Statistical integration of diverse omics data

Bouhaddani, S. el

6

Statistical omics data integration in longitudinal family studies

6.1 Background

In the previous chapters of this thesis, we considered latent variable models to study the relationship between two omics datasets. In Chapter 2, an algorithmic approach to decompose data into sets of joint and specific principal components, called O2PLS [16], was evaluated in omics data from population cohorts with simple random sampling designs, i.e. each member of the population is chosen at random. In clinical biostatistics and epidemiology, usually more complex designs are applied. For example, the longitudinal family study design provides opportunities to investigate changes in time and compare variation within and between families [13]. Nowadays, studies with such sampling designs also include molecular data. In this chapter, we consider the longitudinal family design in omics data integration. The O2PLS framework, discussed in Chapter 2 and 3, is not suitable to analyze omics data from these designs and can yield misleading results [14]. As discussed in the previous chapters, the omics data integration approaches need to be extended to non-standard sampling designs.

In Chapter 4 and 5 of this thesis, a statistical framework for probabilistic omics data integration was developed. We argued that a likelihood-based approach can better handle complex study designs. In this chapter, we take this step forward and show how the PO2PLS framework can be generalized to suit various epidemiological research questions.

Our methodological work is motivated by the Genetic Analysis Workshop (GAW20) [15]. Data from the GOLDN study are available, containing measurements on around 2 million (HapMap imputed) genetic variants, 28285 methylation sites, and triglyceride levels. After sample matching and quality control, data from 717 subjects are available, together with pedigree information. The average sample size within a pedigree is 18 individuals, where most pedigrees are multi-generation. The triglyceride levels were measured on four different time points: twice before taking a triglyceride-altering drug, and twice afterwards, each pair was separately averaged. The methylation data were measured on the second and fourth time point. See Table 6.1 for an overview of the study design. Two centers participated in the study: Minnesota and Utah. In GAW20, the aim was to discover the pharmaco-epigenetics effects on triglycerides [5]; these data are described in [1]. Here, we consider genetic variants, methylation data at both time points, and triglyceride differences as outcome. In this chapter, the first aim is to capture underlying latent pathways between the genetics and methylation data, while modeling the longitudinal aspect of these data. The second aim is to use these pathways to model the association with a response variable, taking into account familial correlations between subjects.

For longitudinal omics data, some approaches have been developed for one dataset based on Principal Components Analysis [18, 17]. These methods project the data onto a set of latent variables, with loadings for each time point. Penalization is used to reduce overfitting. However, these approaches are not suitable for multiple heterogeneous omics data. To the best of our knowledge, no longitudinal method has been proposed for heterogeneous omics data integration.

We propose a three-stage approach to first decompose epigenetic data in joint and specific parts that represent time-stable and time-varying variation, respectively. Then, we capture the genetics underlying these changes by estimating joint genetic-

epigenetic, genetic-, and epigenetic-specific variation. Finally, we regress the outcome on these parts, taking into account familial correlations between subjects. The resulting model captures genetic-epigenetic effects on the outcome over time, while incorporating the kinship correlation structure.

The rest of this chapter is organized as follows. First, the proposed three-stage approach using O2PLS is presented. Then, we apply this method to the genetic, epigenetic and triglyceride data from GAW20 and discuss the findings. Finally, a probabilistic framework for data integration in longitudinal family studies, based on Chapter 4 and 5, is presented and discussed.

6.2 Methods

Let X represent the genetic data matrix, with N samples across the rows and p variables across the columns. Denote by Z_1 and Z_2, the $N \times q$ methylation data matrices at the second and fourth visit, respectively. The triglyceride outcome vector of length N is denoted by y. To integrate genetic and epigenetic data over time, we consider Two-way Orthogonal Partial Least Squares (O2PLS) [16, 3]. The analysis consists of three steps.

