,…, p_4 can be a number between 0 and 1. p_0 controls the correlation between X and G. p_3 controls the effect of G on Y. We can fix all other parameters and while varying these two. We also note that the true correlation between X and G can be calculated as

$$
\begin{aligned}
\rho_{G,X} &= \frac{E\left[\left(G-\mu_G\right)\left(X-\mu_X\right)\right]}{\sqrt{E\left[\left(G-\mu_G\right)^2\right]E\left[\left(X-\mu_X\right)^2\right]}} \\
&= \frac{E\left[GX\right]-p_G p_X}{\sqrt{p_G\left(1-p_G\right)p_X\left(1-p_X\right)}}, \text{ since X and G are Bernoulli } p_G, p_G, \text{ respectively} \\
&= \frac{p_{GX}-p_G p_X}{\sqrt{p_G\left(1-p_G\right)p_X\left(1-p_X\right)}} \\
\text{where } p_{GX} &= P\left(G=1, X=1\right) = P\left(G=1 \mid X=1\right)P\left(X=1\right) \\
\& \; p_G &= P\left(G=1\right) = \sum_x P(G=1 \mid X=x)P(X=x)
\end{aligned}
$$

Then our simulation parameters were chosen as:

$$
\begin{aligned}
p_X &= 0.5 \\
p_0 &= \{0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95\} \\
p_1 &= 0.95 \\
p_2 &= 0.5 \\
p_3 &= \{0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95\} \\
p_4 &= 0.7
\end{aligned}
$$

The respective correlations between G and X to the chosen values of p_0 is

$$\rho_{G,X} = \{0.94, 0.9, 0.85, 0.67, 0.5, 0.33, 0.09, 0\}$$

The respective effect of G on Y is to the chosen values of p_3 is

$$\beta_G = \{-2.94, -2.2, -0.85, 0, 0.85, 2.2, 2.9\}$$

We compared the methods based on power, type I error rate, bias, standard error, and mean square error in detecting and estimating β_G. Power was the proportion of simulations where β_G was found to be significantly nonzero in the regression when it was truly nonzero. Type I error rate was the proportion of times β_G was significantly nonzero when it was truly zero. Bias was difference between the estimated and the true value of β_G. The standard error was the standard deviation in the estimates of β_G. Mean square error is the sum of the squared bias and the squared standard error.

We denote the unadjusted logistic regression as *crude*, adjusted (multiple) logistic regression as *mLR*, Pearson residual analysis as *prLR,* propensity score adjusted logistic regression as *propensity*, and residual logistic regression as *rLR*. All possible combinations of the two parameters, $(\rho_{G,X}, \beta_G)$, were simulated, but each was observed in its own right by averaging the evaluation measures over the values of the other parameter, except in the case of power and Type I error rate. Power was measured when $\beta_G \neq 0$, while Type I error rate was measure only when $\beta_G = 0$. Full results tables are included in the Supplementary materials.

Multiple logistic regression and propensity score adjusted regression performed similarly with respect to power, while Pearson residual analysis and residual logistic regression performed similarly (Figure 2). Multiple logistic regression and propensity score adjusted regression had higher power. With regards to correlation, although initial interpretation of the power curve over correlation would lead one to

conclude that residual logistic regression and Pearson residual analysis does not perform well when *X* and *G* are highly correlated, one must remember that a high correlation between two variables would result in unstable results when both variables are used in a regression model due to multicollinearity (Farrar, et al. 1967). In other words, the two-step based methods would explain all variability in *Y* due to the variable *X* in the first step and if *X* and *G* are high correlated, no variability due to *G* is left to be discovered in the second step. In this sense, Pearson residual analysis and residual logistic regression is actually the stronger methods of analysis to correctly detect this effect.

Type I error rate for crude logistic regression was removed because it performed much worse in this aspect compared to the other methods, making comparison difficult on a scaled graph. The Type I error rate for crude regression averaged over all correlation values was 0.56. With increasing correlation, Type I error rate decrease to nearly zero for Pearson residual analysis and residual logistic regression (Figure 3). For the other methods, Type I error rate was consistent around 0.05.

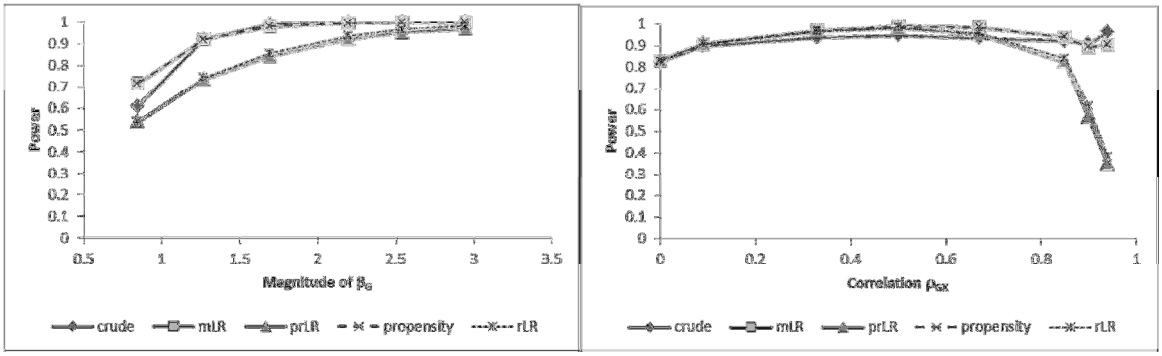


Figure 2 Confounding simulation – Power. Power of each of the regression methods [crude = unadjusted logistic regression, mLR = multiple logistic regression, prLR = Pearson residual logistic regression, rLR = residual logistic regression] to detect an effect of the main variable *G* after controlling for covariate *X*, observed over the effect size of *G* (left) and the correlation between *G* and *X* (right).

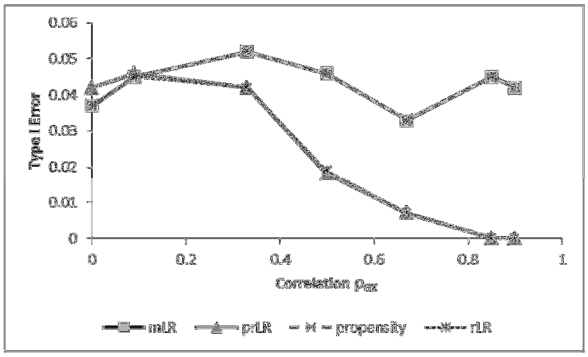


Figure 3 Confounding simulation - Type I Error. Type I error of each of the regression methods [mLR = multiple logistic regression, prLR = Pearson residual logistic regression, rLR = residual logistic regression] to detect an effect of the main variable *G* after controlling for covariate *X*, observed over the correlation between *G* and *X*.

As expected, the higher the correlation between *G* and *X* the more biased the estimate of β_G in the crude estimate (Figure 4). Pearson residual analysis has increasing bias as the magnitude of β_G increases. However, recall that the estimate in the Pearson residual analysis is from a linear regression, which means that bias is not quite correctly defined in this case, since the estimator in the Pearson residual analysis is not an odds ratio. Bias in the residual logistic regression is always in the opposite direct of the β_G, while bias for the remaining methods remained constant.

When observing the bias with respect to correlation, one must use care due to the trend in bias over the β_G for Pearson residual analysis and residual logistic regression, going from positive to negative. Recall that the bias with respect to the correlation is averaged over β_G. Hence, the opposing signs in bias will be cancelled out when averaged over β_G and it would appear that Pearson residual analysis and residual logistic regression have zero bias. To correct for this, we averaged the absolute bias instead.

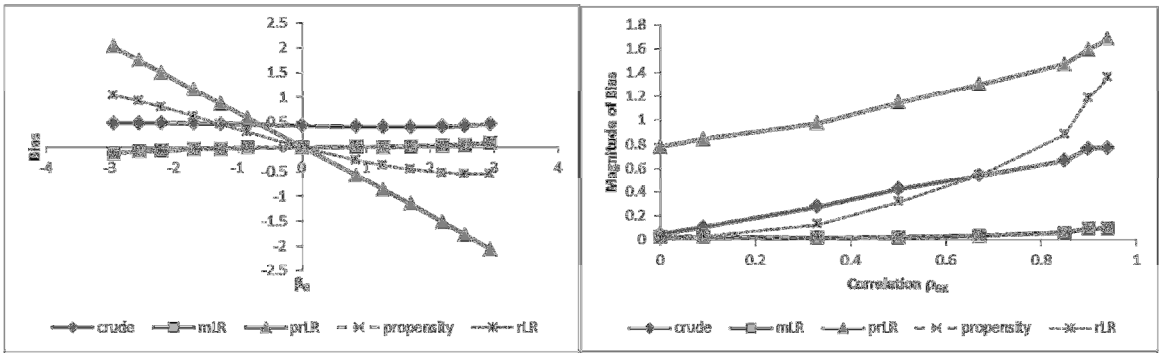


Figure 4 Confounding simulation – Bias. Bias of each of the regression methods (crude = unadjusted logistic regression, mLR = multiple logistic regression, prLR = Pearson residual logistic regression, rLR = residual logistic regression) to estimate the coefficient the main variable *G* after controlling for covariate *X*, observed over the effect size of *G* (left) and the correlation between *G* and *X* (right).

We can see that when *G* and *X* are independent, Pearson residual analysis has high bias while all other methods have nearly zero bias. As correlation increases, bias for residual logistic regression and the crude regression also increases. However, if we interpret the "true" effect of *G* on *Y* as being the sum $\beta_G + \beta_X$ when *X* and *G* are completely correlated, and having been removed when regressed *Y* was regressed on *X* first, then we speculate that the "true" bias is closer to zero for these three methods and all other estimators would have high bias.

The standard error for multiple regression and propensity score adjusted logistic regression increases as the magnitude of β_G increases and correlation increases, indicating erratic behavior in the estimates of β_G in these methods (Figure 5). The remaining methods are fairly consistent. The mean square error basically combines the information about bias and standard error into one evaluation measure. However the problems with interpretations in bias mentioned earlier would also plague these measures.

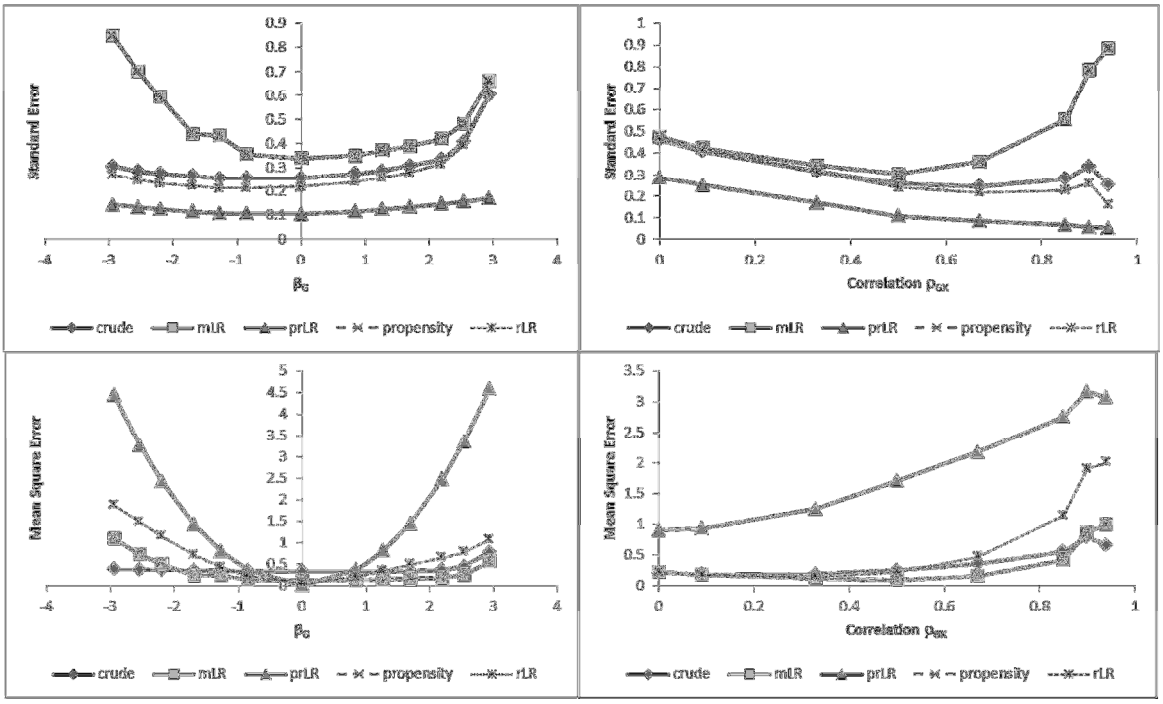


Figure 5 Confounding simulation - Standard Error and Mean Square Error. Standard error and mean square error of each of the regression methods (crude = unadjusted logistic regression, mLR = multiple logistic regression, prLR = Pearson residual logistic regression, rLR = residual logistic regression) to estimate the coefficient for the variable *G* after controlling for covariate *X*, observed over the effect size of *G* (left) and the correlation between *G* and *X* (right).

Propensity score and multiple logistic regression performed ultimately the same way in all aspects, while residual logistic regression and Pearson residual logistic regression were similar. When correlation is high, residual logistic regression and Pearson residual analysis would be better suited to handle the problem of multicollinearity. In this case, if one truly wanted to estimate the effect of *G* then crude regression would suffice, but it should be noted that it is highly correlated with confounder *X*.

5.3 Application to the ADCC study

The stage one regression model with confounders only will be

$$\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 * \text{Age} + \beta_2 * \text{Gender} + \beta_3 * \text{Education} + \beta_4 * \text{LengthofFollowup}$$

where π = P[Decline]. From this regression, we obtain the estimated log odds ratios T to be incorporated in the second stage. Then second stage regression model with variables of interest is

$$\log\left[\frac{\pi}{1-\pi}\right] = \beta_0^* + T + \alpha * Group + \sum_{i=5}^{69} \beta_i X_i \ .$$

Stepwise variable selection will be implemented to reduce the model.

Table 4, Table 5, Table 6, Table 7, Table 8, and Table 9 display results from applying the methods discussed in section 1.1 to the ADCC study in a logistic model, where outcome is either decline to MCI or dementia or no decline (stable). Group was found to have a significant effect for all methods applied (P-values < 0.05), with estimated odds ratios ranging from 5 to 9.8, but confidence interval widths ranged from approximately 15 to 38 units. Pearson residual analysis estimated the effect to be a 0.47 increase in the Pearson residual when in the SCI group and holding all other covariates fixed. HDS11 was found to be consistently significant across all methods applied (P-values < 0.05). All methods except for propensity score adjustment found BEH23 to be significant (P-values < 0.05). All methods except for the Pearson residual analysis found PDS to be significant (P-value < 0.01).

Table 4 Traditional multiple logistic regression with stepwise variable selection. Model coefficient estimates from applying multiple logistic regression to the data by regressing the binary outcome (decline vs. stable) the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and covariates age, gender, education and length of follow-up, and all other variables of interest.

	Beta	Standard Error	OR	95% Wald Confidence Limits		P-value
Intercept	-8.76	2.78				0.002
GROUP	**1.14**	**0.36**	**9.75**	**2.38**	**40.00**	**0.002**
HDS07	0.61	0.30	1.84	1.03	3.29	0.041
HDS11	0.84	0.37	2.32	1.12	4.82	0.024
BEH19	-1.51	0.61	0.22	0.07	0.72	0.012
BEH23	1.46	0.54	4.30	1.48	12.50	0.007
PDS	0.79	0.29	2.19	1.24	3.87	0.007

Table 5 One-at-a-time logistic regression. Model coefficient estimates from applying one-at-a-time logistic regression to the data by regressing the binary outcome (decline vs. stable) on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and covariates age, gender, education and length of follow-up, and including each of the other variables of interest one at a time. *Only significant variables are displayed.

	Beta	Standard Error	OR	95% Wald Confidence Limits		P-value
GROUP	**1.91**	**0.488**	**6.78**	**2.60**	**17.65**	**0.0001**
MMS	-0.39	0.163	0.68	0.49	0.93	0.017
BCRTOT	0.32	0.115	1.38	1.10	1.73	0.005
BCR01	0.66	0.229	1.93	1.23	3.03	0.004
BCR05	0.88	0.379	2.42	1.15	5.08	0.020
HDS08	0.56	0.271	1.75	1.03	2.98	0.038
HDS11	0.79	0.312	2.20	1.20	4.06	0.011
BEH23	0.81	0.405	2.25	1.02	4.97	0.045
PDS	0.78	0.218	2.18	1.43	3.35	<0.0001
PRDD	-0.18	0.072	0.83	0.72	0.96	0.012
PRDI	-0.20	0.082	0.82	0.70	0.96	0.016
DESN	-0.19	0.084	0.83	0.70	0.98	0.026
WASDIGB	-0.33	0.129	0.72	0.56	0.93	0.011
WASV	-0.06	0.020	0.95	0.91	0.98	0.006

Table 6 Pearson residual analysis with stepwise variable selection. Model coefficient estimates from applying Pearson residual analysis to the data by regressing the binary outcome (decline vs. stable) on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and covariates age, gender, education and length of follow-up, and then linearly regressing the Pearson residuals on the other variables of interest.

Variable	Beta	Standard Error	P-value
Intercept	-0.66	0.27	0.018
GROUP	**0.47**	**0.21**	**0.026**
BCR01	0.17	0.10	0.107
HDS08	0.33	0.13	0.011
HDS11	0.33	0.13	0.012
HDS13	-0.26	0.13	0.051
BEH19	-0.41	0.19	0.031
BEH23	0.42	0.15	0.006
PRDI	-0.05	0.03	0.127

Table 7 Logistic regression with raw propensity score adjustment and stepwise variable selection. Model coefficient estimates from applying propensity score adjusted logistic regression to the data by regressing the binary outcome (decline vs. stable) on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and the raw propensity score for the group variable given covariates age, gender, education and length of follow-up, and on the other variables of interest.

	Beta	Standard Error	OR	95% Wald Confidence Limits		P-value
Intercept	-8.33	2.12				<0.0001
Propensity	5.59	2.49	266.786	2.012	>999.999	0.025
GROUP	**0.80**	**0.36**	**4.993**	**1.223**	**20.372**	**0.025**
BCR05	0.92	0.43	2.501	1.083	5.776	0.032
HDS11	0.73	0.35	2.069	1.044	4.103	0.037
PDS	0.73	0.20	2.082	1.395	3.107	<0.0001

Table 8 Logistic regression with logit propensity score adjustment and stepwise variable selection. Model coefficient estimates from applying propensity score adjusted logistic regression to the data by regressing the binary outcome (decline vs. stable) on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and the logit of the propensity score for the group variable given covariates age, gender, education and length of follow-up, and on the other variables of interest.

	Beta	Standard Error	OR	95% Wald Confidence Limits		P-value
Intercept	-5.16	1.00				<0.0001
Logit Propensity	0.92	0.42	2.52	1.10	5.76	0.029
GROUP	**0.81**	**0.36**	**5.02**	**1.23**	**20.50**	**0.025**
BCR05	0.91	0.43	2.48	1.08	5.72	0.033
HDS11	0.69	0.34	2.00	1.02	3.92	0.042
PDS	0.72	0.20	2.06	1.38	3.08	0.0004

Table 9 Residual logistic regression with stepwise variable selection. Model coefficient estimates from applying residual logistic regression to the data. The binary outcome (decline vs. stable) is first logistically regressed on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and covariates age, gender, education and length of follow-up, obtaining the estimated conditional log odds. Then a second logistic regression of the outcome on the other variables of interest is performed offset by the estimated conditional log odds.

	Beta	Standard Error	OR	95% Wald Confidence Limits		P-value
Intercept	-2.82	0.65				<0.0001
GROUP	**1.10**	**0.34**	**9.01**	**2.34**	**34.68**	**0.001**
HDS11	0.81	0.36	2.25	1.12	4.51	0.022
BEH19	-1.14	0.50	0.32	0.12	0.86	0.023
BEH23	1.28	0.44	3.60	1.51	8.62	0.004
PDS	0.72	0.23	2.05	1.32	3.20	0.002

Since the ADCC study is longitudinal with censoring occurring in the data, we also apply the more appropriate Cox PH models and AFT models to the data, with time to decline as the event of interest.

Table 10, Table 11, Table 12, and Table 13 respectively display results from traditional multiple regression, one-at-a-time regression, raw propensity score, and logit propensity score regression for the Cox PH model. Estimated hazard rates ranged from 4.7 to 6.1, with confidence interval widths ranging from approximately 9 to 20 units. Group was also significant for each method applied to the regression (P-values < 0.005). Other variables of interest found to be consistently significant among the different methods were HDS11, and PDS (P-values < 0.05).

Table 10 Traditional multiple Cox PH regression with stepwise variable selection. Model coefficient estimates from applying multiple Cox PH regression to the data by regressing the binary outcome (decline vs. stable) the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and covariates age, gender, education and length of follow-up, and all other variables of interest.

	Beta	Standard Error	HR	95% Confidence Limits		P-value
GROUP	**1.86**	**0.60**	**6.40**	**1.97**	**20.79**	**0.002**
HDS11	0.49	0.17	1.63	1.16	2.28	0.005
PDS	0.51	0.14	1.67	1.25	2.21	0.0004

Table 11 One-at-a-time Cox PH regression*. Model coefficient estimates from applying one-at-a-time Cox PH regression to the data by regressing the binary outcome (decline vs. stable) on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and covariates age, gender, education and length of follow-up, and including each of the other variables of interest one at a time. *Only significant variables are displayed.

	Beta	Standard Error	HR	95% Confidence Limits		P-value
GROUP	**1.56**	**0.43**	**4.77**	**2.07**	**11.00**	**0.0002**
BCRTOT	0.16	0.07	1.18	1.04	1.34	0.012
BCR01	0.40	0.14	1.50	1.14	1.97	0.004
HDS08	0.33	0.17	1.40	1.00	1.95	0.050
HDS11	0.46	0.16	1.59	1.16	2.18	0.004
PDS	0.52	0.12	1.69	1.34	2.14	<0.0001
PARD	-0.08	0.04	0.93	0.86	0.99	0.029
PRDI	-0.14	0.05	0.87	0.79	0.97	0.008
DESN	-0.17	0.05	0.84	0.76	0.94	0.002
WASDIGB	-0.17	0.08	0.85	0.73	0.99	0.032
WASV	-0.03	0.01	0.97	0.96	0.99	0.0004

Table 12 Cox PH regression with raw propensity score adjustment and stepwise variable selection. Model coefficient estimates from applying propensity score adjusted Cox PH regression to the data by regressing the binary outcome (decline vs. stable) on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and the raw propensity score for the group variable given covariates age, gender, education and length of follow-up, and on the other variables of interest.

	Beta	Standard Error	HR	95% Confidence Limits		P-value
propensity	3.50	1.53	33.06	1.66	659.55	0.022
GROUP	**1.77**	**0.60**	**5.88**	**1.80**	**19.19**	**0.003**
HDS08	0.38	0.18	1.46	1.03	2.07	0.033
HDS11	0.44	0.18	1.56	1.10	2.21	0.012
PDS	0.59	0.12	1.81	1.43	2.31	<.0001

Table 13 Cox PH regression with logit propensity score adjustment and stepwise variable selection. Model coefficient estimates from applying propensity score adjusted Cox PH regression to the data by regressing the binary outcome (decline vs. stable) on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and the logit of the propensity score for the group variable given covariates age, gender, education and length of follow-up, and on the other variables of interest.

	Beta	Standard Error	HR	95% Confidence Limits		P-value
logit_propensity	0.55	0.25	1.73	1.06	2.82	0.027
GROUP	**1.78**	**0.60**	**5.91**	**1.81**	**19.32**	**0.003**
HDS08	0.39	0.18	1.48	1.04	2.09	0.029
HDS11	0.43	0.18	1.54	1.09	2.18	0.015
PDS	0.59	0.12	1.80	1.42	2.29	<0.0001

Table 14, Table 15, Table 16, and Table 17 display results from traditional multiple regression, one-at-a-time regression, raw propensity score, and logit propensity score regression for the AFT model. Although group was found to be significant for the first two methods (P-values < 0.01), propensity score adjustment failed to find group to be significant (P-values = 0.07). Percent change in expected survival time for SCI over NCI ranged between -5.9 to -5.1. Common variables of interest that were found to significant were PARD and PRDI (P-values < 0.05).

Table 14 Traditional multiple AFT regression (no variable selection)*. Model coefficient estimates from applying multiple AFT regression to the data by regressing the binary outcome (decline vs. stable) the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and covariates age, gender, education and length of follow-up, and all other variables of interest. *Only significant variables are displayed.

	Beta	Standard Error	95% Confidence Limits		P-value
GROUP	**-0.59**	**0.21**	**-1.01**	**-0.17**	**0.006**
BCR03	0.34	0.12	0.11	0.57	0.003
HDS11	-0.46	0.20	-0.85	-0.06	0.023
BEH20	-0.58	0.19	-0.95	-0.21	0.002
BEH23	-0.58	0.18	-0.93	-0.22	0.001
PDS	0.48	0.23	0.03	0.94	0.036
PARD	0.08	0.03	0.03	0.13	0.002
PRDI	0.14	0.04	0.06	0.22	0.0005
DESN	0.09	0.04	0.01	0.16	0.020
WASDIGB	0.12	0.04	0.03	0.20	0.008

Table 15 One-at-a-time AFT regression*. Model coefficient estimates from applying one-at-a-time AFT regression to the data by regressing the binary outcome (decline vs. stable) on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and covariates age, gender, education and length of follow-up, and including each of the other variables of interest one at a time. *Only significant variables are displayed.

	Beta	Standard Error	95% Confidence Limits		P-value
GROUP	**-0.51**	**0.15**	**-0.80**	**-0.22**	**0.0005**
MMS	0.05	0.03	0.00	0.10	0.054
BCRTOT	-0.06	0.02	-0.11	-0.02	0.006
BCR01	-0.14	0.05	-0.24	-0.05	0.004
BCR05	-0.19	0.09	-0.35	-0.02	0.029
HDS08	-0.13	0.06	-0.25	-0.01	0.031
HDS11	-0.16	0.06	-0.27	-0.04	0.007
PDS	-0.19	0.04	-0.27	-0.11	<0.0001
PARD	0.03	0.01	0.00	0.05	0.035
PRDI	0.05	0.02	0.01	0.08	0.005
PRDD	0.05	0.02	0.01	0.08	0.004
DESN	0.06	0.02	0.03	0.10	0.001
WASDIGB	0.07	0.03	0.02	0.12	0.012
DSST	0.01	0.00	0.00	0.01	0.047
WASV	0.01	0.00	0.00	0.01	0.001

Table 16 AFT regression with raw propensity score (no variable selection)*. Model coefficient estimates from applying propensity score adjusted Cox PH regression to the data by regressing the binary outcome (decline vs. stable) on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and the raw propensity score for the group variable given covariates age, gender, education and length of follow-up, and on the other variables of interest. *Only significant variables are displayed.

	Beta	Standard Error	95% Confidence Limits		P-value
Intercept	-0.23	3.08	-6.27	5.80	0.940
propensity	-3.10	0.80	-4.66	-1.53	0.0001
GROUP	**-0.51**	**0.29**	**-1.07**	**0.05**	**0.073**
BCR03	0.36	0.15	0.08	0.65	0.013
BEH20	-0.63	0.27	-1.16	-0.11	0.018
BEH23	-0.59	0.26	-1.11	-0.07	0.025
PARD	0.08	0.03	0.01	0.15	0.017
PRDI	0.13	0.06	0.02	0.24	0.022

Table 17 AFT regression with logit propensity score (no variable selection)*. Model coefficient estimates from applying propensity score adjusted Cox PH regression to the data by regressing the binary outcome (decline vs. stable) on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and the logit of the propensity score for the group variable given covariates age, gender, education and length of follow-up, and on the other variables of interest. *Only significant variables are displayed.

	Beta	Standard Error	95% Confidence Limits		P-value
Intercept	-1.71	3.14	-7.86	4.45	0.587
logit_propensity	-0.42	0.13	-0.68	-0.16	0.001
GROUP	**-0.54**	**0.30**	**-1.12**	**0.04**	**0.069**
BCR03	0.32	0.15	0.03	0.61	0.032
BEH20	-0.62	0.28	-1.16	-0.07	0.026
BEH23	-0.53	0.27	-1.06	0.00	0.051
PARD	0.08	0.04	0.01	0.15	0.020
PRDI	0.13	0.06	0.01	0.24	0.028

It should be noted that no variable selection procedure was implemented in the AFT regressions for the tables above due to software limitations. Without variable selection, all variables of interests remained in the regression model, resulting in a model with 51 to 55 covariates (some variables had no data) fitted to 154 observations for the traditional multiple regression and the propensity score adjusted models. Hence the results for those methods are unreliable. Correlation analysis was run in an attempt to reduce the model by removing repetitive variables based on significant Pearson correlations with magnitude 0.50 or higher (P-values < 0.0005). Selected correlations are shown in Table 18.

Table 18 Selected Pearson correlations. Correlation analysis was run in an attempt to reduce the model by removing repetitive variables based on significant Pearson correlations with magnitude 0.50 or higher (P-values < 0.0005). Bold correlations indicate significance and magnitude 0.58 or higher.

