

Integrative analysis of time course metabolic data and biomarker discovery

Takoua Jendoubi
 School of public health
 Imperial College London

Timothy M.D. Ebbels
 Departement of surgery and cancer
 Imperial College London

Abstract

Metabolomics time-course experiments provide the opportunity to understand the changes to an organism by observing the evolution of metabolic profiles in response to internal or external stimuli. Along with other omic longitudinal profiling technologies, these techniques have great potential to complement the analysis of complex relations between variations across diverse omic variables and provide unique insights into the underlying biology of the system. However, many statistical methods currently used to analyse short time-series omic data are i) prone to overfitting or ii) do not take into account the experimental design or iii) do not make full use of the multivariate information intrinsic to the data or iv) unable to uncover multiple associations between different omic data. The model we propose is an attempt to i) overcome overfitting by using a weakly informative Bayesian model, ii) capture experimental design conditions through a mixed-effects model, iii) model interdependencies between variables by augmenting the mixed-effects model with a conditional autoregressive (CAR) component and iv) identify potential associations between heterogeneous omic variables by using a horseshoe prior.

Results: We assess the performance of our model on synthetic and real datasets and show that it outperforms existing models for metabolomic longitudinal data analysis. Our proposed method is able to identify metabolic biomarkers related to treatment, infer perturbed pathways as a result of treatment and

find significant associations with additional omic variables. We also show through simulation that an informative beta like prior compares better than a non-informative uniform prior in inferring significant pathways. On real data, we investigate how the number of profiled metabolites can affect the predictive ability of the model.
 Supplementary material for this article are available online.

1 INTRODUCTION

Over the past years, there has been a significant development in high-throughput omics technologies e.g. metabolomics, transcriptomics, genomics, epigenomics and proteomics along with a growing interest into joint modelling of multi-omic data [11, 13]. In metabolomics, several approaches are used to understand the response of a biological system as a function of an internal or external perturbation by monitoring “the chemical fingerprints that specific cellular processes leave behind” [9]. These chemical fingerprints are most commonly interrogated in terms of metabolite (i.e. low weight molecules) concentration, structure and transformation pathways (i.e set of chemical reactions) in order to identify a biomarker related to the studied process. Biomarker discovery consists of identifying a metabolite that has significant association patterns with a particular phenotype (disease, clinical variables, physical traits, etc) and that can be thus used as an indicator of that specific phenotype. Typical experimental platforms use analytical techniques such as nuclear magnetic resonance spectroscopy (NMR) [23] and mass spectrometry (MS) [10] to generate appropriate spectral metabolomic profiles of the studied biological system.

Metabolomic data sets are characterized by high correlation structures in that many spectral peaks can arise from the same metabolite and metabolites operate within networks of chemical reactions. In addition,

further correlation structure is present in longitudinal metabolomic studies due to repeated measurements of observations over time. Additional challenges include not only the low number of time points and samples compared to the number of profiled metabolic variables, but also integration of a different omic data to the metabolomic data.

First, metabolomic time series are often short due to experimental costs or ethical considerations. Typically, less than 10 time points are available compared to a large number of metabolic variables profiled at each time point e.g. hundreds of metabolic variables for targeted experiments and thousands of metabolic variables for untargeted experiments. Taking into account the small number of time points compared to the large number of metabolic variables profiled, the number of temporal patterns that will arise is limited (due to the limited number of degrees of freedom). Some temporal patterns will be repeated and thus these patterns can be induced by randomness. Second, models fitted to a small number of data points are often prone to overfitting i.e. the model is very sensitive to small fluctuations. This can lead to a poor fitting to unseen data and a high generalisation error. Third, it is also important to consider the number of parameters of the statistical model and make use of the simplest models in order to avoid overparametrisation. Finally, monitoring heterogeneous omic variables can substantially enhance the understanding of the underlying biological mechanism and provide a systems biology approach as these omic variables represent entities that are often involved in related cellular processes [11, 13].

For all these reasons, metabolomics scientists need cautious estimates model parameters in order to ensure robust interpretation of the results. Hence, appropriate models are needed to integrate heterogeneous omic data and take into account both experimental conditions and biological variations in order to extract important information from the data. In this paper, we are interested in dose-response time course experiments where additional omic variables (bacteria, genes, transcripts, etc) are monitored along with metabolites in the context of biomarker discovery.

The main contribution of our work is a *single* probabilistic generative model that i) can overcome overfitting via the use of weakly informative priors ii) makes use of mixed effects models to model the experimental design iii) models metabolite interactions by using pathway information through a conditional autoregressive (CAR) component and iv) uncovers multiple associations between metabolites and other omic variables by using a horseshoe prior. An additional benefit of our approach is that it naturally yields a list of perturbed metabolic pathways since it is based both on a

mixed effects component and a CAR component.

2 RELATED WORK

There is a growing interest in longitudinal experiments for heterogeneous omics data and statistical models to infer biomarkers of a particular treatment or disease over time. Different approaches attempt to infer influential or significant metabolites using dynamic metabolomic data under the assumption that metabolites are independant. These models include fitting smooth splines mixed effects models (SME) to time curves [3] and linear mixed effects models augmented with a variable selection approach [18]. However between-metabolite correlation is richly structured and biologically relevant and should be modelled.

Seemingly unrelated regression accounts for metabolite correlation by using correlated regression errors and can be used to identify biologically significant metabolites [6, 7]. In gene expression data analysis, [21] recently proposed to use confirmatory factor analysis to capture gene-pathway relationship and a conditional autoregressive model to capture relationships between a set of pathways where pathway network has been constructed based on KEGG [14] pathways. The latter accounts for biological variation in the data and allows ease of interpretation.

In the metabolomics literature, traditional frameworks for metabolomic data analysis use dimensionality reduction techniques, namely principal component analysis (PCA), partial least squares (PLS) [27] and PLS derived models (OPLS [26], O2PLS [25], OnPLS [17]) to take into account high correlation between metabolites. Extension of PLS to O2PLS and OnPLS allows for integrative analysis of heterogeneous omic data. PCA and PLS models are very popular among metabolomicists as tools for exploratory data analysis, visualisation purposes and ease of interpretation. One of the interests of PCA (PLS) derived models is to be able to visually assess whether or not there is a time effect in the data and identify metabolites that change over time by looking into time trajectories of each metabolite [**antti2002**].

