Latent Variable Analysis of Atherosclerosis Data with PO2PLS

Sanne Meijering

UMC Utrecht

[*]

13 December 2019

[*] Under the supervision of Said el Bouhaddani, Hae-won Uh and René Eijkemans

Introduction

Atherosclerosis is the (partial) blockage of blood vessels by plaques, thereby limiting blood flow. This can eventually lead to a heart attack or stroke[1]. Several studies have investigated the effect of genetics on the plaque formation[2–4]. It was found that up to 60% of atherosclerosis is heritable. If it is known which genes are involved in atherosclerosis, researchers will gain a better understanding of the disease and may be able to develop new treatments.

A commonly used method to investigate heritability is the assessment of single nucleotide polymorphisms (SNPs). SNPs are nucleotide substitutions that occur in at least 1% of the population. There are millions of known SNPs, and often a selection is made to represent the genome. The SNPs that are not included in the study can still be studied through a phenomenon called linkage disequilibrium (LD). This phenomenon entails that SNPs that are close in position on the genome are usually inherited together. One can thus be used to predicts the other. Even with selection, datasets remain large, with millions of SNP variables, and thus require a fast analysis method or further data reduction.

Current methods for analysis of SNP data are generally univariate; they measure the association between one histological trait and one or more genes[5,6]. One common method is the genome-wide association study (GWAS). GWAS calculates a p-value for each combination of an SNP and a histological trait. The method is relatively fast, but has low power due to the high number of tests performed. Gene set enrichment analysis (GSEA)[7], VEGAS[8] and

MAGMA[5] attempt to increase power by aggregating GWAS results into gene set scores, but in doing so is dependent on the GWAS results. GSEA has the additional issue of often relying on an arbritrary cut off value. As can be noted, all of the methods above are univariate methods, or multivariate methods based on univariate results. Multivariate approaches are more powerful when the outcomes are correlated, but usually cannot handle large datasets. In atherosclerosis, histological traits are correlated with each other and many genes are involved in the disease. Thus, a multivariate method that can relate multiple correlated outcomes to a high-dimensional dataset is required.

One method that allows for multivariate analysis of high-dimensional data is latent variable analysis. Latent variables reduce the dimensionality of the dataset further, while retaining the ability to interpret results by investigating the weights assigned to each variable. Two-way Orthogonal Partial Least Squares (O2PLS)[9] is a latent variable method capable of relating multiple outcomes to a multivariate set of predictors. It is reliant on matrix multiplication and is thus capable of fast multivariate analysis of large datasets. To achieve this, O2PLS splits both datasets into a joint part and structured noise specific to one dataset. This 'structured noise' consists of specific effects in one dataset that are not included in the other dataset. The number of joint components and specific components per dataset is determined by the researcher. The joint components represent the covariance between the two datasets, while the specific components represent variance unique to one dataset. While O2PLS is fast and gives reasonably good results, it is known to be prone to overfitting and cannot easily be extended to handle complex study designs.

Probabilistic Two-way Orthogonal Partial Least Squares (PO2PLS) is an implementation of O2PLS in a probabilistic framework that combats both of these issues. It uses the EM-algorithm to maximize the likelihood and obtain estimates for the joint and specific parts simultaneously. It has the same capabilities as O2PLS but is less prone to overfitting, at the cost of being more computationally intensive due to its iterative nature. It is viable to analyze large datasets with PO2PLS as the time needed to run PO2PLS increases in a linear fashion (O(n)) as the number of participants or variables increases. To estimate p-values for the estimates of the joint parts, bootstrapping procedures are considered.

The goal of this paper is to analyze a large atherosclerosis dataset that consists of 80 million genetic traits, 165 thousand after data reduction, and seven histological traits (outcomes). To this end, I will first analyze the effectiveness of different bootstrapping strategies for estimating p-values using PO2PLS. The second step is to use the best performing bootstrapping method to analyze the atherosclerosis dataset. Then the results will be interpreted in a biological perspective. To this end, gene set enrichment analysis will be applied to the genes with significant p-values.