	GROUP	BCRTOT	BCR01	BCR02	BCR03	BCR04
BCRTOT	**0.67**					
BCR01	0.47	**0.76**				
BCR02	**0.66**	**0.70**	0.39			
BCR03	0.34	**0.70**	0.38	0.36		
BCR04	0.24	0.46	0.16	0.13	0.24	
BCR05	**0.52**	**0.68**	0.32	0.48	0.30	0.30

	BEHTOT	BEH06	BEH15	BEH16	BEH18	BEH19
BEH06	0.30					
BEH15	0.19	**0.81**				
BEH16	**0.66**	0.05	0.05			
BEH18	0.32	-0.02	-0.01	**0.59**		
BEH19	**0.53**	0.10	-0.04	0.06	-0.04	
BEH20	**0.64**	0.03	-0.04	0.24	-0.05	0.35
BEH21	**0.58**	0.10	-0.02	0.22	-0.02	0.44
BEH23	**0.61**	0.11	0.01	0.19	0.15	0.27
HDS05	0.31	0.15	0.04	-0.01	-0.07	**0.61**
HDS20	0.36	**0.82**	**0.72**	0.06	-0.03	0.31

	PDS	PARI	PARD	PRDI
PARI	**-0.58**			
PARD	**-0.63**	**0.77**		
PRDI	**-0.72**	0.33	0.46	
PRDD	**-0.74**	0.29	0.44	**0.80**
DESN	**-0.69**	0.30	0.30	0.42
DSST	**-0.65**	0.26	0.30	0.31
WASV	**-0.59**	0.34	0.31	0.18

Table 19, Table 20, and Table 21 are the results of traditional multiple regression and propensity score adjustment on the AFT model with the reduced set of covariates. Although it is still questionable whether there was sufficient reduction in the set of covariates, we can see that the consistency in the estimated effect of group in the previous models is beginning to appear here, with significance appearing for the propensity adjusted models (P-values < 0.05). Aside from group, no other common variables are significant for all the methods. Perhaps with further model reduction, PDS would be significant for all possible methods.

Table 19 Traditional multiple AFT regression (reduced set of VOIs). Model coefficient estimates from applying multiple AFT regression to the data by regressing the binary outcome (decline vs. stable) the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and covariates age, gender, education and length of follow-up, and on a subset of the variables of interest. *Only significant variables and PDS are displayed.

	Beta	Standard Error	95% Confidence Limits		P-value
Intercept	2.73	1.37	0.05	5.41	0.046
GROUP	**-0.55**	**0.19**	**-0.91**	**-0.18**	**0.003**
BCR03	0.22	0.09	0.05	0.39	0.010
BEH18	0.87	0.42	0.05	1.69	0.038
BEH23	-0.25	0.10	-0.45	-0.06	0.010
PDS	-0.12	0.06	-0.25	0.01	0.061

Table 20 AFT regression with raw propensity score (reduced set of VOIs)*. Model coefficient estimates from applying propensity score adjusted AFT regression to the data by regressing the binary outcome (decline vs. stable) on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and the raw propensity score for the group variable given covariates age, gender, education and length of follow-up, and on a subset of the variables of interest. *Only significant variables and PDS are displayed.

	Beta	Standard Error	95% Confidence Limits		P-value
Intercept	5.26	1.70	1.92	8.59	0.002
propensity	-2.61	0.81	-4.20	-1.01	0.001
GROUP	**-0.58**	**0.23**	**-1.04**	**-0.13**	**0.012**
BCR04	-0.30	0.14	-0.58	-0.02	0.036
PDS	-0.20	0.08	-0.35	-0.05	0.011

Table 21 AFT regression with logit propensity score (reduced set of VOIs)*. Model coefficient estimates from applying propensity score adjusted AFT regression to the data by regressing the binary outcome (decline vs. stable) on the main group variable (subjective cognitive impairment vs. no subjective cognitive impairment) and the logit of the propensity score for the group variable given covariates age, gender, education and length of follow-up, and on a subset of the variables of interest. *Only significant variables and PDS are displayed.

	Beta	Standard Error	95% Confidence Limits		P-value
Intercept	3.91	1.73	0.52	7.29	0.024
logit_propensity	-0.38	0.13	-0.63	-0.12	0.004
GROUP	**-0.58**	**0.24**	**-1.05**	**-0.11**	**0.015**
BCR04	-0.29	0.15	-0.58	-0.01	0.043
PDS	-0.19	0.08	-0.35	-0.03	0.017

6 Partial correlation analysis

The goal is to discern the relationship between variable after removing the common variability between the two variables due to a third variable.

Consider a multivariate normal distribution $(X,Y,Z) \sim N(\mu,\Sigma)$ with

$$\mu = \begin{bmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{bmatrix}, \Sigma = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_{XX} & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_{XX} \end{pmatrix} = \begin{pmatrix} \sigma_{XX} & \rho_{XY}\sqrt{\sigma_{XX}\sigma_{YY}} & \rho_{XZ}\sqrt{\sigma_{XX}\sigma_{ZZ}} \\ \rho_{XY}\sqrt{\sigma_{XX}\sigma_{YY}} & \sigma_{YY} & \rho_{YZ}\sqrt{\sigma_{YY}\sigma_{ZZ}} \\ \rho_{XZ}\sqrt{\sigma_{XX}\sigma_{ZZ}} & \rho_{YZ}\sqrt{\sigma_{YY}\sigma_{ZZ}} & \sigma_{ZZ} \end{pmatrix}$$

For convenience, we partition the variance-covariance matrix.

$$\Sigma = \begin{bmatrix} \Sigma_{XY} & \Sigma_{XY.Z} \\ \Sigma_{Z.XY} & \Sigma_{ZZ} \end{bmatrix}, \Sigma'_{Z.XY} = \Sigma_{XY.Z} = \begin{pmatrix} \rho_{XZ}\sqrt{\sigma_{XX}\sigma_{ZZ}} \\ \rho_{YZ}\sqrt{\sigma_{YY}\sigma_{ZZ}} \end{pmatrix}$$

One way to measure the relationship between the two population variables *X* and *Y* after accounting for *Z* is to consider the conditional joint distribution of *X* and *Y* given *Z*. It can be shown that such a distribution can be derived from the multivariate distribution and expressed as $(X,Y \mid Z = z) \sim N(\bar{\mu},\bar{\Sigma})$ (Draper, et al. 1998, Kutner, et al. 2004) with

$$\bar{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} + \Sigma_{XY.Z}\Sigma_{ZZ}^{-1}(z-\mu_Z) = \begin{bmatrix} \mu_X + \dfrac{\sigma_{XZ}}{\sigma_{ZZ}}(z-\mu_Z) \\ \mu_X + \dfrac{\sigma_{XZ}}{\sigma_{ZZ}}(z-\mu_Z) \end{bmatrix} = \begin{bmatrix} \mu_X + \dfrac{\rho_{XZ}}{\sigma_{ZZ}^{3/2}\sqrt{\sigma_{XX}}}(z-\mu_Z) \\ \mu_X + \dfrac{\rho_{YZ}}{\sigma_{ZZ}^{3/2}\sqrt{\sigma_{YY}}}(z-\mu_Z) \end{bmatrix}$$

$$
\begin{aligned}
\bar{\Sigma} &= \Sigma_{XY} - \Sigma_{XY.Z}\Sigma_{ZZ}^{-1}\Sigma_{Z.XY} \\
&= \begin{bmatrix} \sigma_{XX} & \rho_{XY}\sqrt{\sigma_{XX}\sigma_{YY}} \\ \rho_{XY}\sqrt{\sigma_{XX}\sigma_{YY}} & \sigma_{YY} \end{bmatrix} - \frac{1}{\sigma_{ZZ}} \begin{bmatrix} \rho_{XZ}\sqrt{\sigma_{XX}\sigma_{ZZ}} \\ \rho_{YZ}\sqrt{\sigma_{YY}\sigma_{ZZ}} \end{bmatrix} \begin{bmatrix} \rho_{XZ}\sqrt{\sigma_{XX}\sigma_{ZZ}} & \rho_{YZ}\sqrt{\sigma_{YY}\sigma_{ZZ}} \end{bmatrix} \\
&= \begin{bmatrix} \sigma_{XX} & \rho_{XY}\sqrt{\sigma_{XX}\sigma_{YY}} \\ \rho_{XY}\sqrt{\sigma_{XX}\sigma_{YY}} & \sigma_{YY} \end{bmatrix} - \frac{1}{\sigma_{ZZ}} \begin{bmatrix} \rho_{XZ}^2\sigma_{XX}\sigma_{ZZ} & \rho_{XY}\rho_{YZ}\sigma_{ZZ}\sqrt{\sigma_{XX}\sigma_{YY}} \\ \rho_{XY}\rho_{YZ}\sigma_{ZZ}\sqrt{\sigma_{XX}\sigma_{YY}} & \rho_{XZ}^2\sigma_{XX}\sigma_{ZZ} \end{bmatrix} \\
&= \begin{bmatrix} \sigma_{XX} & \rho_{XY}\sqrt{\sigma_{XX}\sigma_{YY}} \\ \rho_{XY}\sqrt{\sigma_{XX}\sigma_{YY}} & \sigma_{YY} \end{bmatrix} - \begin{bmatrix} \rho_{XZ}^2\sigma_{XX} & \rho_{XZ}\rho_{YZ}\sqrt{\sigma_{XX}\sigma_{YY}} \\ \rho_{XZ}\rho_{YZ}\sqrt{\sigma_{XX}\sigma_{YY}} & \rho_{XZ}^2\sigma_{XX} \end{bmatrix} \\
&= \begin{bmatrix} \sigma_{XX}\left(1-\rho_{XZ}^2\right) & \left(\rho_{XY}-\rho_{XZ}\rho_{YZ}\right)\sqrt{\sigma_{XX}\sigma_{YY}} \\ \left(\rho_{XY}-\rho_{XZ}\rho_{YZ}\right)\sqrt{\sigma_{XX}\sigma_{YY}} & \sigma_{YY}\left(1-\rho_{YZ}^2\right) \end{bmatrix}
\end{aligned}
$$

Hence, the population conditional correlation is

$$
\rho_{XY|Z} = \frac{\left(\rho_{XY}-\rho_{XZ}\rho_{YZ}\right)\sqrt{\sigma_{XX}\sigma_{YY}}}{\sqrt{\sigma_{XX}\left(1-\rho_{XZ}^2\right)}\sqrt{\sigma_{YY}\left(1-\rho_{YZ}^2\right)}} = \frac{\rho_{XY}-\rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}}
$$

To say that the conditional correlation between *X* and *Y* is zero is to say that there is no linear relationship between *X* and *Y* after controlling for *Z*. Our goal is to be able to estimate and test such a measure.

6.1 Partial correlation

The partial correlation was formulated during the early nineteen century by (Yule 1897, 1907) and its distribution was derived by (Fisher 1924). Suppose there are a random sample of *n* observations for two variables *X* and *Y* and a vector of control variables *Z*. The partial correlation between *X* and *Y* controlling for *Z* is obtained by correlating the residual from regressing *X* and *Y* on *Z*. First, regress *X* on *Z* and *Y* on *Z* via least squares regression, and then calculate the correlation between the two residuals.

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 Z + \varepsilon_Y \rightarrow \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 Z \rightarrow e_Y = Y - \hat{Y} \\
X &= \gamma_0 + \gamma_1 Z + \varepsilon_X \rightarrow \hat{X} = \hat{\gamma}_0 + \hat{\gamma}_1 Z \rightarrow e_X = X - \hat{X}
\end{aligned}
$$

$$
\begin{aligned}
r_{xy(z)} &= \frac{\sum\left(e_X-\frac{\sum e_X}{n}\right)\left(e_Y-\frac{\sum e_Y}{n}\right)}{\sqrt{\sum\left(e_X-\frac{\sum e_X}{n}\right)^2}\sqrt{\sum\left(e_Y-\frac{\sum e_Y}{n}\right)^2}} = \frac{\sum(e_X e_Y)}{\sqrt{\sum(e_X)^2}\sqrt{\sum(e_Y)^2}} = \frac{\sum\left(X-\hat{X}\right)(e_Y)}{\sqrt{\sum\left(X-\hat{X}\right)e_X}\sqrt{\sum\left(Y-\hat{Y}\right)e_Y}} \\
&= \frac{\sum\left(Xe_Y-\left(\hat{\beta}_0+\hat{\beta}_1 Z\right)e_Y\right)}{\sqrt{\sum\left(Xe_X-\left(\hat{\beta}_0+\hat{\beta}_1 Z\right)e_X\right)}\sqrt{\sum\left(Ye_Y-\left(\hat{\gamma}_0+\hat{\gamma}_1 Z\right)e_Y\right)}} = \frac{\sum(Xe_Y)}{\sqrt{\sum(Xe_X)}\sqrt{\sum(Ye_Y)}} \\
&= \frac{\sum X\left(Y-\hat{\gamma}_0-\hat{\gamma}_1 Z\right)}{\sqrt{\sum X\left(X-\hat{\beta}_0-\hat{\beta}_1 Z\right)}\sqrt{\sum Y\left(Y-\hat{\gamma}_0-\hat{\gamma}_1 Z\right)}} = \frac{\sum X\left(Y-\left(\bar{Y}-\hat{\gamma}_1\bar{Z}\right)-\hat{\gamma}_1 Z\right)}{\sqrt{\sum X\left(X-\left(\bar{X}-\hat{\beta}_1\bar{Z}\right)-\hat{\beta}_1 Z\right)}\sqrt{\sum Y\left(Y-\left(\bar{Y}-\hat{\gamma}_1\bar{Z}\right)-\hat{\gamma}_1 Z\right)}} \\
&= \frac{\sum X\left(Y-\bar{Y}-\hat{\gamma}_1(Z-\bar{Z})\right)}{\sqrt{\sum X\left(X-\bar{X}-\hat{\beta}_1(Z-\bar{Z})\right)}\sqrt{\sum Y\left(Y-\bar{Y}-\hat{\gamma}_1(Z-\bar{Z})\right)}} = \frac{\sum\left[X(Y-\bar{Y})-\hat{\gamma}X(Z-\bar{Z})\right]}{\sqrt{\sum\left[X(X-\bar{X})-\hat{\beta}_1 X(Z-\bar{Z})\right]}\sqrt{\sum\left[Y(Y-\bar{Y})-\hat{\gamma}Y(Z-\bar{Z})\right]}}
\end{aligned}
$$

Recall that via least squares regression

$$
\hat{\beta}_1 = \frac{\sum(X-\bar{X})(Z-\bar{Z})}{\sum(Z-\bar{Z})^2} \text{ and } \hat{\gamma}_1 = \frac{\sum(Y-\bar{Y})(Z-\bar{Z})}{\sum(Z-\bar{Z})^2}
$$

Hence

$$
\begin{aligned}
r_{XY(Z)} &= \frac{\sum\left[X(Y-\bar{Y})-\hat{\gamma}X(Z-\bar{Z})\right]}{\sqrt{\sum\left[X(X-\bar{X})-\hat{\beta}_1 X(Z-\bar{Z})\right]}\sqrt{\sum\left[Y(Y-\bar{Y})-\hat{\gamma}Y(Z-\bar{Z})\right]}} \\
&= \frac{\sum\left[X\left(Y-\bar{Y}\right)-\frac{\sum(Y-\bar{Y})(Z-\bar{Z})}{\sum(Z-\bar{Z})^2}X\left(Z-\bar{Z}\right)\right]}{\sqrt{\sum\left[X\left(X-\bar{X}\right)-\frac{\sum(X-\bar{X})(Z-\bar{Z})}{\sum(Z-\bar{Z})^2}X\left(Z-\bar{Z}\right)\right]}\sqrt{\sum\left[Y\left(Y-\bar{Y}\right)-\frac{\sum(Y-\bar{Y})(Z-\bar{Z})}{\sum(Z-\bar{Z})^2}Y\left(Z-\bar{Z}\right)\right]}} \\
&= \frac{\frac{\sum X(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}}-\left[\frac{\sum(Y-\bar{Y})(Z-\bar{Z})}{\sqrt{\sum(Y-\bar{Y})^2}\sqrt{\sum(Z-\bar{Z})^2}}\right]\left[\frac{\sum X(Z-\bar{Z})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Z-\bar{Z})^2}}\right]}{\sqrt{\frac{\sum X(X-\bar{X})}{\sum(X-\bar{X})^2}-\left[\frac{\sum(X-\bar{X})(Z-\bar{Z})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Z-\bar{Z})^2}}\right]^2}\sqrt{\frac{\sum Y(Y-\bar{Y})}{\sum(Y-\bar{Y})^2}-\left[\frac{\sum(Y-\bar{Y})(Z-\bar{Z})}{\sqrt{\sum(Y-\bar{Y})^2}\sqrt{\sum(Z-\bar{Z})^2}}\right]^2}} \\
&= \frac{r_{xy}-r_{xz}r_{yz}}{\sqrt{1-r_{xz}^2}\sqrt{1-r_{yz}^2}}
\end{aligned}
$$

This expression is similar to the population conditional correlation. Note that the Pearson correlation is an estimate of the population correlation. Hence this partial correlation estimates the conditional correlation. We can test the null hypothesis that the partial correlation is zero via an *F*-test, with *k* being the number of variables being controlled.

$$
F = \frac{r_{xy(z)}^2}{1-r_{xy(z)}^2}(N-k-2) \sim F_{1,N-k-2}
$$

For the remaining partial correlations discussed below, Z is a single variable, not a vector of variables.

6.2 Partial phi coefficient, point-biserial and point-polyserial correlation

Suppose X, Y and Z are binary variables. The most common extension of Pearson correlation formula applied to such data is the phi coefficient. It is only applicable when these variables are assumed to categorize an underlying continuous normal distribution. Consider the following contingency table

		X	
		0	1
Y	0	n_{11}	n_{12}
	1	n_{21}	n_{22}

Direct calculation of the components used in the Pearson correlation coefficient results in

$$\sum x = \sum x^2 = n_{12} + n_{22}$$
$$\sum y = \sum y^2 = n_{21} + n_{22}$$
$$\sum xy = n_{22}$$

Plugging those expressions into the Pearson correlation formula and we obtain the phi coefficient.

$$\begin{aligned} r_{xy} &= \frac{N\sum xy - \sum x \sum y}{\sqrt{N\sum x^2 - \left(\sum x\right)^2}\sqrt{N\sum y^2 - \left(\sum y\right)^2}} \\ &= \frac{Nn_{22} - \left(n_{12}+n_{22}\right)\left(n_{21}+n_{22}\right)}{\sqrt{N\left(n_{12}+n_{22}\right) - \left(n_{12}+n_{22}\right)^2}\sqrt{N\left(n_{21}+n_{22}\right) - \left(n_{21}+n_{22}\right)^2}} \\ &= \frac{\left(n_{11}+n_{12}+n_{21}+n_{22}\right)n_{22} - \left(n_{12}n_{21}+n_{12}n_{22}+n_{22}n_{21}+n_{22}^2\right)}{\sqrt{\left(n_{12}+n_{22}\right)\left[N - \left(n_{12}+n_{22}\right)\right]}\sqrt{\left(n_{21}+n_{22}\right)\left[N - \left(n_{21}+n_{22}\right)\right]}} \\ &= \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{\left(n_{12}+n_{22}\right)\left(n_{11}+n_{21}\right)\left(n_{21}+n_{22}\right)\left(n_{11}+n_{21}\right)}} = \phi_{xy} \end{aligned}$$

If Y is continuous and X is dichotomous, coded 0, 1, then the data can be summarized by stratifying the Y variables by the X variables. Direct calculation of the components used in the Pearson correlation coefficient results in

$$
\begin{aligned}
&N = N_0 + N_1 \\
&\sum x = \sum x^2 = N_1 \\
&\sum y = \sum_{x=1} y + \sum_{x=0} y = Y_1 + Y_0 \\
&\sum xy = \sum_{x=1} y = Y_1
\end{aligned}
$$

The direct application of the Pearson's product-moment correlation will results in the point-biserial correlation.

$$
\begin{aligned}
r_{pbis} &= \frac{N\sum xy - \sum x \sum y}{\sqrt{N\sum x^2 - \left(\sum x\right)^2}\sqrt{N\sum y^2 - \left(\sum y\right)^2}} \\
&= \frac{NY_1 - N_1(Y_1 + Y_0)}{\sqrt{NN_1 - N_1^2}\sqrt{N\sum y^2 - \left(\sum y\right)^2}} = \frac{N_0 Y_1 - N_1 Y_0}{\sqrt{N_1}\sqrt{N_0}\sqrt{N\sum y^2 - \left(\sum y\right)^2}} \\
&= \frac{\bar{y}_{X=1} - \bar{y}_{X=0}}{Ns_y}\sqrt{N_0 N_1}
\end{aligned}
$$

The direct application of the Pearson's product-moment correlation to the scenario when X is multi-categorical will result in the point-polyserial correlation.

Because these correlations are direct applications of Pearson's product-moment correlation, statistical inferences on partial correlations constructed from such measures is carried out a similar way. If the data are from a multivariate normal distribution whose true partial correlation is equal to zero, then an F-test can be applied.

If X, Y, and Z are binary, then the partial correlation between X and Y controlling for Z can be defined as the partial phi coefficient.

$$
r_{xy(z)} = \phi_{xy(z)} = \frac{\phi_{xy} - \phi_{xz}\phi_{yz}}{\sqrt{1-\phi_{xz}^2}\sqrt{1-\phi_{yz}^2}}
$$

If X and Z are binary, but Y is continuous, then the partial correlation between X and Y controlling for Z can be defined as the partial mixed correlation.

$$
r_{mixed.p[xy(z)]} = \frac{r_{pbis[xy]} - \phi_{xz} r_{pbis[yz]}}{\sqrt{1-\phi_{xz}^2}\sqrt{1-r_{pbis[yz]}^2}}
$$

And we could test the null hypothesis that the population partial correlation is zero using the *F* test for the regular partial correlation.

$$F = \frac{r_{xy(z)}^2}{1 - r_{xy(z)}^2}(N - 1 - 2) \sim F_{1,N-1-2}$$

6.3 Partial tetrachoric, polychoric correlation

Suppose that in addition to *X* and *Y* being categorical variables, they each have underlying continuous distributions. The actual correlation between the underlying continuous distributions can be estimated using the tetrachoric correlation (Pearson, et al. 1900, Ekstrom 2008). The tetrachoric correlation was originally derived as a mathematical formula involving the tetrachoric series. However, recent developments have led to maximum likelihood methods that are more commonly used to estimate this correlation (Olsson 1979). Suppose *X* and *Y* are the observed binary variables on bivariate normal distribution of (*X**, *Y**) with thresholds:

$$X^* \to \{a_0, a_1, a_2\}, X = \begin{cases} 0, \text{ if } a_0 < X^* \le a_1 \\ 1, \text{ if } a_1 < X^* < a_2 \end{cases}$$

$$Y^* \to \{b_0, b_1, b_2\}, Y = \begin{cases} 0, \text{ if } b_0 < Y^* \le b_1 \\ 1, \text{ if } b_1 < Y^* < b_2 \end{cases}$$

$$a_0 = b_0 = -\infty$$

$$a_2 = b_2 = +\infty$$

The likelihood function to be maximized would be

$$L = C\prod_{i=1}^{2}\prod_{j=1}^{2}\pi_{ij}^{n_{ij}} \to l = \ln L = \ln C + \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} \ln \pi_{ij}$$

where C is a constant

$$\pi_{ij} = \Phi_2(a_i, b_j) - \Phi_2(a_{i-1}, b_j) - \Phi_2(a_i, b_{j-1}) + \Phi_2(a_{i-1}, b_{j-1})$$

$$\Phi_2(h, k; \rho) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \int_{-\infty}^{h}\int_{-\infty}^{k} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right] dxdy$$

Suppose X and Y are multi-categorical with r and s categories, respectively. Continuing with the assumption that there is an underlying continuous distribution, the extension of the tetrachoric correlation is the polychoric correlation. Maximum likelihood procedures have also been developed to estimate the polychoric correlation (Martinson, et al. 1972, Olsson 1979, Drasgow 1986). With categorization thresholds denoted as

$$X^* \to \{a_0, a_1, ..., a_s\}$$
$$Y^* \to \{b_0, b_1, ..., b_r\}$$
$$a_0 = b_0 = -\infty, a_s = b_r = +\infty$$

the likelihood function to be maximized is

$$L = C\prod_{i=1}^{s}\prod_{j=1}^{r}\pi_{ij}^{n_{ij}} \to l = \ln L = \ln C + \sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij} \ln \pi_{ij}$$

where C is a constant

$$\pi_{ij} = \Phi_2\left(a_i, b_j\right) - \Phi_2\left(a_{i-1}, b_j\right) - \Phi_2\left(a_i, b_{j-1}\right) + \Phi_2\left(a_{i-1,j-1}\right)$$

$$\Phi_2\left(h, k; \rho\right) = \frac{1}{2\pi\left(1-\rho^2\right)^{1/2}} \int_{-\infty}^{h}\int_{-\infty}^{k} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2\left(1-\rho^2\right)}\right] dxdy$$

There are two general procedures to maximize such likelihoods to obtain the correlation estimates. The full maximum likelihood method solves for thresholds and correlation simultaneous (Olsson 1979, Poon, et al. 1987). The two-step method (Martinson, et al. 1972) assumes thresholds are fixed and hence only need to solve for the correlation; thresholds are estimated by the cumulative marginal proportions of the variables. While the full maximum likelihood method is more accurate, the two-step method is computationally more efficient, since it removes the threshold parameters from the optimization problem.

These estimates can be used to estimate the partial correlation of the underlying continuous distribution by using them in the partial correlation expression.

$$r_{tet[xy(z)]} = \frac{r_{tet(xy)} - r_{tet(xz)} r_{tet(yz)}}{\sqrt{1 - r_{tet(xz)}^2}\sqrt{1 - r_{tet(yz)}^2}}$$

$$r_{pch[xy(z)]} = \frac{r_{polychoric(xy)} - r_{polychoric(xz)} r_{polychoric(yz)}}{\sqrt{1 - r_{polychoric(xz)}^2}\sqrt{1 - r_{polychoric(yz)}^2}}$$

However, statistical inference on these measures have not been derived. If we construct an analogous *F* statistic as

$$F = \frac{r^2_{tet[xy(z)]}}{1 - r^2_{tet[xy(z)]}}(N - 1 - 2)$$

$$F = \frac{r^2_{pch[xy(z)]}}{1 - r^2_{pch[xy(z)]}}(N - 1 - 2)$$

We can use simulations to see how these test statistics would performing assuming they are approximately *F* distributed with degrees of freedom 1 and *N*-1-2.

6.4 Biserial, polyserial correlation

Suppose *X* is a binary variable representing an underlying continuous distribution, but *Y* is a continuous variable. In this case, the proper correlation between *Y* and the underlying distribution of *X* is the biserial correlation. If *X* is multi-categorical then the extension of the biserial correlation is the polyserial correlation. The estimation of the polyserial correlation via maximum likelihood is shown here (Olsson, et al. 1982, Drasgow 1986); application to biserial correlation is the same since the biserial correlation is only a special case of the polyserial correlation.

Let *X* have *r* categories, defined by thresholds on a normally distributed variable *X**.

$$X^* \to \{a_0, a_1, ..., a_r\}$$
$$a_0 = -\infty, a_s = +\infty$$

Then the likelihood function to be maximized is

$$L = \prod_{i=1}^{N} p(x_i, y_i) = \prod_{i=1}^{N} p(y_i)\, p(x_i \mid y_i)$$

$$l = \ln L = \sum_{i=1}^{N} \ln p(y_i) + \sum_{i=1}^{N} \ln p(x_i \mid y_i)$$

$$p(y_i) = \tfrac{1}{\sqrt{2\pi}\sigma} \exp\left[-\tfrac{1}{2}\left(\tfrac{y_i - \mu}{\sigma}\right)^2 \right]$$

$$p(x_i \mid y_i) = \Phi(a_i^*) - \Phi(a_{i-1}^*), a_i^* = \tfrac{a_i - \rho\left(\frac{y_i - \mu}{\sigma}\right)}{\sqrt{1-\rho^2}}$$

As in the case of the polychoric correlation, the polyserial correlation can be estimated either by using the full maximum likelihood, estimating all parameters simultaneously, or by a two-step procedure in which the thresholds are estimated by observed cumulative marginal proportions of *X*.

In the case of mixed data, the proper marginal bivariate correlations can be used in the partial correlation expression. For example, let *X* and *Z* are multi-categorical manifestations of continuous variables *X** and *Z** respectively. Furthermore *Y* is continuous. The partial correlation of *X** and *Y*, controlling *Z** would be

$$r_{XY(Z)} = \frac{r_{polyserial[xy]} - r_{polychoric[xz]} r_{polyserial[yz]}}{\sqrt{1 - r^2_{polychoric[xz]}}\sqrt{1 - r^2_{polyserial[yz]}}}$$

6.5 Controlling for more than one variable

For the above definitions of the partial correlation derived from the basic expression

$$r_{xy(z)} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

Only one variable can be controlled; *Z* cannot be a vector of variables. In order to control for more variables, higher order definitions of the partial correlation must be used (Blalock 1972, Wherry 1984). For example, the partial correlation of *X* and *Y* controlling for the two variables Z_1 and Z_2 is

$$r_{XY(Z_2 Z_1)} = \frac{r_{XY(Z_2)} - r_{XZ_1(Z_2)} r_{YZ_1(Z_2)}}{\sqrt{1 - r^2_{XZ_1(Z_2)}}\sqrt{1 - r^2_{YZ_1(Z_2)}}}$$

The process of controlling for more variables is iterative; as the number of variables that need to be controlled for increases, the computational labor also increases.

For continuous variables, the partial correlation is defined as the correlation of residuals. Hence, all variables that need to be controlled for can be included in the regression step.

6.6 Partial phi coefficient for multi-categorical data

We use higher order definitions of partial correlation to extend the partial phi coefficient to multi-categorical data. Suppose *X*, *Y*, *Z* are trichotomous data coded with dummy variables $X^{(1)}, X^{(2)}, Y^{(1)}, Y^{(2)}, Z^{(1)}, Z^{(2)}$. Then we propose that that partial phi coefficient be the maximum partial phi coefficient out of all pairwise partial phi coefficients for the dummy variables of *X* and *Y*, controlling for the dummy variable of *Z*.

$$\phi_{XY(Z)} = \max\left\{\phi_{X^{(i)}Y^{(j)}\left(Z^{(1)}Z^{(2)}\right)}\right\}$$

6.7 Equivalence of Partial and Conditional Correlation

Unfortunately, the partial correlation is not always equivalent to the conditional correlation (Baba, et al. 2004). In probability theory, we know that although when two variables are independent the correlation must equal zero, the converse is not true except in the case of normally distributed variables. Similarly, rarely is the partial correlation equal to the conditional correlation except for the known case of the normal distribution.

