Proteogenomics: statistical issues in data integration and prediction

Júlia M Pavan Soler
Statistics Department, University of São Paulo, São Paulo, SP

Abstract

Proteogenomics inaugurates a new phase of multi-omics research in Molecular Biology, seeking to integrate information of large datasets from the genome, transcriptome and proteome to clinical traits. The promise is to identify patient-specific biomarkers, which can be used on the prognostic in precision medicine. However, the expected contribution of this area depends on overcoming several interdisciplinary challenges, ranging from the design of experiments for samples preparation, storage, processing and integration of data to its analysis and interpretation. Brazil, as other countries, starts to dedicate efforts to the proteogenomic analysis of many diseases in order to identify specific and common biomarkers among world populations. Specifically, the Baependi Family Heart Study is one of the largest ongoing efforts for molecular mapping in cardiovascular diseases in our country, which includes Brazilian family information. Statistical approaches in proteogenomics are typically formulated assuming unrelated individuals, and if family structure is present and ignored, such substructures may induce to misleading results. In this talk, in the context of proteogenomics, we will consider flexible methodologies for dimensionality reduction, variable selection and structure learning taking in account sparsity, dependent observations and missing information.

Keywords

Matrix factorization; Varying coefficients; Multi-omics data; Family based design; Complex data

1. Introduction

A relevant issue that is becoming increasingly important in the big and complex data age is data integration. An early version of that trend can be seen in the multi-omics studies, as exemplified by proteogenomics studies, seeking to integrate many sources of information from large datasets measured on a common set of experimental subjects to clinical traits. In general, the integration scope in these studies try to cover the central dogma of the Molecular Biology including data from genome (such as, SNP and CNV platforms), epigenome (such as, methylation data), transcriptome (such as, Microarrays and RNA-seq data) and proteome (such as, LC-MS/MS data) to phenome (phenotype dataset). The Cancer Genome Atlas (TCGA, Weinstein et al. 2013) project provides a powerful source of such set of data blocks. These cross-platform datasets share common information, but individually contain distinctive patterns. Disentangling between common and distinctive patterns, and also between the noise component, is critically important to perform integrative, discriminative and predictive analysis of these datasets (Smilde et al., 2017; Shu et al., 2018). Whilst single omics analyses, under an unsupervised or supervised scope, are commonly used for dimensionality reduction and selection of relevant features for specific analytical frameworks, the integration of multi-omics information is required to more fully unravel the complexities of biological systems.

The first step on the omics data analysis involves detection and adjustments for undesirable variable effects, which will tend to appear in addition to the measured variable(s) of interest among most, if not all, high-throughput technologies (Leek et al., 2010). Failing for correction of these sources of heterogeneity into the analysis can have widespread and detrimental effects on the study, not only reducing power and inducing unwanted dependence across genes, but it can also introduce sources of spurious signals. This phenomenon is true even for well-designed and randomized studies. For instance, considering whole-genome SNP platforms, Price et al. (2006) applied singular value decomposition to the genotype called data in order to account for systematic sources of variation due to population substructure. In addition, for batch effects correction and normalization in gene expression data, there are many methods based on nonparametric and parametric approaches (Wolfinger et al., 2001; Irizarry et al., 2003; Leek and Storey, 2007; Chen et al., 2011). Further, undesirable effects in mass spectrometry-based proteomics data have been treated by using smooth curves and ANOVA-Simultaneous Component Analysis (Clough et al., 2012; Mitra et al., 2016). Although all these tools are available, for database integration there is no consensus whether the normalization step should be done through uni or researchers need to work directly with the raw data, making normalization and integration in a unique step, which is both statistically and computationally challenging and a topic of current research.

Data integration can be performed through N-integration (variables integration), which consider different omics platforms evaluated on the same samples, or P-integration (sample unities integration), i.e., concatenation across studies on the same variables. Typical techniques for database N-integration use multivariate projection-based methods, as low-rank models, that embed both the sample unities and features of the data blocks into the same low dimensional vector space (Lê Cao et al., 2009; Tenenhaus et al., 2011, 2014; Ray et al., 2017). These low dimensional vectors enable effective data analytics, such as clustering, visualization and missing value imputation. In addition, these vectors are latent variables or scores possibly representing biologically relevant molecular signatures and their analysis can suggest novel biological hypotheses.

