5

A Retrospective Likelihood Approach for Efficient Integration of Multiple Omics Factors in Case-Control Association Studies [1]

Summary

Integrative omics, the joint analysis of outcome and multiple types of omics data, such as genomics, epigenomics and transcriptomics data, constitutes a promising approach for powerful and biologically relevant association studies. These studies often employ a case-control design, and often include non-omics covariates, such as age and gender, that may modify the underlying omics risk factors. An open question is how to best integrate multiple omics and non-omics information to maximize statistical power in case-control studies that ascertain individuals based on the phenotype. Recent work on integrative omics have used prospective approaches, modeling case-control status conditional on omics and non-omics risk factors. Compared to univariate approaches, jointly analyzing multiple risk factors with a prospective approach increases power in non-ascertained cohorts. However, these prospective approaches often lose power in case-control studies. In this article, we propose a novel statistical method for integrating multiple omics and non-omics factors in case-control association studies. Our method is based on a retrospective likelihood function that models the joint distribution of omics and non-omics factors conditional on case-control status. The new method provides accurate control of Type I error rate and has increased efficiency over prospective approaches in both simulated and real data. The method is publicly available at `https://github.com/BrunildaBalliu/IntegrativeOmics`.

[1] Published in *Genetic Epidemiology*.

5.1 Introduction

Recent advances in technology have made it possible to collect multiple types of omics data, such as genomics, transcriptomics, and epigenomics in the same individuals. Genome-, transcriptome-, and epigenome-wide association studies have led to the identification of genetic variants, transcripts, and methylation sites associated with many complex diseases [Edgar et al., 2002; Hindorff et al., 2009; Lv et al., 2012]. However, due to the lack of integrative statistical approaches, these associations were mainly identified through their marginal effects on disease risk. As a result, underlying disease mechanisms through which omics factors affect phenotypes, e.g. joint effects or mediation effects, remain unknown for most complex diseases. Integrative omics studies, the joint analysis of outcome and multiple omics data, have emerged as a promising alternative to more powerful and biologically informative association studies [Chen et al., 2008; Li, 2013; Zhao et al., 2014; Huang et al., 2014].

Here, we are interested in leveraging integrative omics approaches to identify associations between a genetic variant G, a transcript E or a methylation site M and a binary outcome Y, accounting for environmental or clinical factors X. In randomly ascertained studies, when E, M, G and X have independent effects on Y, modeling them jointly can increase power to detect associations between Y and any of E, M, and G [Robinson and Jewell, 1991; Neuhaus and Jewell, 1993; Neuhaus, 1998]. However, E, M, and G can be correlated, e.g. genetic variants can alter gene expression and DNA methylation [Schadt et al., 2003; Zhang et al., 2010] and DNA methylation can regulate gene expression [Gutierrez-Arcelus et al., 2013]. In such scenarios, E and M can act as mediators of G, and testing for their joint effect on Y can be more powerful than testing only for genetic associations [Huang et al., 2014; Zhao et al., 2014]. Moreover clinical covariates X, such as age and gender, can be associated with M and/or E [Richardson, 2003; Horvath et al., 2012; Liu et al., 2010; Dimas et al., 2012; Glass et al., 2013]. Consistent with previous approaches, we make the assumption of independence between X and G [Umbach and Weinberg, 1997; Chatterjee and Carroll, 2005]. In addition to increasing power for G, clinical covariates can confound the effect of E and M on Y, thus including them in the analysis is necessary in order to control bias and prevent false discoveries. Figure 5.1.a illustrates the relationships between E, M, G, X, and Y in a randomly ascertained population cohort.

Integrative omics studies typically employ a case-control design. Since cases are enriched for all risk factors, ascertainment will induce additional correlation between E, M, G and X (Figure 5.1.b). Existing methods for integrative omics analysis use prospective approaches to model the distribution of the case-control status conditional on the risk factors, in our case $P(Y|E,M,G,X)$ [Huang et al., 2014; Zhao et al., 2014]. In these ascertained studies, prospective approaches will not account for the sampling scheme, potentially resulting in severe power loss relative to univariate analyses of each risk factor [Chatterjee and Carroll, 2005; Xing and Xing, 2010; Zaitlen et al., 2012a,b; Pirinen et al., 2012; Mefford and Witte, 2012]. The main reason for this power loss is that, in studies for which ascertainment is based on outcome, the distribution of the risk factors $P(E,M,G,X)$ contains information about the parameters in $P(Y|E,M,G,X)$ [Scott and Wild, 2001]. In a prospective approach $P(E,M,G,X)$ is ignored when making inference and as a result such methods will be less efficient than methods that explicitly make use of $P(E,M,G,X)$.

In this article, we propose a novel integrative omics approach that addresses these issues of power loss in case-control association studies. Our approach is based on a retrospective likelihood function that models the joint distribution of the omics and non-omics factors conditional on the case-control status, $P(E,M,G,X|Y) = P(Y|E,M,G,X) \times P(E,M,G,X)/P(Y)$. In order to model $P(E,M,G,X)$ as efficiently as possible, we exploit knowledge about the correlation structure between risk factors and the distribution of