In the first step, we decompose the methylation datasets into time-stable and time-varying parts using the O2PLS decomposition given by

$$\begin{aligned} Z_1 &= T_1 W_1^{\mathrm{T}} + U_1 C_1^{\mathrm{T}} + E, \\ Z_2 &= T_2 W_2^{\mathrm{T}} + U_2 C_2^{\mathrm{T}} + F. \end{aligned} \tag{6.1}$$

Here, T_1 and T_2 are the time-stable parts (second resp. fourth visit). The time-varying parts are given by U_1 (second visit) and U_2 (fourth visit). The time points are linked by a linear regression given by,

$$T_2 = T_1 B_1 + H_2. \tag{6.2}$$

In the second step, we are interested in the time-varying parts U_τ, where $\tau = 1, 2$, and its overlap with the genetic data X. An O2PLS model linking these parts for

Table 6.1: **Longitudinal design of the GOLDN study.** Triglyceride levels were measured (denoted by an X) twice before (v1 and v2) and twice after treatment (v3 and v4). Methylation was measured before and after treatment (v2 and v4). Genotypes were measured before treatment (v2), and assumed to be constant for all time points (denoted by a star). The time period between v1 and v2, and between v3 and v4 is 1 day. The time period between v2 and v3 is 3 weeks. The triglyceride-altering drug was administered after time point 1, v2.

Measurements	**Time point 1 (v1)**	**Time point 1 (v2)**	**Time point 2 (v3)**	**Time point 2 (v4)**
Triglycerides	X	X	X	X
Methylation		X		X
Genetics	*	X	*	*

each time point is given by

$$X = T_{G\tau} W_{G\tau}^{\mathrm{T}} + U_{G\tau} C_{G\tau}^{\mathrm{T}} + E_\tau, \tag{6.3}$$
$$U_\tau = T_{U\tau} W_{U\tau}^{\mathrm{T}} + U_{U\tau} C_{U\tau}^{\mathrm{T}} + F_\tau,$$

with inner model

$$T_{U\tau} = T_{G\tau} B_{U\tau} + H_{U\tau}. \tag{6.4}$$

The joint parts are $T_{G\tau}$ (genetic-joint) and $T_{U\tau}$ (U_τ-joint), the specific parts are $U_{G\tau}$ (genetic-specific) and $U_{U\tau}$ (methylation-specific).

The last step includes a model for y, the response variable, given the components in each part separately. A linear mixed model [7] is used to estimate the association of each variable in each part with triglyceride outcome, corrected for age, sex and center. Random effects for each family are included to account for familial correlations. The model is given by

$$y_{ij} = \alpha_y + X_{\mathrm{cov}} \beta_{\mathrm{cov}} + \tilde{X}_{ij} \beta + b_j + \epsilon_{ij}, \tag{6.5}$$

where y_{ij} is the outcome for subject i in family j. The intercept is given by α_y, and $X_{\mathrm{cov}}\beta_{\mathrm{cov}}$ is the contribution of the fixed covariates. Further, b_j is a normally distributed random effect for family j with zero mean and correlation matrix equal to two times the kinship matrix K_j. The matrix $\tilde{X}_{ij}$ represents each part in the O2PLS decompositions. For example, the first part consists of T_1, H_2, U_1 and U_2. Note here that since T_1 and T_2 are correlated, H_2 is used instead of T_2 to include additional information in T_2 not captured by T_1. The model for y given the first part is then given by (6.5), with $\tilde{X} = [T_1, H_2, U_1, U_2]$. In total, twelve parts are considered, four parts from step one ((6.1) and (6.2)), and eight parts from step two ((6.3) and (6.4)). Note that by including the joint parts from the second step, joint effects of genetics and methylation on the outcome is captured. Also, the genetic-specific parts may provide information about familial variation associated to the outcome, but unrelated to methylation changes.

6.3 Data analysis results

Pre-processing the data. Prior to analysis, the data were pre-processed. Firstly, we applied a log transformation to all triglyceride measurements to achieve approximate normality. Then, we averaged the log transformed measurements from visit one and two, and separately from visit three and four. The difference of the two averages was taken as response variable y. Secondly, we summarized the 2 million genotypes, coded as $\{0, 1, 2\}$, per gene. To this end, we considered for each gene all SNPs lying within 50 kilobase (kb) distance from that gene. We took as many principal components of these SNPs as needed to explain at least 80% of the total variation of the SNPs in the proximity of that gene.