Baba, et al. (2005) give a sufficient condition which when satisfied implies that the partial correlation and the conditional correlation is equivalent. They also derived two classes of distributions that satisfy the condition and hence have equal partial and conditional correlation. We present their findings here and give the expression of the conditional correlation for binary data based on their theorems.

6.7.1 Condition C

Let M be a subset of the index set $\{1,2,...,m\}$ of a set of random variables $Y=(Y_1,...,Y_m)$, where $m \geq 3$ and $|M| \geq 2$, and let M^c be its non-empty complement. Partition the variance-covariance matrix of Y into a 2x2 block matrix

$$\mathrm{var}(Y)=\begin{pmatrix} \Sigma_M & \Sigma_{MM^c} \\ \Sigma_{M^cM} & \Sigma_{M^c} \end{pmatrix}.$$

The partial covariance of a set of components Y_M given Y_{M^c} is

$$\mathrm{var}(Y_M ; Y_{M^c}) = \Sigma_M - \Sigma_{MM^c}\Sigma_{M^c}^{-1}\Sigma_{M^cM}.$$

This is the variance-covariance matrix of $Y_M - \hat{Y}_M$ where $\hat{Y}_M$ is the least squares linear estimate of Y_M by Y_{M^c}. The correlation matrix derived from the partial covariance of Y_M given Y_{M^c} is the partial correlation matrix of Y_M given Y_{M^c}. The following condition, called Condition C, if satisfied is sufficient for the partial correlation to be equal to the conditional correlation.

Condition C

$E\left[Y_M \mid Y_{M^c}\right] = a + BY_{M^c}$ for constant vector a and a constant matrix B.

The conditional correlation matrix $cor\left(Y_M \mid Y_{M^c}\right)$ is independent of Y_{M^c}.

Baba provided further details on how the natural exponential family satisfies Condition C by introducing "partial sums *Y* of *X*". Let $X = \left(X_1, ..., X_n\right)$ be a random variable with $n \geq 3$ fixed. Partition the index set $\{1, ..., n\}$ into m parts $L_1, ..., L_m$ where $|L_j| = v_j > 0$ $\left(\sum_{j=1}^{m} v_j = n\right)$ and define $Y_j = \sum_{i=1}^{n} I\left[i \in L_j\right] X_i$.

Lemma

Assume that *X* has the following conditional moments given $T = \sum_{j=1}^{n} X_i = t$

$$E\left[X_j \mid T = t\right] = \frac{t}{n}$$
$$\operatorname{var}\left(X_j \mid T = t\right) = \sigma_t^2$$
$$\operatorname{cov}\left(X_i, X_j \mid T = t\right) = \kappa_t$$
$$(i \neq j = 1, ..., n)$$

The conditional expectation, variance-covariance matrix and correlations of partial sums *Y* of *X* given *T*=*t* are

$$E\left[Y \mid T = t\right] = t\xi$$
$$\operatorname{var}\left(Y \mid T = t\right) = -n^2 \kappa_t \left(diag\left(\xi\right) - \xi\xi'\right)$$
$$cor\left(Y_i, Y_j \mid T = t\right) = -\sqrt{\frac{\xi_i \xi_j}{\left(1 - \xi_i\right)\left(1 - \xi_j\right)}}$$
$$\xi = \left(\xi_1, ..., \xi_m\right)'$$
$$\xi_j = \frac{v_j}{n} \quad \left(j = 1, ..., m\right)$$

Theorem

Let $Y=\left(Y_M, Y_{M^c}\right)$ be a partition of partial sums Y of X. Suppose $T=\sum_{j=1}^{m} Y_j$ and Y_{M^c} are given and the first and second order conditional moments of $\left\{X_i; i \in \bigcup_{j \in M} L_j\right\}$, the original components of Y_M, are all the same for i. Let $y_* = \sum_{i \in M^c} y_i$ and $v_* = \sum_{i \in M^c} v_i$. Then the conditional expectation, variance-covariance matrix and correlations of partial sums Y_M of X given $T=t$ and $Y_{M^c} = y_{M^c}$ are

$$E\left[Y_M \mid T=t, Y_{M^c}=y_{M^c}\right]=\left(t-y_*\right)\tilde{\xi}_M$$

$$\operatorname{var}\left(Y_M \mid T=t, Y_{M^c}=y_{M^c}\right)=-\left(n-v_*\right)^2 \kappa_{t, y_{M^c}}\left(diag\left(\tilde{\xi}_M\right)-\tilde{\xi}_M \tilde{\xi}_M'\right)$$

$$\tilde{\xi}_M=\frac{v_M}{n-v_*},\ v_M=\left(v_j; j \in M\right)$$

$$\kappa_{t, y_{M^c}}=\operatorname{cov}\left(Y_i, Y_j \mid T=t, Y_{M^c}=y_{M^c}\right)$$

Thus $\left(Y_M, Y_{M^c} \mid T\right)$ satisfies Condition C.

The above can be applied to the natural exponential family by taking a random sample from the distribution under the condition that the sum of the observations is given.

Let $X=\left(X_1, \ldots, X_n\right)$ be a random sample from

$$p(x, \theta)=a(x) \exp\left(\theta x-\psi(\theta)\right),\ \theta \in \Theta \subset \Re .$$

This is the probability density function of the univariate natural exponential family with cumulant function $\psi(\theta)$. Since it can be specified by the mean $\mu(\theta)=\psi'(\theta)$, it is denoted as $NEF\left(\mu(\theta)\right)$.

$T=\sum_{j=1}^{n} X_j$ is a sufficient statistic and is $NEF\left(n\mu(\theta)\right)$ with density

$$p(t; \theta)=b(t; n) \exp\left(\theta t-n\psi(\theta)\right)$$

where $b(t; n)$ is the n-convolution of $a(t)$. Let $Y = (Y_1,...,Y_m)$ be independent and $Y_j \sim NEF(v_j \mu(\theta))$. $T = \sum_{j=1}^{m} Y_j$ is a sufficient statistic, and $T \sim NEF(v\mu(\theta))$, $v = \sum_{j=1}^{m} v_j$. The conditional density of Y given $T=t$ is

$$\frac{\prod_{j=1}^{m} b(y_j; v_j)}{b(t; v)}$$

By the Lemma above, $E[Y_j \mid t] = \frac{t v_j}{v}$ and the correlations of $(Y_1,...,Y_m) \mid t$ is given by

$$cor(Y_i, Y_j \mid T = t) = -\sqrt{\frac{\xi_i \xi_j}{(1-\xi_i)(1-\xi_j)}}$$

with $\xi = \left(\frac{v_1}{v},...,\frac{v_m}{v}\right)$.

Let $M \bigcup M_c$ be a partition of $\{1,...,m\}$. The conditional density of $Y_M = (Y_j, j \in M)$ given $T=t$ and $Y_{M^c} = y_{M^c}$ is

$$\frac{\prod_{j \in M} b(y_j; v_j)}{b(t - y_*; v - v_*)}$$
$$y_* = \sum_{j \in M^c} y_j$$
$$v_* = \sum_{j \in M^c} v_j$$

When the variance function, $V(\mu) = \psi''(\theta)$, is quadratic, then we have the special class of natural exponential family with quadratic variance function $NEF - QVF(\mu(\theta))$. In this case, the variance –covariance matrix can be obtained explicitly.

Proposition

Assume that the variance function of $NEF(\mu(\theta))$ is $V(\mu) = v_0 + v_1\mu + v_2\mu^2$. The variance-covariance matrix of $Y = (Y_1,...,Y_m)$, $Y_j \sim NEF - QVF(v_j \mu(\theta))$, given $\sum_{j=1}^{m} Y_j = t$ is

$$\operatorname{var}(Y \mid t) = c(t,v)\left(diag(\xi) - \xi\xi'\right)$$
$$v = \sum_{j=1}^{m} v_j$$
$$\xi = (\xi_1, ..., \xi_m)$$
$$\xi_j = \tfrac{v_j}{v},\ j = 1, ..., m$$
$$c(t,v) = \tfrac{v^2}{v+v_2} V\left(\tfrac{t}{v}\right)$$

6.7.2 Derivations for Binary Data

Using the theory from 6.3.1 and distributions stated explicitly in the Baba paper, we present a suitable distribution where the conditional correlation is equal to the partial correlation. Suppose that $X = (X_1, ..., X_n)$ is a random sample from a Bernoulli distribution with mean $\mu = \pi$ and variance $\operatorname{var}(X) = \mu(1-\mu) = \pi(1-\pi)$. Note that the probability mass function of *X* can be written as

$$\begin{aligned} p(x) &= \pi^x (1-\pi)^{1-x} \\ &= \exp\left[x \ln \pi + (1-x) \ln(1-\pi) \right] \\ &= \exp\left[x \ln \pi - x \ln(1-\pi) + \ln(1-\pi) \right] \\ &= \exp\left[x \ln\left(\tfrac{\pi}{1-\pi}\right) - \left(-\ln(1-\pi)\right) \right] \end{aligned}$$

Hence *X* belongs to the natural exponential family with quadratic variance function

$$\theta = \ln\left(\tfrac{\pi}{1-\pi}\right)$$
$$\psi(\theta) = -\ln(1-\pi) = \ln\left(1+\exp(\theta)\right)$$
$$\mu(\theta) = \psi'(\theta) = \tfrac{\exp(\theta)}{1+\exp(\theta)} = \tfrac{\frac{\pi}{1-\pi}}{1+\frac{\pi}{1-\pi}} = \pi$$
$$V(\theta) = \tfrac{(1+\exp(\theta))\exp(\theta) - \exp(2\theta)}{\left[1+\exp(\theta)\right]^2} = \tfrac{\exp(\theta)}{\left[1+\exp(\theta)\right]^2} = \tfrac{\frac{\pi}{1-\pi}}{\frac{1}{(1-\pi)^2}} = \pi(1-\pi) = \pi - \pi^2 = \mu - \mu^2$$

Suppose that $n=m$ such that $Y_j = X_j, v_j = 1, v = \sum_{j=1}^{m} v_j = m$. Then $Y_j \sim NEF - QVF\left(\mu(\theta)\right)$ and $T = \sum_{j=1}^{m} Y_j$ is a sufficient statistic with $T \sim NEF - QVF\left(m\mu(\theta)\right)$. Then distribution of Y given $T = \sum_{j=1}^{m} Y_j = t$ is multivariate hypergeometric with density distribution

$$\frac{\prod_{j=1}^{m} b\left(y_j; v_j\right)}{b(t; v)} = \frac{\prod_{j=1}^{m} \binom{v_j}{y_j}}{\binom{v}{t}} = \frac{\prod_{j=1}^{m} \binom{1}{y_j}}{\binom{n}{t}} = \frac{1}{\binom{n}{t}}$$

With mean and variance-covariance matrix

$$\begin{aligned}
E\left[Y_j \mid t\right] &= \frac{tv_j}{v} = \frac{t}{m} \\
\operatorname{var}\left(Y \mid t\right) &= \frac{v^2}{v+(-1)} V\left(\frac{t}{v}\right)\left(diag\left(\xi\right) - \xi\xi'\right) \\
&= \frac{m^2}{m-1}\left[\frac{t}{m} - \frac{t^2}{m^2}\right]\left(diag\left(\xi\right) - \xi\xi'\right) \\
&= \frac{tm - t^2}{m-1}\left(diag\left(\xi\right) - \xi\xi'\right) \\
&= \frac{t(m-t)}{m-1}\left(diag\left(\xi\right) - \xi\xi'\right)
\end{aligned}$$

Where $\xi = \left(\frac{v_1}{v}, ..., \frac{v_m}{v}\right) = \left(\frac{1}{m}, ..., \frac{1}{m}\right)$. Hence

$$\operatorname{var}\left(Y \mid t\right) = \frac{t(m-t)}{m-1}\begin{bmatrix} \frac{1}{m} - \frac{1}{m^2} & -\frac{1}{m^2} & \cdots & -\frac{1}{m^2} \\ -\frac{1}{m^2} & \frac{1}{m} - \frac{1}{m^2} & \cdots & -\frac{1}{m^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{m^2} & -\frac{1}{m^2} & \cdots & \frac{1}{m} - \frac{1}{m^2} \end{bmatrix}$$

And the correlation for $\left(Y_1, ..., Y_m\right) \mid T$ is $-\frac{1}{m-1}$.

Now consider a partition of $\{1, ..., m\}, M \bigcup M_c$. The conditional density of $Y_M = \left(Y_j, j \in M\right)$ given T=t and $Y_{M^c} = y_{M^c}$ is still hypergeometric

$$\frac{\prod_{j\in M} b(y_j; v_j)}{b(t-y_*; v-v_*)} = \frac{\prod_{j\in M}\binom{v_j}{y_j}}{\binom{v-v_*}{t-y_*}} = \frac{1}{\binom{m-v_*}{t-y_*}}$$

$$y_* = \sum_{j\in M^c} y_j$$

$$v_* = \sum_{j\in M^c} v_j$$

with mean and variance covariance matrix

$$E\left[Y_j; j\in M \mid T=t, Y_{M^c}=y_{M^c}\right] = \frac{t-y_*}{m-v_*} = \frac{t}{m-v_*} - \frac{1}{m-v_*}y_*$$

$$\operatorname{var}\left(Y_M \mid T=t, Y_{M^c}=y_{M^c}\right) = \frac{(t-y_*)\left((m-v_*)-(t-y_*)\right)}{m-v_*-1}\left(diag\left(\tilde{\xi}_M\right) - \tilde{\xi}_M\tilde{\xi}_M'\right)$$

$\tilde{\xi}_M = \left(\xi_j; j\in M\right), \xi_j = \frac{1}{m-v_*}$. Hence

$$\operatorname{var}\left(Y_M \mid T=t, Y_{M^c}=y_{M^c}\right) = \frac{(t-y_*)\left((m-v_*)-(t-y_*)\right)}{m-v_*-1}\begin{bmatrix} \frac{1}{m-v_*}-\frac{1}{(m-v_*)^2} & -\frac{1}{(m-v_*)^2} & \cdots & -\frac{1}{(m-v_*)^2} \\ -\frac{1}{(m-v_*)^2} & \frac{1}{m-v_*}-\frac{1}{(m-v_*)^2} & \cdots & -\frac{1}{(m-v_*)^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{(m-v_*)^2} & -\frac{1}{(m-v_*)^2} & \cdots & \frac{1}{m-v_*}-\frac{1}{(m-v_*)^2} \end{bmatrix}$$

And the conditional correlation for $\left(Y_M \mid T, Y_{M^c}\right)$ is $-\frac{1}{m-v_*-1}$. Thus, $\left(Y_M, Y_{M^c} \mid T\right)$ satisfies Condition C and the partial correlation and the conditional correlation is equivalent.

6.7.3 Comments

Baba and Sibuya gave a sufficient condition which when satisfied implies that the partial correlation and the conditional correlation is equivalent. They also proved that the class of NEF, when conditioned on the sufficient statistic, satisfies the condition and thus their partial and conditional correlations are equivalent and the exact expression for the partial correlation is known.

However it requires fixing the value of the sufficient statistic, which is likely to be unreasonable for most real world data. Regardless, since the expression for the partial correlation is unknown to us in the situation when the sufficient statistic is not fixed, we suggest a simulation should be done based on Baba and Sibuya's assumption so as to explore the statistical properties of our sample partial correlation.

We believe that applying a partial correlation measure to other distributions could still be informative about the relationship between two variables after controlling for an outside variable due to its intuitive interpretation, even if the data are not normal and partial correlation is not equal to conditional correlation. The residuals from any regression represent the variance in the data that cannot be explained by the variables in the model. To take the residuals of two different variables regressed on the same set of covariates and then correlate them would provide the information how the two variables co-vary, independent of the predictor variables.

7 New partial correlation for categorical and mixed data

The problem with the partial phi coefficient is that it can only be applied to binary variables. The partial tetrachoric correlation, and its general form, the polychoric correlation, are good alternatives for non-dichotomous variables, but would require further assumptions about the underlying multivariate distribution. Furthermore, these measures are unable to control for more than one variable. We introduce a recently proposed partial correlation measure for categorical data which can be applied to multi-categorical data and easily control for more than one variable. In addition, it can also measure the partial correlation between categorical and continuous variables. Chen (2011) proposed obtaining a "partial correlation" for categorical data in the spirit of the mechanical way in which partial correlation is obtained for continuous variables using the residuals from the regressions.

7.1 Binary Case

Suppose we have binary variables *X*, *Y*, and *Z*. We can perform logistic regression with *X* and *Y* as outcomes and *Z* as a dependent variable to obtain the Pearson residuals.

$$\log\left[\frac{\pi_x}{1-\pi_x}\right] = \beta_0 + \beta_1 Z \rightarrow \hat{\pi}_x = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z\right)} \rightarrow \hat{r}_x = \frac{x-\hat{\pi}_x}{\hat{\pi}_x(1-\hat{\pi}_x)}$$

$$\log\left[\frac{\pi_y}{1-\pi_y}\right] = \gamma_0 + \gamma_1 Z \rightarrow \hat{\pi}_y = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z\right)}{1+\exp\left(\hat{\gamma}_0 + \hat{\gamma}_1 z\right)} \rightarrow \hat{r}_y = \frac{y-\hat{\pi}_y}{\hat{\pi}_y\left(1-\hat{\pi}_y\right)}$$

For large sample sizes, the Pearson residual is normally distribution with mean zero and variance equal to one. Chen took advantage of the asymptotic nature of the Pearson residual to develop a new novel partial correlation measure.

$$r_{xy(z)} = \frac{\sum\left(\hat{r}_x - \frac{\sum \hat{r}_x}{N}\right)\left(\hat{r}_y - \frac{\sum \hat{r}_y}{N}\right)}{\sqrt{\sum\left(\hat{r}_x - \frac{\sum \hat{r}_x}{N}\right)^2}\sqrt{\sum\left(\hat{r}_y - \frac{\sum \hat{r}_y}{N}\right)^2}}$$

The regular Pearson correlation test is applied to test the null hypothesis that the partial correlation is equal to zero against the alternative that it is nonzero.

$$\frac{r_{xy(z)}^2}{1-r_{xy(z)}^2}(N-2) \sim F_{1,N-2}$$

Whether or not this is the true null distribution will be explored by simulation study.

7.2 Extension to multi-categorical and mixed data via canonical correlations

Chen also proposed a direct extension of this new novel method to categorical data with more than two categories. We will demonstrate the method with trichotomous variables, but the application to more categories is similar. Suppose that *X*, *Y* and *Z* are trichotomous variables, with Z coded as

$$Z^{(1)} = \begin{cases} 1 \text{ if Z is of category 1} \\ 0 \text{ otherwise} \end{cases}$$
$$Z^{(2)} = \begin{cases} 1 \text{ if Z is of category 2} \\ 0 \text{ otherwise} \end{cases}$$

Then one could perform multinomial logistic regression with *X* and *Y* as the dependent variables and *Z* as the predictor.

$$\log\left[\frac{\pi_{x1}}{\pi_{x0}}\right] = \beta_{10} + \beta_{11}Z^{(1)} + \beta_{12}Z^{(2)}, \log\left[\frac{\pi_{x2}}{\pi_{x0}}\right] = \beta_{20} + \beta_{21}Z^{(1)} + \beta_{22}Z^{(2)}$$
$$\log\left[\frac{\pi_{y1}}{\pi_{y0}}\right] = \gamma_{10} + \gamma_{11}Z^{(1)} + \gamma_{12}Z^{(2)}, \log\left[\frac{\pi_{y2}}{\pi_{y0}}\right] = \gamma_{20} + \gamma_{21}Z^{(1)} + \gamma_{12}Z^{(2)}$$
$$\pi_{x0} = P(X=0), \pi_{x1} = P(X=1), \pi_{x2} = P(X=2), \sum_{j=0}^{3}\pi_{xi} = 1$$
$$\pi_{y0} = P(Y=0), \pi_{y1} = P(Y=1), \pi_{y2} = P(Y=2), \sum_{j=0}^{3}\pi_{yi} = 1$$

$$\hat{\pi}_{x1} = \frac{\exp\left(\hat{\beta}_{10}+\hat{\beta}_{11}Z^{(1)}+\hat{\beta}_{12}Z^{(2)}\right)}{1+\exp\left(\hat{\beta}_{10}+\hat{\beta}_{11}Z^{(1)}+\hat{\beta}_{12}Z^{(2)}\right)+\exp\left(\beta_{20}+\hat{\beta}_{21}Z^{(1)}+\hat{\beta}_{22}Z^{(2)}\right)}, \hat{\pi}_{x2} = \frac{\exp\left(\hat{\beta}_{20}+\hat{\beta}_{21}Z^{(1)}+\hat{\beta}_{22}Z^{(2)}\right)}{1+\exp\left(\hat{\beta}_{10}+\hat{\beta}_{11}Z^{(1)}+\hat{\beta}_{12}Z^{(2)}\right)+\exp\left(\beta_{20}+\beta_{21}Z^{(1)}+\beta_{22}Z^{(2)}\right)}$$
$$\hat{\pi}_{y1} = \frac{\exp\left(\hat{\gamma}_{10}+\hat{\gamma}_{11}Z^{(1)}+\hat{\gamma}_{12}Z^{(2)}\right)}{1+\exp\left(\hat{\gamma}_{10}+\hat{\gamma}_{11}Z^{(1)}+\hat{\gamma}_{12}Z^{(2)}\right)+\exp\left(\hat{\gamma}_{20}+\hat{\gamma}_{21}Z^{(1)}+\hat{\gamma}_{12}Z^{(2)}\right)}, \hat{\pi}_{y2} = \frac{\exp\left(\hat{\gamma}_{20}+\hat{\gamma}_{21}Z^{(1)}+\hat{\gamma}_{22}Z^{(2)}\right)}{1+\exp\left(\hat{\gamma}_{10}+\hat{\gamma}_{11}Z^{(1)}+\hat{\gamma}_{12}Z^{(2)}\right)+\exp\left(\hat{\gamma}_{20}+\hat{\gamma}_{21}Z^{(1)}+\hat{\gamma}_{12}Z^{(2)}\right)}$$

$$\hat{r}_{x1} = \frac{x_1-\hat{\pi}_{x1}}{\hat{\pi}_{x1}(1-\hat{\pi}_{x1})},\ \hat{r}_{x2} = \frac{x_2-\hat{\pi}_{x2}}{\hat{\pi}_{x2}(1-\hat{\pi}_{x2})}$$
$$\hat{r}_{y1} = \frac{y_1-\hat{\pi}_{y1}}{\hat{\pi}_{y1}(1-\hat{\pi}_{y1})},\ \hat{r}_{y2} = \frac{y_2-\hat{\pi}_{y2}}{\hat{\pi}_{y2}(1-\hat{\pi}_{y2})}$$

Now we have two sets of two residuals for each of *X* and *Y*. Furthermore, each set has a bivariate normal distribution (Seber, et al. 2000). We can extend the concept of partial correlation via regression residuals using canonical correlations.

Consider all possible linear combinations of each residual set:

$$U = a_1\hat{r}_{x1} + a_2\hat{r}_{x2}$$
$$V = b_1\hat{r}_{y1} + b_2\hat{r}_{y2}$$

The weights a_1, a_2, and b_1, b_2, are chosen such that the correlation between *U* and *V* is maximized and the variances of *U* and *V* are equal to one. *U* and *V* are then known as canonical variates. The maximum correlation possible from such weights is the first canonical correlation.

The partial correlation of multi-categorical variables *X* and *Y* controlling for *Z* will be defined to be the first canonical correlation between the corresponding Pearson residuals obtained from multinomial regressions. It is important to note that the correlation between two individual variables is a special case of the canonical correlation. When using canonical correlation to correlate two sets of variables, the first canonical correlation is always greater than or equal to the absolute value of the correlation between any two variables taken from each set (Johnson, et al. 2002). In the case that each set has only one variable, or, in the context of partial correlation, each regression results in only one residual, the canonical correlation between the two residuals is equal to the absolute value of the correlation between the two residuals. Hence the partial correlation via regression residuals proposed for binary variables is a special case of the extension presented here.

We can test the partial correlation using the Bartlett test (Bartlett 1941). Let *p* be the number of categories in the first variables and *q* be the number of categories in the second variable. Letting ρ_i^* be the i-th canonical correlation, for null hypothesis $H_0 : \boldsymbol{\Sigma}_{12} = \underset{(p\times q)}{\mathbf{0}} \left(\rho_1^* = \rho_2^* = \cdots = \rho_p^* = 0\right)$ against alternative hypothesis $H_1 : \boldsymbol{\Sigma}_{12} \neq \underset{(p\times q)}{\mathbf{0}} \left(\rho_i^* \neq 0 \text{ for some } i = 1,..., p\right)$, we use a test statistic that has Chi-square distribution with *pq* degrees of freedom when *n* is large and the null hypothesis is true.

$$-2\ln\Lambda = -\left(n - 1 - \frac{1}{2}(p + q + 1)\right)\ln\prod_{i=1}^{p}\left(1 - \hat{\rho}_i^{*2}\right) \approx \chi_{pq}^2$$

This method can easily be extended to mixed data. If *Y* were continuous instead, we can perform a linear regression and obtain the regular residuals.

$$Y = \beta_0 + \beta_1 Z + \varepsilon_Y \rightarrow e_Y = Y - \hat{\beta}_0 - \hat{\beta}_1 Z$$

Then we can take the first canonical correlation between the residual of *Y* from linear regression and the set of residuals of *X* from multinomial logistic regression, which is essential the multiple correlation from regression the residuals of *Y* onto the residuals of *X*.

$$U_1 = a_1 \hat{r}_{x1} + a_2 \hat{r}_{x2}, V_1 = b_1 e_Y$$
$$r_{xy(z)} = \operatorname{corr}(U_1, V_1) \leftrightarrow e_y = a_1 \hat{r}_{x1} + a_2 \hat{r}_{x2} + \varepsilon \rightarrow R_{e_y \cdot \hat{r}_{x1} \hat{r}_{x2}}$$

However, if *X* is only binary, then the regular correlation maybe applied to the two residuals.

$$\log\left[\frac{\pi_x}{1-\pi_x}\right] = \beta_0 + \beta_1 Z \rightarrow \hat{\pi}_x = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 z)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 z)} \rightarrow \hat{r}_x = \frac{x - \hat{\pi}_x}{\hat{\pi}_x(1-\hat{\pi}_x)}$$
$$Y = \beta_0 + \beta_1 Z + \varepsilon_Y \rightarrow e_Y = Y - \hat{\beta}_0 - \hat{\beta}_1 Z$$
$$r_{xy(z)} = \operatorname{corr}(\hat{r}_x, e_Y)$$

8 Simulations to compare partial correlation measures

An attempt to analytically derive the statistical properties of the new partial correlation led to a dead end expression. Suppose we have binary variables *X*, *Y*, and *Z*. We can perform logistic regression with *X* and *Y* as outcomes and *Z* as a dependent variable to obtain the Pearson residuals and our new partial correlation.

$$\log\left[\frac{\pi_x}{1-\pi_x}\right] = \beta_0 + \beta_1 Z \rightarrow \hat{\pi}_x = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z\right)} \rightarrow \hat{r}_x = \frac{x-\hat{\pi}_x}{\hat{\pi}_x\left(1-\hat{\pi}_x\right)}$$

$$\log\left[\frac{\pi_y}{1-\pi_y}\right] = \gamma_0 + \gamma_1 Z \rightarrow \hat{\pi}_y = \frac{\exp\left(\hat{\gamma}_0 + \hat{\gamma}_1 z\right)}{1+\exp\left(\hat{\gamma}_0 + \hat{\gamma}_1 z\right)} \rightarrow \hat{r}_y = \frac{y-\hat{\pi}_y}{\hat{\pi}_y\left(1-\hat{\pi}_y\right)}$$

$$r_{xy(z)} = \frac{\sum\left(\hat{r}_x - \frac{\sum \hat{r}_x}{n}\right)\left(\hat{r}_y - \frac{\sum \hat{r}_y}{n}\right)}{\sqrt{\sum\left(\hat{r}_x - \frac{\sum \hat{r}_x}{n}\right)^2}\sqrt{\sum\left(\hat{r}_y - \frac{\sum \hat{r}_y}{n}\right)^2}}$$

$$= \frac{\sum\left[\frac{x-\hat{\pi}_x}{\hat{\pi}_x(1-\hat{\pi}_x)}\frac{y-\hat{\pi}_y}{\hat{\pi}_y(1-\hat{\pi}_y)}\right] - n\sum\frac{x-\hat{\pi}_x}{n\hat{\pi}_x(1-\hat{\pi}_x)}\sum\frac{y-\hat{\pi}_y}{n\hat{\pi}_y(1-\hat{\pi}_y)}}{\sqrt{\sum\left[\frac{x-\hat{\pi}_x}{\hat{\pi}_x(1-\hat{\pi}_x)}\right]^2 - n\left[\sum\frac{x-\hat{\pi}_x}{n\hat{\pi}_x(1-\hat{\pi}_x)}\right]^2}\sqrt{\sum\left[\frac{y-\hat{\pi}_y}{\hat{\pi}_y(1-\hat{\pi}_y)}\right]^2 - n\left[\sum\frac{y-\hat{\pi}_y}{n\hat{\pi}_y(1-\hat{\pi}_y)}\right]^2}}$$

Due to $\hat{\pi}$ being an exponential form of the estimated coefficients, we have not been able to reduce this expression nor take expectation of the expression. Furthermore, although the coefficients are estimated via maximum likelihood, the normal equations do not have a general closed form solution; they are obtained numerically through iterative methods (Agresti 2007). Alternatively, we commence exploration of the new partial correlation's properties by simulations using two possible underlying models.

8.1 Categorized multivariate normal model

The first model is a multivariate normal distribution that is categorized based on percentile cutoffs of the distribution. The advantage in categorizing a multivariate normal distribution is the available explicit form of the bivariate correlations, and hence the partial correlation. In addition, we can

compare the results of the new partial correlation to the partial phi coefficient and the partial tetrachoric (polychoric) correlations. However, the weakness in using the multivariate normal distribution is that the resulting variables are not truly nominal, so the results are not applicable for all categorical variables. Even if the variables were ordered, they may not be from a latent multivariate normal distribution.