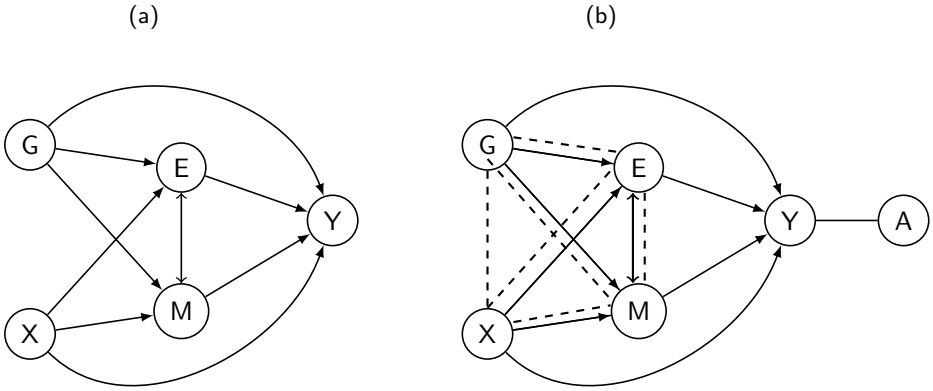

Figure 5.1: Example to illustrate possible correlation structures among risk factors and a trait in (a) a random sample and (b) a case-control sample. G: genetic variant, E: gene expression, M: DNA methylation, X: non-omics covariate, Y: trait/disease, A: ascertainment of cases and controls. Continuous arrows between two nodes connect variables that could be correlated in the population while dashed lines represent induced correlations due to ascertainment.

the risk factors in the population by making parametric assumptions about $P(E, M, G, X)$. When these distributional assumptions hold, the corresponding maximum likelihood estimates are unbiased and statistically efficient, in that they have the smallest variances among all valid estimators, and the corresponding association tests are the most powerful among all valid tests [Chatterjee and Carroll, 2005; Lin and Zeng, 2009].

The use of a retrospective approach to exploit the gene-environment independence assumption in case-control genetic associations studies was originally proposed by Chatterjee and Carroll [2005]. The method accommodates genetic and environmental covariates that are independent in the underlying population or that are independent conditional on some other factors. Our work is an extension of this method to accommodate situations in which independence or conditional independence assumptions for genetic and additional omics risk factors do not hold. Moreover, our approach can accommodate continuous risk factors by using parametric distributions.

The rest of the paper is organized as follows. In Section 5.2, we introduce the method, the assumptions about the distribution of omics and non-omics risk factors and describe the statistical testing. In Section 5.3, we evaluate the finite sample performance of the proposed method using an extensive simulation study. We compare our method with a prospective likelihood approach and show that our method has increased efficiency and power under many realistic disease models while maintaining a properly controlled type I error. In Section 5.4 we demonstrate that our approach is more efficient than the prospective approach when analyzing omics data from a multiple sclerosis study [Huynh et al., 2014]. We also describe how the models can be modified when not all types of data are available. We close with a discussion in Section 5.5.

5.2 Material and Methods

5.2.1 The Statistical Model

Consider a case-control study of N subjects, N_1 cases and N_0 controls, where for each subject, information on genetic variation, DNA methylation, expression, and one or more clinical or environmental covariates is available. If one or more of the data sources is not available, as is the case in our real data example, the following models can be modified accordingly. Here we focus on a single genetic, epigenetic and transcriptional measurement per subject. In the Discussion section, we consider extensions for the high dimensional setting. Let $\mathbf{Y} = (Y_1, \ldots, Y_N)$ be the vector of phenotypes for the subjects in the study, with Y_i a binary indicator of disease status, i.e. $Y_i = 1$ if i is affected and 0 if i is unaffected. Similarly, let $\mathbf{G}$, $\mathbf{M}$, and $\mathbf{E}$ be vectors of a genetic, an epigenetic and a transcriptional factor of the N subjects, respectively. Last, let $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_J})$ denote the matrix of J clinical covariates for the N subjects.

The prospective likelihood models the distribution of the disease status conditional on the potential risk factors and is given as follows,

$$\mathcal{PL}(\boldsymbol{\alpha}) = P\left(\mathbf{Y}|\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right), \tag{5.1}$$

where $\boldsymbol{\alpha}$ is the parameter vector of the effect of risk factors on disease risk. Moreover, the prospective risk model for subject i, given its risk factors, is given by the logistic regression model

$$P\left(Y_i = 1|G_i, M_i, E_i, \mathbf{X}_i\right) = logit^{-1}\{\alpha_0 + \alpha_G G_i + \alpha_M M_i + \alpha_E E_i + \mathbf{X}_i\boldsymbol{\alpha}_X\}, \tag{5.2}$$

where $logit^{-1}\left(x\right) = \frac{e^x}{1+e^x}$, α_0 the intercept and α_G, α_M, α_E and $\boldsymbol{\alpha}_{\mathbf{X}}$ the effect of G, E, M and X on disease risk. In this work we consider only main effects of the risk factors. However, more general models, with interaction of different orders between the risk factors, could also be used.

On the other hand, the retrospective likelihood models the distribution of risk factors conditional on the disease status and is given as follows,

$$\mathcal{RL}(\boldsymbol{\theta}) = P\left(\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}|\mathbf{Y}\right) = \frac{P\left(\mathbf{Y}|\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right) \times P\left(\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right)}{P\left(\mathbf{Y}\right)}, \tag{5.3}$$

where $\boldsymbol{\theta}$ is the parameter vector containing the effect of risk factors on disease risk and parameters for characterizing the distribution of risk factors. The numerator in (5.3) is a product of the prospective risk model and the joint distribution of the risk factors. The denominator represents the marginal disease probability in the population.

The challenge in maximizing the retrospective likelihood (5.3) with respect to $\boldsymbol{\theta}$ is due to the unknown covariate distribution $P\left(\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right)$. It is well known that if no assumption is made about the form of the covariate distribution, $P\left(\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right)$ is not identifiable from case-control data [Prentice and Pyke, 1979]. Furthermore, Rabinowitz [1997] and Breslow et al. [2000] showed that, if $P\left(\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right)$ is treated fully non-parametrically, the efficiencies of (5.1) and (5.3) are equivalent. To optimally model $P\left(\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right)$, and increase efficiency for estimating the parameters of interest, we exploit knowledge about the correlation structure between risk factors. Specifically, we assume that E and M can be correlated, G and X can be associated with E and M, and that G and X are independent of each other in the population.