Results for chromosome 1. The analysis scheme consists of three steps, see Figure 6.1. First, we estimated time-stable and time-varying parts in the methylation data (the blue pentagons in the figure). We retained 8 joint components in both T_1 and T_2 and 16 time-specific components in both U_1 and U_2, based on visual inspection of eigenvalue plots. Then, we estimated the joint part between genetics and the

time-varying methylation parts, together with the genetic-specific and time-varying methylation-specific parts (the green circles in the figure). We retained 5 joint, 10 gene-specific and 5 methylation-specific components. Finally, we regressed the triglyceride outcome on each part individually (all colored blocks in the figure), followed up with a regression of triglycerides on the significant components within each part (denoted by a star in the figure). Here, significance is defined by a t-value greater than two.

We considered, in total, 12 separate fits with the time-stable (T_1 and T_2) and time-varying (U_1 and U_2) methylation components from step one in (6.1), and genetic-epigenetic joint (T_{G1}, T_{G2}, T_{U1}, T_{U2}) and specific (U_{G1}, U_{G2}, U_{U1}, U_{U2}) components from step two in (6.3), see also Figure 6.1. The variance in the methylation data explained by the time-specific parts was around 21%. The genetic data explained around 38% of these time-specific parts in the methylation data. For each fit, we took the significant variables and combined them in a single final fit using (6.5), in total with eight latent variables as covariates: $\tilde{X} = [(U_2)_{11}, (T_{G1})_3, (U_{G1})_{1,6,9,10}, (U_{G2})_{1,10}]$. The results are shown in Table 6.2.

The estimates for the time-varying methylation part from visit two (U_2) and the genetic-specific parts (U_{G1}) had absolute t-values between 2 and 2.5, indicating significant effects of these components on triglyceride changes. The time-stable components did not have any significant association with the outcome.

The variance of the random effect was 0.0057 (standard deviation 0.075), the residual variance was 0.079 (s.d. 0.28), thus the residual familial correlation was about 7%. Furthermore, in Figure 6.2, scatterplots of each component with y are shown. Each family is colored if there is a member that has a value on the horizontal axis of more than four times the standard deviation of the corresponding variable. We observed some families in the genetic-specific components with extraordinary high scores, possibly indicating that the genetic-specific components pick up familial stratification among participants. The analysis was repeated for other chromosomes, similar results were obtained.

6.4 Discussion of the results

Previously in this chapter, we described a three-stage approach to decompose epigenetics data in time-stable and time-varying parts, integrate these parts with genetics, and used the resulting common and specific parts in a linear mixed model with changes in triglyceride levels as response variable and a random effect for each family member. The data integration steps were performed with O2PLS.

In the original article [5], an epigenome wide association study is described, where the triglyceride outcome was regressed on each methylation variable separately. Significant sites were then inspected and interpreted. Most notably, the most significant sites were located on the *CPT1A* gene on chromosome 11. The O2PLS analysis described here addresses a different question, namely how changes in triglycerides associate with joint genetic-epigenetic variation over time, while also considering associations specific to the genetic and epigenetic data, respectively. These effects are additionally modeled as multivariate, where combinations of genes and CpG simultaneously explain triglyceride changes. Therefore, these findings do not necessarily

