Liu et al. BMC Proceedings 2014, 8(Suppl 1):S73
http://www.biomedcentral.com/1753-6561/8/S1/S73

BMC Proceedings

PROCEEDINGS Open Access

Penalized multivariate linear mixed model for longitudinal genome-wide association studies

Jin Liu[1†], Jian Huang[2†], Shuangge Ma[3,4*]

From Genetic Analysis Workshop 18
Stevenson, WA, USA. 13-17 October 2012

Abstract

We consider analysis of Genetic Analysis Workshop 18 data, which involves multiple longitudinal traits and dense genome-wide single-nucleotide polymorphism (SNP) markers. We use a multivariate linear mixed model to account for the covariance of random effects and multivariate residuals. We divide the SNPs into groups according to the genes they belong to and score them using weighted sum statistics. We propose a penalized approach for genetic variant selection at the gene level. The overall modeling and penalized selection method is referred to as the penalized multivariate linear mixed model. Cross-validation is used for tuning parameter selection. A resampling approach is adopted to evaluate the relative stability of the identified genes. Application to the Genetic Analysis Workshop 18 data shows that the proposed approach can effectively select markers associated with phenotypes at gene level.

Background

The Genetic Analysis Workshop 18 (GAW18) data consists of multiple longitudinal traits and dense genome-wide single-nucleotide polymorphism (SNP) markers. A commonly used approach for identifying markers associated with traits is to conduct single-variant analysis and then adjust for multiple comparisons on each trait. However, for complex polygenic traits, single-variant analysis methods may not be appropriate as they fail to take into account the accumulated and/or joint effects of multiple genetic variants on the traits. In addition, analyzing each trait separately does not take into account the correlation among traits, and thus can be ineffective. To overcome these limitations, we developed a joint analysis approach referred to as the penalized multivariate linear mixed model (PMLMM). This approach takes into account covariance of both random effects and residuals and uses a group minimax concave penalty (MCP) approach [1] for variant selection at the gene level. A resampling approach is adopted to evaluate the relative stability of the identified genes. Our analysis of the GAW18 data indicates that the proposed approach can effectively select markers associated with multiple traits at the gene level.

Methods

Consider a genetic association study with longitudinal measurements on N subjects, p genetic variants, and q environmental exposure covariates. Here a genetic variant can be a single SNP marker or a score representing a group of SNPs. For subject i, suppose that there are n_i longitudinal measurements on m traits. Let Y_i be the $n_i \times m$ trait matrix for subject i. Let Y be the $n \times m$ trait matrix for all the N subjects, where $n = \sum_{i=1}^{N} n_i$. The transpose of Y is $Y\prime = (Y_1', \cdots, Y_N')$. Let X_i be the $n_i \times p$ matrix consisting of the genetic variant scores of subject i. Let Z_i be the $n_i \times q$ covariate matrix. We center all the measurements to have sample means equal to zero. When $m = 1$, this setting simplifies to that in Schelldorfer et al [2].

Consider the multivariate linear mixed model

$$Y_i = X_i B + Z_i C_i + E_i, \quad i = 1, \ldots, N, \tag{1}$$

where B is a $p \times m$ matrix representing the effects of p genetic variants on m traits, and C_i is a $q \times m$ matrix

* Correspondence: shuangge.ma@yale.edu
† Contributed equally
[3]School of Public Health, Yale University, 60 College Street, New Haven, CT 06520, USA
Full list of author information is available at the end of the article

BioMed Central

© 2014 Liu et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

representing the subject specific effects of the covariates Z_i for the i^{th} subject. We treat C_i as random effects. Assume that (a) $E_i \sim \text{MN}_{n_i \times m}(0, \sum_1, I_{n_i})$, that is, E_i is row-independent with column covariance matrix Σ_1, and each E_i is independent for $i = 1, ..., N$; (b) $C_i \sim \text{MN}_{q \times m}(0, \sum_2, D)$, where Σ_2 is the column covariance matrix and D is the row covariance matrix, and each C_i is independent for $i = 1, ..., N$; (c) each E_i and C_i is independent; and (d) $\Sigma_1 = \Sigma_2 = \Sigma$.