Thus $\mathcal{RL}$ is further factorized as follows

$$\mathcal{RL}(\boldsymbol{\theta}) = \frac{P\left(\mathbf{Y}|\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right) \times P\left(\mathbf{M}, \mathbf{E}|\mathbf{G}, \mathbf{X}\right) \times P\left(\mathbf{G}\right) \times P\left(\mathbf{X}\right)}{P\left(\mathbf{Y}\right)}. \tag{5.4}$$

Explicitly imposing independence between G and X will result in efficiency gain for estimating α_G and $\boldsymbol{\alpha}_X$, compared to approaches that ignore their distribution or do not exploit this assumption [Chatterjee and Carroll, 2005].

To further increase efficiency, we exploit knowledge about the distribution of the omics factors. Specifically, we make the HWE assumption to model $P(G)$. Under this assumption, $G \sim Binomial(2, p)$ with p the minor allele frequency so that $P(G)$ is characterized by a single parameter. This will increase efficiency to estimate α_G. Moreover, we assume that, after proper transformations and normalization procedures, the epigenetic and transcriptional factors are normally distributed and therefore use a multivariate normal for their joint distribution [Calza and Pawitan, 2010; Yousefi et al., 2013]. This parametric model will result in efficiency gain for estimating α_M and α_E, compared to methods that treat the distribution of E and M non-parametrically. We model the conditional distribution of M and E using a multivariate linear regression model:

$$\begin{aligned} M_i &= \beta_{GoM} G_i + \mathbf{X}_i^T \boldsymbol{\beta}_{\mathbf{XoM}} + \epsilon_{i1} \\ E_i &= \beta_{GoE} G_i + \mathbf{X}_i^T \boldsymbol{\beta}_{\mathbf{XoE}} + \epsilon_{i2} \end{aligned} \tag{5.5}$$

where β_{GoM}, $\boldsymbol{\beta}_{\mathbf{XoM}}$, β_{GoE} and $\boldsymbol{\beta}_{\mathbf{XoE}}$ are the effects of the genetic and clinical factors on the epigenetic and transcriptional factor. We assume that the errors $(\epsilon_{i1}, \epsilon_{i2})^T$ follow a bi-variate normal distribution, $\mathbf{MVN}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}\sigma^2_{\epsilon_1} & \sigma_{\epsilon_1\epsilon_2}\\ \sigma_{\epsilon_1\epsilon_2} & \sigma^2_{\epsilon_2}\end{bmatrix}\right)$, with $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$, and $\sigma_{\epsilon_1\epsilon_2}$ the variances and co-variance of ϵ_1 and ϵ_2. Notice that in the model we center M, E, G and $\mathbf{X}$ around zero such that there is no need for intercepts in (5.5).

For parametrization of the distribution of $\mathbf{X}$, we assume that $\mathbf{X_1}, \ldots, \mathbf{X_J}$ are mutually independent in the population and factorize their distribution as $P(\mathbf{X}) = \prod_{j=1}^{J} P(\mathbf{X_j})$. This assumption will increase efficiency to estimate each α_X and can be relaxed when independence is not plausible. For simplicity of exposition, we focus on binary $\mathbf{X}$ and model them using binomial distributions. In the Discussion section, we consider problems arising from deviations from the assumed correlation structure and distributions of the risk factors, and propose solutions to address them.

Last, we specify the marginal distribution of Y. Since $P(Y)$ is usually not known, we compute it by marginalizing the joint distribution $P(Y, G, M, E, \mathbf{X})$ over all possible values of G, M, E and $\mathbf{X}$, denoted by G^*, M^*, E^* and $\mathbf{X}^*$. Therefore, we need to compute a two dimensional integral over M and E for all possible values of G and $\mathbf{X}$:

$$P(\mathbf{Y}) = \sum_{G^*, \mathbf{X}^*} \int_{E^*, M^*} P(\mathbf{Y}|G^*, E^*, M^*, \mathbf{X}^*) P(E^*, M^*|G^*, \mathbf{X}^*) P(G^*) P(\mathbf{X}^*) d_{E^*, M^*}$$

There exists no closed form solutions for this integral, thus numerical methods need to be employed to compute the integral and maximize the likelihood. Here, we use the Gauss-Hermite Quadrature for numerical integration and the R package `optim` for numerical optimization.

5.2.2 Statistical Testing

We wish to test the null hypothesis of no omics effect on disease risk. The null hypothesis can be written using the regression coefficients in (5.2) as:

$$\begin{aligned} H_0 &: \alpha_G = \alpha_M = \alpha_E = 0 \text{ versus} \\ H_1 &: \text{at least one of } \alpha_G, \alpha_M, \alpha_E \neq 0. \end{aligned} \tag{5.6}$$

Likelihood-based statistics can be used to make inference about the parameters of main interest. Here, a likelihood ratio test (LRT) is used to test the null hypothesis. Following standard likelihood theory, the LRT statistic under the null hypothesis asymptotically follows a χ^2_3 distribution for a correctly specified model.

In the simulation study below, we examine the impact of model misspecification on the distribution of the test statistic under the null and alternative hypothesis.

5.3 Simulation Study

We wish to compare the relative performance of our proposed $\mathcal{RL}$ approach with the $\mathcal{PL}$. We present results on type I error rate, bias, efficiency and power. We also study the performance of the methods under the null hypothesis when the joint distribution of M and E deviates from normality. In each scenario described below, 1000 replication data sets were simulated. In each replication, we generated data for 500 cases and 500 controls by sampling the cases and controls from a larger random sample of subjects.