1000 datasets of $n = \{100, 200, 500\}$ observations and 50 additional datasets of 5000 observations were generated for three variables *X**, *Y**, *Z** from a standardized trivariate normal distribution with varying covariance matrices using the *mvrnorm* function from the MASS package in R. Since the distribution is standardized, the covariance matrix is equivalent to the correlation matrix and off-diagonal entries will determine the value of the true partial correlation.

$$(X^*, Y^*, Z^*) \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}'$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho_{X^*Y^*} & \rho_{XZ^*} \\ \rho_{X^*Y^*} & 1 & \rho_{Y^*Z^*} \\ \rho_{X^*Z^*} & \rho_{Y^*Z^*} & 1 \end{bmatrix}$$

A combination of values for the correlations was set such that the covariance matrix is positive definite and a range of partial correlations could be observed. The population partial correlation between *X** and *Y** controlling for *Z** is calculated as

$$\rho_{X^*Y^*(Z^*)} = \frac{\rho_{X^*Y^*} - \rho_{X^*Z^*}\rho_{Y^*Z^*}}{\sqrt{\left(1-\rho_{X^*Z^*}^2\right)}\sqrt{\left(1-\rho_{Y^*Z^*}^2\right)}}$$

The sample partial correlation based on the sampled continuous data was calculated for baseline comparisons. Then the data was categorized into categorical variables *X*, *Y*, and *Z* based on percentile cutoffs in the normal distribution (Table 23). Scenarios A and C represent the best possible break-up of the distribution: categorizations equally divide the normal distribution, thus balancing the data. Scenarios B and D are based on less favorable skewed cutoffs. The cutoffs for D at 0.5625 and 0.9375 are loosely based on SNP theory. If 25% of the population has the risk allele *a*, then the genotype distribution would be

Genotype	Code	Probability	Cumulative Probability
AA	0	$(1-0.25)^2 = 0.5625$	0.5625
Aa	1	$2(0.25)(1-0.25) = 0.375$	0.9375
aa	2	$0.25^2 = 0.0625$	1.000

It should be noted that the new partial correlation and the partial phi coefficient estimate the partial correlation after categorization (manifest partial correlation), while the partial tetrachoric (polychoric) correlation estimates the partial correlation before categorization (latent partial correlation, as defined above). The true manifest partial correlation in the case where all three variables are dichotomized based on median splits has an explicit formula (Vargha, et al. 1996).

$$\rho_{XY(Z)} = \frac{0.637 \arcsin \rho_{X^*Y^*} - 0.405\left(\arcsin \rho_{X^*Z^*}\right)\left(\arcsin \rho_{Y^*Z^*}\right)}{\sqrt{\left(1-0.405\left(\arcsin \rho_{X^*Z^*}\right)^2\right)}\sqrt{\left(1-0.405\left(\arcsin \rho_{X^*Z^*}\right)^2\right)}}$$

We also use the multivariate normal distribution to create mixed variable scenarios. Only *X* and *Z* are categorized from X* and Z*; *Y* is allowed to be the continuous *Y** (Table 24). *Pcor.pch* was modified in the following way: the biserial (polyserial) correlation estimated the correlation between *Y* and the categorical variables instead of the tetrachoric (polychoric) correlation. These correlations are used to estimate the *pcor.pch*.

$$pcor.pch = \begin{cases} \dfrac{r_{biserial(XY)} - r_{tetrachoric(XZ)} r_{biserial(YZ)}}{\sqrt{1-r^2_{tetrachoric(XZ)}}\sqrt{r^2_{biserial(YZ)}}} & \text{(binary data)} \\ \dfrac{r_{polyserial(XY)} - r_{polychoric(XZ)} r_{polyserial(YZ)}}{\sqrt{1-r^2_{polychoric(XZ)}}\sqrt{r^2_{polyserial(YZ)}}} & \text{(trichotomous data)} \end{cases}$$

Then this measure, to be referred to as the partial mixed correlation, is compared to the new partial correlation.

The true manifest partial correlation in this case where only *X** and *Z** are dichotomized based on median splits (Vargha, et al. 1996).

$$\rho_{XY^*(Z)} = \frac{0.798\rho_{X^*Y^*} - 0.508\left(\arcsin \rho_{X^*Z^*}\right)\rho_{Y^*Z^*}}{\sqrt{\left(1-0.405\left(\arcsin \rho_{X^*Z^*}\right)^2\right)}\sqrt{\left(1-0.637\rho^2_{Y^*Z^*}\right)}}$$

Unfortunately, we do not have calculations of the manifest partial correlation for other situations

Table 22 shows the correlation combinations and their corresponding partial correlations.

Table 22 Multivariate Normal Model correlations and corresponding partial correlations. Each row represents one model setting for the generated multivariate normal data *X*, Y*, Z** which are transformed into categorical variables *X, Y, Z*. The first three columns determine the correlation matrix and hence the variance-covariance matrix. The fourth column is the (latent) partial correlation between *X** and *Y** controlling for *Z**. The third column is the (manifest) partial correlation between *X* and *Y* controlling for *Z* when *X**, *Y**, *Z** are dichotomized by median splits of their marginal normal distributions. The last column is the (manifest) partial correlation between *X* and *Y** controlling for *Z*, when only *X** and *Z** are dichotomized by median splits of their marginal normal distributions.

$\rho_{X^*Y^*}$	$\rho_{X^*Z^*}$	$\rho_{Y^*Z^*}$	$\rho_{X^*Y^*(Z^*)}$	$\rho_{XY(Z)}$ (median splits)	$\rho_{XY^*(Z)}$ (median splits)
-0.8	-0.5	0.0	-0.92	-0.63	-0.68
-0.5	-0.5	-0.2	-0.71	-0.40	-0.49
-0.5	0.0	0.0	-0.50	-0.33	-0.40
-0.2	-0.2	-0.2	-0.25	-0.15	-0.18
0.0	0.0	0.0	0.00	0	0
0.2	0.2	-0.2	0.25	0.15	0.18
0.5	0.0	0.0	0.50	0.33	0.40
0.5	0.5	-0.2	0.71	0.40	0.49
0.8	0.5	0.0	0.92	0.63	0.68

Table 23 Multivariate Normal Model based simulation settings. Details of how the continuous data *X*, Y*, Z** are transformed into dichotomous and trichotomous variables *X, Y, Z*. Each scenario will be denoted by their corresponding column letter heading in the remainder of section 8.1. A and B are dichotomous scenarios; C and D are trichotomous settings. A and C are when the thresholds divide the distribution equally; B and D are when the thresholds skew the distribution of the categorical data.

Dichotomous			Trichotomous		
	A	**B**		**C**	**D**
p_x	0.5	0.9	p_{x_1}, p_{x_2}	0.33, 0.67	0.5625, 0.9375
p_y	0.5	0.9	p_{y_1}, p_{y_2}	0.33, 0.67	0.5625, 0.9375
p_z	0.5	0.9	p_{z_1}, p_{z_2}	0.33, 0.67	0.5625, 0.9375
$X = \begin{cases} 1 \text{ if } X^* > x, P(X^* \le x) = p_x \\ 0 \text{ otherwise} \end{cases}$ $Y = \begin{cases} 1 \text{ if } Y^* > y, P(Y^* \le y) = p_y \\ 0 \text{ otherwise} \end{cases}$ $Z = \begin{cases} 1 \text{ if } Z^* > z, P(Z^* \le z) = p_z \\ 0 \text{ otherwise} \end{cases}$			$X = \begin{cases} 2 \text{ if } X^* > x_2, P(X^* \le x_2) = p_{x_2} \\ 1 \text{ if } x_1 < X^* \le x_2, P(X^* \le x_1) = p_{x_1} < p_{x_2} \\ 0 \text{ otherwise} \end{cases}$ $Y = \begin{cases} 2 \text{ if } Y^* > y_2, P(Y^* \le y_2) = p_{y_2} \\ 1 \text{ if } y_1 < Y^* \le y_2, P(Y^* \le y_1) = p_{y_1} < p_{y_2} \\ 0 \text{ otherwise} \end{cases}$ $Z = \begin{cases} 2 \text{ if } Z^* > z_2, P(Z^* \le z_2) = p_{z_2} \\ 1 \text{ if } z_1 < Z^* \le z_2, P(Z^* \le z_1) = p_{z_1} < p_{z_2} \\ 0 \text{ otherwise} \end{cases}$		

Table 24 Multivariate Normal Model based simulation settings for mixed data. Details of how the continuous data *X*, Y*, Z** are transformed into mixed data containing two categorical variables *X* and *Z* and one continuous variable *Y*. Each scenario will be denoted by their corresponding column letter heading in the remainder of section 8.1. E and F are dichotomous scenarios; G and H are trichotomous settings. E and G are when the thresholds divide the distribution equally; F and H are when the thresholds skew the distribution of the categorical data.

Dichotomous			Trichotomous		
	E	**F**		**G**	**H**
p_x	0.5	0.9	p_{x_1}, p_{x_2}	0.33, 0.67	0.5625, 0.9375
p_y	NA	NA	p_{y_1}, p_{y_2}	NA	NA
p_z	0.5	0.9	p_{z_1}, p_{z_2}	0.33, 0.67	0.5625, 0.9375
$X = \begin{cases} 1 \text{ if } X^* > x, P(X^* \le x) = p_x \\ 0 \text{ otherwise} \end{cases}$ $Y = Y^*$ $Z = \begin{cases} 1 \text{ if } Z^* > z, P(Z^* \le z) = p_z \\ 0 \text{ otherwise} \end{cases}$			$X = \begin{cases} 2 \text{ if } X^* > x_{2,} P(X^* \le x_2) = p_{x_2} \\ 1 \text{ if } x_1 < X^* \le x_2, P(X^* \le x_1) = p_{x_1} < p_{x_2} \\ 0 \text{ otherwise} \end{cases}$ $Y = Y^*$ $Z = \begin{cases} 2 \text{ if } Z^* > z_2, P(Z^* \le z_2) = p_{z_2} \\ 1 \text{ if } z_1 < Z^* \le z_2, P(Z^* \le z_1) = p_{z_1} < p_{z_2} \\ 0 \text{ otherwise} \end{cases}$		

100 observations turned out to be too small for scenarios C and D, resulting in sparse tables with a cutoff at 0.9375, so we have discarded that sample size for trichotomous variables.

The following partial correlation measures for *X* and *Y* controlling for *Z* were calculated:

- *Pcor.est* - partial correlation based on the continuous data before categorization; this serves as a baseline for the categorical data analysis
- *Pcor.new* - new partial correlation for categorical data
- *Pcor.phi* - partial phi coefficient (dichotomous data)
- *Pcor.p* – partial correlation for continuous data applied after categorization (trichotomous data)
- *Pcor.tet* - partial tetrachoric correlation (dichotomous data)
- *Pcor.pch* - partial polychoric correlation (trichotomous data) or partial mixed correlation (mixed data)

8.1.1 The null distribution of the test statistics

Although the Pearson residuals from individual logistic regression are normally distributed, it is unknown whether the bivariate distribution of the combined Pearson residuals from two logistic regressions is normal. The *F*-test applied for the new partial correlation is a natural starting place for statistical inference, given the use of correlation, but the appropriateness of applying the test needs to be examined. The same could be said for the application of the *F*-tests for partial correlations made up of the tetrachoric, polychoric, and polyserial correlations. Since the phi coefficient is the direct application of Pearson's product-moment correlation, we expect to be able to apply the usual *F*-test without problems.

Out of the simulation settings mentioned earlier, we consider the case when the true partial correlation is equal to zero. The goal here is to examine, from 1000 simulations, the empirical null distribution of the test statistic for each partial correlation and compare it against what has been assumed to be the null distribution. The null distribution is the distribution of the test statistic when the null hypothesis, that the partial correlation is zero, is true.

For scenarios A and B, the following tests were applied:

$$\text{new partial correlation} \rightarrow \frac{r_{xy(z)}^2}{1-r_{xy(z)}^2}(N-2) \sim F_{1,N-2}$$

$$\text{partial phi coefficient} \rightarrow \frac{\phi_{xy(z)}^2}{1-\phi_{xy(z)}^2}(N-3) \sim F_{1,N-3}$$

$$\text{partial tetrachoric correlation} \rightarrow \frac{r_{tet[xy(z)]}^2}{1-r_{tet[xy(z)]}^2}(N-3) \sim F_{1,N-3}$$

Table 25 and Table 26 show the density curves of the test statistics in binary scenarios A and B, respectively. The distribution of the test statistic for each partial correlation is graphed against their assumed null distribution. In general, there are anomalies in the distribution of the test statistic for the new partial correlation; the density tends to verge on two modes, indicating that the *F*-test may not be appropriate. Not surprisingly, the partial phi coefficient does match its corresponding null distribution well. Again, the phi coefficient is the direct application of Pearson's product-moment correlation to dichotomous variables. The partial phi coefficient could have been obtained by correlating the residuals of linear regression of the data; hence the *F*-test is perfectly reasonable here. The partial tetrachoric correlation does not resemble its assumed null distribution at all when the data are categorized with

balanced distribution over the categorical variables. When the data are skewed with cutoff threshold at 0.9 as in scenario B, the distribution begins to resemble a familiar unimodal distribution, but still not the presumed *F* distribution.

Table 25 Density curves of test statistics for Scenario A (dichotomous data, split at 0.5). Comparison of the empirical distribution of the test statistic for each of the partial correlation measures against their presumed null distribution.

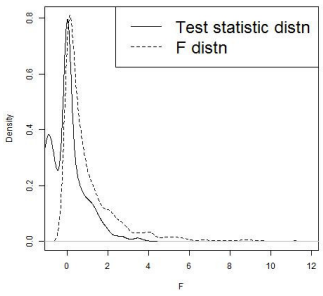
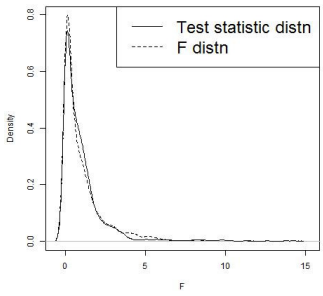
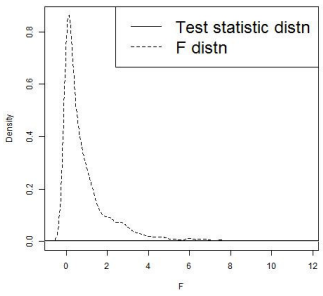
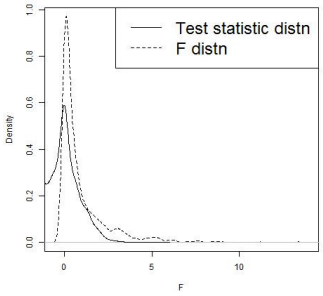
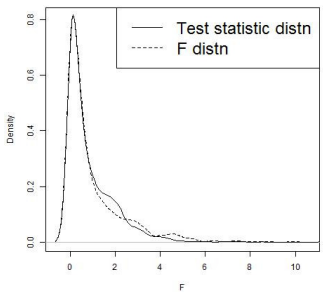
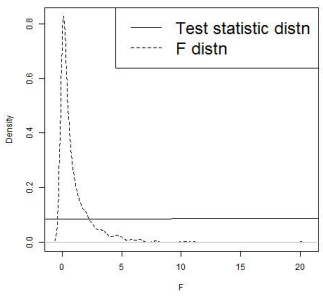
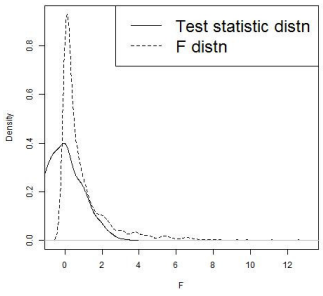
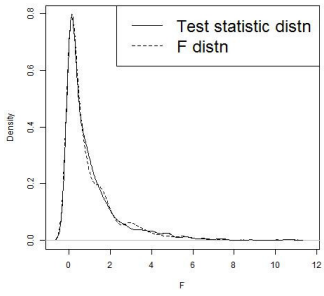
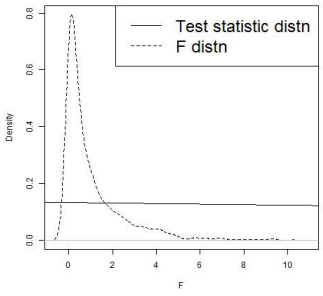
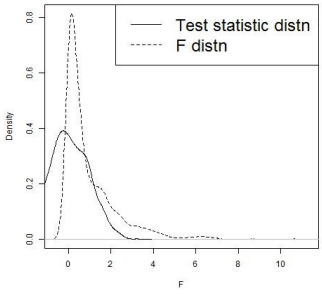
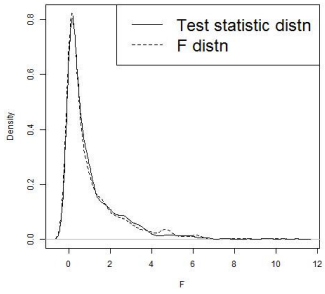
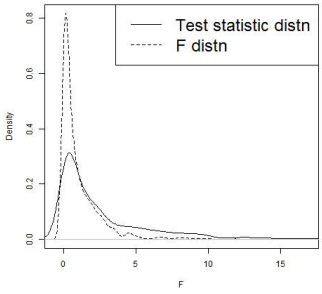
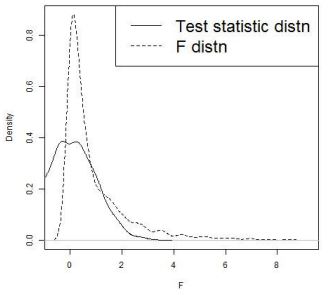
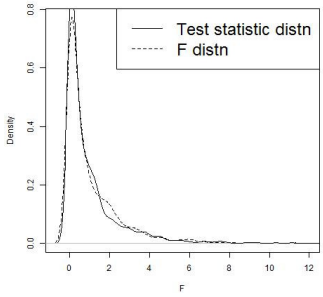
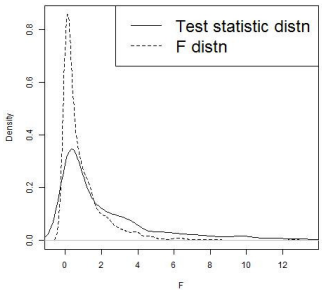
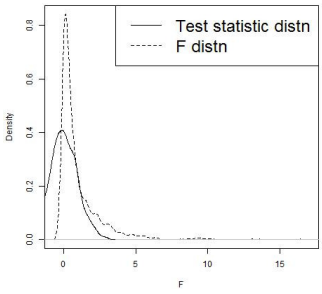
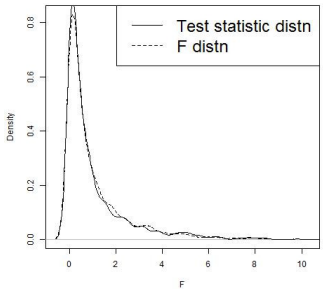
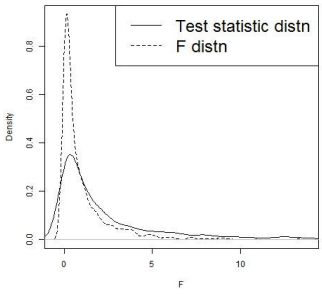
N	New partial correlation	Partial phi coefficient	Partial tetrachoric correlation
100			
200			
500			

Table 26 Density curves of test statistics for scenario B (dichotomous data split at 0.9). Comparison of the empirical distribution of the test statistic for each of the partial correlation measures against their presumed null distribution.

N	New partial correlation	Partial phi coefficient	Partial tetrachoric correlation
100			
200			
500			

For scenarios C and D, we used the following tests

$$\text{new partial correlation } \rightarrow -2\ln\Lambda = -\left(n-1-\frac{1}{2}(2+2+1)\right)\ln\prod_{i=1}^{2}\left(1-\hat{\rho}_i^{*2}\right) \approx \chi^2_{2(2)}$$

$$\text{partial correlation}\left(\text{continuous version applied to 0,1,2 coded data}\right)$$

$$\rightarrow F = \frac{r^2_{XY(Z)}}{1-r^2_{XY(Z)}}(N-1-2) \sim F_{1,N-1-2}$$

$$\text{partial polychoric correlation} \rightarrow F = \frac{r^2_{pch[XY(Z)]}}{1-r^2_{pch[XY(Z)]}}(N-1-2) \sim F_{1,N-1-2}$$

For scenarios E and F, we used the following tests

$$\text{new partial correlation } \rightarrow -2\ln\Lambda = -\left(n-1-\frac{1}{2}(2+1+1)\right)\ln\prod_{i=1}^{2}\left(1-\hat{\rho}_i^{*2}\right) \approx \chi^2_{2(1)}$$

$$\text{partial correlation}\left(\text{continuous version applied to 0,1,2 coded data}\right)$$

$$\rightarrow F = \frac{r^2_{XY(Z)}}{1-r^2_{XY(Z)}}(N-1-2) \sim F_{1,N-1-2}$$

$$\text{partial mixed correlation} \rightarrow F = \frac{r^2_{mixed[XY(Z)]}}{1-r^2_{mixed[XY(Z)]}}(N-1-2) \sim F_{1,N-1-2}$$

Table 27 shows the density curves of the test statistics in trichotomous scenarios C and D; Table 28 and Table 29 show density curves of test statics in the mixed scenarios of E and F, G and H respectively. In general, the distribution of the test statistic for the new partial correlation matches the chi-square distribution very well for multi-categorical data. The partial correlation matches its corresponding null distribution, as expected. The partial polychoric correlation maintains the familiar unimodal shape, but has a lower peak and longer tail to the right than its presumed null distribution. The partial mixed correlation has strange distribution for n=200 in the binary case (E), but otherwise matches its null distribution well when the data are categorized with thresholds at 0.33, 0.67 (G). It is slightly shifted the left. With skewed thresholds (H), the peak drops below what is expected.

Table 27 Density curves of test statistics for scenarios C (trichotomous data, split at 0.33, 0.67) and D (trichotomous data split at 0.5625, 0.9375). Comparison of the empirical distribution of the test statistic for each of the partial correlation measures against their presumed null distribution. Partial correlation is just the applying the continuous method to the data.

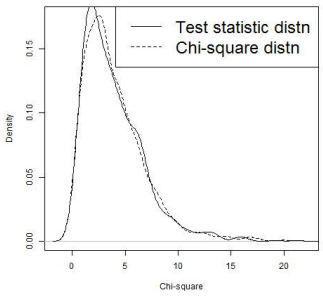
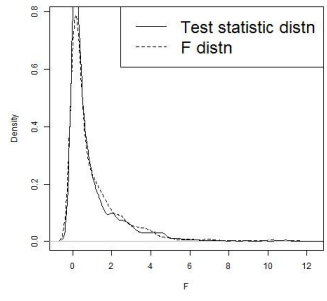
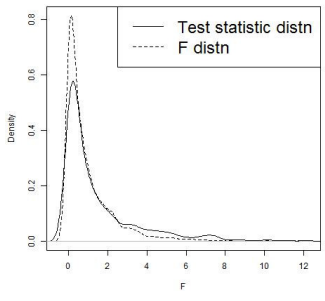
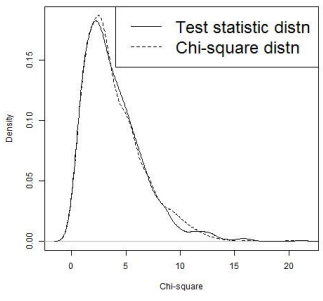
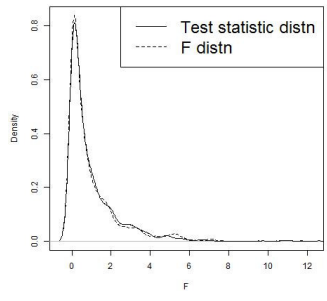
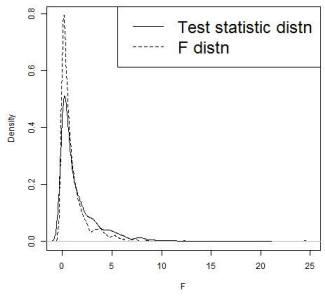
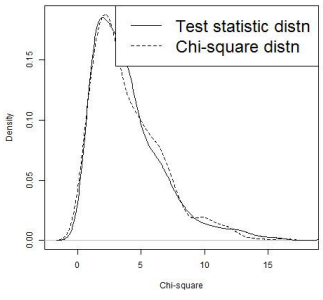
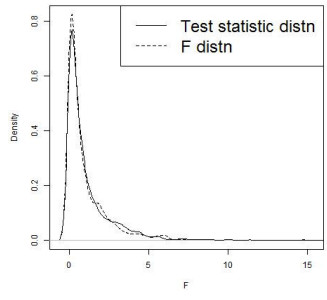
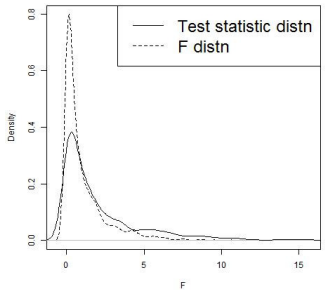
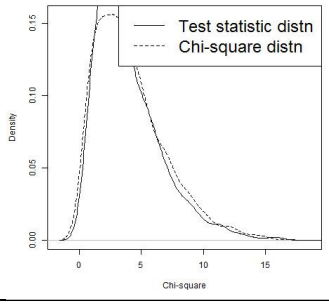
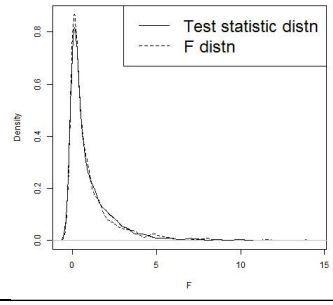
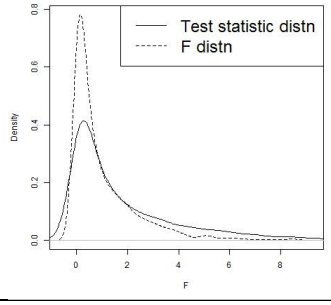
C			
N	New partial correlation	Partial correlation	Partial polychoric correlation
200			
500			
D			
N	New partial correlation	Partial correlation	Partial polychoric correlation
200			
500			

Table 28 Density curves of test statistics for mixed data scenarios E (*X*, *Z* dichotomous, split at 0.5) and F (*X, Z* dichotomous split at 0.9). Comparison of the empirical distribution of the test statistic for each of the partial correlation measures against their presumed null distribution. Partial correlation is just the applying the continuous method to the data.