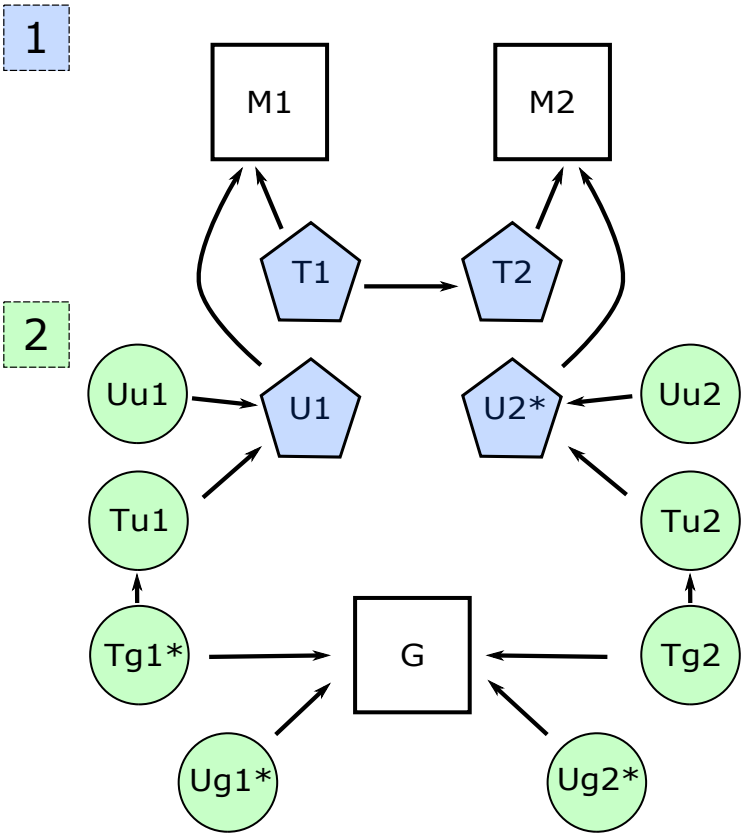

Figure 6.1: **A schematic representation of the O2PLS analyses.** In the first step, denoted by blue pentagons, the two methylation datasets are decomposed in time-stable (T_1 and T_2) and time-varying (U_1 and U_2) parts. In the second step, denoted by green circles, the time-varying parts are integrated with genetic data, yielding genetic-epigenetic joint (T_{G1}, T_{G2}, T_{U1}, T_{U2}) and specific (U_{G1}, U_{G2}, U_{U1}, U_{U2}) parts. All colored parts are used separately in a linear mixed model for the triglyceride outcome. Components in the parts that are marked with a star had a significant t-value (more than 2).

Component	**Part**	**Estimate** (S.E.)	***t*-value**
$(U_2)_{11}$	Time-varying methylation v4	-0.003 (0.001)	**-2.491**
$(T_{G1})_3$	Genetic-joint v2	0.004 (0.003)	1.369
$(U_{G1})_1$	Genetic-specific v2	0.007 (0.003)	**2.227**
$(U_{G1})_6$	Genetic-specific v2	0.006 (0.003)	**2.018**
$(U_{G1})_9$	Genetic-specific v2	-0.007 (0.003)	**-2.379**
$(U_{G1})_{10}$	Genetic-specific v2	0.007 (0.003)	**2.109**
$(U_{G2})_1$	Genetic-specific v4	0.003 (0.003)	1.266
$(U_{G2})_{10}$	Genetic-specific v4	-0.001 (0.003)	-0.128

Table 6.2: **Fixed effects estimates for the final model.** The components from this final single fit represent (from top to bottom): time-varying part in methylation (visit four), genetic-joint (visit two) and genetic-specific parts (first four: visit two, last two: visit four). The *t*-statistics larger than 2 in absolute value are boldfaced.