Since in our real data set we have information on age (A) and gender (S) of the subjects, we also consider these two clinical covariates. Thus, in all the formulas above, $\mathbf{X} = (\mathbf{S}, \mathbf{A})$, $\boldsymbol{\alpha}_X = (\alpha_S, \alpha_A)^T$, $\boldsymbol{\beta}_{XoM} = (\beta_{SoM}, \beta_{AoM})$, and $\boldsymbol{\beta}_{XoE} = (\beta_{SoE}, \beta_{AoE})$. Age was treated as binary, with 0 indicating an individual with age younger than or equal to the median age in the population, i.e. 35, and 1 otherwise. In all scenarios below, binary age and gender were considered to be mutually independent. Thus $P(\mathbf{S}, \mathbf{A}) = P(\mathbf{S}) \times P(\mathbf{A})$. Age was generated from a $Binomial(1, .5)$, gender was generated from a $Binomial(1, .5)$, with zero indicating a male and one indicating a female, and the genetic variant G was generated from a $Binomial(2, .20)$. In all scenarios presented in the main text, in order to speed up the computation time, $\sigma^2_{\epsilon_1}$, $\sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1,\epsilon_2}$ were fixed to their sample estimates, $\hat{\sigma}^2_{\epsilon_1}$, $\hat{\sigma}^2_{\epsilon_2}$ and $\hat{\sigma}_{\epsilon_1,\epsilon_2}$, and they no longer were part of the optimization procedure. In the simulation scenarios presented in the Appendix, $\sigma^2_{\epsilon_1}$, $\sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1,\epsilon_2}$ were part of the optimization procedure.

5.3.1 Type I Error

First we studied the performance of the two methods in terms of type I error rate, when the distribution of the errors in (5.5) was properly specified. M and E were generated from (5.5) with no genetic, age or gender effect and normally distributed errors with $\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = 1$ and $\sigma_{\epsilon_1\epsilon_2} = 0$. Other values were also tested but results are similar and are not shown. The binary disease outcome was generated from (5.2) with no genetic, epigenetic or transcriptional effect. The effect of age and gender was also set to zero, although this is not necessary and different values can be chosen. We set the intercept $\alpha_0 = logit(1e-03)$, such that the marginal disease probability in the population would be approximately $P(Y = 1) = .1\%$, reflecting a common disease with relatively low prevalence, such as multiple sclerosis. The type I error rate for $\mathcal{PL}$ and $\mathcal{RL}$ was 5.4% and 5.6%, respectively.

Next, we studied the performance of the methods when the normality assumption for the distribution of the errors in (5.5) was violated. To mimic situations in which outliers are present, we simulated the errors from a bi-variate t-distribution with 10 degrees of freedom and same location and scale parameters as the normal case. All other parameters remain the same. All methods properly control for the type I error rate; type I error rate for $\mathcal{PL}$ and $\mathcal{RL}$ was 4.7% and 5.1%, respectively.

5.3.2 Bias and Efficiency

Two different scenarios were considered. In the first scenario, the risk factors had moderate effect on disease risk. Specifically, parameters in (5.2) were set to $\alpha_E = \alpha_M = .18$, corresponding to an OR of 1.2, $\alpha_G = \alpha_S = \alpha_A = .26$, corresponding to an OR of 1.3. Moreover, parameters in (5.5) were set to values similar to our real data example, that is $\beta_{GoE} = \beta_{SoE} = \beta_{AoE} = \beta_{GoM} = \beta_{bSoM} = .1$, $\beta_{bAoM} = .3$, $\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = 1$ and $\sigma_{\epsilon_1\epsilon_2} = .3$. In the second scenario, we considered effect sizes for E, M, S and A on disease risk that were closer to our real data example. Specifically, parameters in (5.2) were set to $\alpha_E = \alpha_M = 1$, corresponding to an OR of 3, $\alpha_G = .26$, corresponding to an OR of 1.3, and $\alpha_S = \alpha_A = -.69$, corresponding to an OR of .5. α_0 was set to $logit(9e-04)$, such that the marginal disease probability in the population would again be $.1\%$. Results on bias and efficiency of parameter estimates, for both scenario and likelihood approaches, are listed in Table 5.1. To study the impact of fixing $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1\epsilon_2}$ to their sample estimates, we repeat the analysis in both the scenarios described above, but this time we estimate also $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1\epsilon_2}$. Results on bias and efficiency for this case are listed in Table 5.1 of the Appendix.

Based on these simulation results we make the following key observations. First, as expected from theory, both $\mathcal{PL}$ and the proposed $\mathcal{RL}$ estimators provide essentially unbiased estimators of all regression parameters. For scenario 2, the bias for both likelihood is slightly larger than the bias for Scenario 1. This small increase in bias stems from the fact that in scenario 2 the effect sizes are larger, as compared to scenario 1. As a consequence, the impact of the ascertainment is stronger and thus the information available to estimate the parameters of interest is more limited. As expected, part of the bias of the $\mathcal{RL}$ also comes from fixing $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1\epsilon_2}$ to their sample estimates. This can be seen by comparing the bias of the $\mathcal{RL}$ in Table 5.1 with the bias in Table 5.1 of the Appendix; bias for α_E decreases from 4.6% to 3% and bias for α_A decreases from 4.5% to 2.1%.

Secondly, ratios of variance estimates of the parameter estimates from $\mathcal{RL}$ and $\mathcal{PL}$ estimators show that, when the information on the distribution of covariates is exploited correcting for ascertainment in case-control data, there is a major efficiency gain for the estimation of the regression coefficients. The gain is larger for the scenario with larger effect sizes, as compared to smaller effect sizes; and for continuous, as compared to discrete covariates. Results for the LRT for testing (5.6) also agree with the efficiency results; the test based on $\mathcal{RL}$ offers a mean increase of 9.5 $\%$ in χ^2 test statistic for the first scenario and 22.2% for the second scenario (results not shown). Moreover, the gain is larger when $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1\epsilon_2}$ are estimated rather than fixed to their sample estimates. Third, comparison of the empirical standard errors (SE) with the estimated SE of the $\mathcal{RL}$ shows that the numerical approximation of the integral using Gauss-Hermite Quadrature and the numerical optimization algorithm perform well for realistic parameter values and modest sample sizes. Finally, the estimated SE when $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1\epsilon_2}$ are estimated are smaller, compared to the estimated SE when $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1\epsilon_2}$ are fixed to their sample estimates.