Extensions to PCA (PLS) for longitudinal analysis include lagged PCA (PLS) and dynamic PCA (PLS) where a backshift matrix is introduced to take into account time dependancy. Similarly, [22] used a set piecewise orthogonal projections latent structures to describe changes between neighboring time points. PARAFAC [4] is a multi-linear unsupervised decomposition method that can account for the multi-way variation seen in dynamic metabolomics data. Recently, dynamic probabilistic PCA (DPPCA) was proposed in

[19] as a generative probabilistic model of longitudinal metabolomic data where a stochastic volatility model is used for the latent variables. The main inconvenience of these approaches is that further techniques such as multiple testing correction have to be separately applied to the data in order to identify biomarkers and they do not take heterogeneous data (i.e. data from different omics techniques) into account.

In contrast, we here provide a *single* model that takes into account metabolite interactions, time variation and experimental design, infers influential metabolites and also quantifies relationships between metabolites and additional omic variables (if any).

3 MODEL FRAMEWORK

Given a metabolomics data $\mathbf{X} \in \mathbb{R}^{N \times T \times M}$ where N is the number of observations, T the number of time points and M the number of metabolites. $\mathbf{Y} \in \mathbb{R}^{N \times T \times K}$ is an additional continuous omic data measured along with \mathbf{X} where K is the number of associated omic variables. The set of N observations consists of a set of cases and controls. Throughout the paper, index i always runs through observations, index t runs through time points, index m runs through metabolites and index k runs through \mathbf{Y} variables. Vector quantities are written in bold. Matrices are written in bold capitals. Our goal is to build a simple model that can identify biomarkers relative to a specific treatment in time taking into account the multiple sources of variations in the data.

The model is built on three levels: First, a CAR component to capture interaction between metabolites. Second, a variable selection model to uncover associations between metabolites and \mathbf{Y} data. Third, a mixed effects component to model experimental design. We give more details about each level of our model respectively in each of the following sections.

3.1 Metabolite interactions

In a similar fashion to [21], we model metabolite interactions via the CAR model. In fact, incorporation of pathway information in the CAR model through the variance matrix helps provide chemists with an easily interpretable model. First, we assume that the concentration of each metabolite is linearly influenced by concentration levels of metabolites in the same pathway. Let $\mathbf{C} \in \mathbb{R}^{M \times M}$ be the design matrix quantifying metabolite interactions such that matrix elements $c_{mm} = 0$, $c_{mj} \neq 0$ if metabolites m and j are in the same pathway and 0 otherwise. Thus, metabolite in-

tensity levels can be expressed as:

$$x_{itm} | \mathbf{x}_{it,-m}, \boldsymbol{\mu}_{it}, \mathbf{C}, \sigma \sim N(\mu_{itm} + \sum_{\substack{j=1 \\ j \neq m}}^M c_{mj}(x_{itj} - \mu_{itj}), \sigma^2) \quad (1)$$

where $\mathbf{x}_{it,-m}$ represents measurements of metabolites of sample i at time point t excluding metabolite m , μ_{itm} is a function of covariates of sample i for metabolite m at time point t . If we define \mathbf{I}_M the M th order identity matrix, the joint distribution of \mathbf{x}_{it} can be explicitly written as [8]:

$$\mathbf{x}_{it} | \boldsymbol{\mu}_{it}, \mathbf{C}, \sigma \sim N(\boldsymbol{\mu}_{it}, (\mathbf{I}_M - \mathbf{C})^{-1} \sigma^2) \quad (2)$$

Chemists are most interested by identifying which pathways are “on” or “off” as an effect of treatment. In the CAR literature, the design matrix \mathbf{C} can be modeled as a scaled product of a diagonal weight matrix and an adjacency matrix. In order to allow for pathway perturbation inference we construct the distance matrix based on the individual contribution of each pathway. To be precise, we define $\mathbf{C}(\phi) = \sum_{p=1}^P \phi_p \mathbf{G}_p \mathbf{A}_p$ where P is the number of pathways. The distance matrices \mathbf{A}_p are zero-diagonal symmetric adjacency matrices with elements a_{mj}^p equal to the inverse of the length of the shortest path between metabolites m and j if they are in pathway p and 0 otherwise. A path between two metabolites consists in the number of reactions that lead from one metabolite to the other. The shortest path is the path that contains the smallest number of reactions. In the diagonal matrices \mathbf{G}_p we use the reciprocal of the number of neighbors of each metabolite in pathway p i.e $(g_{mm}^p)^{-1} = \sum_{j=1}^M (a_{mj} > 0)$ so that the squared partial correlation $\text{cor}(x_{itm}, x_{itj} | \mathbf{x}_{it,-(m,j)})^2 \propto \phi_p g_{mm}^p g_{jj}^p$ is reduced when more metabolites from the same pathway are profiled [8]. The model parameter $\phi = \{\phi_p\}_{p=1}^P$ is estimated. It quantifies pathway contribution and is referred to as *spatial*-dependence parameter in the CAR literature.

The model needs to comply with the condition that $\mathbf{I}_M - \mathbf{C}(\phi)$ is positive definite. If we assume that pathways are *a priori* equally perturbed, ϕ_p must fall in the interval $(\frac{1}{P\xi_p^1}, \frac{1}{P\xi_p^2})$ where ξ_p^1 and ξ_p^2 are the minimum and maximum eigenvalues of $\mathbf{G}_p \mathbf{A}_p$, respectively. In practice, strong interaction between observed metabolites of pathway p is reproduced in CAR models only when the scaling parameter ϕ_p is quite close to one of the boundaries $\frac{1}{P\xi_p^1}, \frac{1}{P\xi_p^2}$. Hence, we use a beta-type prior for ϕ_p that places substantial mass on large values of $|\phi_p|$ [2] :

$$p(\phi_p) = \frac{1}{B(\frac{1}{2}, \frac{1}{2})} \left(\phi_p - \frac{1}{P\xi_p^1} \right)^{-\frac{1}{2}} \left(\frac{1}{P\xi_p^2} - \phi_p \right)^{-\frac{1}{2}} \quad (3)$$