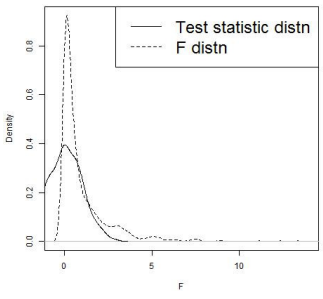
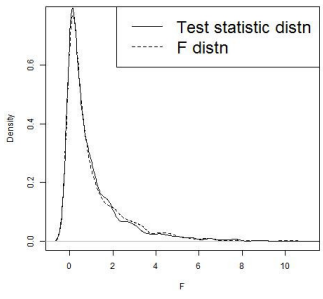
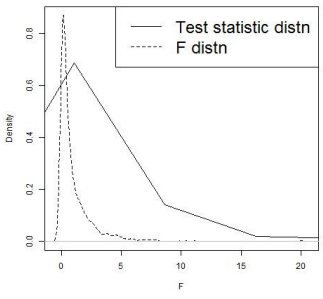
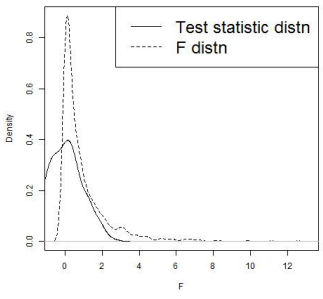
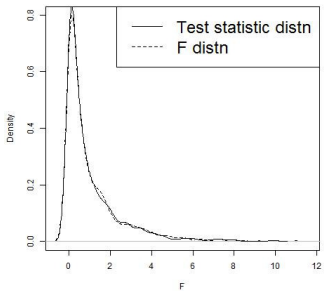
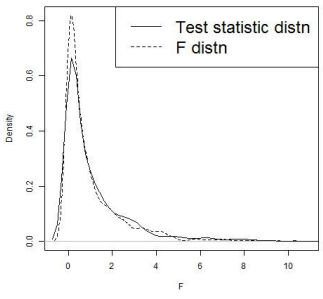
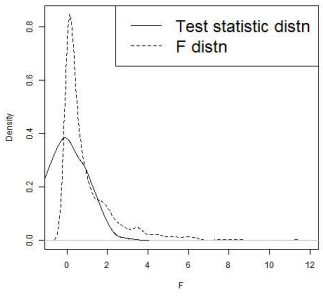
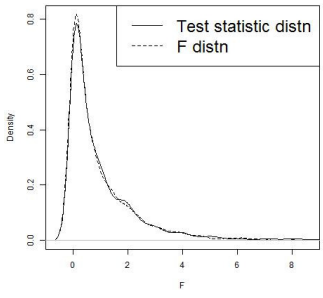
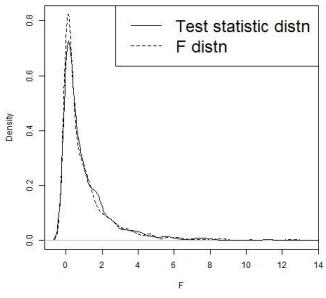
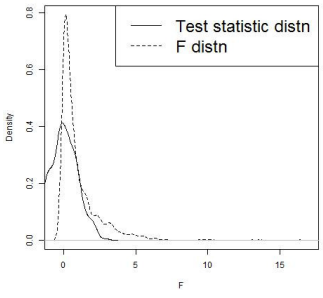
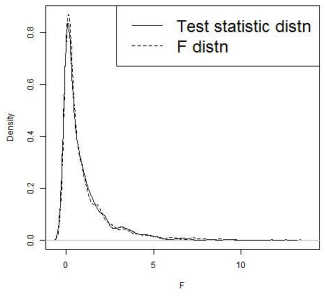
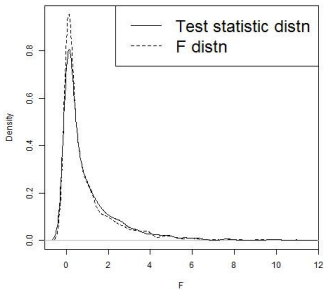
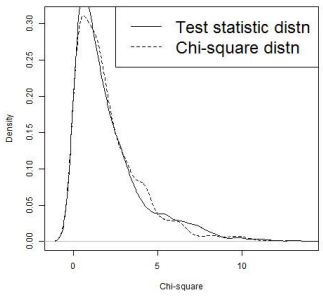
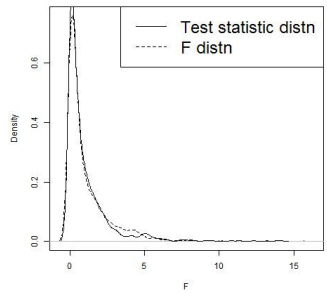
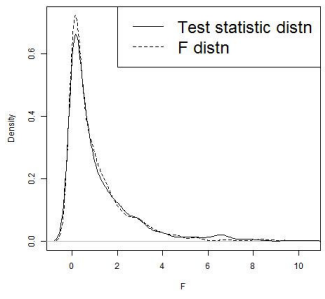
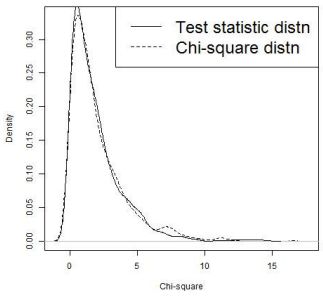
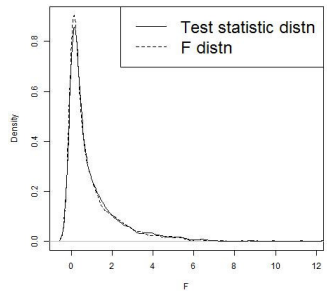
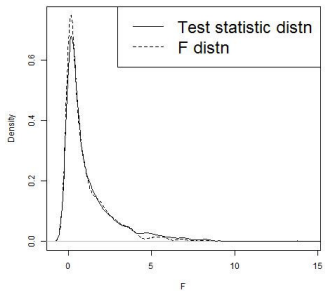
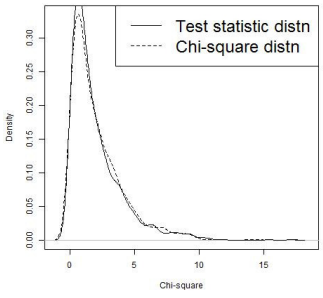
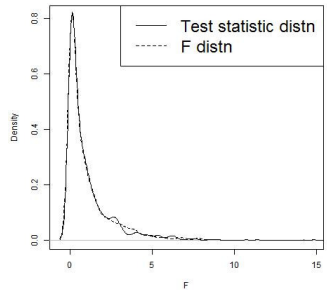
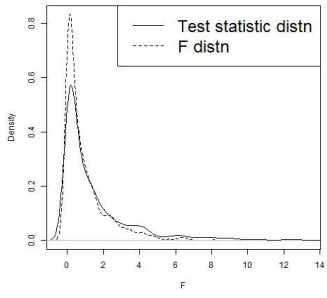
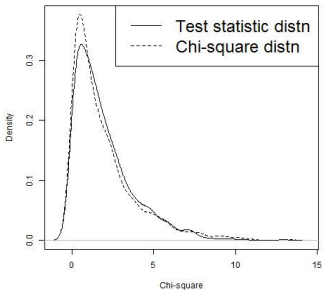
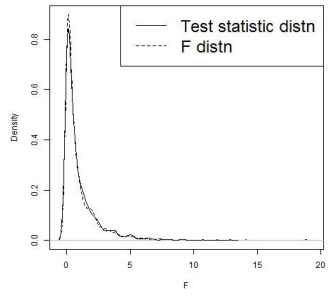
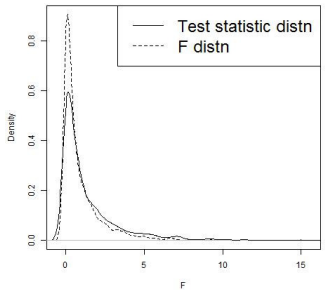
E			
N	New partial correlation	Partial correlation	Partial mixed correlation
200			
500			
F			
N	New partial correlation	Partial correlation	Partial mixed correlation
200			
500			

Table 29 Density curves of test statistics for mixed data scenarios G (*X, Z* trichotomous, splits at 0.33, 0.67) and H (*X, Z* trichotomous, splits at 0.5625, 0.9375). Comparison of the empirical distribution of the test statistic for each of the partial correlation measures against their presumed null distribution. Partial correlation is just the applying the continuous method to the data.

G			
N	New partial correlation	Partial correlation	Partial mixed correlation
200			
500			
H			
N	New partial correlation	Partial correlation	Partial mixed correlation
200			
500			

8.1.2 The performance of the partial correlations

From the earlier discussion of what each partial correlation measure is estimating (manifest, latent partial correlation), bias and standard error for the new partial correlation and partial phi coefficient can only be interpretable in the case when the data are binary and threshold splits are at 0.5 (A, E). Bias and standard error for partial tetrachoric (polychoric) correlation can be calculated based on the always available latent partial correlation (A-H). Bias and standard error for partial mixed correlation can be analyzed for scenario E only.

From the simulation study of the distribution of the test statistics, we can analyze power and type I meaningfully for the new partial correlation when dealing with multi-categorical or mixed data (C, D, G, H) and the partial phi coefficient in general. The power and type I error rates of partial tetrachoric and polychoric correlations cannot be analyzed because they do not have the null *F* distribution. The power and type I error rates for the partial mixed correlation can be analyzed for scenarios F and G.

Results shown are for n=500 and partial correlations are for *X* and *Y* controlling for *Z*. Full results are available in supplementary materials.

We reiterate the use of notation:

- *Pcor.est* - partial correlation based on the continuous data before categorization; this serves as a baseline for the categorical data analysis
- *Pcor.new* - new partial correlation for categorical data
- *Pcor.phi* - partial phi coefficient (dichotomous data)
- *Pcor.p* – partial correlation for continuous data applied after categorization (trichotomous data)
- *Pcor.tet* - partial tetrachoric correlation (dichotomous data)
- *Pcor.pch* - partial polychoric correlation (trichotomous data) or partial mixed correlation (mixed data)

Table 30 Multivariate Normal Model based simulation – Bias and Standard Error for scenarios A (dichotomous, split at 0.5) and E (mixed. *X*, *Z* dichotomous, split at 0.5). Here pcor.true (latent) is the partial correlation for the continuous data that was originally generated; pcor.true (manifest) is the partial correlation for the data after categorization. Pcor.est is the partial correlation applied to the original continuous data. Pcor.new is the new partial correlation. Pcor.phi is the partial phi coefficient (applied to A only). pcor.p is the partial correlation applied to the categorized data (applied to E only). Pcor.tet is the partial tetrachoric correlation (applied to A only). Pcor.pch is the partial mixed correlation (applied to E only). Scenarios A and E are the only scenarios for which the true partial correlation for the categorical data is known.

			pcor.est		pcor.new		pcor.phi (pcor.p)		pcor.tet (pcor.pch)	
	pcor.true (latent)	pcor.true (manifest)	Bias	SE	Bias	SE	Bias	SE	Bias	SE
A	-0.92	-0.63	0.00	0.01	0.03	0.03	0.00	0.03	0.00	0.02
	-0.71	-0.40	0.00	0.02	0.01	0.04	0.00	0.04	0.00	0.06
	-0.50	-0.33	0.00	0.03	0.00	0.04	0.00	0.04	0.00	0.06
	-0.25	-0.15	0.00	0.04	0.00	0.04	0.00	0.04	0.00	0.07
	0.00	0.00	0.00	0.04	0.00	0.04	0.00	0.04	0.00	0.07
	0.25	0.15	0.00	0.04	-0.01	0.04	0.00	0.04	0.00	0.07
	0.50	0.33	0.00	0.03	0.00	0.04	0.00	0.04	0.00	0.06
	0.71	0.40	0.00	0.02	-0.04	0.04	0.00	0.04	0.00	0.05
	0.92	0.63	0.00	0.01	-0.03	0.03	0.00	0.03	0.00	0.02
E	-0.92	-0.68	0.00	0.01	0.00	0.02	0.00	0.02	0.14	0.02
	-0.71	-0.49	0.00	0.02	0.00	0.03	0.00	0.03	0.11	0.04
	-0.50	-0.40	0.00	0.03	0.00	0.04	0.00	0.04	0.09	0.04
	-0.25	-0.18	0.00	0.04	0.00	0.04	0.00	0.04	0.04	0.05
	0.00	0.00	0.00	0.04	0.00	0.04	0.00	0.04	0.00	0.05
	0.25	0.18	0.00	0.04	0.00	0.04	0.00	0.04	-0.04	0.04
	0.50	0.40	0.00	0.03	0.00	0.04	0.00	0.04	-0.09	0.04
	0.71	0.49	0.00	0.02	0.00	0.03	0.00	0.03	-0.11	0.04
	0.92	0.68	0.00	0.01	0.00	0.02	0.00	0.02	-0.14	0.02

Bias and standard error for scenarios A and E is shown in Table 30. Almost all partial correlation measures have little to no bias, with the exceptions of the tetrachoric correlation. Upon closer examination of the analysis, we found that whenever the partial tetrachoric correlation obtained magnitude greater than 1, the marginal tetrachoric correlation was between -5 and -1 for all pairs of variables; if all correlations used in a variance-covariance matrix was in this range, the resulting matrix would not be positive-definite.

Bias and standard error for the remaining scenarios are shown in Table 31. The partial tetrachoric correlation has high biases for binary scenarios (B, F) with little consistency in the estimates, but has little to no bias and good consistency for trichotomous scenarios.

Table 31 Multivariate Normal Model based simulation – Bias and Standard Error (Scenarios B-D, F-H). Here pcor.true (latent) is the partial correlation for the continuous data that was originally generated. Pcor.est is the partial correlation applied to the original continuous data. Pcor.tet is the partial tetrachoric correlation (applied to B, C only). Pcor.pch is the partial polychoric correlation (applied to D) or the partial mixed correlation (applied to F, G, H). Because the true partial correlation for the data after categorization is unknown for these scenarios, the new partial correlation and partial phi coefficient is not presented here.

		pcor.est		pcor.tet (pcor.pch)	
	pcor.true (latent)	Bias	SE	Bias	SE
B	-0.92	0.00	0.01	-1.78	1.57
	-0.71	0.00	0.02	-2.89	4.82
	-0.50	0.00	0.03	-0.30	0.39
	-0.25	0.00	0.04	-0.35	1.13
	0.00	0.00	0.04	-0.03	0.20
	0.25	0.00	0.04	0.07	0.31
	0.50	0.00	0.03	0.00	0.15
	0.71	0.00	0.02	0.25	0.94
	0.92	0.00	0.01	0.03	0.28
C	-0.92	0.00	0.01	0.00	0.02
	-0.71	0.00	0.02	0.00	0.04
	-0.50	0.00	0.03	0.00	0.05
	-0.25	0.00	0.04	0.00	0.06
	0.00	0.00	0.04	0.00	0.06
	0.25	0.00	0.04	0.00	0.06
	0.50	0.00	0.03	0.00	0.05
	0.71	0.00	0.02	0.00	0.04
	0.92	0.00	0.01	0.00	0.02
D	-0.92	0.00	0.01	0.00	0.04
	-0.71	0.00	0.02	0.00	0.06
	-0.50	0.00	0.03	0.00	0.05
	-0.25	0.00	0.04	0.00	0.06
	0.00	0.00	0.04	0.00	0.06
	0.25	0.00	0.04	0.00	0.06
	0.50	0.00	0.03	0.00	0.05
	0.71	0.00	0.02	0.00	0.05
	0.92	0.00	0.01	0.00	0.02

		pcor.est		pcor.pch	
	pcor.true (latent)	Bias	SE	Bias	SE
F	-0.92	0.00	0.01	-0.27	0.56
	-0.71	0.00	0.02	-0.51	0.66
	-0.50	0.00	0.03	0.21	0.05
	-0.25	0.00	0.04	0.02	0.24
	0.00	0.00	0.04	0.00	0.05
	0.25	0.00	0.04	-0.09	0.05
	0.50	0.00	0.03	-0.21	0.06
	0.71	0.00	0.02	-0.26	0.06
	0.92	0.00	0.01	-0.42	0.04
G	-0.92	0.00	0.01	0.00	0.02
	-0.71	0.00	0.02	0.00	0.03
	-0.50	0.00	0.03	0.00	0.04
	-0.25	0.00	0.04	0.00	0.05
	0.00	0.00	0.04	0.00	0.05
	0.25	0.00	0.04	0.00	0.05
	0.50	0.00	0.03	0.00	0.04
	0.71	0.00	0.02	0.00	0.03
	0.92	0.00	0.01	0.00	0.02
H	-0.92	0.00	0.01	0.00	0.03
	-0.71	0.00	0.02	0.00	0.04
	-0.50	0.00	0.03	0.00	0.04
	-0.25	0.00	0.04	0.00	0.05
	0.00	0.00	0.04	0.00	0.05
	0.25	0.00	0.04	0.00	0.05
	0.50	0.00	0.03	0.00	0.04
	0.71	0.00	0.02	0.00	0.04
	0.92	0.00	0.01	0.00	0.02

Power for all scenarios are shown in Table 32. The test based on the new partial correlations has very high power close to 1.0 for scenarios C, D, G, H. The test based on the partial phi coefficient is able to detect a nonzero partial correlations with power in the range of 0.8-1.0 for almost all scenarios. Power for the test using the partial phi coefficient drops in scenario B when the magnitude of the partial correlation is less than 0.25. The test based on the partial mixed correlation mostly good power, except in scenario E, when the partial correlation is strongly negative.

Type I error rates for all scenarios are shown in Table 33. Almost all methods hovered around an error rate of 0.05, except for the test based on the partial mixed correlation; type I error rates were inflated to 0.08.

Table 32 Multivariate Normal Model based simulation – Power. Here pcor.true (latent) is the partial correlation for the continuous data that was originally generated; pcor.true (manifest) is the partial correlation for the data after categorization. Pcor.est is the partial correlation applied to the original continuous data. Pcor.new is the new partial correlation. Pcor.phi is the partial phi coefficient. pcor.p is the partial correlation applied to the categorized data. Pcor.tet is the partial tetrachoric correlation. Pcor.pch is either the partial polychoric correlation or the partial mixed correlation depending on the scenario. Some power results were removed based on the findings of section 8.1.1.

	pcor.true (latent)	pcor.true (manifest)	pcor.est	pcor.new	pcor.phi (pcor.p)		pcor.true (latent)	pcor.true (manifest)	pcor.est	pcor.new	pcor.phi (pcor.p)	pcor.tet (pcor.pch)
A	-0.92	-0.63	1.00		1.00	E	-0.92	-0.68	1.00		1.00	
	-0.71	-0.40	1.00		1.00		-0.71	-0.49	1.00		1.00	
	-0.50	-0.33	1.00		1.00		-0.50	-0.40	1.00		1.00	
	-0.25	-0.15	1.00		0.91		-0.25	-0.18	1.00		0.98	
	0.25	0.15	1.00		0.91		0.25	0.18	1.00		0.98	
	0.50	0.33	1.00		1.00		0.50	0.40	1.00		1.00	
	0.71	0.40	1.00		1.00		0.71	0.49	1.00		1.00	
	0.92	0.63	1.00		1.00		0.92	0.68	1.00		1.00	
B	-0.92		1.00		0.98	F	-0.92		1.00		1.00	0.31
	-0.71		1.00		0.89		-0.71		1.00		1.00	0.33
	-0.50		1.00		0.82		-0.50		1.00		1.00	1.00
	-0.25		1.00		0.22		-0.25		1.00		0.79	0.89
	0.25		1.00		0.49		0.25		1.00		0.83	0.93
	0.50		1.00		0.99		0.50		1.00		1.00	1.00
	0.71		1.00		1.00		0.71		1.00		1.00	1.00
	0.92		1.00		1.00		0.92		1.00		1.00	1.00
C	-0.92		1.00	1.00	1.00	G	-0.92		1.00	1.00	1.00	1.00
	-0.71		1.00	1.00	1.00		-0.71		1.00	1.00	1.00	1.00
	-0.50		1.00	1.00	1.00		-0.50		1.00	1.00	1.00	1.00
	-0.25		1.00	1.00	0.94		-0.25		1.00	0.99	1.00	1.00
	0.25		1.00	1.00	0.94		0.25		1.00	0.99	0.99	1.00
	0.50		1.00	1.00	1.00		0.50		1.00	1.00	1.00	1.00
	0.71		1.00	1.00	1.00		0.71		1.00	1.00	1.00	1.00
	0.92		1.00	1.00	1.00		0.92		1.00	1.00	1.00	1.00
D	-0.92		1.00	1.00	1.00	H	-0.92		1.00	1.00	1.00	
	-0.71		1.00	1.00	1.00		-0.71		1.00	1.00	1.00	
	-0.50		1.00	1.00	1.00		-0.50		1.00	1.00	1.00	
	-0.25		1.00	1.00	0.83		-0.25		1.00	0.98	0.99	
	0.25		1.00	1.00	0.88		0.25		1.00	0.98	0.99	
	0.50		1.00	1.00	1.00		0.50		1.00	1.00	1.00	
	0.71		1.00	1.00	1.00		0.71		1.00	1.00	1.00	
	0.92		1.00	1.00	1.00		0.92		1.00	1.00	1.00	

Table 33 Multivariate Normal Model based simulation – Type I Error. Here pcor.true (latent) is the partial correlation for the continuous data that was originally generated; pcor.true (manifest) is the partial correlation for the data after categorization. Pcor.est is the partial correlation applied to the original continuous data. Pcor.new is the new partial correlation. Pcor.phi is the partial phi coefficient. pcor.p is the partial correlation applied to the categorized data. Pcor.tet is the partial tetrachoric correlation. Pcor.pch is either the partial polychoric correlation or the partial mixed correlation depending on the scenario. Some power results were removed based on the findings of section 8.1.1.

	pcor.true (latent)	pcor.true (manifest)	pcor.est	pcor.new	pcor.phi (pcor.p)	pcor.tet (pcor.pch)
A	0.00	0.00	0.05		0.05	
B	0.00		0.05		0.05	
C	0.00		0.05	0.05	0.04	
D	0.00		0.05	0.05	0.04	
E	0.00	0.00	0.05		0.05	
F	0.00		0.05		0.05	0.07
G	0.00		0.05	0.04	0.05	0.08
H	0.00		0.05	0.04	0.05	

8.1.3 New partial correlation versus partial phi coefficient

Because the new partial correlation and the partial phi coefficient are meant to measure the partial correlation of the manifest variables, we also ran comparison tests to determine how similar these two estimates are asymptotically (n=5000, 50 simulations) (Table 34). When the data are balanced, as in the threshold determining the categorization is at 0.5 (A), the mean of the new partial correlation and the phi coefficient are quite similar, When the data are categorized in a skewed manner, the mean of the new partial correlation and partial phi coefficient are slightly different by a marginal amount (0.04-0.3) when the true latent partial correlation is less than -0.5. Tests of the measures found that they are still significantly different measures with the exception of when the partial correlation is close to zero. This indicates that the two estimators may have different distributions.

Table 34 Comparing the new partial correlation to the partial phi coefficient. Here pcor.true (latent) is the partial correlation for the continuous data that was originally generated; pcor.true (manifest) is the partial correlation for the data after categorization. Pcor.new is the new partial correlation. Pcor.phi is the partial phi coefficient. Scenario A is when data are dichotomized at 0.5; scenario B is when data are dichotomized at 0.9.

			pcor.new		pcor.phi		difference		p-value	
	pcor.true (latent)	pcor.true (manifest)	mean	SE	mean	SE	mean	SD	Paired t	Signed Rank
A	-0.92	-0.63	-0.59	0.01	-0.63	0.01	0.03	0.00	0.000	0.000
	-0.71	-0.40	-0.39	0.01	-0.40	0.01	0.01	0.00	0.000	0.000
	-0.50	-0.33	-0.34	0.01	-0.34	0.01	0.00	0.00	0.000	0.000
	-0.25	-0.15	-0.15	0.01	-0.15	0.01	0.00	0.00	0.032	0.039
	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.175	0.359
	0.25	0.15	0.14	0.01	0.15	0.01	0.00	0.00	0.000	0.000
	0.50	0.33	0.34	0.01	0.34	0.01	0.00	0.00	0.000	0.000
	0.71	0.40	0.37	0.01	0.40	0.01	-0.04	0.00	0.000	0.000
	0.92	0.63	0.59	0.01	0.63	0.01	-0.03	0.00	0.000	0.000
B	-0.92		-0.07	0.02	-0.11	0.00	0.05	0.02	0.000	0.000
	-0.71		-0.08	0.03	-0.11	0.01	0.03	0.03	0.000	0.000
	-0.50		-0.10	0.00	-0.10	0.00	0.00	0.00	0.008	0.001
	-0.25		-0.06	0.01	-0.06	0.01	0.00	0.00	0.000	0.000
	0.00		0.00	0.01	0.00	0.01	0.00	0.00	0.623	0.962
	0.25		0.09	0.02	0.09	0.02	0.00	0.00	0.943	0.817
	0.50		0.24	0.02	0.24	0.02	0.00	0.00	0.111	0.334
	0.71		0.27	0.02	0.27	0.02	0.00	0.00	0.000	0.000
	0.92		0.54	0.02	0.53	0.02	0.01	0.00	0.000	0.000

8.2 The Ising model

The second model to be considered in evaluating the new partial correlation is the Ising model, an exponential model for multivariate binary variables (Holland, et al. 1981). In general, the Ising model for $\left(X_1, X_2, ..., X_p\right)$ is

$$f\left(X_1, X_2, ..., X_p\right) = \frac{1}{Z(\Theta)} \exp\left(\sum_{i=1}^{p} \theta_{ii} X_i + \sum_{i<j} \theta_{ij} X_i X_j\right)$$

where $Z(\Theta)$ is a normalizing constant such that the distribution sums to one. Only pairwise interaction effects can be considered, since higher order interactions can be converted to pairwise ones through the introduction of additional variables (Ravikumar, et al. 2010). The Ising model can be reduced to a logistic model (Wasserman, et al. 1996). For example, consider two binary variables *X* and *Y*. Their bivariate distribution can be expressed as an Ising model.

$$f(X,Y) = \frac{1}{Z(\Theta)} \exp\left(\theta_{XX} X + \theta_{YY} Y + \theta_{XY} XY\right)$$

If we derive the conditional distributions of *X* given *Y*, we can obtain the logistic model for each one.

$$\begin{aligned} P(X = x \mid Y = y) &= \frac{f(X=x,Y=y)}{\sum_x f(X=x,Y=y)} \\ &= \frac{\frac{1}{Z(\Theta)} \exp\left(\theta_{XX} x + \theta_{YY} y + \theta_{XY} xy\right)}{\frac{1}{Z(\Theta)} \exp\left(\theta_{YY} y\right) + \frac{1}{Z(\Theta)} \exp\left(\theta_{XX} + \theta_{YY} y + \theta_{XY} y\right)} \\ &= \frac{\exp\left(\theta_{XX} x + \theta_{XY} xy\right)}{1 + \exp\left(\theta_{XX} + \theta_{XY} y\right)} \end{aligned}$$

$$\begin{aligned} \ln\left[\frac{P(X=1|Y=y)}{P(X=0|Y=y)}\right] &= \ln\left[\frac{\frac{\exp(\theta_{XX}+\theta_{XY} y)}{1+\exp(\theta_{XX}+\theta_{XY} y)}}{\frac{1}{1+\exp(\theta_{XX}+\theta_{XY} y)}}\right] \\ &= \ln\left[\exp\left(\theta_{XX} + \theta_{XY} y\right)\right] \\ &= \theta_{XX} + \theta_{XY} y \end{aligned}$$

And symmetrically, we obtain analogous results for the conditional distribution of *Y* given *X*.

$$\begin{aligned} \ln\left[\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right] &= \ln\left[\frac{\frac{\exp(\theta_{YY}+\theta_{XY} x)}{1+\exp(\theta_{YY}+\theta_{XY} x)}}{\frac{1}{1+\exp(\theta_{YY}+\theta_{XY} x)}}\right] \\ &= \ln\left[\exp\left(\theta_{YY} + \theta_{XY} x\right)\right] \\ &= \theta_{YY} + \theta_{XY} x \end{aligned}$$

We note that the intercept in each model is the corresponding parameter in the main effect of the bivariate model. In particular, the slope coefficient, or the effect of the predictor on the outcome, is the interaction term in the bivariate model. Hence, the advantage of using the Ising model is the symmetric property of the interaction term. In other words, regardless of which variable is the conditioned variable, its effect will always be the same.

Consider three binary variables *X*, *Y*, *Z*. Their multivariate distribution as an Ising model is

$$f(X,Y,Z) = \frac{1}{Z(\Theta)} \exp\left(\theta_{XX} X + \theta_{YY} Y + \theta_{ZZ} Z + \theta_{XY} XY + \theta_{XZ} XZ + \theta_{YZ} YZ\right)$$

From this model, a conditional correlation measure can be derived. The marginal distribution of *Z* is

$$\begin{aligned} f(Z) &= \sum_x \sum_y \frac{1}{Z(\Theta)} \exp\left(\theta_{XX} x + \theta_{YY} y + \theta_{ZZ} Z + \theta_{XY} xy + \theta_{XZ} xZ + \theta_{YZ} yZ\right) \\ &= \frac{1}{Z(\Theta)} \exp\left(\theta_{ZZ} Z\right) + \frac{1}{Z(\Theta)} \exp\left(\theta_{XX} + \theta_{ZZ} Z + \theta_{XZ} Z\right) + \frac{1}{Z(\Theta)} \exp\left(\theta_{YY} + \theta_{ZZ} Z + \theta_{YZ} Z\right) \\ &\quad + \frac{1}{Z(\Theta)} \exp\left(\theta_{XX} + \theta_{YY} + \theta_{ZZ} Z + \theta_{XY} + \theta_{XZ} Z + \theta_{YZ} Z\right) \\ &= \frac{1}{Z(\Theta)} \exp\left(\theta_{ZZ} Z\right)\left[1 + \exp\left(\theta_{XX} + \theta_{XZ} Z\right) + \exp\left(\theta_{YY} + \theta_{YZ} Z\right) + \exp\left(\theta_{XX} + \theta_{YY} + \theta_{XY} + \theta_{XZ} Z + \theta_{YZ} Z\right)\right] \end{aligned}$$

The conditional bivariate distribution of *X*, *Y* given *Z* is

$$\begin{aligned} f(X,Y \mid Z) &= \frac{f(X,Y,Z)}{f(Z)} \\ &= \frac{\frac{1}{Z(\Theta)} \exp\left(\theta_{XX} X + \theta_{YY} Y + \theta_{ZZ} Z + \theta_{XY} XY + \theta_{XZ} XZ + \theta_{YZ} YZ\right)}{\frac{1}{Z(\Theta)} \exp\left(\theta_{ZZ} Z\right)\left[1 + \exp\left(\theta_{XX} + \theta_{XZ} Z\right) + \exp\left(\theta_{YY} + \theta_{YZ} Z\right) + \exp\left(\theta_{XX} + \theta_{YY} + \theta_{XY} + \theta_{XZ} Z + \theta_{YZ} Z\right)\right]} \\ &= \frac{\exp\left(\theta_{XX} X + \theta_{YY} Y + \theta_{XY} XY + \theta_{YZ} XZ + \theta_{YZ} YZ\right)}{1 + \exp\left(\theta_{XX} + \theta_{XZ} Z\right) + \exp\left(\theta_{YY} + \theta_{YZ} Z\right) + \exp\left(\theta_{XX} + \theta_{YY} + \theta_{XY} + \theta_{XZ} Z + \theta_{YZ} Z\right)} \end{aligned}$$

The marginal bivariate distribution of X, Z

$$\begin{aligned} f(X,Z) &= \sum_y f(X, Y = y, Z) \\ &= \frac{1}{Z(\Theta)}\left[\exp\left(\theta_{XX} X + \theta_{ZZ} Z + \theta_{XZ} XZ\right) + \exp\left(\theta_{XX} X + \theta_{YY} + \theta_{ZZ} Z + \theta_{XY} X + \theta_{XZ} XZ + \theta_{YZ} Z\right)\right] \\ &= \frac{1}{Z(\Theta)} \exp\left(\theta_{XX} X + \theta_{ZZ} Z + \theta_{XZ} XZ\right)\left[1 + \exp\left(\theta_{YY} + \theta_{XY} X + \theta_{YZ} Z\right)\right] \end{aligned}$$