Method

Data

The atherosclerosis dataset used in this study contains 7 histological traits of the plaques of 1358 patients as well as approximately 8 million SNPs spread over 22 chromosomes. The sex chromosomes were excluded. METC approval was obtained for the use of this dataset. GWAS was previously performed on this dataset in an unpublished study. This yielded no significant results.

In the analysis, the first data reduction on the SNPs is performed by running a PCA on all SNPs within 20 kilobase distance of a gene. For each gene, the number of components with a cumulative explained variance of >80% is determined and the scores of that number of components are included in the dataset. The data are then centered and scaled. Finally, the data is split into a training set and a test set. The training set consists of a thousand randomly selected participants. The remainder of the participants (n=358) constitute the test set.

PO2PLS Method

The PO2PLS method is implemented in the R-package "PO2PLS"[10]. This method is based on the O2PLS method, and like the O2PLS method decomposes the data into a joint component and a dataset-specific component. To this end, the following model is used:

$$\begin{aligned} X &= tW^T + t_sW_s^T + e, \\ y &= uC^T + u_sC_s^T + f, \\ u &= tB + h. \end{aligned} \tag{1}$$

A visualization of this model can be found in Figure 1. In this model, x and y are the datasets. W and C represent the joint loadings, while W_s and C_s represent the data-specific loadings. e and f are the residuals of each dataset, and have a normal distribution with a mean of zero and covariance matrices $\sigma_e^2 I_p$ and $\sigma_f^2 I_q$ respectively. t, t_s, u and u_s are the latent variables. t, t_s, u_s and h all have a multivariate normal distribution with a mean of zero and respective diagonal covariance matrices of Σ_t, Σ_{ts}, Σ_{us} and Σ_h, and u having a covariance matrix of $\Sigma_u = B^T\Sigma_t B + \Sigma_h$, with B being a diagonal matrix.

In PO2PLS, the joint and specific parts are estimated simulataneous. The log-likelihood function associated with this model is

$$L(\theta|x, y) = -\frac{1}{2}\{(p+q)\log|\Sigma_\theta + (x, y)\Sigma_\theta^{-1}(x, y)^T\}, \tag{2}$$

with θ being the collection of all parameters and p and q being the sizes of x and y respectively. As the log of the likelihood is not linear and requires the calculation of a covariance matrix of size $(p+q)^2$, direct optimization of the likelihood is not viable. The EM-algorithm is used to find the maximum likelihood estimates as it allows for the decomposition of the loglikehood into terms, with each term being optimized seperately. This turns the computation of the maximum likelihood into multiple calculations that are computationally viable.

Simulation Study

Extraction of Data Traits

The traits of the training set are used for the creation of the datasets for the simulation study. For each of the 22 chromosomes, the optimal number of joint and specific components is determined with scree plots.

For each chromosome, PO2PLS is then run with the optimal number of components and 2000 EM-steps. For the simulation datasets, the median number of components is chosen. As the number of components must be an integer, non-integer medians are rounded down. For the signal-to-noise ratio, the means of the explained variance of the specific and joint components and the mean noise ratio is used. The dataset size was set at 1000 participants and the mean number of features per chromosome, as each chromosome will be assessed separately.

Dataset Creation

Four datasets are created to assess the effectiveness of the different bootstrapping methods. Two datasets contain no joint component and thus should not yield significant p-values. A joint component was included in the first hundred genes of the other two datasets. Specific components and noise are added to both datasets, with the noise ratio being either the mean observed value in the dataset or the lower noise ratio of 0.2. This noise ratio was included to assess the effect of pre-selection of genes that might affect plaque formation on the performance of the bootstrapping method.

Bootstrapping and Comparison to other methods

On each dataset, parametric bootstrapping, non-parametric bootstrapping and a permutation test are performed to estimate the p-value per gene. The sensitivity and specificity of each method are then calculated. They will be compared to the performance of O2PLS and MAGMA. The method with the highest specificity, while respecting the correct type I error rate, is chosen to analyze the atherosclerosis dataset.