3.2 Integrative analysis

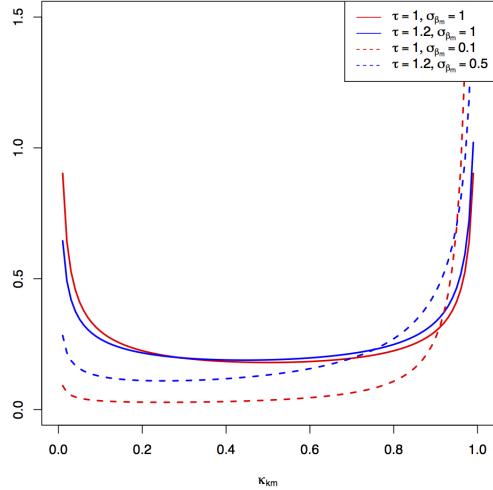


Figure 1: Prior $p(\kappa_{mk}|\tau, \sigma_{\beta_m})$ on κ_{mk} for different values of σ_{β_m} and τ . The prior distribution skews towards 1 if τ increases or σ_{β_m} decreases.

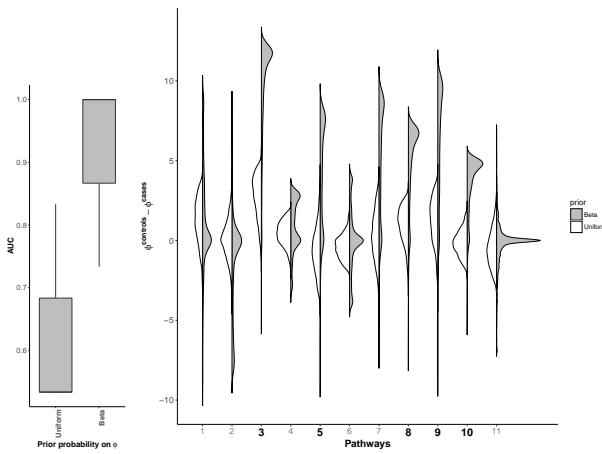


Figure 2: Left : Boxplots of Area Under the Curve (AUC) for pathway perturbation inference with uniform prior on ϕ compared to a beta like prior on ϕ for 10 simulated datasets. We infer perturbations based on the posterior distribution of $\phi^{\text{controls}} - \phi^{\text{cases}}$. Right : Distribution of $\phi^{\text{controls}} - \phi^{\text{cases}}$ under the uniform prior (white) and the beta like prior (grey). True perturbed pathways are printed in bold on the x axis.

where \mathbf{B} is the beta function. The parameter σ^2 captures variance heterogeneity in metabolite intensities and is given an inverse gamma prior $\mathbf{G}(\psi, \psi - 1)$. This prior provides 2ψ pseudo-observations in addition to NT available observations. In order to build a reasonably informative prior we set $\psi = N \times T/4$.

Interactions between heterogeneous omic variables such as transcripts and metabolites or gene expression and metabolites is modeled via the following hierarchical shrinkage model:

$$\mu_{itm} = \alpha_m + \gamma_{im} + \mathbf{y}_{it}\boldsymbol{\beta}_m + \nu_{itm} \quad (4)$$

$$\beta_{mk} | \lambda_{mk}, \sigma_{\beta_m} \sim N(0, \lambda_{mk}^2 \sigma_{\beta_m}^2) \quad (5)$$

$$\lambda_{mk} | \tau \sim \text{St}^+(\tau, 0, 1) \quad (6)$$

where St^+ denotes the half Student-t distribution with τ degrees of freedom, α_m represents treatment effect for metabolite m , $\gamma_{im} \sim N(0, \sigma_{\gamma_m}^2)$ represents individual perturbations for metabolite m , $\nu_{itm} | \nu_{i,t-1,m} \sim N(\theta_m \nu_{i,t-1,m}, \sigma_{\nu_m}^2)$ follows an auto-regressive process and represents temporal effects for metabolite m of individual i at time point t . Finally, $\boldsymbol{\beta}_m$ quantifies interactions between metabolite m and other omic variables. λ_{mk} is called local shrinkage parameter whilst $\sigma_{\beta_m}^2$ is the global shrinkage parameter. For $\tau = 1$, this prior reduces to the horseshoe prior. Intuitively, for small values of λ_{mk} the coefficient β_{mk} is very close to 0 while for relevant variables λ_{mk} will be large. In addition, σ_{β_m} controls the overall shrinkage level i.e sparsity of the vector $\boldsymbol{\beta}_m$ is more important for small values of σ_{β_m} .

Define $\kappa_{mk} = \frac{1}{1 + \lambda_{mk}^2 \sigma_{\beta_m}^2 / \tau}$ a random shrinkage coefficient such that $\kappa_{mk} \approx 0$ when λ_{mk} is large and $\kappa_{mk} \approx 1$ when λ_{mk} is small. This transformation implies the following prior distribution on κ_{mk} :

$$p(\kappa_{mk}|\tau, \sigma_{\beta_m}) = \frac{1}{2\sqrt{\pi} \mathbf{B}(\frac{\tau}{2}, \frac{1}{2})} \frac{\sigma_{\beta_m}^\tau \kappa_{mk}^{\tau/2-1} (1-\kappa_{mk})^{-1/2}}{(1-\kappa_{mk} + \kappa_{mk} \sigma_{\beta_m}^2)^{\frac{\tau+1}{2}}} \quad (7)$$