The conditional distribution of *X* given *Z*

$$\begin{aligned} f(X \mid Z) &= \frac{f(X,Z)}{f(Z)} \\ &= \frac{\frac{1}{Z(\Theta)} \exp\left(\theta_{XX} X + \theta_{ZZ} Z + \theta_{XZ} XZ\right)\left[1 + \exp\left(\theta_{YY} + \theta_{XY} X + \theta_{YZ} Z\right)\right]}{\frac{1}{Z(\Theta)} \exp\left(\theta_{ZZ} Z\right)\left[1 + \exp\left(\theta_{XX} + \theta_{XZ} Z\right) + \exp\left(\theta_{YY} + \theta_{YZ} Z\right) + \exp\left(\theta_{XX} + \theta_{YY} + \theta_{XY} + \theta_{XZ} Z + \theta_{YZ} Z\right)\right]} \\ &= \frac{\exp\left(\theta_{XX} X + \theta_{XZ} XZ\right)\left[1 + \exp\left(\theta_{YY} + \theta_{XY} X + \theta_{YZ} Z\right)\right]}{1 + \exp\left(\theta_{XX} + \theta_{XZ} Z\right) + \exp\left(\theta_{YY} + \theta_{YZ} Z\right) + \exp\left(\theta_{XX} + \theta_{YY} + \theta_{XY} + \theta_{XZ} Z + \theta_{YZ} Z\right)} \end{aligned}$$

$$\mu_{X|Z} = f\left(X=1|Z\right)$$
$$= \frac{\exp\left(\theta_{XX}+\theta_{XZ}Z\right)\left[1+\exp\left(\theta_{YY}+\theta_{XY}+\theta_{YZ}Z\right)\right]}{1+\exp\left(\theta_{XX}+\theta_{XZ}Z\right)+\exp\left(\theta_{YY}+\theta_{YZ}Z\right)+\exp\left(\theta_{XX}+\theta_{YY}+\theta_{XY}+\theta_{XZ}Z+\theta_{YZ}Z\right)}$$
$$\sigma^2_{X|Z} = f\left(X=1|Z\right)\left[1-f\left(X=1|Z\right)\right] = f\left(X=1|Z\right)f\left(X=0|Z\right)$$
$$= \frac{\exp\left(\theta_{XX}+\theta_{XZ}Z\right)\left[1+\exp\left(\theta_{YY}+\theta_{XY}+\theta_{YZ}Z\right)\right]\left[1+\exp\left(\theta_{YY}+\theta_{YZ}Z\right)\right]}{\left[1+\exp\left(\theta_{XX}+\theta_{XZ}Z\right)+\exp\left(\theta_{YY}+\theta_{YZ}Z\right)+\exp\left(\theta_{XX}+\theta_{YY}+\theta_{XY}+\theta_{XZ}Z+\theta_{YZ}Z\right)\right]^2}$$

By symmetry

$$f\left(Y,Z\right) = \tfrac{1}{Z(\Theta)}\exp\left(\theta_{YY}Y+\theta_{ZZ}Z+\theta_{YZ}YZ\right)\left[1+\exp\left(\theta_{XX}+\theta_{XY}Y+\theta_{XZ}Z\right)\right]$$

$$f\left(Y|Z\right) = \frac{\exp\left(\theta_{YY}Y+\theta_{YZ}YZ\right)\left[1+\exp\left(\theta_{XX}+\theta_{XY}Y+\theta_{XZ}Z\right)\right]}{1+\exp\left(\theta_{XX}+\theta_{XZ}Z\right)+\exp\left(\theta_{YY}+\theta_{YZ}Z\right)+\exp\left(\theta_{XX}+\theta_{YY}+\theta_{XY}+\theta_{XZ}Z+\theta_{YZ}Z\right)}$$

$$\mu_{Y|Z} = \frac{\exp\left(\theta_{YY}+\theta_{YZ}Z\right)\left[1+\exp\left(\theta_{XX}+\theta_{XY}+\theta_{XZ}Z\right)\right]}{1+\exp\left(\theta_{XX}+\theta_{XZ}Z\right)+\exp\left(\theta_{YY}+\theta_{YZ}Z\right)+\exp\left(\theta_{XX}+\theta_{YY}+\theta_{XY}+\theta_{XZ}Z+\theta_{YZ}Z\right)}$$
$$\sigma^2_{Y|Z} = \frac{\exp\left(\theta_{YY}+\theta_{YZ}Z\right)\left[1+\exp\left(\theta_{XX}+\theta_{XY}+\theta_{XZ}Z\right)\right]\left[1+\exp\left(\theta_{XX}+\theta_{XZ}Z\right)\right]}{\left[1+\exp\left(\theta_{XX}+\theta_{XZ}Z\right)+\exp\left(\theta_{YY}+\theta_{YZ}Z\right)+\exp\left(\theta_{XX}+\theta_{YY}+\theta_{XY}+\theta_{XZ}Z+\theta_{YZ}Z\right)\right]^2}$$

The conditional covariance of *X* and *Y* given *Z*

$$\sigma_{XY|Z} = E\left[XY|Z\right]-\mu_{X|Z}\mu_{Y|Z}$$
$$= f\left(X=1,Y=1|Z\right)-\mu_{X|Z}\mu_{Y|Z}$$
$$= \frac{\exp\left(\theta_{XX}+\theta_{YY}+\theta_{XZ}Z+\theta_{YZ}Z\right)\left[\exp\left(\theta_{XY}\right)-1\right]}{\left[1+\exp\left(\theta_{XX}+\theta_{XZ}Z\right)+\exp\left(\theta_{YY}+\theta_{YZ}Z\right)+\exp\left(\theta_{XX}+\theta_{YY}+\theta_{XY}+\theta_{XZ}Z+\theta_{YZ}Z\right)\right]^2}$$

Hence, the conditional correlation between *X* and *Y* given *Z*

$$\rho_{XY|Z} = \frac{\sigma_{XY|Z}}{\sigma_{X|Z}\sigma_{Y|Z}}$$
$$= \frac{\sqrt{\exp\left(\theta_{XX}+\theta_{YY}+\theta_{XZ}Z+\theta_{YZ}Z\right)}\left[\exp\left(\theta_{XY}\right)-1\right]}{\sqrt{\left[1+\exp\left(\theta_{XX}+\theta_{XY}+\theta_{XZ}Z\right)\right]\left[1+\exp\left(\theta_{YY}+\theta_{XY}+\theta_{YZ}Z\right)\right]}\sqrt{\left[1+\exp\left(\theta_{XX}+\theta_{XZ}Z\right)\right]\left[1+\exp\left(\theta_{YY}+\theta_{YZ}Z\right)\right]}}$$

If $\theta_{XY} = 0$ then $\rho_{XY|Z} = 0$ and *X* and *Y* are independent given *Z*. On the other hand, if $\theta_{XY} < 0$ then $\rho_{XY|Z} < 0$ and *X* and *Y* are negatively correlated when conditioned on *Z*. If $\theta_{XY} > 0$ then $\rho_{XY|Z} < 0$ and *X* and *Y* are positively correlated when conditioned on *Z*. This property is true no matter the number of variables in the distribution. In other words, each interaction parameter represents the relationship between two variables conditioned all other variables in the model; when it is equal to zero, the two variables involved are conditionally independent given all other variables. Note however, that actual value of the conditional correlation is dependent on the controlled variable *Z*, so the conditional correlation cannot be equal to the partial correlation. Regardless, we still know when the conditional correlation is equal to zero, so we can examine the performance of the methods under such conditions.

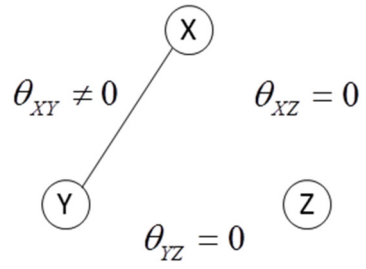
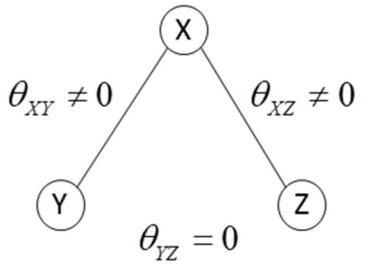
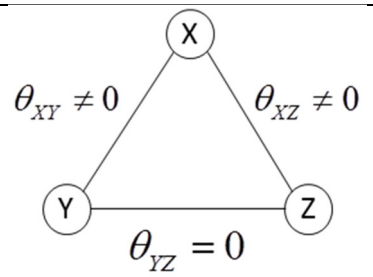
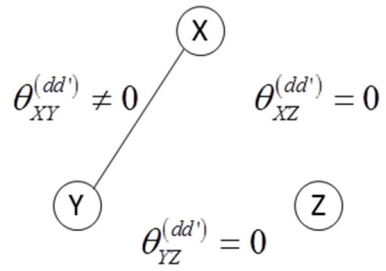
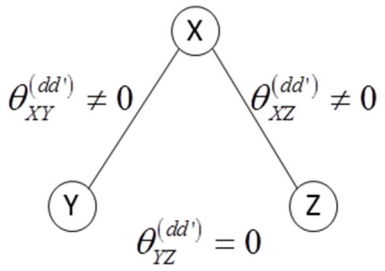
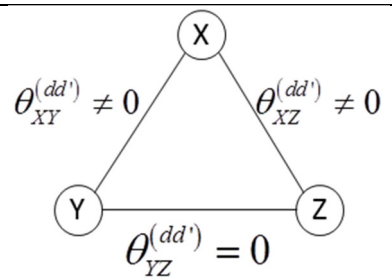
A modified form of the Ising model to more than two categories uses dummy variables to code the multi-categorical data (Guo, et al. 2010). Let $X_{1,}X_2,..,X_p$ be *p* categorical variables with $X_j, 1 \le j \le p$ belonging to one of D_j categories denoted by the set $\{1,2,...,D_j\}$. Denote $X_j^{(1)},...,X_j^{(D_j-1)}$ as the dummy variables associated with X_j, i.e. $X_j^{(d)} = I\left(X_j = d\right), 1 \le d \le D_j - 1$. Then the joint distribution of $X_{1,}X_2,..,X_p$ as an Ising model is

$$f\left(X_1,...,X_p\right) = \frac{1}{Z(\Theta)} \exp\left(\sum_{j=1}^{p} \sum_{d=1}^{D_j-1} \left(\theta_{jj}^{(d)} X_j^{(d)} + \sum_{k \ne j} \sum_{d'=1}^{D_j-1} \theta_{jk}^{(dd')} X_j^{(d)} X_k^{(d')} \right) \right)$$

Similar to the dichotomous case, if the interaction effect $\theta_{jk}^{(dd')} = 0$ for all *d* and *d'*, then X_j and X_k are conditionally independent given all other variables, and hence their nodes in a network diagram would not be connected.

The simulation study based on the Ising model for three variables, *X*, *Y*, and *Z*, was designed similar to that of Guo, et al. (2010). 1000 datasets with $n = \{100, 200, 500\}$ observations plus an additional 50 datasets of 5000 observations were simulated for three scenarios (Table 35): (A&D) *X* and *Y* are conditional dependent, all other relationships are conditionally independent, (B&E) *X* and *Y*, *X* and Z are conditionally dependent, *Y* and *Z* are conditionally independent, and (C&F) all pairs are conditionally dependent.

Table 35 Ising Model based simulation settings. Parameter values for the Ising model and their corresponding network structure. Scenarios A-C are when *X, Y, Z* are dichotomous variables. Scenarios D-F are when *X, Y*, Z are trichotomous variables.

Dichotomous		
A	**B**	**C**
		
Trichotomous *(for all d, d')*		
D	**E**	**F**
		

All main effects were also nonzero. To start, all nonzero parameters were fixed to one (scenarios A1-F1). Due to the additive nature of the distribution function and all nonzero thetas are fixed to be equal to one, the bulk of the distribution is concentrated where all variables are equal to one. The fixed theta was then adjusted to 0.5 to try to mitigate this effect and create a more balanced distribution (A2-F2). To be able to further generalized our results, we followed previous simulation study designs (Guo, et al. 2010) and allowed all nonzero parameters to be randomly selected from a uniform distribution in the domain $(-1,-0.5)\cup(0.5,1)$ (A3-F3). In this way, the analysis of the performance the new partial correlation will not be affected by any special cases of the joint distribution.

From the joint distribution specified in the Ising model, conditional distributions were derived and used in a Gibbs sampling procedure (Casella, et al. 1992, Hogg, et al.).

Gibbs Sampling Algorithm:

Step 0: Generate initial values for X, Y and Z from marginal distributions derived from Ising model.

Step 1: Generate first observation of X given zero-th observations of Y and Z from its conditional distribution. Generate first observation of Y given zero-th observations of X and Z from its conditional distribution. Generate first observation of Z given zero-th observations of X and Z from its conditional distribution.

…

Step *k*: Generate k-th observation of X given (k-1)-th observations of Y and Z from its conditional distribution. Generate k-th observation of Y given (k-1)-th observations of X and Z from its conditional distribution. Generate k-th observation of Z given (k-1)-th observations of X and Z from its conditional distribution.

The Gibbs sampler allows the distributions of the variables to converge to their marginal distributions as k increases. Hence, the first 10^6 observations were discarded (burn-in point). To produce independent observations, only every hundred-th observation was kept.

The following partial correlations were calculated:

- *Pcor.new* - new partial correlation for categorical data
- *Pcor.phi* - partial phi coefficient (dichotomous data)
- *Pcor.p* - partial correlation for continuous data (trichotomous data)
- *Pcor.tet* - partial tetrachoric correlation (dichotomous data)
- *Pcor.pch* - partial polychoric correlation (trichotomous data)

All results are provided in the Supplementary materials. Results for n=500 are shown here. Again, the goal is to be able to detect and measure the partial correlation between X and Y after controlling for Z. The major disadvantage of the Ising model is that the conditional partial correlation is dependent on the value of the Z, as shown above, so the true partial correlation is unknown. Hence evaluating the estimated partial correlation based on bias and mean square error is impossible. However, the consistency of the estimates can still be checked across different sample sizes and between different methods. Furthermore, the power and specificity (true negative) of the three methods can be measured using the tests as in the case of the multivariate normal distribution, for lack of an alternative test in all

scenarios. However, a simulation on the distribution of the test statistics used here will be studied in the future.

With regards to the estimates, in the fixed theta scenarios with binary data (A1-C1, A2-C2), the new partial correlation and the partial phi coefficient are identical (Table 36). The partial tetrachoric correlation is double those estimates. The joint distribution of these settings place much more density on the 1 category than the 0 categories; in other words, the distribution is skewed for A1-C1, A2-C2. Based on our findings from the multivariate normal model, the partial tetrachoric is likely to be the bias estimate in this case.

In the trichotomous scenarios, it is the new partial correlation that is able to estimate a value that is far from nonzero. Note that the estimate decreases together with the decrease in the theta value. On the other hand, the partial correlation (*pcor.p*) and partial polychoric correlation increase, which is counterintuitive to what we would expect.

Table 36 Ising Model based simulation - Means and Standard Errors. Scenarios A1-F1: nonzero parameters fixed to one; A2-F2: nonzero parameters fixed to 0.5; A3-F3: nonzero parameters randomly generated from uniform distribution in the domain $(-1,-0.5)\cup(0.5,1)$. Refer to **Table 35**.

	pcor.new		pcor.phi (pcor.p)		pcor.tet (pcor.pch)	
Scenario	Mean	SE	Mean	SE	Mean	SE
A1	0.15	0.05	0.15	0.05	0.30	0.10
B1	0.11	0.06	0.11	0.06	0.27	0.13
C1	0.08	0.05	0.08	0.06	0.20	0.34
D1	0.15	0.06	0.04	0.07	0.05	0.10
E1	0.17	0.08	0.02	0.08	0.03	0.12
F1	0.30	0.15	0.00	0.12	0.00	0.17
A2	0.11	0.05	0.11	0.05	0.18	0.08
B2	0.10	0.05	0.10	0.05	0.18	0.09
C2	0.08	0.04	0.08	0.05	0.14	0.10
D2	0.11	0.04	0.04	0.05	0.05	0.06
E2	0.11	0.04	0.04	0.05	0.04	0.07
F2	0.12	0.05	0.03	0.06	0.04	0.07
A3	0.16	0.06	0.01	0.17	0.01	0.28
B3	0.15	0.05	0.00	0.16	0.00	0.28
C3	0.15	0.06	0.00	0.16	0.00	0.28
D3	0.26	0.09	0.02	0.12	0.03	0.17
E3	0.28	0.10	0.02	0.12	0.03	0.17
F3	0.32	0.13	0.01	0.12	0.01	0.17

Power for the partial correlation measures is low in the fixed theta scenarios (Table 37), but this may be due to the specific distribution resulting from fixing all the thetas to the same positive value (1 or 0.5). The randomized thetas of A3-F3 would be more informative to observe the power of the partial correlation measures, with regards to the question of detection, regardless of effect size. In this situation, the partial phi coefficient and partial tetrachoric correlation outperforms the new partial correlation when the data are binary. When the data are multi-categorical, the new partial correlation outperforms the others. This can be explained by the nature of the data and the nature of the partial correlation measures. The data are truly nominal, but aside from the new partial correlation, the other measures were designed for ordered data. Hence, it has less power to detect the relationships between the nominal variables.

Since all edges are present in scenarios C and F, they are not included in Table 38. Type I error rates for the partial tetrachoric correlation is inflated throughout all scenarios.

Table 37 Ising Model based simulation – Power for each corresponding test of each partial correlation measure. Scenarios A1-F1: nonzero parameters fixed to one; A2-F2: nonzero parameters fixed to 0.5; A3-F3: nonzero parameters randomly generated from uniform distribution in the domain $(-1,-0.5)\cup(0.5,1)$. XY edge exists in all scenarios. XZ edge exists in scenarios B, C, E, F. YZ edge exists in scenarios C, F. Refer to **Table 35**.

	XY edge			XZ edge			YZ edge		
Scenario	pcor.new	pcor.phi (pcor.p)	pcor.tet (pcor.pch)	pcor.new	pcor.phi (pcor.p)	pcor.tet (pcor.pch)	pcor.new	pcor.phi (pcor.p)	pcor.tet (pcor.pch)
A1	0.85	0.85	0.98						
B1	0.65	0.66	0.93	0.68	0.68	0.93			
C1	0.38	0.42	0.80	0.41	0.43	0.79	0.41	0.43	0.80
D1	0.57	0.28	0.43						
E1	0.67	0.32	0.46	0.66	0.30	0.45			
F1	0.92	0.48	0.63	0.92	0.49	0.64	0.91	0.47	0.64
A2	0.66	0.66	0.90						
B2	0.60	0.60	0.82	0.60	0.58	0.84			
C2	0.40	0.42	0.72	0.56	0.54	0.80	0.52	0.54	0.74
D2	0.30	0.19	0.31						
E2	0.33	0.17	0.28	0.33	0.18	0.30			
F2	0.41	0.18	0.29	0.40	0.17	0.27	0.41	0.17	0.29
A3	0.88	0.89	0.97						
B3	0.88	0.88	0.97	0.87	0.87	0.97			
C3	0.86	0.87	0.96	0.86	0.86	0.96	0.85	0.85	0.94
D3	0.94	0.56	0.68						
E3	0.96	0.56	0.69	0.95	0.56	0.67			
F3	0.96	0.56	0.68	0.96	0.56	0.68	0.97	0.55	0.68

Table 38 Ising Model based simulation - Type I Error for each corresponding test for each partial correlation measure. Scenarios A1-F1: nonzero parameters fixed to one; A2-F2: nonzero parameters fixed to 0.5; A3-F3: nonzero parameters randomly generated from uniform distribution in the domain $(-1,-0.5)\cup(0.5,1)$. XZ edge nonexistent in scenarios A, D. YZ edge nonexistent in scenarios A, B, D, E. Refer to **Table 35**.

	Type I Error (XZ edge)			Type I Error (YZ edge)		
scenario	pcor.new	pcor.phi (pcor.p)	pcor.tet (pcor.pch)	pcor.new	pcor.phi (pcor.p)	pcor.tet (pcor.pch)
A1	0.06	0.05	0.39	0.05	0.06	0.39
B1				0.04	0.05	0.51
D1	0.21	0.08	0.19	0.23	0.10	0.20
E1				0.48	0.19	0.34
A2	0.06	0.04	0.24	0.08	0.06	0.32
B2				0.00	0.04	0.24
D2	0.11	0.05	0.13	0.11	0.04	0.13
E2				0.13	0.07	0.15
A3	0.05	0.05	0.27	0.05	0.05	0.26
B3				0.05	0.05	0.32
D3	0.18	0.05	0.17	0.17	0.04	0.13
E3				0.73	0.08	0.22

9 Partial correlation network analysis

Given a large set of variables one could be interested in the relationship between all possible pairs of variables while controlling for all other variables. With the proper statistic, one can a) measure the strength of the relationship between any two variables while controlling for all other network variables and b) test whether the relationship is significant given some assumptions about the distribution of the data. From such information, an overall structural network can be constructed. Common hypothesis-driven data modeling tools for such analysis include Structural Equation Modeling (SEM) and Dynamic Causal Modeling (DCM). However without any prior knowledge about the structure of the network, such methods would be inappropriate. Partial Correlation Network Analysis (PCNA) is a purely data-driven analysis approach that does not require a priori information (Fransson, et al. 2008).

Partial correlation network analysis (PCNA) generates an undirected graph, $G = \{V, E\}$; V represents a set of nodes, or variables, and E represents a set of edges that convey the conditional relationships between pairs of nodes. If an edge does not exist between two nodes, then the two nodes are conditionally independent given all other nodes in the network. As mentioned in Chapter 6, the conditional correlation is equal to the partial correlation for multivariate normal distributions; hence, the existence of an edge can be determined by the significance of the partial correlation; if the partial correlation is significantly nonzero, then the edge will appear between the two nodes. Correspondingly, the strength of that relationship can be measure by the magnitude of the partial correlation.

Partial correlation analysis in its continuous form has been used in brain imagining analysis and genetic studies (De La Fuente, et al. 2004, Marrelec, et al. 2006, Marrelec, et al. 2009). However, such application in genetic studies is unsuitable due to the categorical nature of the data. Hence, the new partial correlation measure developed in Chapter 7 provides a suitable alternative.

In this chapter, we consider covariate in network analysis and propose an extension of the two-level regression that was previously developed in the continuous case of PCNA to the categorical PCNA based on the new partial correlation.

9.1 Covariate partial correlation network analysis: Two-level regression

Suppose are interested on the effect of a covariate $G=\{0,1\}$ on the partial correlation between *X* and *Y* after controlling for *Z*. Pradhan (2009) proposed a method to analyze such a situation that draws on the fact that testing the correlation between two variables being equal to zero is equivalent to testing the slope coefficient in a regression between the two variables being equal to zero. Consider the continuous partial correlation. First residuals are obtained from linear regression. The test of the correlation between the two residuals equal to zero would be equivalent to a test on the coefficient b_1 equal to zero.

$$Y = \beta_0 + \beta_1 Z + \varepsilon_Y \rightarrow \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 Z \rightarrow e_Y = Y - \hat{Y}$$
$$X = \gamma_0 + \gamma_1 Z + \varepsilon_X \rightarrow \hat{X} = \hat{\gamma}_0 + \hat{\gamma}_1 Z \rightarrow e_X = X - \hat{X}$$
$$e_Y = b_0 + b_1 e_X + \varepsilon$$

We integrate *G* into the coefficient b_1 and rewrite the regression model.

$$b_1 = a_0 + a_1 G$$
$$e_Y = b_0 + (a_0 + a_1 G) e_X + \varepsilon = b_0 + a_0 e_X + a_1 e_X G + \varepsilon$$

Thus the significance of a_1 tells whether or not *G* has an effect on the partial correlation between *X* and *Y*. If a_1 is nonsignificant, then the partial correlation between *X* and *Y* for groups *G*=1 and *G*=0 are the same. If a_1 is significant, the partial correlation between *X* and *Y* for groups *G*=1 and *G*=0 are different.

9.2 Extension to categorical and mixed variables

Consider the first canonical variates U_1 and V_1 found by canonical correlation analysis on the residuals of *X* and *Y*, regardless of whether they are both categorical or if one is continuous. The canonical correlation analysis can be expressed as a linear regression

$$U_1 = b_0 + b_1 V_1 + \varepsilon$$

The test of the correlation between the two canonical variates equal to zero would be equivalent to a test on the coefficient b_1 equal to zero. We propose a two-level regression at this step incorporating G into the coefficient of V.

$$b_1 = a_0 + a_1 G$$

So we have

$$U_1 = b_0 + (a_0 + a_1 G) V_1 + \varepsilon = b_0 + a_0 V_1 + a_1 V_1 G + \varepsilon$$

Analogously, we can test a_1 to measure the effect of G on the partial correlation between X and Y. As noted by Pradhan, reversing the independent and dependent variables in the second regression will produce different results, so we applied the regression in both directions and average the P-value.

$$V_1 = b_0^* + a_0^* U_1 + a_1^* U_1 G + \varepsilon^*$$

Thus we have effectively extended two-level regression to the case of categorical variables.

9.3 Application to COGEND

We apply the new partial correlation measure to the COGEND data to construct an overall SNP network and then use two-level regression to compare SNP networks between subjects with nicotine addiction (cases) and subjects who did not have nicotine addiction by testing each edge. Results are illustrated using network plotting software Cytoscape 2.8.2 (Smoot, et al. 2011) and the heatmap.2 function in the *gplots* package in Rv2.14.1 (Team 2011) (Warnes 2011).

Figure 6 shows the overall SNP network structure, with edges appearing when FDR adjusted p-values were less than 0.05. The opacity of the edge indicates magnitude of the significance of the partial correlation; the more opaque the edge, the more significant the partial correlation.

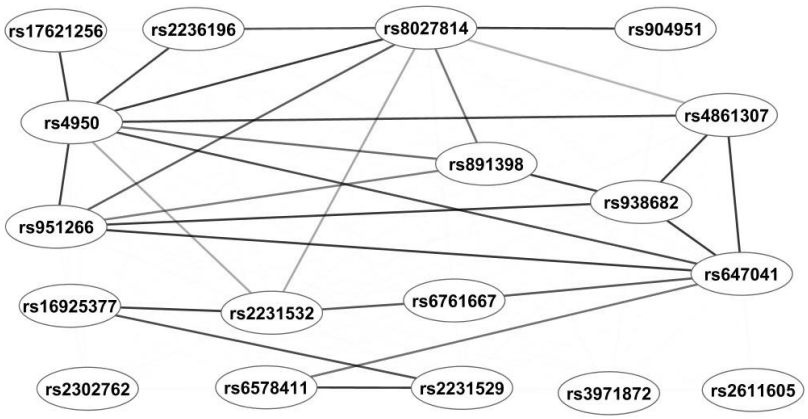


Figure 6 COGEND SNP network using the new partial correlation for categorical data (FDR=0.05). Opacity of edges is determined by significance of the partial correlation controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly nonzero the partial correlation between the two connected SNPs.

rs4950, *rs8027814*, and *rs647041* came out with the most number of edges; *rs4950* topped the three with eight significant partial correlations with other SNPs. Figure 7 shows a heat map of the partial correlations.

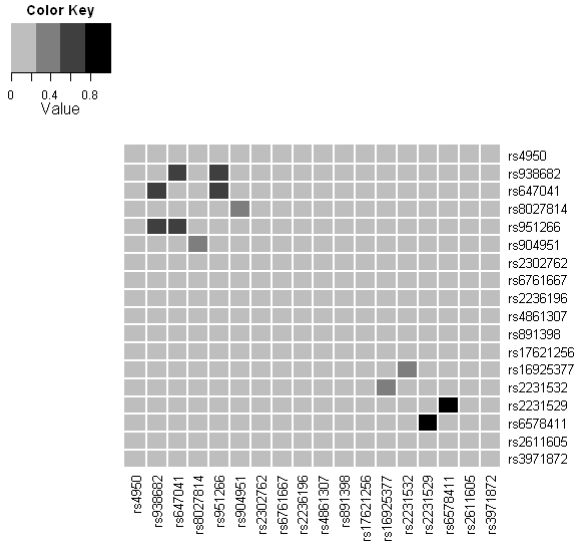
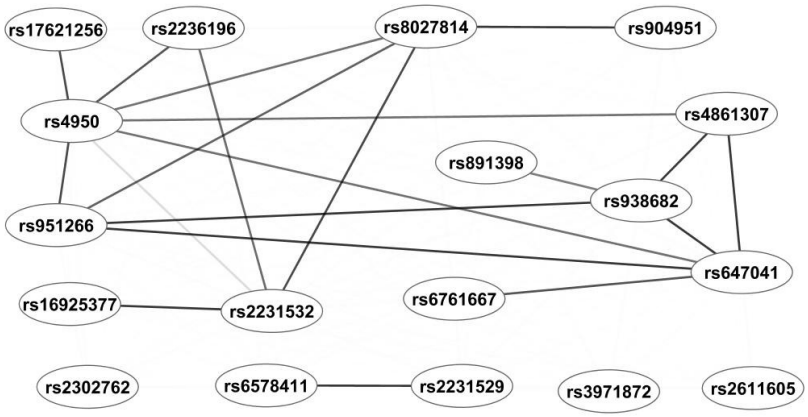


Figure 7 COGEND heat map of new partial correlations between SNPs. Measures the strength of the relationship between pairs of SNPs after controlling for all other SNPs.

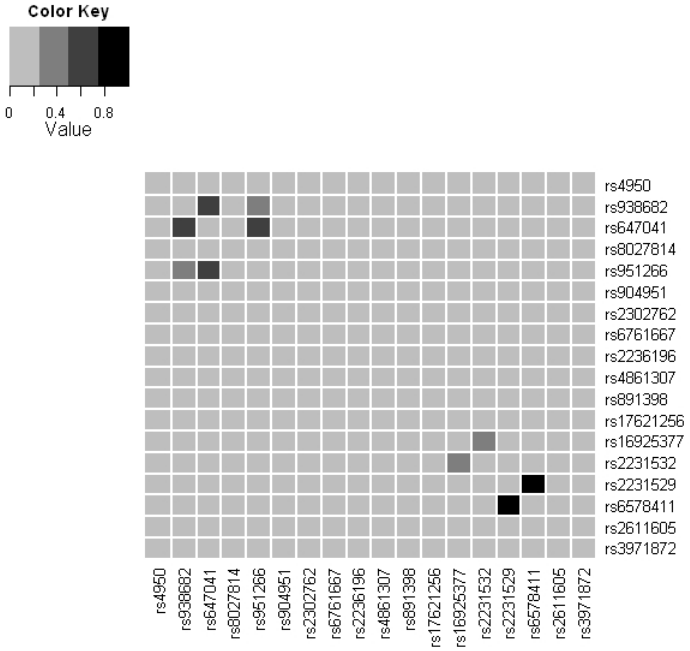
Despite having many significant partial correlations, the effect size of the partial correlations involving *rs4950* is quite low at less than 0.25. On the other hand, *rs6578411* and *rs2231529* are very highly correlated at more than 0.75. *rs647041, rs938682* and *rs951266* are all highly correlated with each other with partial correlations between 0.5 and 0.75. Partial correlations between pairs *rs2231532*, *rs16925377* and *rs8027814*, *rs904951* are at low to moderate levels between 0.25 and 0.4.