Then $Z_iC_i + E_i \sim \text{MN}_{n_i \times m}(0, \sum, Z_iDZ_i' + I_{n_i})$ and $Y \sim \text{MN}_{n \times m}(XB, \sum, V)$ where $V = Diag(V_1, ..., V_N)$ and $V_i = Z_iDZ_i' + \text{I}_{n_i}$, where I_{n_i} is an $n_i \times n_i$ identity matrix. A more detailed description of this model can be found in Liu et al [3].

From Dawid [4], the negative log-likelihood function is:

$$-\ell(B, V, \sum) = \text{constant} + \frac{n}{2}\log|\sum| + \frac{m}{2}\log|V| + \frac{1}{2}\text{tr}(\sum^{-1}(\gamma - XB)'V^{-1}(\gamma - XB)) \quad (2)$$

Hastie et al. [5] suggest using $\hat{\Sigma} = \gamma'\gamma/n$ for estimating $\sum$. We estimate D by using the estimates from m univariate linear mixed models and subsequently get the estimate $\hat{V}$ of V as $\hat{V}_i = Z_i\hat{D}Z_i' + \text{I}_{n_i}$. Given $\hat{\sum}$ and $\hat{V}$, we can transform the negative likelihood function into a weighted least squares criterion for estimating B, which is $\text{tr}(\hat{\sum}^{-1}(\gamma - XB)'\hat{V}^{-1}(\gamma - XB))$. For variant selection, we adopt the group MCP approach [6]. The overall penalized objective function is

$$Q(B) = \text{tr}(\hat{\sum}^{-1}(\gamma - XB)'\hat{V}^{-1}(\gamma - XB)) + \sum_{j=1}^{p}\rho(||B_j||_2; \lambda, \gamma), \quad (3)$$

where B_j is the j^{th} row of B and $\rho(t; \lambda, \gamma) = \lambda\int_0^{|t|}(1 - x/(\gamma\lambda))_+dx$ is the MCP with tuning parameter λ and regularization parameter γ [7].

For computation, we use a group coordinate descent algorithm [1]. The group MCP involves a regularization parameter and a tuning parameter. Generally speaking, smaller values of γ are better at retaining the unbiasedness of the MCP penalty for large coefficients, but they have the risk of creating objective functions that have problems with nonconvexity [8], are difficult to optimize, and yield solutions that are discontinuous with respect to λ. Simulation studies by Breheny and Huang [8] suggest that $\gamma = 6$ is a reasonable choice. Therefore, we fix it to be 6 in our analyses. We search for the optimal value of λ using 5-fold cross-validation.

Results

The GAW18 data set consists of dense genome-wide markers with longitudinal measurements on systolic and diastolic blood pressure (SBP and DBP) and other covariates. Other measurements include gender, age, year of examination, use of antihypertensive medications, and tobacco smoking at up to 4 time points. In this study, we analyze the 157 unrelated individuals using SBP and DBP as traits and other medical and demographic covariates as random effects. Gene annotations for SNP data are obtained from http://www.scandb.org. SNPs in each gene are scored using weighted sum statistics to generate gene-level measurements [9]. After quality control, we have the genetic scores of 10,400 genes for further analysis. SBP, DBP, and genetic scores are standardized to have zero means and unit variances. This procedure removes the estimation of intercepts and makes the genes comparable.

We apply the proposed PMLMM to identify genetic variants that are associated with both SBP and DBP at the gene level. As a benchmark, we also analyze each trait separately using a penalized linear mixed model (PLMM) approach. Table 1 shows the genes identified using PMLMM. Table 2 summarizes the overlaps of genes selected using the different approaches. Although there is overlap, PMLMM and PLMM identify significantly different sets of genes. We evaluate the relative stability of identification of each gene using a resampling approach and calculate the observed occurrence index (OOI) [10]. A larger value of OOI indicates that the corresponding identified gene is more stably identified. Table 1 also shows OOI results. The identified genes have reasonably high OOIs.