5.4 Data Example

In this section, we re-analyze data from a case-control study of 28 patients of multiple sclerosis and 19 controls. In the initial study, quantile-normalized DNA methylation, log_2 normalized gene expression data, as well as information on age and gender, was available for each subject. In the initial study, Huynh et al. [2014] analyzed DNA methylation and gene expression data sets separately, correcting for age and gender. Significant results from each analysis were compared and several genes showed overlapping signals in both DNA methylation and gene expression analysis.

Table 5.1: Simulation study for studying bias and efficiency of the prospective likelihood ($\mathcal{PL}$) and retrospective likelihood ($\mathcal{RL}$). The frequency of the genetic variant was .20, the frequency of category 0 for gender and age was .5. The disease prevalence in the population was .1%, corresponding to a common disease with low prevalence. VR, variance ratio; SE, standard error; Emp: Empirical, Est: Estimated. Results are based on the average over 1000 simulated data set. $\sigma^2_{\epsilon_1}$, $\sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1,\epsilon_2}$ in (5.5) were fixed to their sample estimates and were no longer part of the optimization procedure.

True	Bias		Emp SE		Est SE		MSE		VR
Values	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\frac{\mathcal{RL}}{\mathcal{PL}}$
				Scenario 1: Moderate effect sizes.					
$\alpha_E = .18$	.000	-.004	.068	.062	.069	.062	.005	.004	.910
$\alpha_M = .18$	.000	-.004	.068	.062	.070	.063	.005	.004	.911
$\alpha_G = .26$	.009	.009	.111	.109	.112	.111	.012	.012	.983
$\alpha_S = .26$	.004	.004	.130	.129	.136	.133	.017	.017	.986
$\alpha_A = .26$	.006	.007	.132	.129	.137	.133	.017	.017	.983
				Scenario 2: Large effect sizes.					
$\alpha_E = 1$	.013	-.046	.103	.091	.102	.088	.011	.010	.885
$\alpha_M = 1$	.011	-.045	.103	.091	.104	.090	.011	.010	.885
$\alpha_G = .26$	.011	-.030	.148	.137	.150	.138	.022	.020	.927
$\alpha_S = -.69$	-.008	-.027	.181	.169	.185	.178	.033	.029	.935
$\alpha_A = -.69$	-.005	-.025	.181	.169	.184	.173	.033	.029	.935

Here, we study one of the significant genes identified from the original analysis, SLC47A22, and apply both the prospective and the proposed retrospective likelihood approach. Age was treated as a binary variable, with 0 indicating an individual younger than or equal to 60 years old, which was the median age in our sample. The binary age and gender were considered to be independent in the population. For a binary age, this assumption is realistic, since, in 2010 in the United States, where our sample comes from, 83% of males were younger than 60 as opposed to 81 % of females [Howden and Meyer, 2011].

Since we do not have information for the genetic covariates, the two likelihoods are modified as follows:

$$\mathcal{PL}(\boldsymbol{\alpha}) = P\left(\mathbf{Y}|\mathbf{M},\mathbf{E},\mathbf{A},\mathbf{S}\right) \tag{5.7}$$

$$\mathcal{RL}(\boldsymbol{\theta}) = \frac{P\left(\mathbf{Y}|\mathbf{M},\mathbf{E},\mathbf{A},\mathbf{S}\right) P\left(\mathbf{M},\mathbf{E}|,\mathbf{A},\mathbf{S}\right) P\left(\mathbf{A}\right) P\left(\mathbf{S}\right)}{P\left(\mathbf{Y}\right)} \tag{5.8}$$

and the null hypothesis for the parameters of interest is now the following

$$H_0 : \alpha_M = \alpha_E = 0.$$

We assume that under the null hypothesis the LRT statistic is asymptotically χ^2_2 distributed.

For this gene, DNA methylation is available for 15 sites. Given the small size of our sample, we applied (5.7) and (5.8) 15 times, keeping the same E, A and S and adding a different methylation site in the model each time. Parameter estimates, standard errors and p-value for the LRT test for each model and method used, are listed in Tables A.5.3 - A.5.5

of the Appendix. In Figure 5.2, we plot the parameter estimates with their 95% confidence intervals (CI) for both methods.

Based on these results, we make the following observations. Our approach had smaller standard errors than $\mathcal{PL}$ approach in 11 out of 15 estimates for α_M, with an increase in efficiency of $5-20\%$, comparable standard errors in 3 out of 15 sites, with a $< 5\%$ increase or decrease in efficiency, and larger standard errors in 1 out of 15 sites, with a $4-7\%$ decrease in efficiency. The largest reduction in standard errors, 20%, was for the fourth methylation site. Moreover, site 9 was significant at nominal level when the $\mathcal{RL}$ was used and not when $\mathcal{PL}$ was used. Standard errors of α_E, α_S and α_A for $\mathcal{RL}$ were $3-8\%$ smaller than for $\mathcal{PL}$ when averaging across the 15 models.

5.5 Conclusions and Discussion

In this paper, we have proposed a statistical framework for efficient integration of omics and non-omics factors in case-control association studies. We used a retrospective likelihood approach to model the distribution of the risk factors conditional on the case controls status and performed a LRT for the joint effect of omics factors on disease risk. We demonstrated via simulation studies and real data analysis that the retrospective likelihood approach can be more efficient than the prospective likelihood when integrating data from case-control studies.