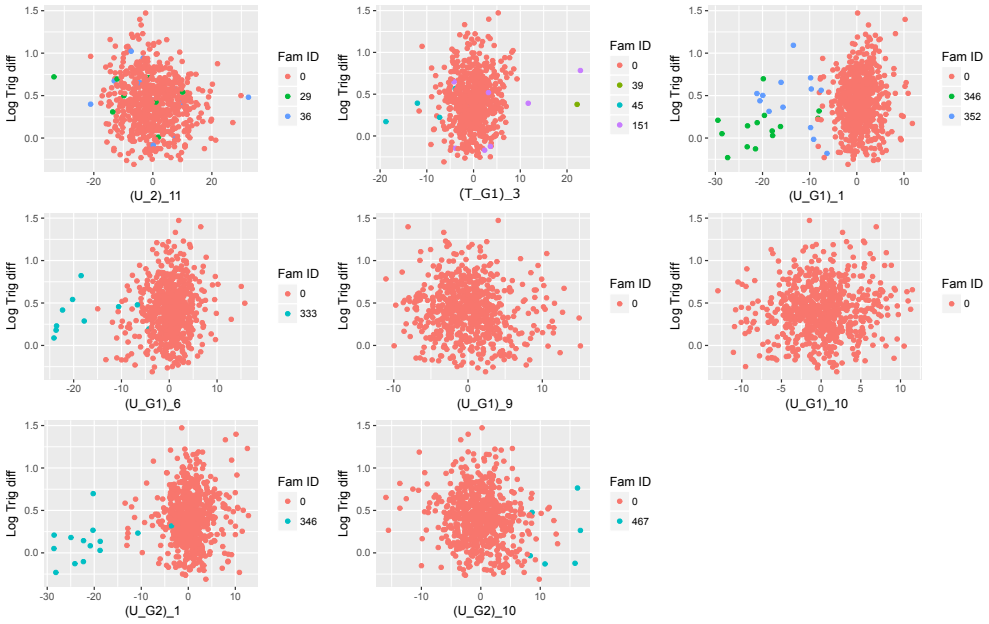

Figure 6.2: **Scatterplots of each included component against** y**.** The components represent (from left to right): time-varying methylation-specific (visit four), genetic-joint (visit two) and genetic-specific parts (first four plots: visit two, last two plots: visit four). They correspond with the components in Table 6.2.

need to overlap with what was reported earlier [5].

In equation (6.5), both a family random effect as well as genetic-specific components were included in the final fit. Note that these components can also capture part of the familial effect. This is suggested by Figure 6.2, where it can be seen that the genetic-specific component scores partially capture segregation between (extreme) families. A mixed model without the genetic-epigenetic joint components can reveal whether some of the variation in the family random effect was indeed captured by the genetic-specific parts.

We found that a functional annotation of the top genes in the joint components of the genetic data (i.e. $T_{G\tau}$), for several chromosomes separately, showed some clusters reflecting immunological pathways, see Table 6.3. Note that these data were measured on CD4+ T-cells, so further research is needed to identify genes that got high weight only due to the tissue type. Compared to the original article [5], the *CPT1A* gene had a relatively high joint genetic-methylation ranking of 57 out of 28285, indicating that it plays an important role in the common part of genetic and epigenetics. As a future step, a follow up on interpreting the top methylation sites can be performed, for example by investigating how they influence the biological processes of genes related to triglyceride levels.

6.5 Methodological future work

The O2PLS analysis that is carried out in this chapter consists of three steps: first the longitudinal aspect of the methylation data is addressed, by considering time-stable and time-varying parts in these datasets. The latter part is of interest, since it represents the change in methylation over time. The output of this step is then used to find the statistical overlap of the time-varying parts with the genetic data. The components from the second step are finally used in a linear mixed model with triglyceride changes as outcome. A random effect per family is used to account for the correlation structure between relatives.

This approach has some disadvantages, mainly in the way the longitudinality and the association with the outcome is addressed. Firstly, when methylation data are available for more than two time points, O2PLS cannot be directly applied. Rather than treating each time point as a distinct source of variation, other methods that impose a functional form for the methylation data over the time points can be used (e.g. Functional PCA [6]). Secondly, the final regression step uses O2PLS components that are agnostic of the outcome. The components that are discarded can still contain information about the outcome. In that case, the last regression step of the outcome on the components does not (fully) represent the association of triglycerides with genetics and methylation. Several methods have been proposed for supervised dimension reduction (e.g. supervised PCA [2] and Collaborative Regression [4]) where the components also explain variation in the outcome. However, even after implementing these adjustments, the overall analysis still consists of at least two steps rather than a joint approach for all data.

From a statistical point of view, a simultaneous approach, rather than three consecutive steps, is expected to yield overall more accurate estimates for the parameters in each step. As argued in Chapter 5, a framework that facilitates such simultaneous

Chr	GO keywords	Top Gene
3	Lympho- and leukocyte apoptotic process	*BCL6*
3	Immune system process	*KIAA0226*
4	Fatty acid metabolic process	*ACSL1*
4	Catabolic processes	*QDPR*
4	Humoral immune response	*RBPJ*
11	Immunological synapse	*PTPRJ*
11	Carboxylic/organic acid binding	*HMBS*
21	Cellular lipid metabolic process	*AGPAT3*
21	Apoptotic process	*IFNG*

Table 6.3: **Functional annotation of top 200 genes with O2PLS.** For each chromosome (denoted by Chr), the genes are ranked according to the loading weights in the joint genetic-epigenetic components that were significantly associated with changes in triglycerides.

approach is probabilistic O2PLS (PO2PLS). The PO2PLS model can be extended to handle more complex situations. In the GAW case study, these extensions are as mentioned above: (1) a supervised regression model for PO2PLS with an outcome variable with random effects to take into account correlations among subjects, (2) a longitudinal model for the methylation changes in time, and (3) a data integration model to capture common and specific variation in genetic and methylation data. While such extension is not proposed yet, the next subsection proposes a model and briefly discusses modeling strategies to correctly reflect characteristics of the longitudinal family design as found in the GAW case study.