This prior density is shown in figure 1 for different values of σ_{β_m} and τ . It reduces to a $\text{Beta}(\tau/2, 1/2)$ distribution if $\sigma_{\beta_m} = 1$ and to a $\text{Beta}(1/2, 1/2)$ which looks like a horseshoe, if in addition $\tau = 1$. When τ increases, $\text{Beta}(\tau/2, 1/2)$ skews towards 1 which increases the global shrinkage power. The expectation of $\boldsymbol{\beta}_m$ given $\mathbf{Y}, \kappa_m, \tau, \mu_{tm}$ can be expressed as:

$$\begin{aligned} \mathbb{E}(\boldsymbol{\beta}_m | \mathbf{Y}, \kappa_m, \tau, \mu_{tm}) &= \left(\sum_{t=1}^T \mathbf{Y}_t^T \Sigma_m^{-1} \mathbf{Y}_t + \tau \Upsilon_m \right)^{-1} \\ &\quad \times \sum_{t=1}^T \mathbf{Y}_t^T \Sigma_m^{-1} \mu_{tm} \end{aligned} \quad (8)$$

where $\Sigma_m = \left(\frac{\sigma_{\nu_m}^2}{1-\theta_m^2} + \sigma_{\gamma_m}^2 \right) \mathbf{I}_N$ and Υ_m is a diagonal matrix of order K with elements $1/\kappa_{mk} - 1$. Equation (8) introduces a penalty term $\tau \Upsilon_m$ where

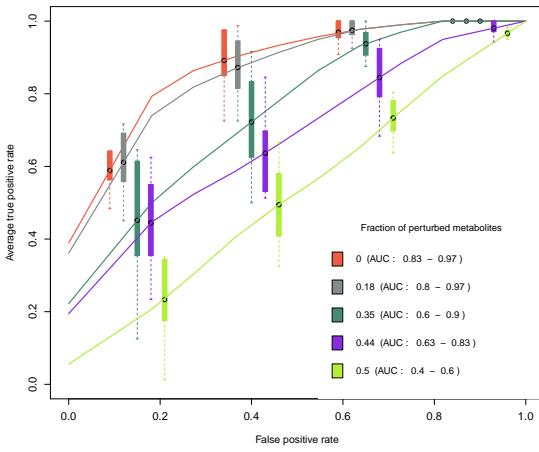


Figure 3: Average Receiver Operating Characteristic (ROC) curves for pathway perturbation inference across 10 datasets for different fractions of “falsely” assigned metabolites.

Υ_m is a metabolite specific penalty term introduced by the horseshoe prior and τ is a global penalty term. Precisely, τ captures the overall sparsity level amongst all metabolites. The expectation of β_m given $\mathbf{Y}, \kappa_m, \tau, \mu_{tm}$ is very similar to the estimate of β_m under ridge regression where $\tau\Upsilon_m$ simply reduces to $\tau\mathbf{I}_N$.

The global sparsity level can be controlled using τ . Increasing the global sparsity level is a desired property in omic studies, as usually we deal with a large number of omic variables where only few are important. Moreover, when there is prior knowledge available, specifying τ *a priori* can optimize the inference and additionally, provide a more informative prior on λ_{mk} . If we fix $p(\sigma_{\beta_m}^2) \propto 1/\sigma_{\beta_m}^2$, integrating over σ_{β_m} gives the expected value of κ_{mk} as :

$$\begin{aligned} \mathbb{E}(\kappa_{km}|\tau) &= \frac{\Gamma(1/2)^{-1}}{2\sqrt{\pi}\Gamma(\tau/2)} \\ &\times \mathbf{G}_{3,3}^{2,3} \left(\begin{matrix} 1, \tau/2, 0 \\ \tau/2, \tau/2 - 1/2, 0 \end{matrix} \middle| 1 - \sigma_{\beta_m}^2 \right) \end{aligned}$$

where $\mathbf{G}_{\cdot,\cdot}^{2,3}$ is Meijer’s G-function [meijer1936]. The equation above can be used to fix τ *a priori* by defining the expected proportion of shrunk coefficients. In practice, different values of τ are plugged into the equation above to get the desired proportion of shrunk coefficients. However, many definite integrals can be obtained using the tables of Meijer functions in [5] for special values of parameters.

3.3 Experimental design

The covariance structure between metabolites might change drastically as a result of treatment if the latter affects metabolic pathways. The model can be extended to take into account the experimental design. As specified in the previous section, α_m captures the treatment effect for metabolite m , γ_{im} represents individual perturbations for metabolite m , $\nu_{itm}|\nu_{i,t-1,m} \sim N(\theta_m\nu_{i,t-1,m}, \sigma_{\nu_m}^2)$ represents temporal effects for metabolite m of individual i at time point t in equation (5). In addition, we allow covariance structures $\mathbf{C}(\phi^e)$ to be different for the control samples and the cases where $e \in \{\text{cases, controls}\}$ designates experimental groups. This yields the overall hierarchical model:

$$\mathbf{x}_{it}^e | \boldsymbol{\mu}_{it}, \mathbf{C}, \sigma \sim N\left(\boldsymbol{\mu}_{it}, (\mathbf{I}_M - \mathbf{C}(\phi^e))^{-1} \sigma^2\right) \quad (9)$$

$$\mu_{itm} = \alpha_m + \gamma_{im} + \mathbf{y}_{it}\boldsymbol{\beta}_m + \nu_{itm} \quad (10)$$

$$\beta_{mk} | \lambda_{mk}, \sigma_{\beta_m} \sim N(0, \lambda_{mk}^2 \sigma_{\beta_m}^2) \quad (11)$$

$$\lambda_{mk} | \tau \sim \text{St}^+(\tau, 0, 1) \quad (12)$$

$$\gamma_{im} | \sigma_{\gamma_m} \sim N(0, \sigma_{\gamma_m}^2) \quad (13)$$

$$\nu_{itm} | \theta_m, \sigma_{\nu_m} \sim N(\theta_m \nu_{i,t-1,m}, \sigma_{\nu_m}^2) \quad (14)$$

Note that by specifying different dependence parameters for metabolite interactions in cases and controls metabolism, the model is able to identify perturbed pathways by comparing ϕ^{cases} and ϕ^{controls} .

4 RESULTS

In this section we perform experiments on both synthetic and real data to investigate whether our algorithm gives reasonable solutions. We first try our method on a simulated dataset in section 4.1 to get an understanding of the performance of our method. In section 4.2, we test our method on a data set using metabolomic and bacterial composition in a drug treatment experiment. In the following, we refer to our model as “iCARH” model for “integrative CAR Horseshoe” model.

4.1 Simulation study

To get better understanding of our method and test its applicability, we first perform our approach on synthetic datasets. We will mainly focus on the ability of our model to infer pathway perturbation.