Another class of N-integration techniques is based on a flexible regression framework. Under an unsupervised approach, probabilistic graphical models (PGMs) can be used for learning relations among multiple variables (Meinshausen et al., 2006). Tenenhaus et al. (2014) proposed a generalized canonical correlation analysis for N-integration with loads in the optimization problem defined in terms of the connections in a PGM. In addition, supervised N-integration can be performed by incorporating varying coefficients (Hastie and Tibshirani, 1993) into the regression model, with the multi-omics integration oriented for prediction of clinical outcomes. In this context, Ni et al. (2018) proposed a Bayesian hierarchical varying-sparsity regression model and apply for genomic and proteomic data integration to be prognostic for the patient's survival time.

Further, the P-integration of independent data sets measured on the same common set of variables (omics data) can be a useful opportunity to increase sample size and gain statistical power. The main challenge in this case is to prevent the analysis from systematic heterogeneities arising from the different sources of variation, as those coming from different protocols. For instance, batch and multi-center effects are unwanted variation, which often acts as strong confounders in the P-integration analysis. Such effects may lead to spurious conclusions if they are not accounted for in the statistical model.

Despite the recent progress made in the area of multi-omics integration, the methods assume independent observations (unrelated individuals), and if family structure is present and ignored in the analysis, such substructures may induce artefactual results for data integration. For instance, in the context of uni-omics data, specifically considering large pedigrees and high dimensional SNP-genotype data, de Andrade et al. (2015) obtained valid principal components estimators and showed that the latent variables taking into account the family structure are more informative than those ignoring such substructure. Ribeiro and Soler (2018), who proposed a probabilistic graphical model for learning relationships among multiple variables from family data, also consider the impact of clustered observation at the analysis. The outline of this work is as follows. First, we will review and discuss unsupervised and supervised multi-omics data integration methods, under the assumption of unrelated samples. Subsequently, we will consider family based designs, incorporate dependence among related individuals and exploit how the covariance matrix among variables is decomposed into genetic and environmental components. Finally, we will discuss the advancement of data integration methods to take into account family structure present on the data.

2. Methodology

A detailed review of multi-omics integration is presented by Huang et al. (2017). All efforts are dedicated to fully account for the uncertainties and heterogeneities in the datasets. Figure 1 shows a schematic representation of the datasets structure involved in Omic's studies. Based on matrix factorization approaches, unsupervised and supervised analysis have been used. In R package, mixOmics (Lê Cao et al., 2009; Rohart et al., 2017) is a powerful resource for integration of multi-omics datasets. In this case, multivariate projection-based methods are proposed to summarise datasets, $X_{n\times p}$, by latent components or scores ($F_{n\times m}$) and loadings ($W_{p\times m}$), such that$X \approx FW'$, m ≤ min(n, p). To properly do data reduction, different optimization problems are formulated to attain objective functions. For unsupervised uni-omics analysis, principal components or its improved version via independent components are used, and for unsupervised multi-omics, generalized canonical correlation can be a useful strategy. Considering supervised contexts, discriminant analysis combined with partial least square have been proposed. In all cases, regularised and sparse solutions are required.

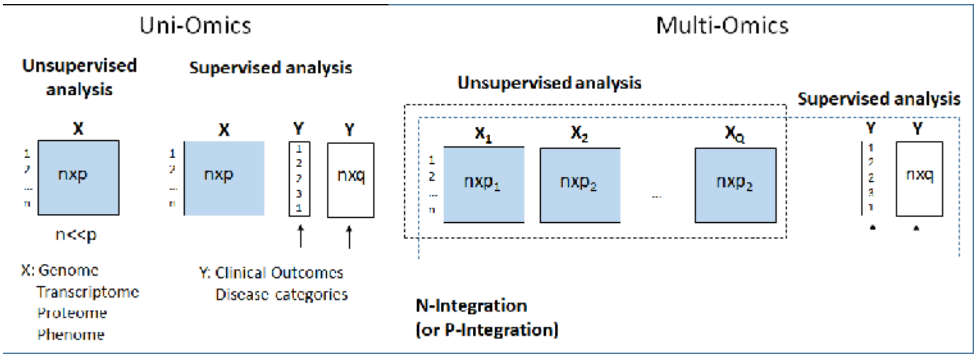

Figure 1. Schematic representation of datasets integration in Omics studies.