6.5.1 A probabilistic model for supervised data integration in longitudinal family studies

Within a probabilistic model, extensions are possible to accommodate longitudinal family designs such as in the data analysis of this chapter. To this end, let x represent the genetic data, and m_τ the methylation data for time point τ. A model resembling the three-stage analysis consists two parts. In the first part, the methylation and genetics data are linked over time via latent variables as follows, with $\tau = 1, \ldots, \mathcal{T}$,

$$
\begin{aligned}
g &= t_g W_g^{\mathrm{T}} + u_g W_{u_g}^{\mathrm{T}} + e_g, && (6.6)\\
m_\tau &= t_\tau W_m^{\mathrm{T}} + u_\tau W_\tau^{\mathrm{T}} + e_\tau, \qquad \tau = 1, \ldots, \mathcal{T}, && (6.7)\\
u_\tau &= t_g B_\tau + h_\tau, \qquad \tau = 1, \ldots, \mathcal{T}, && (6.8)\\
t_{\tau+1} &= t_\tau B_t + h_{t_\tau}, \qquad \tau = 1, \ldots, \mathcal{T} - 1. && (6.9)
\end{aligned}
$$

Here, m_τ is the methylation data for time point τ. The time effect is decomposed, similarly as in the previous analysis, into time-stable and time-varying components t_τ and u_τ, respectively. In the time-stable parts, factorial invariance [10] is assumed; the directions W_m are the same across time points to ensure the same constructs are obtained. For simplicity, the process $\{t_\tau\}$ is assumed to be Markov; the relationships between the t_τ are forward in time only and given t_τ, the process after τ is independent

of the process before τ. The genetic data are denoted by g, and are decomposed in a part joint with methylation per time point, and a genetic-specific part. The same assumptions as in the PO2PLS model are made regarding independence of the latent variables and orthogonality of the loadings.

Denote by y the response variable and let f index families. The second model relates y to the common and specific parts given in the previous model,

$$y_f = \beta_0 + t_g\beta_g + \sum_\tau h_\tau\beta_\tau + \gamma_f + \epsilon_f, \tau = 1, \ldots, \mathcal{T}. \tag{6.10}$$

Here, γ_f is a normally distributed random effect per family with zero mean and covariance matrix $\sigma^2_\gamma K_f$, where K_f is two times the genetic kinship matrix for family f. The residual joint components h_τ, representing information in u_τ not present in t_g, are used in the model for y to avoid collinearity issues. The total model is graphically depicted in Figure 6.3. Note that, for simplicity, in the model for g and m_τ, family information is not taken into account.

Estimating the joint model The complete data likelihood can be written as

$$\prod_\tau f(g, m_\tau, y, t_\tau, u_\tau, t_g, u_g, \gamma) \tag{6.11}$$

where f is a multivariate normal density of proper dimensions. Also here, an ECM algorithm [9] can be deployed to obtain maximum likelihood estimates for all parameters. Note that, since all random variables are normal, the Expectation step involves computing conditional first and second moments of the latent variables, given the observed data g, m_τ and y. Given these quantities, the Maximization step is decomposed in several distinct optimization problems.

The E and M steps can be computed analogous to the computations in Chapter 5. However, an additional challenge is the multiple time points: in the E step, the conditioning is performed over all m_τ, and the M step involves objective functions that share the same parameters (e.g. W_m).