Discussion

In this study, we analyze the GAW18 data and develop a PMLMM approach. A multivariate linear mixed model is used to model variance components among traits and longitudinal measurements. A penalization approach is adopted for variant selection. In the estimation procedure, it can be considered heuristic to use $\hat{\sum}$ and $\hat{V}$ as proposed. Assumptions (a) to (c) are standard in mixed models, but the assumption that $\Sigma_1 = \Sigma_2$ may be restrictive. Because our study is to identify multitrait-associated markers at the gene level, the restriction on variance components does not affect the selection result significantly. We are currently developing a similar approach to update variance components with more relaxed assumptions on Σ_1 and Σ_2. An iterative algorithm can be implemented to solve for B, $\sum$, and V. In variant selection, our method is designed to search for genes associated with all the traits considered. When different sets of genetic variants are suspected to be associated with different phenotypes, the sparse group penalization approach [11] can be applied.

Conclusions

We have presented a penalized multivariate linear mixed model (PMLMM) for detecting pleiotropic genetic

Table 1 Genes identified by PMLMM: estimates for SBP and DBP, and OOI

Gene	SBP	DBP	OOI	Gene	SBP	DBP	OOI
MMEL1	0.002	−0.002	0.333	*TMEM41B*	0.027	0.033	0.403
CD52	0.085	0.060	0.697	*ARNTL*	0.024	0.006	0.247
DPH2	0.071	−0.032	0.323	*SPTY2D1*	0.025	−0.007	0.507
C8A	0.018	0.032	0.563	*CHST1*	−0.008	−0.031	0.540
DNAJB4	−0.028	−0.022	0.333	*MRE11A*	−0.042	−0.007	0.623
HS2ST1	0.002	0.006	0.307	*ENOX1*	−0.068	−0.032	0.647
PROK1	0.006	0.010	0.373	*LOC100132760*	−0.041	0.048	0.693
THBS3	−0.004	0.001	0.337	*SPRY2*	−0.004	−0.005	0.297
C1orf182	2E-04	0.045	0.490	*GABRG3*	−0.027	0.006	0.573
TGFBR2	−0.033	−0.030	0.627	*THBS1*	0.023	0.028	0.353
LOC100129194	0.005	0.011	0.217	*CSPG4*	0.098	−0.012	0.880
LMOD3	−0.013	−0.028	0.493	*C15orf27*	0.047	−0.014	0.490
LOC653712	0.017	0.001	0.450	*LOC100128570*	0.026	0.019	0.283
LAMP3	0.034	0.008	0.627	*HOMER2*	0.024	0.013	0.250
EIF2B5	−0.014	−0.014	0.417	*ADAMTS17*	0.002	0.002	0.270
EHHADH	0.003	−0.052	0.677	*SLC16A11*	0.015	−0.004	0.170
SFRS12	0.005	0.007	0.290	*ALDH3A1*	0.002	0.021	0.657
C5orf32	0.019	−0.029	0.560	*FLJ44815*	0.019	−0.005	0.210
ZNF346	0.001	−0.024	0.553	*TANC2*	0.003	−0.001	0.187
LOC100128901	−0.069	0.006	0.627	*PDE6G*	−0.056	−0.012	0.377
OGDH	0.123	0.038	0.777	*C19orf6*	0.013	0.048	0.577
NSUN5	0.035	−0.014	0.660	*TMEM146*	−0.001	−0.063	0.550
PPP1R3A	−0.034	−0.041	0.453	*STX10*	4E-04	0.015	0.333
MEST	−2E-04	−3E-04	0.197	*RLN3*	0.024	−0.034	0.603
NOM1	0.029	−0.001	0.630	*CYP4F11*	0.018	0.005	0.127
FLJ41200	0.013	0.005	0.333	*LOC728326*	0.048	−0.040	0.547
LRRC19	−0.015	−0.040	0.480	*ZNF585A*	0.028	0.082	0.720
CCIN	0.006	−0.004	0.437	*SUPT5H*	0.020	−0.014	0.233
LOC100130911	0.004	0.024	0.467	*FLJ10490*	0.004	2E-04	0.327
PTCH1	0.004	−1E-04	0.300	*ZNF331*	0.027	0.003	0.330
DFNB31	0.057	−0.008	0.710	*BACE2*	−0.116	−0.074	0.940
OR52D1	0.023	−0.017	0.490	*KRTAP10-12*	0.002	0.003	0.233