Regression models are powerful tools for supervised multi-omics integration. Ni et al. (2018) proposed an interesting varying coefficients regression model, which allow integration of multi-omics datasets driven for prediction of target outcomes. The model is flexible to take in account subject-specific coefficient estimation, i.e, on the patient level. Under regression formulation, regulatory axes given by proteomic (X_1) and genomic (X_2) data are connected to build clinically relevant prognostic through $Y_i \approx \sum X_{1ij}\beta_j(X_{2ij})$, where the varying coefficients $\beta_j(X_{2ij})$ define gene-protein interactions by adopting smooth functions of X_{2ij}.

All of those methods assume independent observations, and are not applied for family-based data, which are very common in genomic studies. Family data are mainly analysed using mixed model approaches that allow including familial dependences among observations. For based family multivariate data, let X_f be a vector for all p variables and all members of the f-th family, with covariance matrix given by $\Omega = 2\phi_f \otimes \Sigma_g + I_f \otimes \Sigma_e$, where $2\phi_f$ is the kinship matrix for family f, Σ_g and Σ_e are $(p \times p)$ covariance matrix associated with polygenic and error component, respectively, and $\otimes$ is the Kronecker product. Oualkacha et al. (2012) obtained MANOVA based estimators for these covariance matrices. De Andrade et al. (2015) obtained principal components of heritability for reduction of genomic dataset in terms of ancestry scores. Different scores can be obtained from family data by operating on the covariance components, i.e., Σ_g, Σ_e, $\sum_e^{-1} \Sigma_g$ as well as $\Sigma = \Sigma_g + \Sigma_e$. Following this idea, Ribeiro and Soler (2018) proposed to learn polygenic, environmental and total graphical models from family dataset exploiting Σ_g, Σ_e and $\Sigma = \Sigma_g + \Sigma_e$. The authors also exploit to learn the multivariate relations among variables based on a univariate polygenic mixed models framework.

For multi-omics integration in family data, we are extending the multivariate projection-based methods available for independent observations to include familial dependences. Under quadratic solutions, in $\Re^{p \times p}$, it is performed considering the factorization of the polygenic and environmental components of the covariance matrix. In addition, for rectangular solutions, in $\Re^{n \times p}$, N-integration can be performed structuring data matrix through ANOVA-simultaneous component analysis (Smilde et al., 2005) and then building the reductions on the components of the data.

3. Results

Figure 2 shows two representations of n observations clustered in family structure. In (a) it is assumed independent observations, where the principal components are extracted from covariance matrix Σ. In (b) familial dependences are taking in account, where the principal components are extracted from matrix $\sum_e^{-1} \Sigma_g$. Different colors are used to discriminate members from different families. The uni-omics dataset correspond to genotype information obtained from SNP platform (Affymetrics 6.0). A detailed description of the dataset is in de Andrade et al. (2015). The figure illustrates the impact of modelling family structure on the reduction analysis. When familial dependence is used more adaptive representation of the data is obtained, allowing discriminate members between the ancestry arms found in the analysis.

Figure 3 shows probabilistic graphical models learned from family data considering multiple phenotypes extracted from the Baependi Heart Study (Oliveira et al., 2006; Egan et al., 2016). In the figure, vertices represent variables and the connections indicate partial correlations between variables. Important differences are found on the relations obtained from patterns coming from the polygenic, environmental or total covariance matrices.