Assessing pathway inference with beta like prior

In the first simulation our objective is to assess how the beta like prior in equation (3) improves the iCARH model. We first fixed the number of pathways P to 11,

then simulate the design matrices \mathbf{A}_p . Specifically, a membership matrix \mathbf{Z} with dimensions $M \times P$ is randomly generated based on the density of the number of KEGG pathways in which a single metabolite is involved. Each design matrix \mathbf{A}_p is then equal to $\mathbf{z}_p \mathbf{z}_p^T$ where \mathbf{z}_p is the p th column of Z . Finally, we generated 10 datasets according to the model below in order to assess how our model infers perturbed pathways:

$$\omega \sim \text{Bernoulli}(\pi_\omega) \quad (15)$$

$$\phi_p^{\text{controls}} | \omega \sim \omega N_{[0,1/P\xi_p^2]} \left(\frac{1}{P\xi_p^2} - \rho, \sigma_\phi^2 \right) + (1 - \omega) N(0, \psi^2) \quad (16)$$

$$\phi_p^{\text{cases}} | \phi_p^{\text{controls}}, \omega \sim \omega N_{[1/P\xi_p^1, 0]} \left(\frac{1}{P\xi_p^1} + \rho, \sigma_\phi^2 \right) + (1 - \omega) \delta_{\phi_p^{\text{controls}}} \quad (17)$$

with $N_{[a,b]}$ denotes the truncated normal distribution with boundaries a , b , ξ_p^1 and ξ_p^2 are the minimum and maximum eigenvalues of $\mathbf{G}_p \mathbf{A}_p$. The rest of the parameters is set as follows : number of bacterial variables $K = 1$, number of metabolites $M = 40$, number of time points $T = 7$, number of samples $N = 22$, global parameter τ fixed to 1.2, parameters ν_{itm} , γ_{im} , μ_{itm} , \mathbf{x}_{it}^e simulated according to equations (14), (13), (10), (9) respectively.

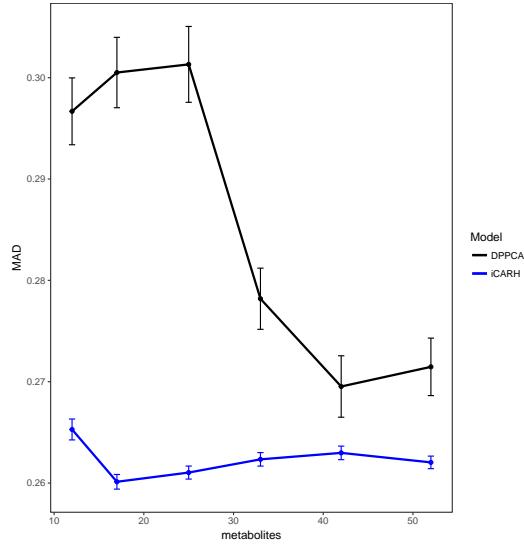


Figure 4: Posterior predictive checks for mean absolute deviation (MAD) compared to DPPCA for different numbers of metabolites included. The MAD decreases as the number of metabolites increases. Our model performs clearly better than the DPPCA model.

We set non-informative uniform priors on α_m , $\sigma_{\gamma_{im}}$, θ_m , $\sigma_{\mu_m}^2$. We set an informative prior on $\sigma_{\gamma_{im}}^2 \sim$ inverse-gamma (1, 0.1) as we expect low variability amongst biological samples of the same group. We fix π_ω to 0.7 the proportion of expected perturbed pathways. ρ is fixed to a value of 0.05, σ_ϕ^2 to 0.2 and ψ to a large value. We compare inference of the model under a uniform prior for ϕ_p^e and the prior in equation (3).

Inference is done using 2000 iterations of Hamiltonian Monte Carlo sampling and 1000 warm-up iterations.

The left plot in figure 2 shows the boxplots of the Area Under the Curve (AUC) for pathway perturbation inference for 10 simulated datasets with uniform prior on ϕ and a beta like prior on ϕ . We infer perturbations based on the posterior probability that ϕ^{controls} and ϕ^{cases} are different i.e. the 95% credible interval of $\phi^{\text{controls}} - \phi^{\text{cases}}$ does not contain zero. The AUC values for the beta like distribution is significantly higher than the AUC values for the uniform distribution. On average pathway perturbation inference under the uniform distribution reduces to a random guess with an average AUC of 0.53. This is likely due to the lack of variance of the uniform distribution. The right plot in figure 2 shows the posterior distributions of $\phi^{\text{cases}} - \phi^{\text{cases}}$ under the uniform prior (white) and the beta like prior (grey) for each pathway. True perturbed pathways are printed in bold on the x axis.

Assessing pathway inference against design inaccuracies

It is very common in metabolomics data to find metabolites that are correlated but not in the same KEGG pathway. In the following simulation we assess how inaccuracies in the covariance structure between metabolites and the design matrices \mathbf{A}_p affect the iCARH model. We used the 10 datasets from the previous simulation and perturbed the design matrices by selecting a random fraction of metabolites in each pathway. We then randomly (falsely) assign these metabolites to no pathway, or to different pathways. We similarly run the model for 2000 iterations of Hamiltonian Monte Carlo sampling and 1000 warm-up iterations for each of the fractions $y = 0, 0.18, 0.35, 0.44, 0.5, 0.62$ of perturbed metabolites. Finally, in the same fashion, we assess perturbations based on the 95% credible interval of $\phi^{\text{controls}} - \phi^{\text{cases}}$. Figure 3 is a series of average Receiver Operating Characteristic (ROC) curves across 10 datasets for each of the fractions y . On average, the performance of our model reduces to a random guess (AUC of 0.5) if 50% of the metabolites in each pathway is perturbed. The AUC of our model reaches 0.97 if no metabolites are perturbed and is about 0.88 if 18% of the metabolites in each pathway are perturbed.