In order to compute the retrospective likelihood, we made certain assumptions about the correlation structure between the risk factors in the population. If evidence about additional independences exists, e.g. independence between E and M or their independence from G and X, our method can be modified accordingly. If G and X are not independent, e.g. due to population stratification, estimates of α_G and α_X could be biased. To address this issue, Chatterjee and Carroll [2005] proposed to model the distribution of G and X conditional on other common measured factors, such as principal components. Alternatively, Mukherjee and Chatterjee [2008] proposed to use an empirical Bayes-type shrinkage estimator that corrects for falsely attributed independence of covariates. For X binary, a multinomial distribution can be used for the joint distribution $P(\mathbf{X}, \mathbf{G})$. In addition, if X is a discrete variable with many levels or a continuous variable, the joint distribution could be factorized as $P(X, G) = P(X|G) \times P(X)$ and a Poisson or linear regression could be used. Last, we assumed the non-omics X's to be mutually independent. For age and gender this assumption can be verified using population registries. If this assumption is violated, methods proposed above to address the violation of G-X independence assumption can be used.

In addition to assumptions about the correlation structure between the risk factors, our method makes assumptions about the distribution of the risk factors. We assume that after proper transformations and normalization procedures, E and M are normally distributed [Calza and Pawitan, 2010; Yousefi et al., 2013]. When this assumption is violated, e.g. heavy tails or skewed distributions, our method could give biased parameter estimates (see Table A.5.2 of the Appendix). To avoid this issue, more flexible or discrete distributions can be considered for the error distributions of E and M, e.g. Laplace or negative binomial distribution [Purdom and Holmes, 2005; Sun, 2012]. Alternatively, quantile normalization techniques can be used to align the quantiles of E and M to a normal distribution. Such techniques can result in the dilution of the effects of the risk factors on disease risk and should therefore be used with caution. Moreover, the interpretation of parameters after the quantile normalization is no longer possible, thus we advice the use of different distributions rather than normalization. In our real data example, E and M were normalized prior to analysis [Huynh et al., 2014]. However, in such a small sample, it is difficult to verify normality and we did not formally test the fit. Among other possible reasons, deviations from normality could explain why the $\mathcal{PL}$ had in some cases similar or smaller standard errors

for α_M and α_E than the $\mathcal{RL}$. In this work, we treated age as binary, which might have decreased the gain in efficiency from exploiting the independence assumption for estimating α_S and α_A. Last, it is known that in small samples the logistic regression can give biased estimates of the OR's [Nemes et al., 2009] thus results in the real data for both methods should be interpreted with caution.

Efficiency of parameter estimation can be further increased using external knowledge about the disease prevalence or distribution of the covariates in the population. This information could be incorporated in several ways. Chatterjee and Carroll [2005] and Tsonaka et al. [2013] show how to incorporate external information about disease prevalence. Huijts et al. [2014] show how to increase efficiency for estimating genetic effects using existing genotype data from controls. Zaitlen et al. [2012a] show how to incorporate external information about the distribution of the risk factors based on the liability threshold model with parameters informed by external epidemiological information. The latter approach provides increased efficiency not only for phenotype based ascertainment but also for phenotype and covariate based ascertainment. Our approach could be modified in a similar way to include such information and this is among our future extensions.

In this article, we have considered a simple association approach by focusing on the joint association of a single genetic, epigenetic and transcriptional factor per gene with the phenotype. One possible way to accommodate settings with several genetic or epigenetic factors per gene is to use a mixed logistic regression in which the regression coefficient of the main genetic or epigenetic effects are assumed to follow an arbitrary distribution, e.g. the normal distribution [Huang et al., 2014]. Alternatively, penalized likelihood approaches, which put a separate penalty for the genetic and epigenetic factors, could be used. Furthermore, concepts from mediation analysis framework can be used to construct more powerful testing procedures. For example, here we test for the joint effect of the omics factors. This test can be considered as a test for the total effect of G on Y, directly on Y or indirectly via E and M. In the case of complete mediation of the genetic effect via E and M, i.e. α_G=0, a more powerful approach for assessing genetic associations would be to test only for the indirect genetic effect [Kenny and Judd, 2013; Zhao et al., 2014]. These possibilities are among our future research interests.

In summary, the retrospective likelihood based inference can be more efficient than prospective based inference for joint analysis of multiple omics and non-omics risk factors in case-controls association studies. Efficiency gain is a function of the number of parameters used to model the distribution of the risk factors and the effect sizes of risk factors, with increased efficiency gain for continuous factors and for risk factors with large effect sizes.