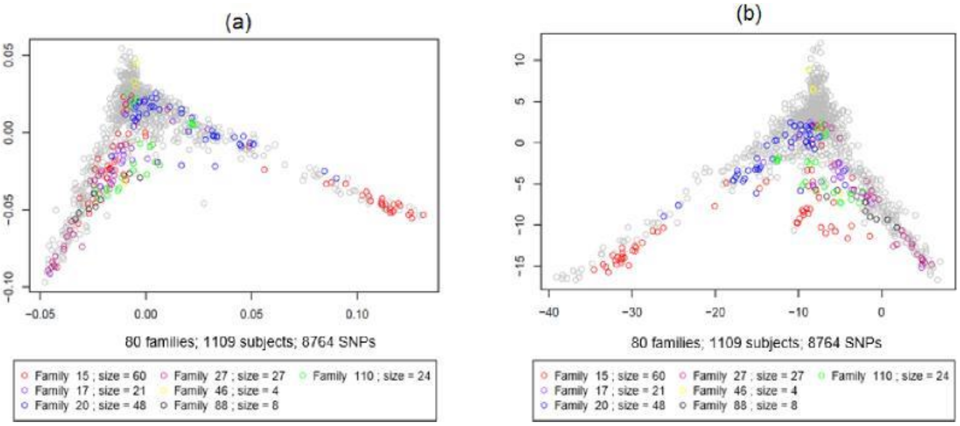

Figure 2: Representation of observations clustered in family structure. In (a), principal components were obtained under independent observations assumption. In (b) principal components are obtained by assuming familial dependences.

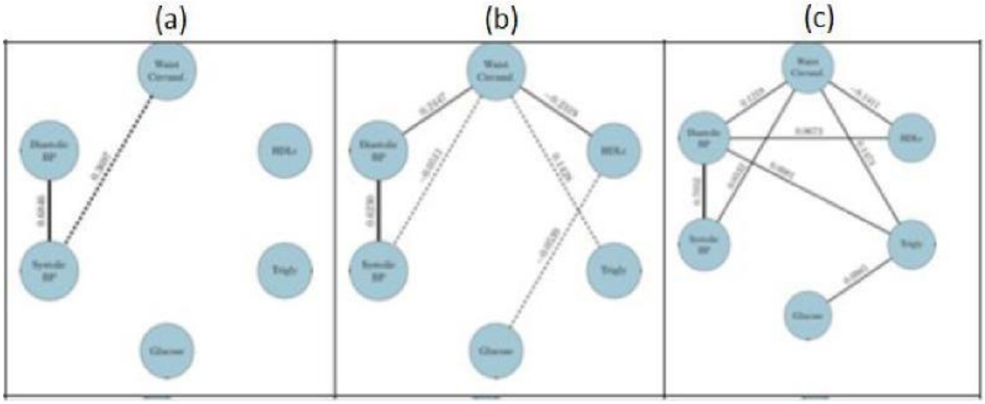

Figure 3. Probabilistic graph models to structure learning from family data. Vertices are metabolic syndrome variables: waist circumference (cm), diastolic blood pressure (mmHg), systolic blood pressure (mmHg), fasting glucose (mg/dL), triglycerides (mg/dL) and HDL-cholesterol (mg/dL). Connections indicate partial correlation between variables. In (a), polygenic covariance matrix, Σ_g, is analyzed. In (b), environmental covariance matrix, Σ_e, is used. In (c), the total covariance matrix, $\Sigma = \Sigma_g + \Sigma_e$, is used.

4. Discussion and Conclusion

It is widely recognized that integrative multi-omics analysis holds an important role for precision medicine. Despite the recent progress in the area, data integration remains a challenge, requiring combination of several software tools, mainly through bioinformatics pre-processing procedures, and extensive statistical expertise to appropriate account for the properties of heterogeneous data. To fully account for the uncertainties, data structure should be taking in account on the analysis, as integration of unsupervised or supervised datasets, N-integration or P-integration, big-n problem, independent versus dependent observations, etc. All of these topics impose challenges for conduction the analysis.

The main expected result in datasets integration is the representation of the observations under a reduced dimension, which is committed to optimizing any objective function that establishes relations among the datasets. Such relations can be based on covariance matrices or prediction functions, according to unsupervised or supervised proposals, respectively. Here we focused mainly in methods derived from matrix factorization and regression models. Most of the analyses available consider independent observations, but several multi-omics studies are based on family data that impose familial dependences among observations.