Table 2 Overlap of selected genes between PMLMM and PLMM

	PMLMM	PLMM*	PLMM†
PMLMM	64	24	16
PLMM[1]		40	0
PLMM[2]			29

*PLMM on SBP.
†PLMM on DBP.

associations among multiple traits in the presence of pedigree structures. The proposed approach combines the advantages of mixed models that allow for elegant correction for pedigree-based family data and integrative analysis of multiple traits. Compared with PLMM which considers one trait at a time, the proposed PMLMM can achieve better performance when the pleiotropic effect is appropriately modeled.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors were involved in study design. JL conducted the numerical work. All authors were involved in manuscript preparation, and read and approved the final manuscript.

Acknowledgements

This study was supported by NIH grants CA142774, CA165923, and CA120988, the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development, and 2012LD001 from National Bureau of Statistics of China. The GAW18 whole genome sequence data were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW18 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants P01 HL045222, R01 DK047482, and R01 DK053889. The Genetic Analysis Workshop is supported by NIH grant R01 GM031575.
This article has been published as part of *BMC Proceedings* Volume 8 Supplement 1, 2014: Genetic Analysis Workshop 18. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcproc/

supplements/8/S1. Publication charges for this supplement were funded by the Texas Biomedical Research Institute.

Authors' details
[1]School of Public Health, University of Illinois at Chicago, 1601 W. Taylor Street, Chicago, IL 60612, USA. [2]Department of Statistics & Actuarial Science, Department of Biostatistics, University of Iowa, 241 Schaeffer Hall, Iowa City, IA 52242, USA. [3]School of Public Health, Yale University, 60 College Street, New Haven, CT 06520, USA. [4]VA Cooperative Studies Program Coordinating Center, West Haven, CT 06516, USA.

Published: 17 June 2014

References

1. Huang J, Wei F, Ma S: **Semiparametric regression pursuit.** *Stat Sin* 2012, **22**:1403-1426.
2. Schelldorfer J, van de Geer S: **Estimation for high-dimensional linear mixed-effects models using L1-penalization.** *Scand Stat Theory Appl* 2011, **38(2)**:197-214.
3. Liu J, Yang C, Shi X, Zhao H, Huang J, Ma S: **A penalized multiple-trait mixed model for association mapping with population structure correction.** *Technical Report (arXiv preprint arXiv:1305.4413)* 2013.
4. Dawid A: **Some matrix-variate distribution theory: notational considerations and a Bayesian application.** *Biometrika* 1981, **68**:265-274.
5. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning: Data Mining, Inference, and Prediction.** New York, Springer-Verlag;, 2 2009.
6. Liu J, Huang J, Ma S: **Analysis of genome-wide association studies with multiple outcomes using penalization.** *PLoS One* 2012, **7**:e51198.
7. Zhang CH: **Nearly unbiased variable selection under minimax concave penalty.** *Ann Stat* 2010, **38**:894-942.
8. Breheny P, Huang J: **Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection.** *Ann Appl Stat* 2011, **5**:232-253.
9. Madsen B, Browning S: **A groupwise association test for rare mutations using a weighted sum statistic.** *PLoS Genet* 2009, **5**:e1000384.
10. Huang J, Ma S: **Variable selection in the accelerated failure time model via the bridge method.** *Lifetime Data Anal* 2010, **16**:176-195.
11. Liu J, Huang J, Xie Y, Ma S: **Sparse group penalized integrative analysis of multiple cancer prognosis datasets.** *Genet Res (Camb)* 2013, **95(2-3)**:68-77.

doi:10.1186/1753-6561-8-S1-S73
Cite this article as: Liu *et al.*: Penalized multivariate linear mixed model for longitudinal genome-wide association studies. *BMC Proceedings* 2014 **8**(Suppl 1):S73.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central