Figure 8 shows SNP network and partial correlation heat map for nicotine addicts. r*s4950* maintains its status as the SNP with most number of significant partial correlations at seven. Compared to the heat map of Figure 7 above, roughly the same partial correlations stand out, with *rs8027814*, *rs904951*disappearing. Using guidelines for the effect size of correlation, *rs2231529* and *rs6578411* are highly correlated with a partial correlation value of 0.75 or more. *rs647041* is highly correlated with *rs938682* and *rs951266* with partial correlation between 0.5 and 0.75. *rs938682* is moderately correlation with *rs951266*.

For the non-nicotine addicts, Figure 9 shows the corresponding SNP network and partial correlation heat map. While a large number of pathways are not significant for the non-nicotine addicts, the heat map shows the high correlations between the same SNPs as in the nicotine addicts

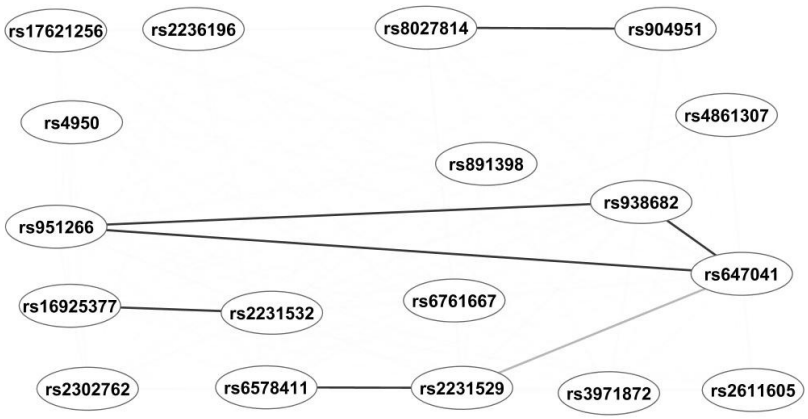


(A)

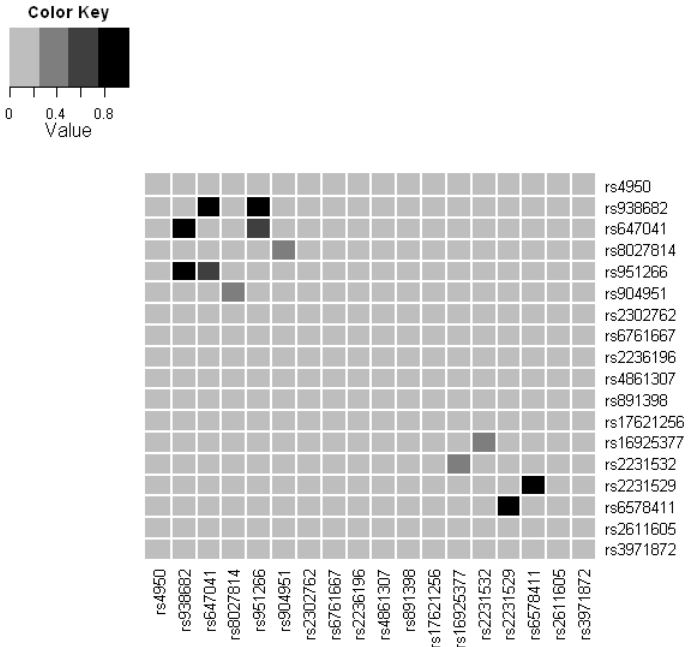


(B)

Figure 8 COGEND SNP new partial correlation (A) network and (B) heat map for nicotine addicts (FDR=0.05). Opacity of edges is determined by significance of the partial correlation controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly nonzero the partial correlation between the two connected SNPs. The heat map of new partial correlations measures the strength of the relationship between pairs of SNPs after controlling for all other SNPs.



(A)



(B)

Figure 9 COGEND SNP network and new partial correlation heat map for non-nicotine addicts (FDR=0.05). Opacity of edges is determined by significance of the partial correlation controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly nonzero the partial correlation between the two connected SNPs. The heat map of new partial correlations measures the strength of the relationship between pairs of SNPs after controlling for all other SNPs.

We apply the two-level regression to test which pathways are significantly different between the two groups by including an interaction term of one canonical variate with the nicotine addiction indicator variable in the regression model. Again, we permute the canonical variates for two regression and take the average of the two p-values.

$$U_1 = b_0 + (a_0 + a_1 \text{NicotineAddict}) V_1 + \varepsilon = b_0 + a_0 V_1 + a_1 V_1 (\text{NicotineAddict}) + \varepsilon$$
$$V_1 = b_0^* + (a_0^* + a_1^* \text{NicotineAddict}) U_1 + \varepsilon^* = b_0^* + a_0^* U_1 + a_1^* U_1 (\text{NicotineAddict}) + \varepsilon^*$$

Figure 10 Results of two-level regression - pathways that are significantly different between nicotine addicts and non-nicotine addicts. Opacity of edges is determined by significance of nicotine addiction on the partial correlation between the two SNPs controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly the effect of nicotine addiction on the partial correlation between the two connected SNPs.

The pathways between *rs951266*, *rs647041*, and *rs938682* were found to be significantly different between nicotine addicts and non-nicotine addicts, even though within each group, the pathways were significant with similar magnitudes (>0.75) (Figure 10). There are two possibilities for this finding; either the differences were truly detectable between the two groups, even though the difference is not of practical value, or the direction of the correlations are different between the two groups, but the new partial correlation is not able to output negative correlations. Either way, the two-level regression is able to detect pathways that could be further studied to provide information on the connection between the SNPs and nicotine addiction.

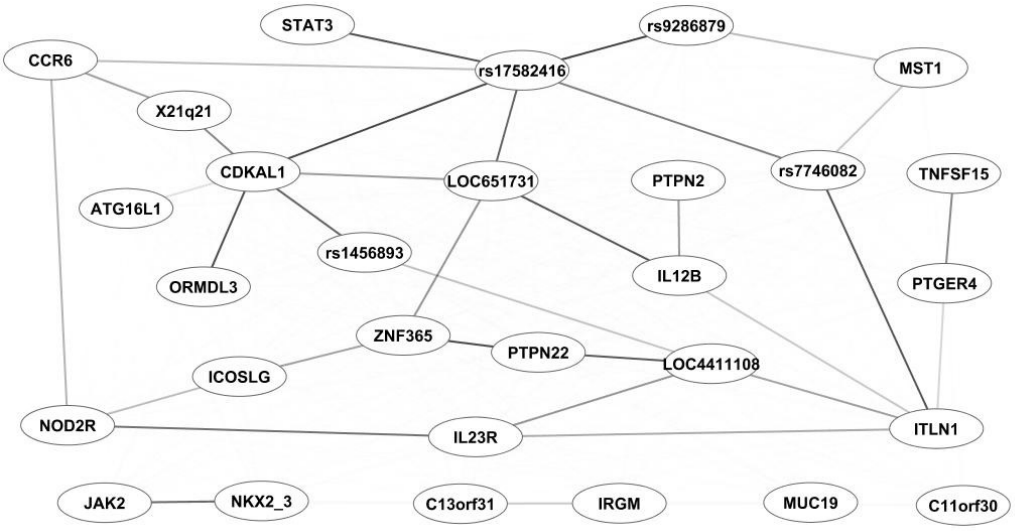
rs951266 and *rs647041* are both located in the region of gene *CHRNA5* (cholinergic receptor, nicotinic, alpha 5). *rs938682* is in gene *CHRNA3* (cholinergic receptor, nicotinic, alpha 3). All three SNPs are found in chromosome 15. The *CHRNA5- CHRNA3- CHRNB4* region has been shown to be a risk factor for age-dependent nicotine addiction (Weiss, et al. 2008) as well as a susceptibility locus for lung cancer (Hung, et al. 2008).

9.4 Application to Crohn's disease

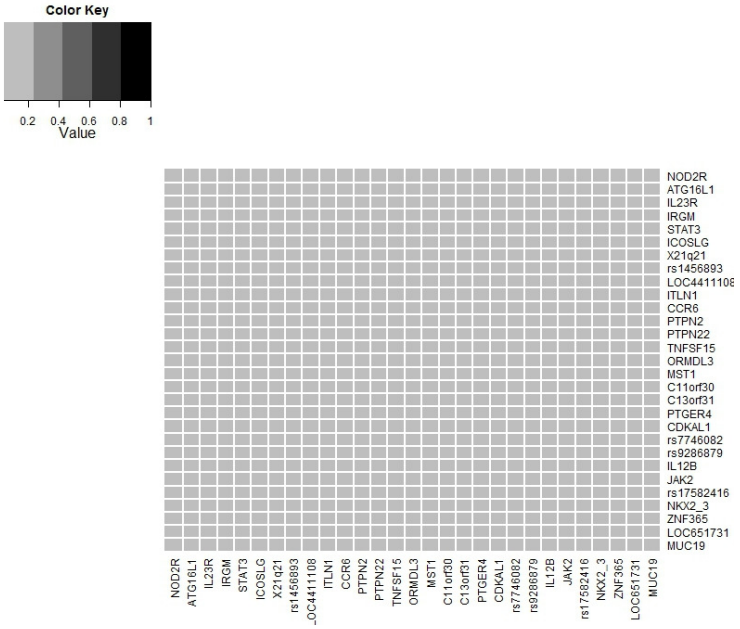
We also applied the method to the Crohn's disease data. If the data are FDR adjusted, no pathways were significant in most of the analysis. With the exception of the control group (non-ileum afflicted), the results were not adjusted for multiple testing.

The overall SNP network and corresponding heat map of partial correlations is displayed in Figure 11. Again, the opacity of an edge indicates its significance: more opaque means more significant. *CDKAL1* has the most number of pathways with connections to six other SNPs: *rs17582416*, *LOC651731*, *ATG16L1*, *rs1456893*, *ORMDL3*, and *X21q21*. Despite the complexity of the network, the actual partial correlations have small effect sizes (<0.2).

The SNP network and heat map of partial correlations for cases are displayed in **Figure 12**. *rs17582416* takes over the title of most number of connections, five, with *X21q21*, *LOC651731*, *PTGER4*, *CCR6*, *STAT3*, and *rs9286879.* But once again, the actual partial correlations have small effect sizes (<0.2).

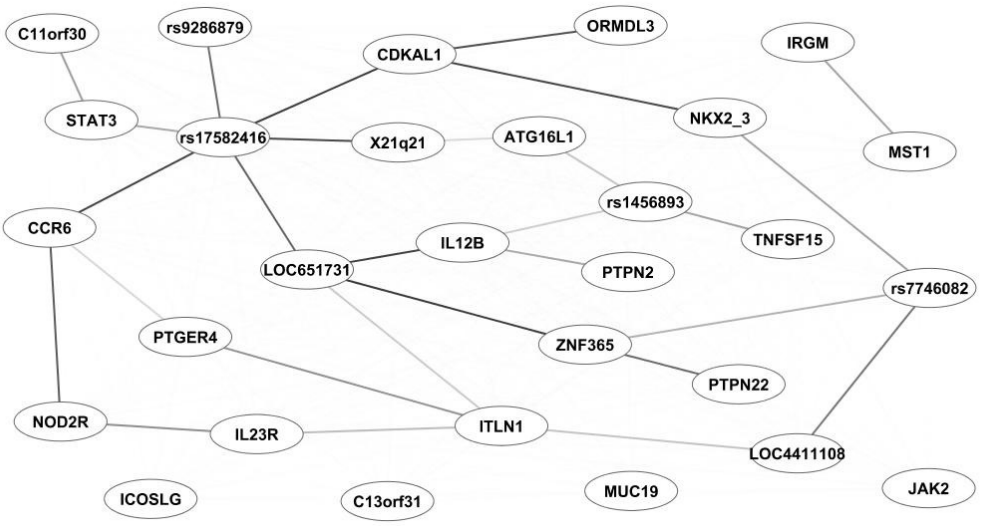


(A)

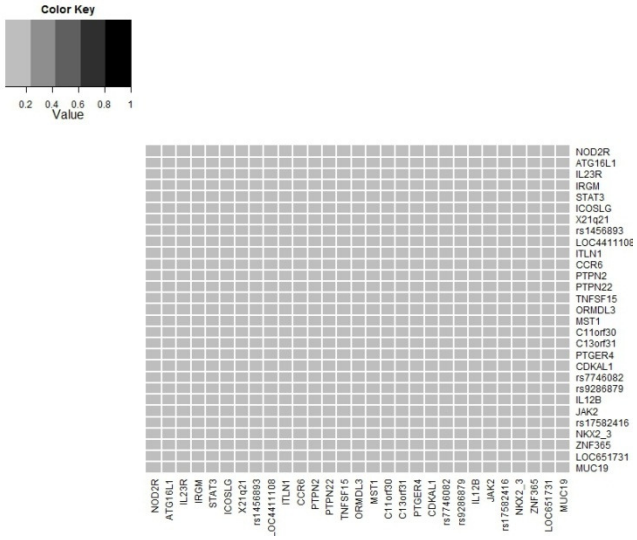


(B)

Figure 11 Crohn's disease SNP network and new partial correlation heat map. Opacity of edges is determined by significance of the partial correlation controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly nonzero the partial correlation between the two connected SNPs. The heat map of new partial correlations measures the strength of the relationship between pairs of SNPs after controlling for all other SNPs.



(A)

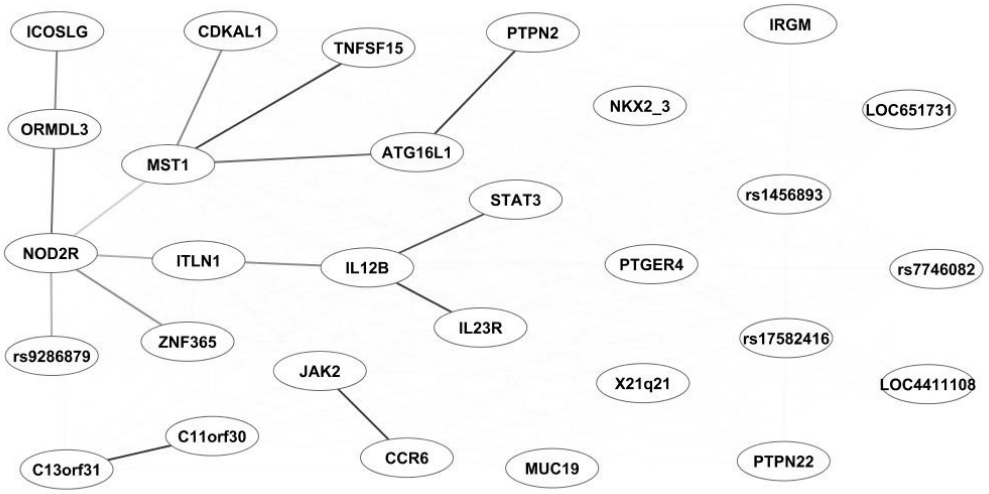


(B)

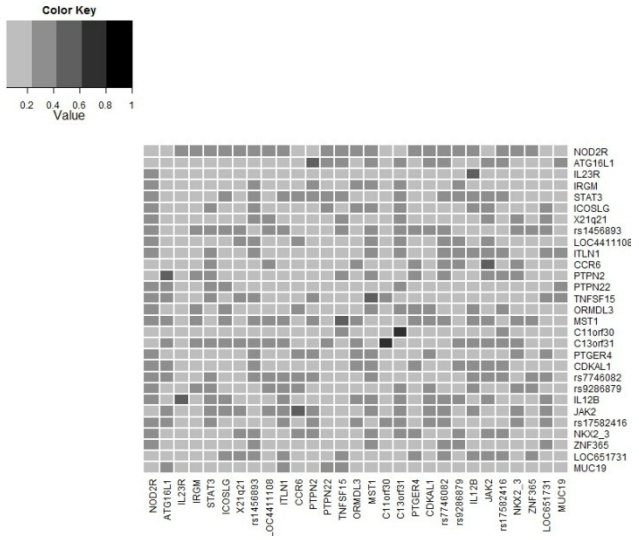
Figure 12 Crohn's disease SNP network and new partial correlation heat map for ileum afflicted subjects. Opacity of edges is determined by significance of the partial correlation controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly nonzero the partial correlation between the two connected SNPs. The heat map of new partial correlations measures the strength of the relationship between pairs of SNPs after controlling for all other SNPs.

For the non-ileum afflicted patients, we found that the unadjusted p-values detection too many pathways to be interpretable, so adjustment for multiple testing was applied with FDR=0.05. Figure 13 shows the SNP network and corresponding heat map of partial correlations. Interestingly, *NOD2R* stands

out with significant connections to *ZNF365*, *rs9286879*, *MST1*, *ORMDL3*, and *ITLN1* and moderately sized partial correlations with each one of them (>0.5).



(A)



(B)

Figure 13 Crohn's disease SNP network and new partial correlation heat map for non-ileum afflicted patients (FDR=0.05). Opacity of edges is determined by significance of the partial correlation controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly nonzero the partial correlation between the two connected SNPs. The heat map of new partial correlations measures the strength of the relationship between pairs of SNPs after controlling for all other SNPs.

To compare the networks between ileum disease patients and non-ileum disease patient, we applied two-level regression.

$$U_1 = b_0 + \left(a_0 + a_1 \text{DiseaseLocation}\right) V_1 + \varepsilon = b_0 + a_0 V_1 + a_1 V_1 \left(\text{DiseaseLocation}\right) + \varepsilon$$
$$V_1 = b_0^* + \left(a_0^* + a_1^* \text{DiseaseLocation}\right) U_1 + \varepsilon^* = b_0^* + a_0^* U_1 + a_1^* U_1 \left(\text{DiseaseLocation}\right) + \varepsilon^*$$

Results are shown in Figure 14. The connection between *NOD2R* and *rs9286879* are significantly different between the two disease locations; in the networks, the ileum afflicted group did not have this relationship.

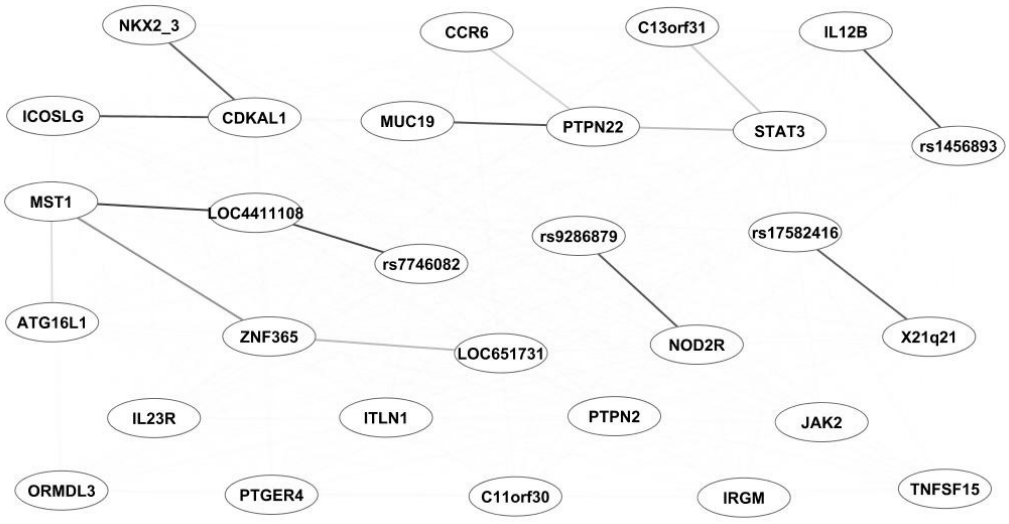
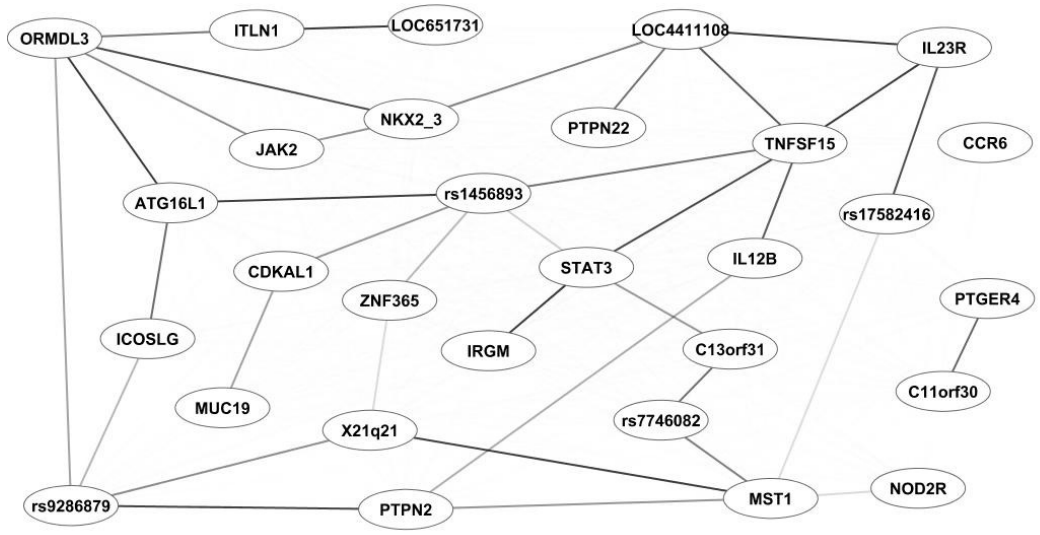
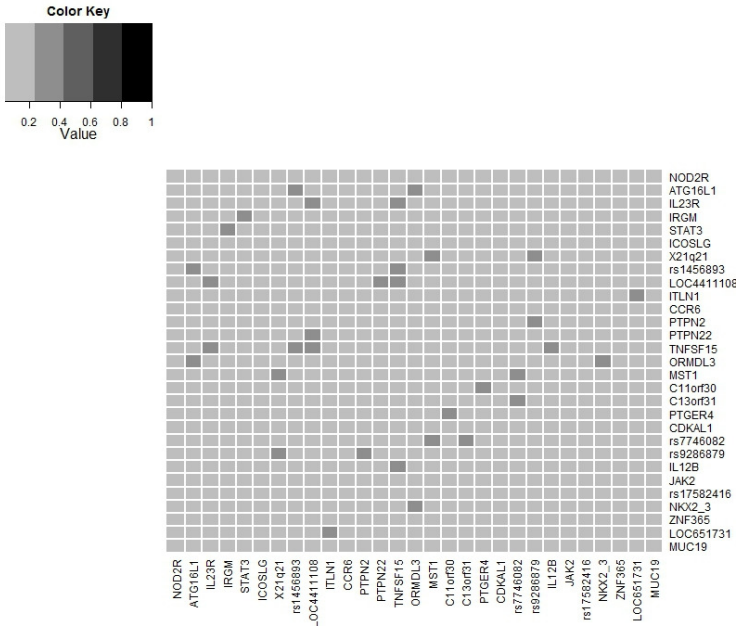


Figure 14 Results of two-level regression - pathways that are significantly different between ileum and non-ileum afflicted. Opacity of edges is determined by significance of disease location on the partial correlation between the two SNPs controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly the effect disease location on the partial correlation between the two connected SNPs.

Smoking status (never versus current/ex) was also significantly associated with disease location (p=0.004), so we stratified on smoking status and disease location to construct four partial correlation networks and their corresponding partial correlation heat maps (Figure 15, Figure 16, Figure 17, Figure 18). Within smokers, *ORMDL3* was correlated the most number of SNPs for ileum afflicted patients, while *NOD2* was correlated with the most number of SNPs for non-ileum afflicted patients. Within nonsmokers, *rs17582416* was correlated with three SNPs for ileum afflicted patients, while PTGER4 was correlated with four SNPs for non-ileum afflicted patients. The magnitude of the partial correlations for non-ileum afflicted patients was generally higher than ileum afflicted patients.

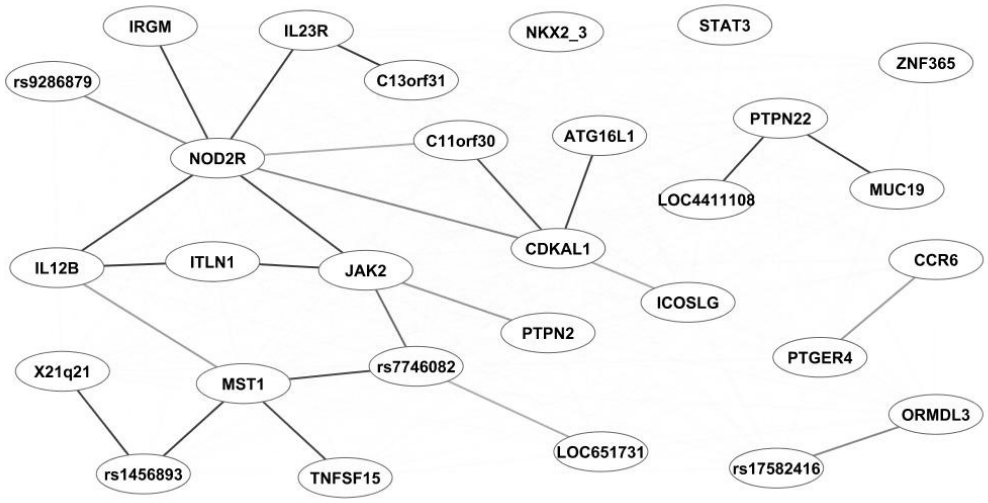


(A)

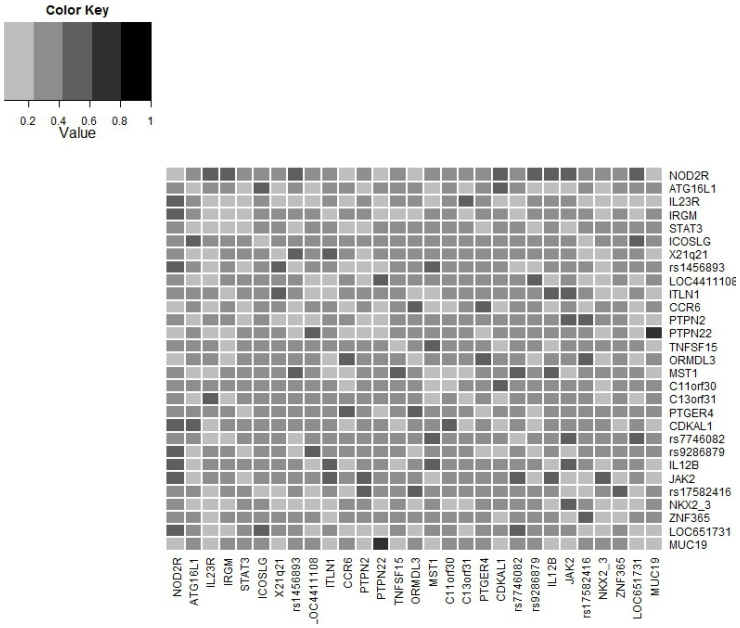


(B)

Figure 15 Crohn's disease SNP network and new partial correlation heat map for ileum afflicted patients who are smokers. Opacity of edges is determined by significance of the partial correlation controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly nonzero the partial correlation between the two connected SNPs. The heat map of new partial correlations measures the strength of the relationship between pairs of SNPs after controlling for all other SNPs.

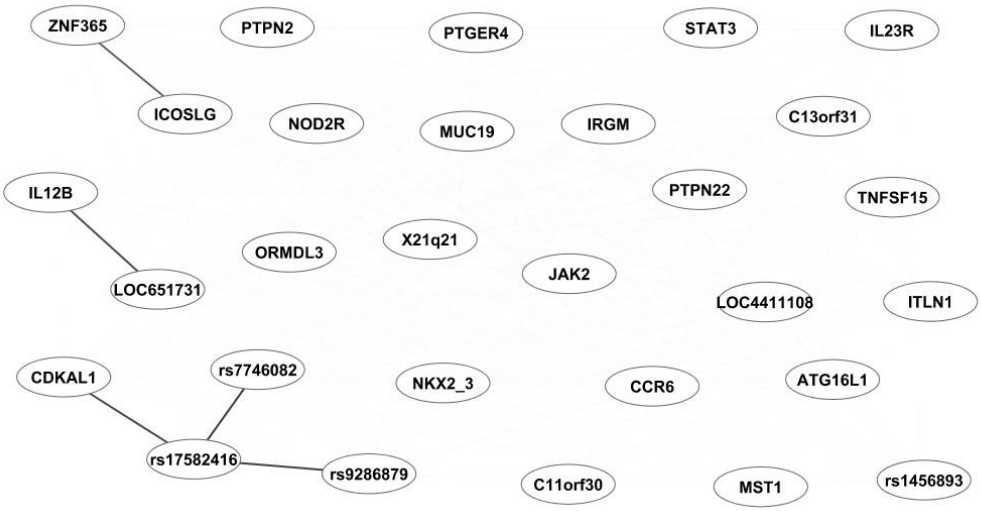


(A)

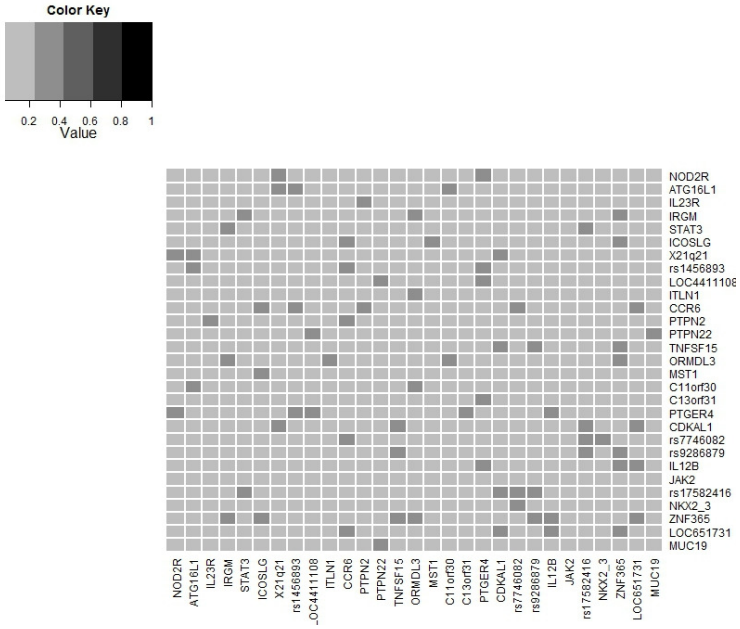


(B)

Figure 16 Crohn's disease SNP network and new partial correlation heat map for non-ileum afflicted patients who are smokers (FDR=0.05). Opacity of edges is determined by significance of the partial correlation controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly nonzero the partial correlation between the two connected SNPs. The heat map of new partial correlations measures the strength of the relationship between pairs of SNPs after controlling for all other SNPs.