Analysis of the Atherosclerosis Dataset

To analyze the atherosclerosis dataset, PO2PLS with bootstrapping is used on the test set. For each chromosome, the number of components is set to be the optimal number of components determined by the cross-validation on the training set. Tukey's HSD is used for correction for multiple testing.

Possible Additions to the Project

The method described above will be the main component of this study. If there is time left, I will first interpret the results of the analysis from a biological perspective. A second possible addition is to test the methods described above with different types of datasets to assess their robustness and find limitations.

References

1. Atherosclerosis — National Heart, Lung, and Blood Institute (NHLBI) https://www.nhlbi.nih.gov/health-topics/atherosclerosis.

2. Fox Caroline S., Polak Joseph F., Chazaro Irmarie, et al. Genetic and environmental contributions to atherosclerosis phenotypes in men and women: Heritability of carotid intima-media thickness in the Framingham heart study *Stroke.* 2003;34:397–401.

3. Lusis Aldons J.. Genetics of atherosclerosis *Annual review of genomics and human genetics.* 2012;28:267–275.

4. Seifi M., Ghasemi A., Khosravi M., et al. Genetic Variants and Atherosclerosis *International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering.* 2009;3:251–259.

5. Leeuw Christiaan A., Mooij Joris M., Heskes Tom, Posthuma Danielle. MAGMA: Generalized Gene-Set Analysis of GWAS Data *PLoS Computational Biology.* 2015;11.

6. Holden Marit, Deng Shiwei, Wojnowski Leszek, Kulle Bettina. GSEA-SNP: Applying gene set enrichment analysis to SNP data from genome-wide association studies *Bioinformatics.* 2008;24:2784–2785.

7. Subramanian Aravind, Tamayo Pablo, Mootha Vamsi K., et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles *Proceedings of the National Academy of Sciences of the United States of America.* 2005;102:15545–15550.

8. Liu Jimmy Z., McRae Allan F., Nyholt Dale R., et al. A versatile gene-based test for genome-wide association studies *American Journal of Human Genetics.* 2010;87:139–145.

9. Trygg Johan, Wold Svante. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter *Journal of Chemometrics.* 2003;17:53–64.

10. Bouhaddani Said. GitHub - selbouhaddani/PO2PLS: Probabilistic O2PLS `https://github.com/selbouhaddani/PO2PLS`.

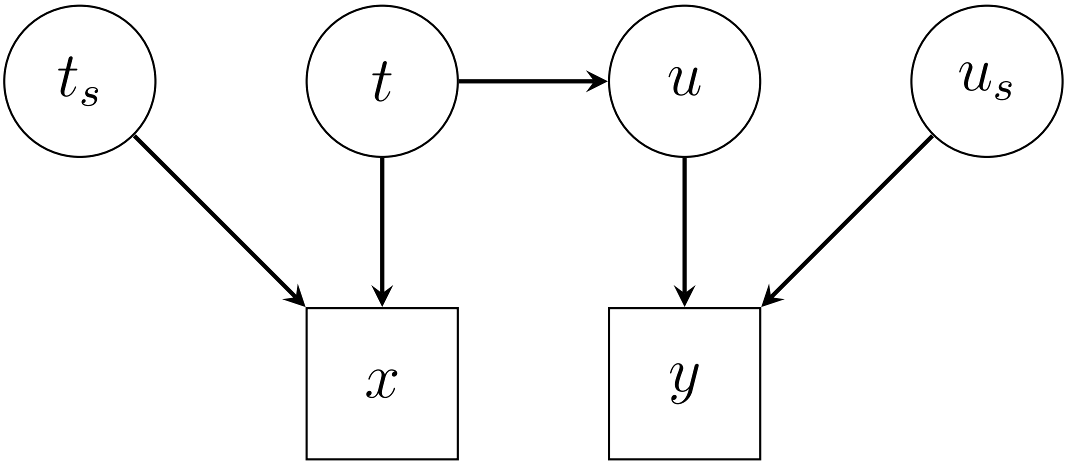

Figure 1. The PO2PLS model. The datasets x and y are decomposed into a specific component (t_s, u_s) and a joint component (t, u). t and u are linearly related to each other.