For multi-omics integration in family data we are considering strategies that decompose the problem to polygenic components integration and environmental components integration. It is a direct extension of the need to include random effect when analysing data with dependencies. Each data block is decomposed into two covariance matrices modelling different types of variation, one due the polygenic random effect, that is sharing among members from the same family and represents among-family variation, and another due the error random effect (environmental), that is the within-family variation. Then, it is performed low-rank approximation of the polygenic variation across the blocks, and low-rank approximations of the environmental variation components. The rational of our approach have been used in other contexts. Feng et al. (2018), addressing the matrices decomposition problem in datasets integration, proposed the angle-based joint and individual variation explained method that allow to compute block scores, block loadings, global loadings and global scores. We are working on the computational implementation of our methods by using the R package facilities.

References

1. Chen, C et al. (2011) Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PloS One* 6(2): e17238.
2. Clough, T et al. (2012). Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. BMC Bioinformatics 13(Suppl 16): S6.
3. de Andrade, M et al. (2015). Global Individual Ancestry Using PCs for Family Data. *Human Heredity* 80: 1-11.
4. Egan, KJ et al. (2016). Cohort profile: the Baependi Heart Study—a family-based, highly admixed cohort study in a rural Brazilian town. *BMJ Open 6*: 1:8.
5. Feng et al. (2018). Angle-based joint and individual variation explained. Journal of Multivariate Analysis 66: 241-265.
6. Hastie, T.; Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society*, Series B (Methodological): 757-796.

7. Huang, S.; Chaudhary, K; Garmire, L.X. (2017). More is better: Recent progress in Multi-Omics data integration methods. Frontiers in Genetics 8, Article 84: 1-12.
8. Irizarry, RA et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249-264.
9. Lê Cao, KA; González I; Déjean, S. (2009). IntegrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* 25: 2855-2856.
10. Leek, JT; Storey, JD. (2007) Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PloS Genetics* 3 (9): e161.
11. Leek. JT et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11(10): 1-15.
12. Meinshausen, N; Bühlmann, P. et al. (2006). Highdimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34: 1436-1462.
13. Mitra, V et al. (2016). Identification of Analytical Factors Affecting Complex Proteomics Profiles Acquired in a Factorial Design Study with Analysis of Variance: Simultaneous Component Analysis. *Analytical Chemistry* 88: 4229-4238.
14. Ni, Y. et al. (2018). Bayesian Hierarchical Varying-sparsity Regression Models with Application to Cancer Proteogenomics. *Journal of the American Statistical Association* 0(0): 1-13, Applications and Case Studies.
15. Oliveira, CM et al. (2008). Heritability of cardiovascular risk factors in a Brazilian population: Baependi Heart Study. *BMC Medical Genetics* 32:1-8.
16. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38: 904-909.
17. Oualkacha, K. et al. (2012). Principal Components of Heritability for High Dimension Quantitative Traits and General Pedigrees. *Statistical Applications in Genetics and Molecular Biology* 11(2), Article 4:
18. Ray, B.; Liu, W.; Fenyö, D. (2017). Adaptive Multiview Nonnegative Matrix Factorization Algorithm for Integration of Multimodal Biomedical Data. *Cancer Informatics* 16: 1-12.
19. Ribeiro, A.H.; Soler, J.M.P. (2018). Learning Genetic and Environmental Graphical Models from Family Data. In Annals of the XXIXth International Biometric Conference, in Barcelona, Spain, July 8-13th, 2018. (article submitted to Statistics in Medicine)
20. Rohart, F. et al. (2017). mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLoS Comput Biol* 13(11): 1-19.
21. Smilde, A.K. et al. (2005). ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics* 21(132005): 3043–48.

22. Tenenhaus A.; Tenenhaus M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika* 76(2): 257-284.
23. Tenenhaus, A. et al. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics* 15(3): 569-83.
24. Wolfinger, RD et al. (2001) Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models. *Journal of Computational Biology* 8(6): 625-637.