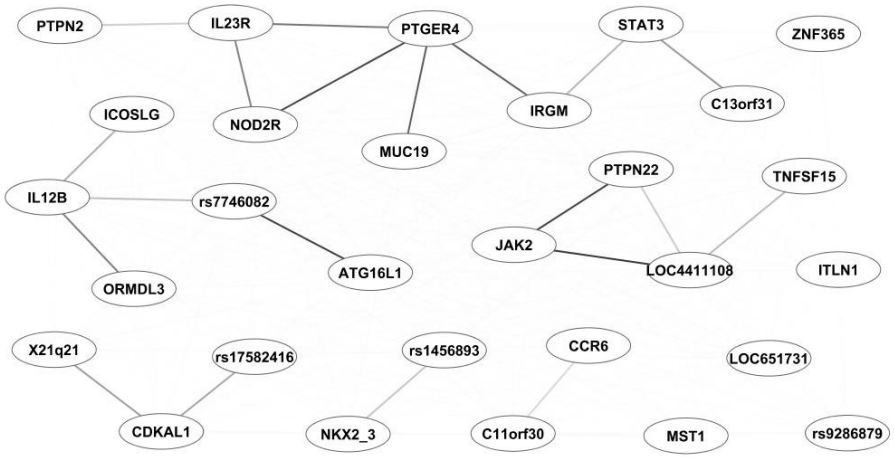


(A)

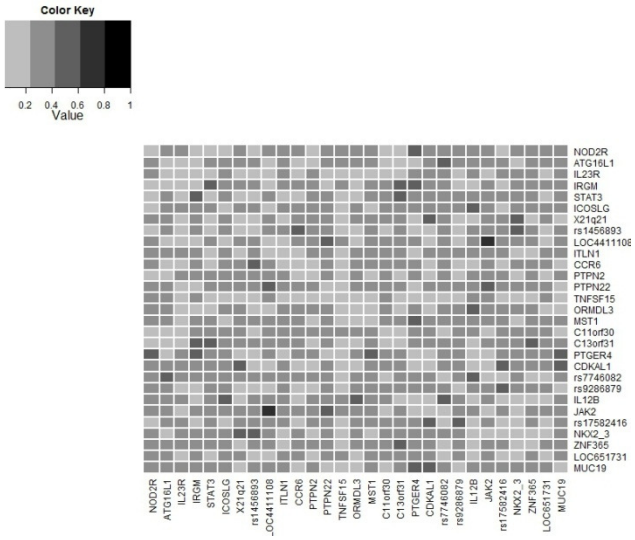


(B)

Figure 17 Crohn's disease SNP network and new partial correlation heat map for ileum afflicted patients who are nonsmokers. Opacity of edges is determined by significance of the partial correlation controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly nonzero the partial correlation between the two connected SNPs. The heat map of new partial correlations measures the strength of the relationship between pairs of SNPs after controlling for all other SNPs.



(A)



(B)

Figure 18 Crohn's disease SNP network and partial correlation heat map for non-ileum afflicted patients who are nonsmokers (FDR=0.05). Opacity of edges is determined by significance of the partial correlation controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly nonzero the partial correlation between the two connected SNPs. The heat map of new partial correlations measures the strength of the relationship between pairs of SNPs after controlling for all other SNPs.

To test for whether or not the effect of disease location on the network edges while controlling for smoking status, another two-level regression was applied. The p-values for a_1, a_1^* were averaged.

$$\begin{aligned} U_1 &= b_0 + \left(a_0 + a_1 \text{DiseaseLocation} + a_2 \text{Smoking}\right) V_1 + \varepsilon \\ &= b_0 + a_0 V_1 + a_1 V_1 \left(\text{DiseaseLocation}\right) + a_2 V_1 \left(\text{Smoking}\right) + \varepsilon \\ V_1 &= b_0^* + \left(a_0^* + a_1^* \text{DiseaseLocation} + a_2^* \text{Smoking}\right) U_1 + \varepsilon^* \\ &= b_0^* + a_0^* U_1 + a_1^* U_1 \left(\text{DiseaseLocation}\right) + a_2^* U_1 \left(\text{Smoking}\right) + \varepsilon^* \end{aligned}$$

After controlling for smoking (Figure 19), we get roughly the same pathways that are significantly different between ileum afflicted and non-ileum afflict patients, even after controlling for smoking status. Only *PTPN22-STAT3* was no longer significantly different between the two groups.

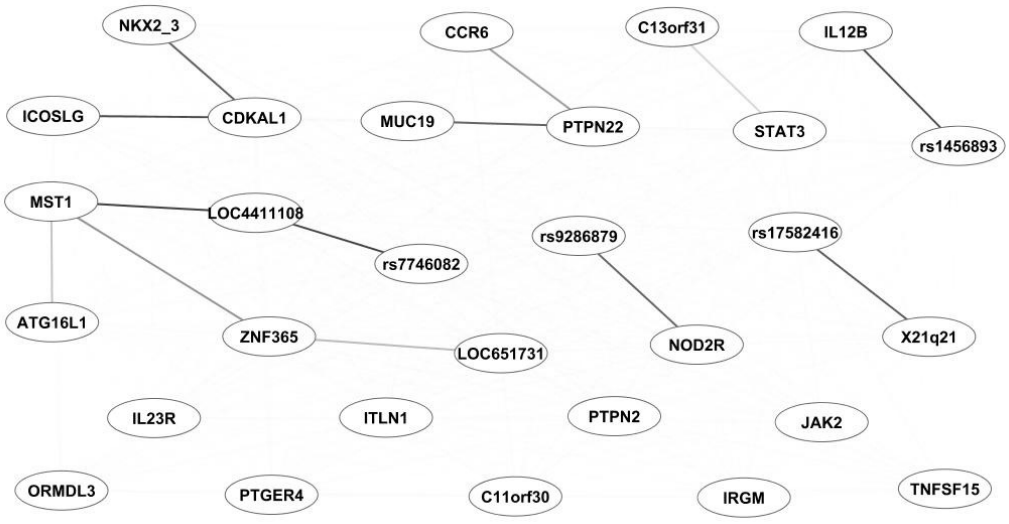


Figure 19 Results of two-level regression - pathways that significantly different between ileum and non-ileum afflicted, controlling for smoking status. Opacity of edges is determined by significance of disease location controlling for smoking status on the partial correlation between the two SNPs controlling for all other SNPs ($p < 0.05$). The more opaque an edge is, the more significantly the effect disease location on the partial correlation between the two connected SNPs, even after controlling for smoking status.

MST1 locus is encoded for protein macrophage stimulating-1. The association between *MST1* and Crohn's disease was identified in Marquez, et al. (2009). That study pointed out that this locus seems to mainly affect the ileal CD subphenotype, although this point awaits further corroboration in independent cohorts. Here our study showed the *MST1* locus has many shared connections with other SNPs that are associated with the ileum and non-ileum afflicted status.

STAT3 (signal transducer and activator of transcription 3), *PTPN22* (protein tyrosine phosphatase non-recptor type 22) and *CCR6* (chemokine C-C motif receptor 6) are found to be significantly associated with Crohn's disease by GWAS studies via biological pathways of T-cell differentiation, Immnune-cell activation and Chemokine signaling, respectively (Zhernakova, et al. 2009). Our results confirmed that the association between these locus with Crohn's disease. We also found the relations among these three loci also associated with disease location even after controlling for smoking status, as smoking is seen as an influential factor for Crohn's disease on location and severity (Picco, et al. 2003, Mahid, et al. 2007).

10 Discussion and future work

10.1 Models and confounders

A few simple observations can be made about the methods to control for confounding implemented here based on the simulation study and empirical findings.

Propensity score adjusted regression is the same as multiple regression. However, propensity scores were developed to deal with a large number of confounders. In the simulation, there was only one confounder in the model, so this may have been the reason by propensity score was unable to stand out.

When the variable of interest and the confounder are highly correlated, multiple regression and propensity score adjusted regression do not truly control for the confounding and multicollinearity will cause problems in the analysis. Residual logistic regression and Pearson residual analysis is able to attribute all effects to the confounder in the first regression and detect that no variance is left to be explained by the variable of interest during the second regression.

Empirically, one-at-a-time regression consistently finds the highest number of significant variables of interest and group effect estimates were most different from that of traditional multiple regression compared to the other methods. Both of these discrepancies may be due to the small number of covariates used in the model in comparison to the traditional multiple regression method. Propensity score adjustment based on the raw propensity score and the logit of the propensity score give almost identical results for variables of interest, but estimated parameters for the logit of the propensity score covariate were more reasonable.

For the logistic model, the estimated odds ratio for group of the residual logistic regression was closest to that estimated by traditional multiple logistic regression and had a smaller standard error. Pearson residual analysis had the smallest standard error for the group parameter estimate, but as the second stage of this method is a linear regression, odds ratio interpretation is lost.

Only two variables, group and HDS11 were consistently significant across the various strategies implemented on the logistic model. Only three variables, group, HDS11, and PDS, were consistently significant across the various strategies implemented on the Cox PH model. When considering a reduced set of covariates, only group was significant across the various methods implements on the AFT model.

These findings signify the importance to compare these methods and to establish a guideline as to which methods are more appropriate under what circumstances.

Theoretical verification of what residual logistic regression does to control for confounding has yet to be derived. Once that has been completed, extensions of residual logistic regression to multinomial outcomes and longitudinal type models can follow. Another set of comparisons in the multinomial outcomes case needs to be executed. Methods controlling for confounders when there are time-dependent confounders have not been thoroughly studied here, but will also need to be considered more in depth.

10.2 Partial correlation and network analysis

The new partial correlation was designed to measure the relationship between two categorical variables, or two mixed variables, after controlling for a third variable, in a way that is analogous to the continuous partial correlation. First logistic regressions are applied then the residuals obtained are correlated; the Pearson residuals were used for their asymptotic properties. Compared to other partial correlation procedures, the new partial correlation is computationally less laborious because additional control variables can be included in the regression step. The other partial correlation procedures are defined on the alternative partial correlation expression

$$r_{xy(z)} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1-r_{xz}^2}\sqrt{1-r_{yz}^2}}$$

Hence, they require higher order definitions of partial correlation in order to control for variables and computing time would increase as the number of control variables increases.

A statistical test for the new partial correlation in the case when either variable is binary needs to be developed. The test based on canonical correlations is adequate for multi-categorical cases, but the actual partial correlation estimate is always positive. Hence, for the multi-categorical case, the nature of the correlation (positive or negative) is unknown.

On average the new partial correlation was similar to the partial phi coefficient, but on a pair wise basis, the two will give different estimates for the same dataset. Since both the new partial correlation and the partial phi coefficient estimate the partial correlation between two binary variables, this property should be further studied to draw comparisons between the new method and the established phi

coefficient. We found that in the binary case, when data are categorized at 0.5 thresholds, the new partial correlation estimates the manifest partial correlation very well with very little bias and standard error. Unfortunately, we do not have formulas of the partial correlation for other situations, such when the threshold is at 0.9 or when the data are categorized into more than two categories.

Although we only explored the use of the new partial correlation using logistic regression, applications of the new partial correlation to other generalized linear regression models, such as Poisson regression and negative binomial regression, could be studied use the Pearson residual.

We used the new partial correlation in network analysis and extended the existing continuous method, two-level regression, to be able to consider covariates in network analysis with categorical or mixed data. We applied this new method to two genetic studies; one studying the relationship between SNP pathways and nicotine addiction and other studying the relationship between SNP pathways and Crohn's disease location. We were able to identify pathways which are different between the groups of interest involving SNPs that have been previously associated with the disease phenotype.

It should be noted that the original SNPs in the COGEND study were selected to be included because of their predisposition to be associated with nicotine dependence. However, we found only three SNPs to have significantly different partial correlations between nicotine addicts and non-addicts. This means that after controlling for all other SNPs, they are still able to exhibit an association with nicotine addiction. These SNP relationships should be studied further to determine the nature of the differences.

Partial correlation network analysis is a data-driven procedure which may help narrow down the focus of a genetic study to a select few SNPs. No assumptions about the SNP network is needed to initiate the analysis, enabling maximum flexibility. With covariate in network analysis, our new procedure can identify SNP pathways that are different between phenotypes for further study.

In the future, we should compare our method to existing methods to analysis covariate in network analysis. However, the use of network analysis on categorical data is still a developing field. Also, evaluation indices measuring the overall network structure should be applied for enhancing the interpretability of the results.

References

Abraham, C., and Cho, J. (2009), "Mechanisms of Disease: Inflammatroy Bowel Disease," *New England Journal of Medicin*, 2066-2078.

Agresti, A. (2007), *An Introduction to Categorical Data Analysis* (Vol. 423), Wiley-Blackwell.

Alley, W. M. (1987), "A Note on Stagewise Regression," *The American Statistician*, 41, 132-134.

Anstrom, K. J., and Tsiatis, A. A. (2001), "Utilizing Propensity Scores to Estimate Causal Treatment Effects with Censored Time-Lagged Data," *Biometrics*, 57, 1207-1218.

Austin, P. C. (2007), "The Performance of Different Propensity Score Methods for Estimating Marginal Odds Ratios," *Statistics in Medicine*, 26, 3078-3094.

Baba, K., Shibata, R., and Sibuya, M. (2004), "Partial Correlation and Conditional Correlation as Measures of Conditional Independence," *Australian & New Zealand Journal of Statistics*, 46, 657-664.

Baba, K., and Sibuya, M. (2005), "Equivalence of Partial and Conditional Correlation Coefficients," *Journal of the Japan Statistical Society*, 35, 1-19.

Baez-Revueltas, F. B. (2009), "*Residual Logistic Regression*," State Univeristy of New York at Stony Brook, Applied Mathematics and Statistics.

Baez-Revultas, F. B. (2009), "*Residual Logistic Regression*," Dissertation, State Univerisity of New York at Stony Brook, Applied Mathematics and Statistics.

Barrett, J. C., et al. (2008), "Genome-Wide Association Defines More Than 30 Distinct Susceptibility Loci for Crohn's Disease," *Nature genetics*, 40, 955-962.

Bartlett, M. S. (1941), "The Statistical Significance of Canonical Correlations," *Biometrika*, 32, 29-37.

Bergstralh, E. J., Kosanke, J. L., and Jacobsen, S. J. (1996), "Software for Optimal Matching in Observational Studies," *Epidemiology*, 7, 331-332.

Blalock, H. M. (1972), *Social Statistics: 2d Ed*, McGraw-Hill.

Casella, G., and George, E. I. (1992), "Explaining the Gibbs Sampler," *American Statistician*, 167-174.

Cepeda, M. S., Boston, R., Farrar, J. T., and Strom, B. L. (2003), "Comparison of Logistic Regression Versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders," *Am. J. Epidemiol.*, 158, 280-287.

Chen, H. (2011), "*Clustering and Network Analysis with Single Nucleotide Polymorphism (Snp)*," Stony Brook University, Applied Mathematics and Statistics.

Cochran, W. G. (1968), "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24, 295-313.

Cochran, W. G., and Chambers, S. P. (1965), "The Planning of Observational Studies of Human Populations," *Journal of the Royal Statistical Society. Series A (General)*, 128, 234-266.

Cuthbert, A. P., et al. (2002), "The Contribution of Nod2 Gene Mutations to the Risk and Site of Disease in Inflammatory Bowel Disease**," *Gastroenterology*, 122, 867-874.

D'Agostino, J., Ralph B. (1998), "Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group," *Statistics in Medicine*, 17, 2265-2281.

De La Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004), "Discovery of Meaningful Associations in Genomic Data Using Partial Correlation Coefficients," *Bioinformatics*, 20, 3565-3574.

Derigs, U. (1988), "Solving Non-Bipartite Matching Problems Via Shortest Path Techniques," *Annals of Operations Research*, 13, 225-261.

Draper, N. R., and Smith, H. (1998), *Applied Regression Analysis*, New York: John Wiley & Sons, Inc.

Drasgow, F. (1986), "Polychoric and Polyserial Correlations," *S. Kotz and N. Johnson*, 68-74.

Ekstrom, J. (2008), "The Phi-Coefficient, the Tetrachoric Correlation Coefficient, and the Pearson-Yule Debate."

Farrar, D. E., and Glauber, R. R. (1967), "Multicollinearity in Regression Analysis: The Problem Revisited," *The Review of Economics and Statistics*, 49, 92-107.

Fisher, R. A. (1924), "The Distribution of the Partial Correlation Coefficient," *Metron*, 3, 329-332.

Fowler, E. V., et al. (2008), "Atg16l1 T300a Shows Strong Associations with Disease Subgroups in a Large Australian Ibd Population: Further Support for Significant Disease Heterogeneity," *The American journal of gastroenterology*, 103, 2519-2526.

Fransson, P., and Marrelec, G. (2008), "The Precuneus/Posterior Cingulate Cortex Plays a Pivotal Role in the Default Mode Network: Evidence from a Partial Correlation Network Analysis," *Neuroimage*, 42, 1178-1184.

Freund, R. J., Richard, W., and Clunies-Ross, C. W. (1961a), "Corrigenda: Residual Analysis," *Journal of the American Statistical Association*, 56, 1005.

Freund, R. J., Vail, R. W., and Clunies-Ross, C. W. (1961b), "Residual Analysis," *Journal of the American Statistical Association*, 56, 98-104.

Geng, Z., Guo, J., and Fung, W.-K. (2002), "Criteria for Confounders in Epidemiological Studies," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 3-15.

Goyette, P., Labbe, C., Trinh, T. T., Xavier, R. J., and Rioux, J. D. (2007), "Molecular Pathogenesis of Inflammatory Bowel Disease: Genotypes, Phenotypes and Personalized Medicine," *Annals of medicine*, 39, 177-199.

Greenland, S., and Morgenstern, H. (2001), "Confounding in Health Research," *Annual Review of Public Health*, 22, 189-212.

Greenland, S., and Robins, J. (1986), "Identifiability, Exchangeability, and Epidemiological Confounding," *Int. J. Epidemiol.*, 15, 413-419.

Greenland, S., Robins, J. M., and Pearl, J. (1999), "Confounding and Collapsibility in Causal Inference," *Statistical Science*, 14, 29-46.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010), "Joint Structure Estimation for Categorical Markov Networks," *Submitted. Available at http://www.stat.lsa.umich.edu\~elevina*.

Hogg, R., McKean, J., and Craig, A. (2004), *Introduction to Mathematical Statistics* (6 ed.), Prentice Hall.

Holland, P. W., and Leinhardt, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs," *Journal of the American Statistical Association*, 76, 33-50.

Hosmer, D., and Lemeshow, S. (2000), *Applied Logistic Regression* (2 ed.), John Wiley & Sons, Inc.

Hung, R. J., et al. (2008), "A Susceptibility Locus for Lung Cancer Maps to Nicotinic Acetylcholine Receptor Subunit Genes on 15q25," *Nature*, 452, 633-637.

Ising, E. (1925), "Beitrag Zur Theorie Des Ferromagnetismus," *Zeitschrift für Physik A Hadrons and Nuclei*, 31, 253-258.

Jaspen, N. (1946), "Serial Correlation," *Psychometrika*, 11, 23-30.

Joffe, M. M., and Rosenbaum, P. R. (1999), "Invited Commentary: Propensity Scores," *Am. J. Epidemiol.*, 150, 327-333.

Johnson, R. A., and Wichern, D. W. (2002), *Applied Multivariate Statistical Analysis* (Vol. 4), Prentice Hall Upper Saddle River, NJ.

Kabe, D. G. (1963), "Stepwise Multivariate Linear Regression," *Journal of the American Statistical Association*, 58, 770-773.

Kelley, T. L. (1923), *Statistical Method.*, New York: MacMillan.

Kleinbaum, D., Kupper, L., and Morgenstern, H. (1982), *Epidemiologic Research: Principles and Quantitative Methods*, Wiley.

Kutner, M. H., Nachtsheim, C., and Neter, J. (2004), *Applied Linear Regression Models*, McGraw-Hill New York, NY.

Lesaffre, E., and Albert, A. (1989), "Multiple-Group Logistic Regression Diagnostics," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 38, 425-440.

Lesage, S., et al. (2002), "*Card15/Nod2* Mutational Analysis and Genotype-Phenotype Correlation in 612 Patients with Inflammatory Bowel Disease," *The American Journal of Human Genetics*, 70, 845-857.

Li, Y. P., Propert, K. J., and Rosenbaum, P. R. (2001), "Balanced Risk Set Matching," *Journal of the American Statistical Association*, 96, 870-882.

Louis, E., et al. (2001), "Behaviour of Crohn's Disease According to the Vienna Classification: Changing Pattern over the Course of the Disease," *Gut*, 49, 777-782.

Lu, B. (2005), "Propensity Score Matching with Time-Dependent Covariates," *Biometrics*, 61, 721-728.

Mahid, S. S., Minor, K. S., Stromberg, A. J., and Galandiuk, S. (2007), "Active and Passive Smoking in Childhood Is Related to the Development of Inflammatory Bowel Disease," *Inflammatory bowel diseases*, 13, 431-438.

Marquez, A., et al. (2009), "Effect of Bsn-Mst1 Locus on Inflammatory Bowel Disease and Multiple Sclerosis Susceptibility," *Genes and immunity*, 10, 631-635.

Márquez, A., et al. (2009), "Role of Atg16l1 Thr300ala Polymorphism in Inflammatory Bowel Disease: A Study in the Spanish Population and a Meta-Analysis," *Inflammatory bowel diseases*, 15, 1697-1704.

Marrelec, G., Kim, J., Doyon, J., and Horwitz, B. (2009), "Large-Scale Neural Model Validation of Partial Correlation Analysis for Effective Connectivity Investigation in Functional Mri," *Human brain mapping*, 30, 941-950.

Marrelec, G., et al. (2006), "Partial Correlation for Functional Brain Interactivity Investigation in Functional Mri," *Neuroimage*, 32, 228-237.

Martens, E. P., et al. (2008), "Comparing Treatment Effects after Adjustment with Multivariable Cox Proportional Hazards Regression and Propensity Score Methods," *Pharmacoepidemiology and Drug Safety*, 17, 1-8.

Martinson, E., and Hamdan, M. (1972), "Maximum Likelihood and Some Other Asymptotically Efficient Estimators of Correlation in Two Way Contingency Tables," *Journal of Statistical Computation and Simulation*, 1, 45-54.

Miettinen, O. S., and Cook, e. F. (1981), "Confounding: Essense and Detection," *Am. J. Epidemiol.*, 114, 593-603.

Newgard, C. D., Hedges, J. R., Arthur, M., and Mullins, R. J. (2004), "Advanced Statistics: The Propensity Score--a Method for Estimating Treatment Effect in Observational Research," *Academic Emergency Medicine*, 11, 953-961.

Olsson, U. (1979), "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient," *Psychometrika*, 44, 443-460.

Olsson, U., Drasgow, F., and Dorans, N. (1982), "The Polyserial Correlation Coefficient," *Psychometrika*, 47, 337-347.

Pearson, K. (1895), "Note on Regression and Inheritance in the Case of Two Parents," *Proceedings of the Royal Society of London*, 58, 240-242.

Pearson, K. (1909), "On a New Method of Determining Correlation between a Measured Character a, and a Character B, of Which Only the Percentage of Cases Wherein B Exceeds (or Falls Short of) a Given Intensity Is Recorded for Each Grade of A," *Biometrika*, 7, 96-105.

Pearson, K. (1920), "Notes on the History of Correlation," *Biometrika*, 13, 25-45.

Pearson, K., and Lee, A. (1900), "Mathematical Contributions to the Theory of Evolution. Viii. On the Inheritance of Characters Not Capable of Exact Quantitative Measurement. Part I. Introductory. Part Ii. On the Inheritance of Coat-Colour in Horses. Part Iii. On the Inheritance of Eye-Colour in Man," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 195, 79-150.

Picco, M. F., and Bayless, T. M. (2003), "Tobacco Consumption and Disease Duration Are Associated with Fistulizing and Stricturing Behaviors in the First 8 Years of Crohn's Disease," *The American journal of gastroenterology*, 98, 363-368.

Poon, W.-Y., and Lee, S.-Y. (1987), "Maximum Likelihood Estimation of Multivariate Polyserial and Polychoric Correlation Coefficients," *Psychometrika*, 52, 409-430.

Pradhan, K. (2009), "*Partial Correlation Analysis in Functional Brain Imaging Studies*," Stony Brook University, Applied Mathematics and Statistics.

Prescott, N. J., et al. (2007), "A Nonsynonymous Snp in Atg16l1 Predisposes to Ileal Crohn's Disease and Is Independent of Card15 and Ibd5," *Gastroenterology*, 132, 1665-1671.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010), "High-Dimensional Ising Model Selection Using ℓ1-Regularized Logistic Regression," *The Annals of Statistics*, 38, 1287-1319.

Reeve, B., Smith, A., Arora, N., and Hays, R. (2008), "Reducing Bias in Cancer Research: Application of Propensity Score Matching," *Health Care Financ Rev*, 29, 69-80.

Reisberg, B., Ferris, S., de Leon, M., and Crook, T. (1982), "The Global Deterioration Scale for Assessment of Primary Degenerative Dementia," *Am J Psychiatry*, 139, 1136-1139.

Reisberg, B., Shulman, M. B., Torossian, C., Leng, L., and Zhu, W. (2010), "Outcome over Seven Years of Healthy Adults with and without Subjective Cognitive Impairment," 6, 11-24.

Ritschard, G., Kellerhals, J., Olszak, M., and Sardi, M. (1996), "Path Analysis with Partial Association Measures," *Quality & Quantity*, 30, 37-60.

Rosenbaum, P. R. (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387-394.

Rosenbaum, P. R. (2002), *Observational Studies* (2nd ed.), New York: Springer-Verlag.

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

Rosenbaum, P. R., and Rubin, D. B. (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516-524.

Rosenbaum, P. R., and Rubin, D. B. (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 39, 33-38.

Saccone, N. L., et al. (2009), "The Chrna5-Chrna3-Chrnb4 Nicotinic Receptor Subunit Gene Cluster Affects Risk for Nicotine Dependence in African-Americans and in European-Americans," *Cancer research*, 69, 6848.

Satsangi, J., Silverberg, M., Vermeire, S., and Colombel, J. (2006), "The Montreal Classification of Inflammatory Bowel Disease: Controversies, Consensus, and Implications," *Gut*, 55, 749-753.

Seber, G., and Nyangoma, S. (2000), "Residuals for Multinomial Models," *Biometrika*, 87, 183-191.

Senn, S., Graf, E., and Caputo, A. (2007), "Stratification for the Propensity Score Compared with Linear Regression Techniques to Assess the Effect of Treatment or Exposure," *Statistics in Medicine*, 26, 5529-5544.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011), "Cytoscape 2.8: New Features for Data Integration and Network Visualization," *Bioinformatics*, 27, 431-432.

Spreeuwenberg, M., et al. (2010), "The Multiple Propensity Score as Control for Bias in the Comparison of More Than Two Treatment Arms: An Introduction from a Case Study in Mental Health," *Medical Care*, 48, 166.

Team, R. D. C. (2011), "R: A Language and Environment for Statistical Computing."

Unkart, J. T., et al. (2008), "Risk Factors for Surgical Recurrence after Ileocolic Resection of Crohn's Disease," *Diseases of the Colon & Rectum*, 51, 1211-1216.

Van Limbergen, J., et al. (2008), "Autophagy Gene Atg16l1 Influences Susceptibility and Disease Location but Not Childhood-Onset in Crohn's Disease in Northern Europe," *Inflammatory bowel diseases*, 14, 338-346.

Vargha, A., Rudas, T., Delaney, H. D., and Maxwell, S. E. (1996), "Dichotomization, Partial Correlation, and Conditional Independence," *Journal of educational and behavioral statistics*, 21, 264-282.

Wang, J., Donnan, P. T., Steinke, D., and MacDonald, T. M. (2001), "The Multiple Propensity Score for Analysis of Dose-Response Relationships in Drug Safety Studies," *Pharmacoepidemiology and Drug Safety*, 10, 105-111.

Warnes, G. R. (2011), "Gplots: Various R Programming Tools for Plotting Data.."

Wasserman, S., and Pattison, P. (1996), "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs Andp," *Psychometrika*, 61, 401-425.

Weiss, R. B., et al. (2008), "A Candidate Gene Approach Identifies the Chrna5-A3-B4 Region as a Risk Factor for Age-Dependent Nicotine Addiction," *PLoS Genetics*, 4, e1000125.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., and Mor, V. (2005), "Weaknesses of Goodness-of-Fit Tests for Evaluating Propensity Score Models: The Case of the Omitted Confounder," *Pharmacoepidemiology and Drug Safety*, 14, 227-238.

Wherry, R. J. (1984), *Contributions to Correlational Analysis*, Academic Press London.

Yule, G. U. (1897), "On the Theory of Correlation," *Journal of the Royal Statistical Society*, 60, 812-854.

Yule, G. U. (1907), "On the Theory of Correlation for Any Number of Variables, Treated by a New System of Notation," *Proceedings of the Royal Society of London. Series A*, 79, 182-193.

Zhernakova, A., Van Diemen, C. C., and Wijmenga, C. (2009), "Detecting Shared Pathogenesis from the Shared Genetics of Immune-Related Diseases," *Nature Reviews Genetics*, 10, 43-55.

Zyskind, G. (1963), "A Note on Residual Analysis," *Journal of the American Statistical Association*, 58, 1125-1132.