6.5.2 Further directions: extending the joint model

In the model for m_τ in (6.9), a latent growth model [11] can be recognized. In our case, multiple instances of a multivariate random vector m_τ is related via an underlying latent system of equations for all t_τ. When data are available for more than two time points, the inner model connecting the latent variables t_τ can be extended using a latent growth model, given in its general form by

$$x(\tau) = \sum_k g_k(\tau) + e(\tau), \tag{6.12}$$

where x is an observable random vector measured on time points indexed by τ, and g_k are basis functions. In our analysis, also time-specific parts are included and actually of more interest, as they indicate the part that is not shared across time points.

In this chapter, the outcome is assumed to be normally distributed, and linearly dependent on the latent variables. In some applications, these assumptions are not

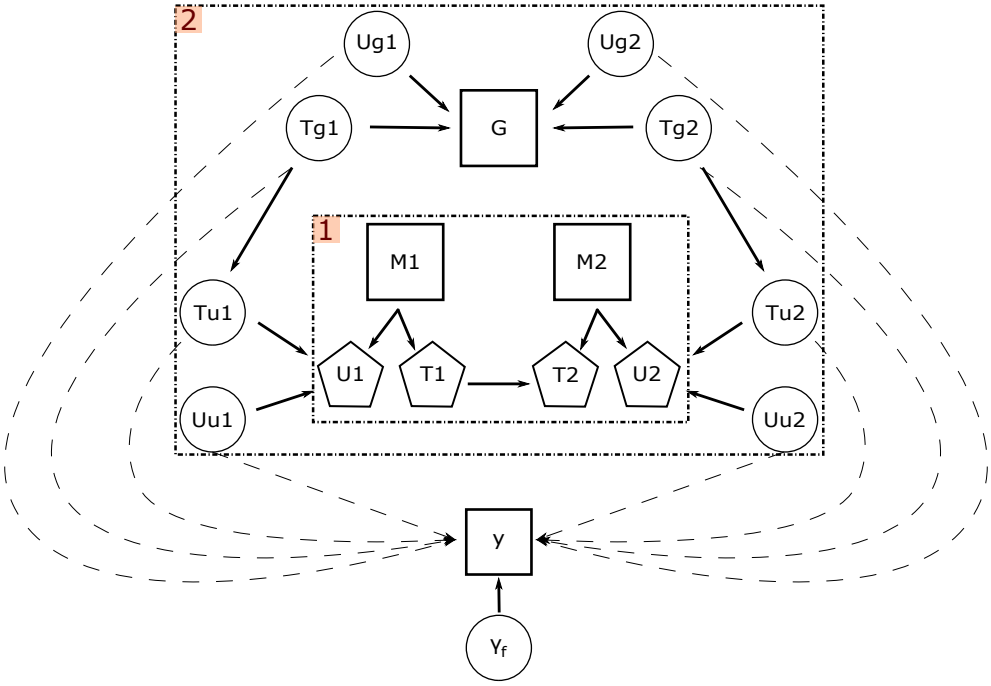

Figure 6.3: **Path diagram of probabilistic data integration in longitudinal family studies.** The observed variables are denoted with a square. They are modeled in three steps, these steps are indicated by a number in the top left corner of the two large rectangles. For the third step, the rectangle and number are omitted. In the first step, the methylation data from time point 1 and 2, M1 resp. M2, are decomposed in time-stable (T1 and T2) and time-varying (U1 and U2) parts. In the second step, the parts from step 1 are related with the genetic data, G, and decomposed in joint (Tg1, Tg2, Tu1 and Tu2) and specific (Ug1, Ug2, Uu1 and Uu2) parts. In the third step, the parts from step 2 are related to the outcome, y, together with a latent variable (γ_f) for the family effect. Error variables are omitted.

suitable for y. In particular, y may be non-normally distributed, or the relationship between y and its predictors is not linear. In the first case, a generalized linear mixed model can be employed to model the expected value of $\eta(y)$, where η is a suitable link function. A similar model has been described in [12]. If a non-linear link function is used, the likelihood cannot be obtained analytically, which is a major drawback. In the second case, a non-linear functional, say F, may be assumed between y and its predictors:

$$y_f = F\left(t_g, (h_\tau)_\tau\right) + \gamma_f + \epsilon_f. \tag{6.13}$$

To estimate F, a non-parametric approach is feasible using kernel or spline-based methods (see [8]). Since the complete data likelihood can be factored in distinct terms (see Chapters 5 and 6), the likelihood contribution of y is decoupled and separately optimized in the M step.