4.2 Case study

In this section, we test our model on an actual metabolomic data and 16S data for bacterial profiles. In this study we are interested in the influence of metformin on a non-diabetic model. Metformin is the first-line medicine to treat type 2 diabetes. It has also been

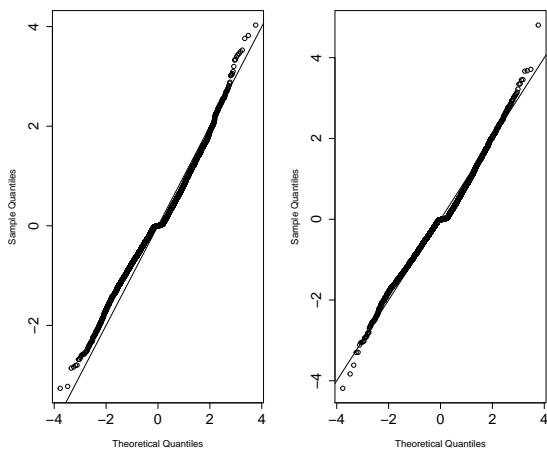


Figure 5: Right and left panels show model fit assessment for controls and cases data. Left : quantile-quantile normal plot of $\Psi_{\text{cases}}^{-1}(\mathbf{x}_{it} - \boldsymbol{\mu}_{it})$. Right : quantile-quantile normal plot of $\Psi_{\text{controls}}^{-1}(\mathbf{x}_{it} - \boldsymbol{\mu}_{it})$.

suggested that metformin has anti-cancer, cardiovascular and anti-aging effects. Because of their very large metabolic capacity, the gut bacteria can influence toxicity and metabolism of drugs. Here, we are particularly looking for metabolic biomarkers indicative of microbiota changes as result of treatment.

The study design is as follows: metabolic profiles of 24 rats are acquired once a week using different mass spectrometry techniques from plasma samples over a period of 9 weeks. Bacterial profiles are acquired using miSeq. The study has allowed for two groups of 12 rats where metformin has been administrated to the second group (weeks 3 to 7) allowing for 2 weeks of acclimatation (weeks 1 and 2) and 2 weeks of recovery (weeks 8 and 9). After data processing and metabolite identification, a total of 56 metabolites and 6 bacteria species are further analysed using our model. Inference is done using 2000 iterations of Hamiltonian Monte Carlo sampling.

We assess performance of our model for different values of τ using the Watanabe-Akaike information criterion (WAIC). Tested values of τ comprise 1, 1.2, 5, 10 with corresponding WAIC values of 7317.296 , 7322.798 , 7317.457 , 7316.476 respectively. WAIC values are very similar for different values of τ which suggests to use the most selective model with $\tau = 10$ as it is the simplest i.e with the smallest number of selected variables.

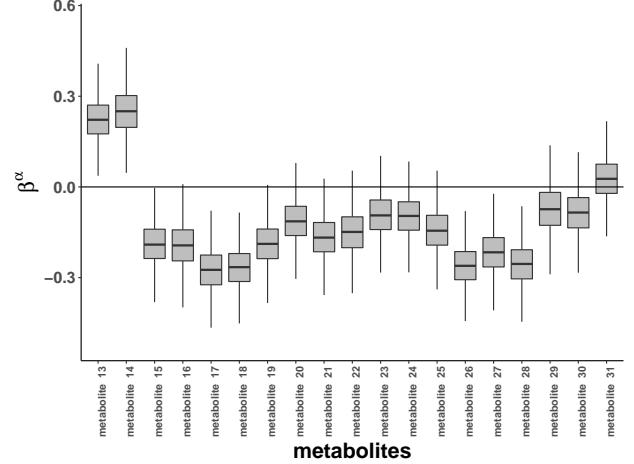


Figure 6: Estimates of effects of treatment on metabolite profiles are captured by β_m^α . Only part of the data is plotted as we are only interested in “metabolite 27”.

Assessing model fit

In order to assess our model fit, we perform posterior predictive checks of our model compared to DPPCA [19]. The DPPCA model is a multivariate model using PCA, where PCA scores are modeled via a stochastic volatility model. In the Bayesian framework, posterior predictive checks consist in comparing data simulated from the posterior predictive distribution with the observed data. We compared the simulated data and the observed data by means of mean absolute deviations (MADs) between the observed and the simulated covariance matrices for different numbers of metabolites included i.e only part of the data corresponding to these metabolites is considered. The same process was repeated for inference using the DPPCA model [19]. Figure 4 shows MADs of our model and the DPPCA model. As expected, MADs for both models decrease when the number of metabolites increases. Overall, our model clearly outperforms the DPPCA model.

In addition to posterior predictive checks, goodness of fit was also checked by using $\Psi_e^{-1}(\mathbf{x}_{it} - \boldsymbol{\mu}_{it}) \sim N(0, \mathbf{I}_M)$ where Ψ_e denotes the Cholesky factor of $(\mathbf{I}_M - \mathbf{C}(\phi^e))^{-1}\sigma^2$. Zero-mean and normality were thus checked for $\Psi_e^{-1}(\mathbf{x}_{it} - \boldsymbol{\mu}_{it})$ (See figure 5).

Data results

Since the administrated drug was also profiled using mass spectrometry, we fit the iCARH model with $\alpha_m = \beta_m^\alpha \mathbf{y}_{\text{drug}}$. Figure 6 is a series of boxplots for β_m^α for metabolites 13 to 31. We are mainly interested in “metabolite 27” as it is associated with some bacteria species.

Figures 7 and 8 show posterior distributions of ϕ^e for

each pathway and estimates of effects of bacteria on metabolites. Results in section 4.1 suggest to compare the covariance structure of metabolites in the observed data with the covariance induced by the design matrices in order to have an *a priori* idea on the robustness of pathway inference (See figure 3). For a correlation threshold of 0.3, about 25% of the metabolites are misspecified in the design matrices which corresponds to an AUC around 0.8 according to figure 3. If we set a higher correlation threshold, a lower number of metabolites will be misspecified. This supports the use of the iCARH model for pathway perturbation inference for this data.

Estimates of effects of bacteria on metabolite profiles are captured by β_m . Some metabolites present significant changes along with the bacterial profiles. For example, “metabolite 27”, a hydroxy fatty acid, is associated with alterations in abundance of 4 bacteria species. Figure 7 shows that, as a result of treatment, KEGG pathways are not significantly altered. However, distributions of ϕ^{controls} for “fatty acids biosynthesis” and “biosynthesis of unsaturated fatty acids” KEGG pathways are remarkably flatter than the distributions of ϕ^{cases} . These pathways involve the previously identified hydroxy fatty acid metabolite. Our analysis confirms previously reported studies that hydroxy fatty acids might be produced by the gut microbiome [kishino2013, kimura2013].