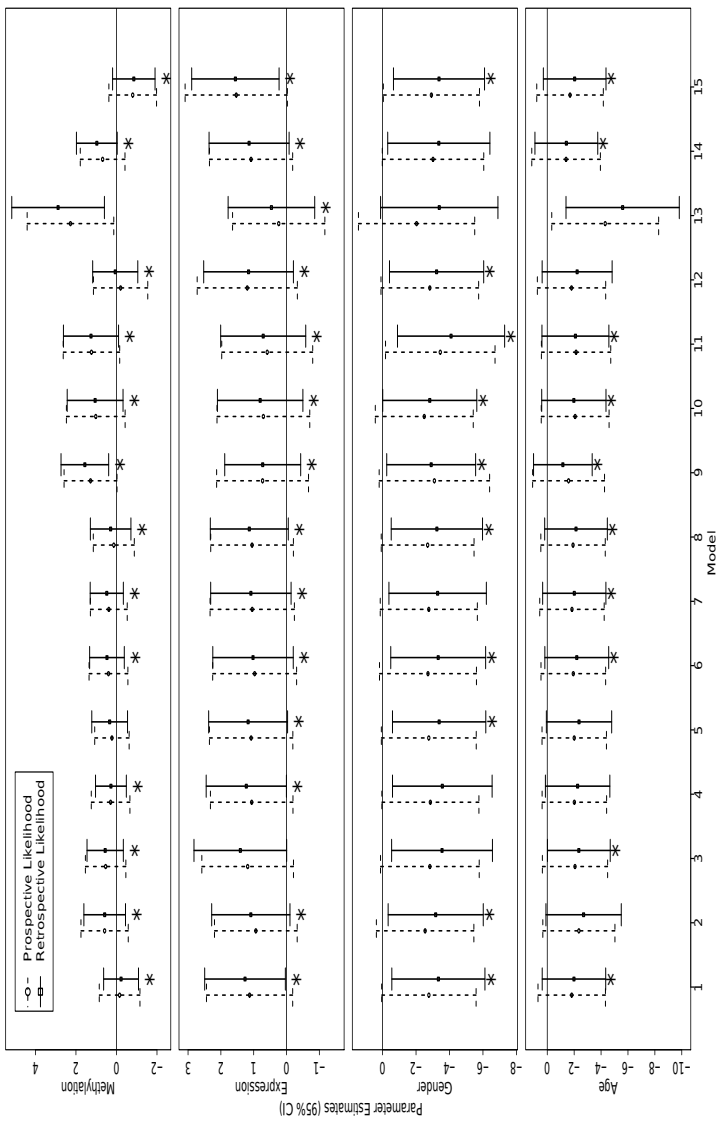

Figure 5.2: Results from applying prospective (5.1) and retrospective (5.4) approaches in the multiple sclerosis data. Each plot (top to bottom) shows estimates of the OR parameters, with their 95% confidence intervals, for the effect of methylation, gene expression, gender and age on multiple sclerosis in each of the 15 models fitted. The asterisk (*) denotes cases in which the confidence intervals of a parameter estimated using the retrospective likelihood are narrower compared to the confidence intervals estimated using the prospective approach.

Appendix

Table A.5.1: Simulation study for studying bias and efficiency of the prospective likelihood ($\mathcal{PL}$) and retrospective likelihood ($\mathcal{RL}$). The frequency of the genetic variant was .20, the frequency of category 0 for gender and age was .5. The disease prevalence in the population was .1%, corresponding to a common disease with low prevalence. VR, variance ratio ; SE, standard error; Emp: Empirical, Est: Estimated. Results are based on the average over 1000 simulated data set. $\sigma^2_{\epsilon_1}$, $\sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1,\epsilon_2}$ in (5.5) were part of the optimization procedure.

True	Bias		Emp SE		Est SE		MSE		VR
Values	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\frac{\mathcal{RL}}{\mathcal{PL}}$
Scenario 1: Moderate effect sizes.									
$\alpha_E = .18$	.003	.006	.068	.065	.068	.051	.005	.004	.948
$\alpha_M = .18$	.003	.000	.068	.065	.070	.055	.005	.004	.952
$\alpha_G = .26$	-.001	.015	.111	.109	.116	.086	.012	.012	.986
$\alpha_S = .26$	.003	.026	.130	.130	.130	.103	.017	.017	.994
$\alpha_A = .26$	-.004	.024	.132	.131	.126	.098	.017	.018	.992
Scenario 2: Large effect sizes.									
$\alpha_E = 1$	.019	.030	.103	.091	.106	.061	.011	.009	.882
$\alpha_M = 1$	.011	.021	.103	.090	.104	.059	.011	.009	.879
$\alpha_G = .26$	.006	.028	.148	.135	.144	.084	.022	.019	.914
$\alpha_S = -.69$	-.009	.032	.181	.162	.190	.098	.033	.027	.893
$\alpha_A = -.69$	-.008	.029	.181	.162	.180	.097	.033	.027	.896

Table A.5.2: Simulation study for studying bias and efficiency of the $\mathcal{PL}$ and $\mathcal{RL}$ when the true error distribution for E and M was a bi-variate (a) normal, (b) t_{10}, (b) t_{50}, (d) $SN^{1,2}$ with slant $a = c(1,1)$ and (e) $SN^{1,2}$ with slant $a = c(2,2)$. Only E, M and G were considered as covariates. The frequency of the genetic variant was .20; disease prevalence was .1%; $\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = 1$ and $\sigma_{\epsilon_1\epsilon_1} = .3$. The location and scale parameters of the t and SN distributions are the same as the normal. SN: skew-normal, VR, variance ratio ; SE, standard error; Emp: Empirical, MSE: mean squared error. Results are based on the average over 1000 simulated data set. $\sigma^2_{\epsilon_1}$, $\sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1,\epsilon_2}$ in (5.5) were part of the optimization procedure.

True	Bias		Emp SE		Est SE		MSE		VR
Values	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\frac{\mathcal{RL}}{\mathcal{PL}}$
				(a) Bi-variate Normal					
$\alpha_E = 1$	.010	.012	.101	.087	.102	.080	.010	.008	.856
$\alpha_M = 1$	.010	.012	.101	.087	.103	.082	.010	.008	.856
$\alpha_G = .26$	.001	.008	.146	.132	.146	.124	.021	.018	.906
				(b) Bi-variate t_{10}					
$\alpha_E = 1$	.012	-.164	.101	.071	.099	.106	.010	.032	.701
$\alpha_M = 1$	.010	-.164	.101	.071	.103	.104	.010	.032	.702
$\alpha_G = .26$	.009	.032	.165	.141	.167	.204	.027	.021	.850
				(c) Bi-variate t_{50}					
$\alpha_E = 1$	.003	-.022	.101	.085	.104	.086	.010	.008	.840
$\alpha_M = 1$	.009	-.023	.101	.084	.099	.079	.010	.008	.836
$\alpha_G = .26$	.003	.011	.149	.133	.152	.139	.022	.018	.897
				(d) Bi-variate Skew-Normal with $a = (1,1)$					
$\alpha_E = 1$	.010	-.012	.099	.087	.100	.093	.010	.008	.873
$\alpha_M = 1$	.007	-.013	.099	.087	.102	.094	.010	.008	.875
$\alpha_G = .26$	-.003	-.006	.133	.124	.136	.127	.018	.015	.931
				(e) Bi-variate Skew-Normal with $a = (2,2)$					
$\alpha_E = 1$	.009	-.014	.099	.087	.098	.101	.010	.008	.876
$\alpha_M = 1$	.004	-.019	.099	.087	.099	.098	.010	.008	.878
$\alpha_G = .26$	-.005	-.007	.131	.122	.134	.131	.017	.015	.933

[1] Notation as Azzalini, A. with the collaboration of Capitanio, A. (2014). The Skew-Normal and Related Families. Cambridge University Press, IMS Monographs series.