Bibliography

[1] S. Aslibekyan, L. Almasy, M. A. Province, D. M. Absher, and D. K. Arnett. Data for GAW20: Genome-wide DNA sequence variation and epigenome-wide DNA methylation before and after fenofibrate treatment in a family study of metabolic phenotypes 06 Biological Sciences 0604 Genetics. *BMC Proc.*, 12(S9):35, sep 2018.

[2] E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by Supervised Principal Components. *J. Am. Stat. Assoc.*, 101(473):119–137, mar 2006.

[3] S. el Bouhaddani, H.-W. Uh, C. Hayward, G. Jongbloed, and J. Houwing-Duistermaat. Probabilistic partial least squares model: Identifiability, estimation and application. *J. Multivar. Anal.*, 167:331–346, sep 2018.

[4] S. M. Gross and R. Tibshirani. Collaborative regression. *Biostatistics*, 16(2):326–338, 2015.

[5] M. R. Irvin, D. Zhi, R. Joehanes, M. Mendelson, S. Aslibekyan, S. A. Claas, K. S. Thibeault, N. Patel, K. Day, L. W. Jones, L. Liang, B. H. Chen, C. Yao, H. K. Tiwari, J. M. Ordovas, D. Levy, D. Absher, and D. K. Arnett. Epigenome-Wide Association Study of Fasting Blood Lipids in the Genetics of Lipid-Lowering Drugs and Diet Network Study. *Circulation*, 130(7):565–572, aug 2014.

[6] G. M. James, T. J. Hastie, and C. A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87(3):587–602, sep 2000.

[7] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

[8] G. Li and S. Jung. Incorporating Covariates into Integrated Factor Analysis of Multi-View Data. *Biometrics*, 73(4):1433–1442, dec 2017.

[9] X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[10] W. Meredith and J. A. Teresi. An essay on measurement and factorial invariance, 2006.

[11] W. Meredith and J. Tisak. Latent curve analysis. *Psychometrika*, 55(1):107–122, 1990.

[12] S. Rabe-Hesketh, A. Skrondal, and A. Pickles. Generalized Multilevel Structural Equation Modeling. *Psychometrika*, 69(2):167–190, 2004.

[13] S. M. Schüssler-Fiorenza Rose, K. Contrepois, K. J. Moneghetti, W. Zhou, T. Mishra, S. Mataraso, O. Dagan-Rosenfeld, A. B. Ganz, J. Dunn, D. Hornburg, S. Rego, D. Perelman, S. Ahadi, M. R. Sailani, Y. Zhou, S. R. Leopold, J. Chen, M. Ashland, J. W. Christle, M. Avina, P. Limcaoco, C. Ruiz, M. Tan, A. J. Butte, G. M. Weinstock, G. M. Slavich, E. Sodergren, T. L. McLaughlin, F. Haddad, and M. P. Snyder. A longitudinal big data approach for precision health. *Nat. Med.*, 25(5):792–804, 2019.

[14] C. J. Skinner, D. J. Holmes, and T. M. Smith. The effect of sample design on principal component analysis. *J. Am. Stat. Assoc.*, 81(395):789–798, 1986.

[15] N. L. Tintle, D. W. Fardo, M. De Andrade, S. Aslibekyan, J. N. Bailey, J. L. Bermejo, R. M. Cantor, S. Ghosh, P. Melton, X. Wang, J. W. MacCluer, and L. Almasy. GAW20: Methods and strategies for the new frontiers of epigenetics and pharmacogenomics. *BMC Proc.*, 12(S9):26, sep 2018.

[16] J. Trygg and S. Wold. O2-PLS, a two-block (X–Y) latent variable regression (LVR) method with an integral OSC filter. *J. Chemom.*, 17(1):53–64, 2003.

[17] Y. Zhang and Z. Ouyang. Joint principal trend analysis for longitudinal high-dimensional data. *Biometrics*, 74(2):430–438, 2018.

[18] Y. Zhang, R. Tibshirani, and R. Davis. Classification of patients from time-course gene expression. *Biostatistics*, 14(1):87–98, 2013.