5 DISCUSSION

Identifying biomarkers in time course metabolic data and inferring significant associations with heterogeneous omic variables is extremely challenging due to the several sources of variations of the data. In addition, existing methods developed to analyze such data are very scarce and have the limitations of i) overfitting to the few available data points or ii) confounding the experimental and longitudinal variation or iii) ignor-

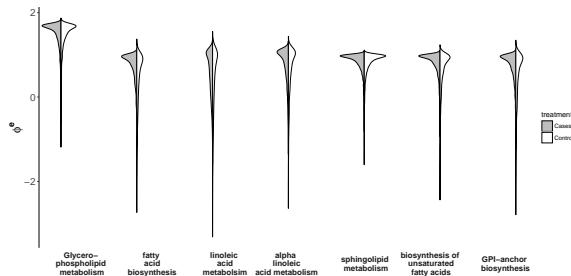


Figure 7: Posterior distributions of ϕ^e for each pathway for each treatment group. Posterior distributions of ϕ^e are very similar for both controls and cases which is indicative of no -or mild- pathway alterations.

ing the metabolite interactions or iv) ignoring effects of other omic variables. In this paper, the model we have developed combines several approaches to take into account the different aspects of the data namely the number of time points, the experimental variation captured by μ_{it} , interactions between metabolites captured by ϕ and interactions with additional omic variables captured by β_m .

Our results demonstrate that our model successfully addresses the main questions of a metabolomic study. Most importantly, our model is able to identify metabolic biomarkers related to treatment, infer perturbed pathways as a result of treatment and find significant associations with additional omic variables. We have shown that providing an informative prior on metabolic pathways and an informative prior over the parameter ϕ is a significant improvement over the DPPCA model. Particularly, our model is more robust to slight variations usually observed in short time series data thanks to the small number of covariance parameters needed to estimate compared to DPPCA. We have also shown through simulation that an informative beta like prior compares better than a non-informative uniform prior in inferring significant pathways. On real data, we have investigated how the number of profiled metabolites can affect the predictive ability of the model.

Several potential extensions arise naturally from our model. In terms of the metabolite interactions component, many research questions can arise. Alternative strategies to modeling metabolite interactions can be examined such as modeling the non-zero elements of the adjacency matrix C of each pathway as random variables. This strategy was adopted in the CAR literature by [16, 24] to take into account step changes in spatial variation. Step changes can potentially be useful to model changes in metabolites correlations as a result of treatment. [15] provide an overview of different CAR models used in spatial modeling. The proposed models can be adapted to fit into the metabolomics literature.

From a practical point of view, the model has been fitted using HMC sampling but takes a large amount of time (about 1 hour) mostly because of the variable selection procedure and metabolites interdependence. This could be addressed by using variational Bayes. In fact, variational Bayes inference procedures offer cost-effective inference by means of principled approximations and appealing computational time for high dimensional data. A variational bayes inference of CAR models was proposed by [12] for high dimensional data, and a variational bayes approach for variable selection was recently proposed by [20].