[2] To generate data from a bi-variate SN, the R package `SN` was used.

Table A.5.3: Results from analysis of multiple sclerosis data using the prospective likelihood ($\mathcal{PL}$) and retrospective liklelihood ($\mathcal{RL}$). Estimates (Est) and standard errors (SE) of the effect of methylation and gene expression on multiple sclerosis α_M, in each of the 15 models. VR, variance ratio. If VR < 1, $\mathcal{RL}$ gives smaller standard errors for the parameter estimates.

Model	$\mathcal{PL}$		$\mathcal{RL}$		VR
	Est	SE	Est	SE	$\frac{\mathcal{RL}}{\mathcal{PL}}$
			Methylation		
1	$-.16$	.51	$-.23$	.44	.86
2	.59	.60	.58	.52	.88
3	.53	.51	.56	.46	.90
4	.29	.49	.27	.39	.80
5	.22	.43	.34	.45	1.04
6	.39	.49	.47	.44	.89
7	.38	.46	.48	.42	.90
8	.13	.52	.29	.51	.99
9	1.28	.67	1.56	.60	.90
10	1.02	.74	1.05	.70	.95
11	1.23	.71	1.26	.69	.98
12	$-.20$	.68	.06	.57	.83
13	2.27	1.09	2.88	1.17	1.07
14	.68	.56	.97	.51	.91
15	$-.80$	.60	$-.86$	.53	.89
			Gene Expression		
1	1.13	.67	1.26	.63	.94
2	.93	.64	1.09	.61	.95
3	1.18	.71	1.41	.72	1.01
4	1.06	.64	1.23	.62	.97
5	1.08	.65	1.17	.61	.95
6	.97	.65	1.02	.63	.96
7	1.04	.66	1.09	.62	.95
8	1.05	.64	1.13	.61	.94
9	.73	.71	.73	.59	.83
10	.70	.72	.80	.66	.92
11	.59	.71	.71	.66	.93
12	1.20	.78	1.16	.70	.90
13	.24	.72	.46	.67	.94
14	1.08	.65	1.14	.62	.96
15	1.53	.79	1.56	.68	.85

Table A.5.4: Results from analysis of multiple sclerosis data using the prospective likelihood ($\mathcal{PL}$) and retrospective liklelihood ($\mathcal{RL}$). Estimates (Est) and standard errors (SE) of the effect of gender and age on multiple sclerosis α_S, in each of the 15 models. VR, variance ratio. If VR < 1, $\mathcal{RL}$ gives smaller standard errors for the parameter estimates.

Model	$\mathcal{PL}$		$\mathcal{RL}$		VR
	Est	SE	Est	SE	$\frac{\mathcal{RL}}{\mathcal{PL}}$
			Gender		
1	-2.77	1.44	-3.34	1.42	.99
2	-2.54	1.49	-3.17	1.45	.98
3	-2.83	1.51	-3.55	1.54	1.02
4	-2.86	1.48	-3.57	1.52	1.03
5	-2.76	1.45	-3.38	1.42	.98
6	-2.71	1.48	-3.32	1.45	.98
7	-2.76	1.48	-3.29	1.49	1.00
8	-2.69	1.42	-3.25	1.39	.98
9	-3.10	1.68	-2.91	1.36	.80
10	-2.49	1.49	-2.82	1.43	.96
11	-3.45	1.67	-4.09	1.63	.98
12	-2.82	1.49	-3.23	1.44	.96
13	-2.03	1.77	-3.39	1.79	1.01
14	-3.02	1.55	-3.36	1.56	1.01
15	-2.92	1.47	-3.37	1.39	.95
			Age		
1	-1.82	1.28	-1.99	1.20	.94
2	-2.35	1.37	-2.70	1.43	1.04
3	-2.06	1.24	-2.35	1.19	.96
4	-2.02	1.22	-2.26	1.22	1.00
5	-2.01	1.22	-2.36	1.24	1.01
6	-1.94	1.23	-2.19	1.21	.99
7	-1.84	1.22	-2.01	1.20	.98
8	-1.92	1.22	-2.14	1.19	.97
9	-1.58	1.36	-1.16	1.11	.82
10	-2.07	1.29	-1.98	1.22	.95
11	-2.14	1.32	-2.09	1.27	.96
12	-1.80	1.30	-2.22	1.33	1.02
13	-4.31	2.03	-5.60	2.15	1.06
14	-1.40	1.30	-1.42	1.19	.92
15	-1.69	1.26	-2.04	1.19	.94

Table A.5.5: Results from analysis of multiple sclerosis data using the prospective likelihood ($\mathcal{PL}$) and retrospective liklelihood ($\mathcal{RL}$). Pvalues from the two degrees of freedom likelihood ratio test for testing the null hypothesis of no methylation and expression effect on multiple sclerosis, in each of the 15 models.

Model	$\mathcal{PL}$	$\mathcal{RL}$
1	1.8e-01	9.4e-02
2	1.1e-01	5.4e-02
3	1.0e-01	4.5e-02
4	1.6e-01	8.2e-02
5	1.7e-01	8.1e-02
6	1.4e-01	5.7e-02
7	1.4e-01	5.5e-02
8	1.9e-01	9.1e-02
9	2.1e-02	2.3e-03
10	5.9e-02	2.5e-02
11	2.8e-02	1.1e-02
12	1.8e-01	1.1e-01
13	5.0e-03	3.0e-04
14	8.6e-02	1.8e-02
15	6.7e-02	2.4e-02