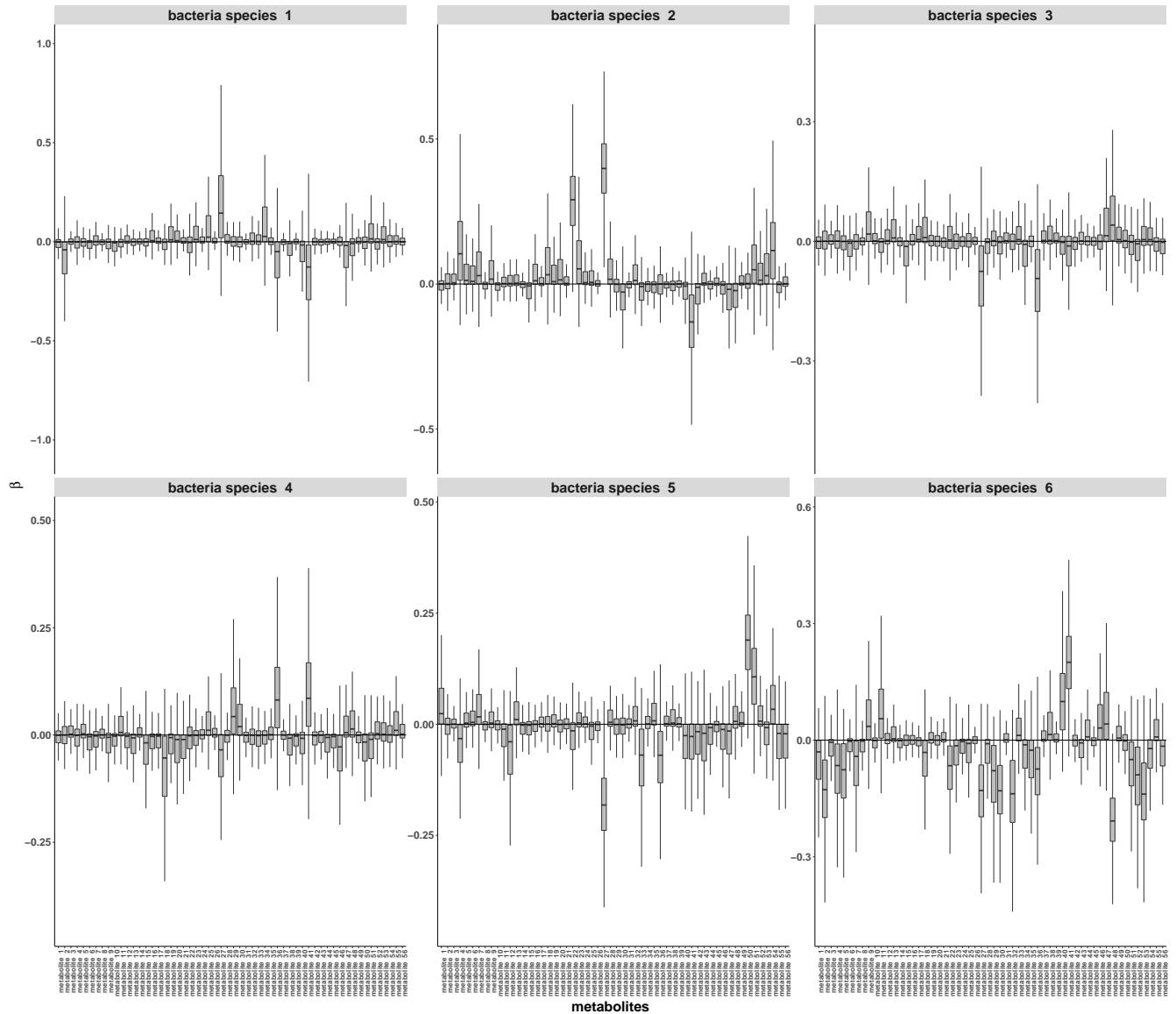


Figure 8: Estimates of effects of bacteria on metabolite profiles are captured by β_m .

6 CONCLUSION

Metabolomics longitudinal profiling techniques are imperative to understand the effect of a drug or a disease across time and can provide enhanced understanding of the underlying biology of the system. In a data integration framework, we have illustrated the use of the CAR model to incorporate metabolites interactions in the model and the horseshoe prior to identify association with heterogeneous omic variables obtained by other omic techniques. The combination of the CAR and horseshoe levels yields the “integrative CAR Horseshoe” (iCARH) model which we presented in this article.

Although, it is computationally expensive, the iCARH model has various appealing features such that it is able to identify metabolic biomarkers related to treatment, infer perturbed pathways as a result of treatment and identify potential associations between heterogeneous omic variables. Clearly, these appealing features open up further research topics.

Acknowledgements

Thanks are due to Panagiotis Vorkas for providing the metabolic and 16S data. Infrastructure support for this work was provided by the NIHR Imperial Biomedical Research Centre.

References

- Appel. Sur les fonctions hypergomtriques de deux variables. *fre. Journal de Mathmatiques Pures et Appliques*, 173–216 (1882).
- Banerjee, S., Carlin, B. P. & Gelfand, A. E. *Hierarchical modeling and analysis for spatial data* (Crc Press, 2014).
- Berk, M., Ebbels, T. & Montana, G. A statistical framework for biomarker discovery in metabolomic time course data. *Bioinformatics* **27**, 1979–1985 (2011).
- Bro, R. PARAFAC. Tutorial and applications. *Chemometrics and intelligent laboratory systems* **38**, 149–171 (1997).
- Brychkov, Y. A. *Handbook of special functions: derivatives, integrals, series and other formulas* (CRC Press, 2008).
- Chen, C. *et al.* Altered metabolite levels and correlations in patients with colorectal cancer and polyps detected using seemingly unrelated regression analysis. *Metabolomics* **13**, 125. ISSN: 1573-3882 (2017).
- Chen, C. *et al.* Exploring metabolic profile differences between colorectal polyp patients and controls using seemingly unrelated regression. *Jour-**nal of Proteome Research* **14**, 2492–2499. ISSN: 15353907 (2015).
- Cressie, N. & Wikle, C. K. *Statistics for spatio-temporal data* (John Wiley & Sons, 2015).
- Daviss, B. Growing pains for metabolomics: the newest’omic science is producing results—and more data than researchers know what to do with. *The Scientist* **19**, 25–29 (2005).
- Dettmer, K., Aronov, P. A. & Hammock, B. D. Mass spectrometry-based metabolomics. *Mass spectrometry reviews* **26**, 51–78 (2007).
- Ebrahim, A. *et al.* Multi-omic data integration enables discovery of hidden biological regularities. *Nature communications* **7** (2016).
- Harrison, L. M. & Green, G. G. A Bayesian spatiotemporal model for very large data sets. *NeuroImage* **50**, 1126–1141 (2010).
- Joyce, A. R. & Palsson, B. Ø. The model organism as a system: integrating’omics’ data sets. *Nature reviews Molecular cell biology* **7**, 198–210 (2006).
- Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30 (2000).
- Lee, D. A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology* **2**, 79–89 (2011).
- Lee, D. & Mitchell, R. Locally adaptive spatial smoothing using conditional auto-regressive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**, 593–608 (2013).
- Löfstedt, T. & Trygg, J. OnPLSa novel multi-block method for the modelling of predictive and orthogonal variation. *Journal of Chemometrics* **25**, 441–455 (2011).
- Mei, Y., Kim, S. B. & Tsui, K. L. Linear-mixed effects models for feature selection in high-dimensional NMR spectra. *Expert Systems with Applications* **36**, 4703–4708. ISSN: 09574174 (2009).
- Nyamundanda, G., Gormley, I. C. & Brennan, L. A dynamic probabilistic principal components model for the analysis of longitudinal metabolomics data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **63**, 763–782 (2014).
- Ormerod, J. T., You, C., Müller, S., *et al.* A variational Bayes approach to variable selection. *Electronic Journal of Statistics* **11**, 3549–3594 (2017).

21. Pham, L. M., Carvalho, L., Schaus, S. & Kolaszyk, E. D. Perturbation Detection Through Modeling of Gene Expression on a Latent Biological Pathway Network: A Bayesian hierarchical approach. *Journal of the American Statistical Association* **145**, 1–61. ISSN: 0162-1459 (2015).
22. Rantalainen, M. *et al.* Piecewise multivariate modelling of sequential metabolic profiling data. *BMC bioinformatics* **9**, 105 (2008).
23. Reo, N. V. NMR-based metabolomics. *Drug and chemical toxicology* **25**, 375–382 (2002).
24. Rushworth, A., Lee, D. & Saran, C. An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk. *Journal of the Royal Statistical Society. Series C: Applied Statistics* **66**, 141–157. ISSN: 14679876 (2017).
25. Trygg, J. & Wold, S. O₂-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics* **17**, 53–64 (2003).
26. Trygg, J. & Wold, S. Orthogonal projections to latent structures (O-PLS). *Journal of chemometrics* **16**, 119–128 (2002).
27. Wold, S., Martens, H. & Wold, H. The multivariate calibration problem in chemistry solved by the PLS method. *Matrix pencils*, 286–293 (1983).