BIOINFORMATICS

Vol. 30 ECCB 2014, pages i461–i467
doi:10.1093/bioinformatics/btu455

Stronger findings for metabolomics through Bayesian modeling of multiple peaks and compound correlations

Tommi Suvitaival[1], Simon Rogers[2] and Samuel Kaski[1,3,*]

[1]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, FI-00076 Espoo, Finland, [2]School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK and [3]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland

ABSTRACT

Motivation: Data analysis for metabolomics suffers from uncertainty because of the noisy measurement technology and the small sample size of experiments. Noise and the small sample size lead to a high probability of false findings. Further, individual compounds have natural variation between samples, which in many cases renders them unreliable as biomarkers. However, the levels of similar compounds are typically highly correlated, which is a phenomenon that we model in this work.

Results: We propose a hierarchical Bayesian model for inferring differences between groups of samples more accurately in metabolomic studies, where the observed compounds are collinear. We discover that the method decreases the error of weak and non-existent covariate effects, and thereby reduces false-positive findings. To achieve this, the method makes use of the mass spectral peak data by clustering similar peaks into latent compounds, and by further clustering latent compounds into groups that respond in a coherent way to the experimental covariates. We demonstrate the method with three simulated studies and validate it with a metabolomic benchmark dataset.

Availability and implementation: An implementation in R is available at http://research.ics.aalto.fi/mi/software/peakANOVA/.

Contact: samuel.kaski@aalto.fi.

1 INTRODUCTION

Changes in metabolite concentrations provide insights into disturbances in biological processes that take place in organisms. Changes in the metabolome are informative, especially about nutrition and metabolism (Orešič, 2009), and about the immune system (Kau *et al*., 2011). Chromatography-coupled mass spectrometry (Plumb *et al*., 2004) is the standard measurement technology for the untargeted quantification of metabolites and other small molecules.

The spectral data from the measurement device are known to be noisy with various sources of uncertainty (Katajamaa and Orešič, 2007), starting from sample preparation and compound ionization, and ending at peak identification, annotation and summarization. However, the spectra also have structure (Steuer, 2006; Rogers *et al*., 2009) that is useful for the inference of differences between groups of samples.

Because of the high level of noise, excessive false discovery has been highlighted among the main risks in the analysis of metabolomic data (Broadhurst and Kell, 2006). On the other hand, weak changes are likely to go undetected from observations of individual compounds (Saccenti *et al*., 2014).

Singular value decomposition (SVD)-based dimensionality reduction techniques, such as analysis of variance (ANOVA)-simultaneous component analysis (ASCA; Smilde *et al*., 2005), have been proposed to identifying interpretable associations between experimental covariates and multivariate changes in the metabolome. However, as the decomposition in ASCA operates on the covariate effects of the standard ANOVA model, the method does not improve the quantification of the covariate effects compared with the standard model. Further, SVD has been applied to interpreting changes in the variance of the samples in association to the covariates (Jansen *et al*., 2012), again, building on the standard ANOVA model.

Outside metabolomics, structured ANOVA-type models have been proposed to improve the inference of covariate effects: a Gaussian process-based ANOVA model for spatial data (Kaufman and Sain, 2010) enables the inference of smooth covariate effects for nearby data points, and a dependent Dirichlet process mixture of ANOVA models (De Iorio *et al*., 2004) can identify substructure in a designed experiment with low-dimensional observations of the outcome.

For metabolomics, a Bayesian clustering model (Suvitaival *et al*., 2014) has recently been proposed for improving the inference of covariate effects through the integration of multiple same-source spectral peaks. Individual spectral peaks have been argued to be unreliable for the statistical analysis because of their high level of noise. Although the mass spectrometer produces multiple peaks that arise from one compound, there so far are only few methods to integrate these additional observations: Kuhl *et al*. (2012), Rogers *et al*. (2009) and Tikunov *et al*. (2012) used multiple peaks to enhance peak annotation, addressing a major source of error in the analysis of metabolomic data. The recently proposed multipeak model for the inference of covariate effects (Suvitaival *et al*., 2014) is, to our knowledge, the first systematic approach for using additional peaks in the statistical analysis of intensity data.

In this work, we aim at reducing the risk of false associations between experimental covariates and the observed metabolome. We propose a structured ANOVA-type model that benefits both from the multiple spectral peaks produced by the mass spectrometer and from the collinear structure (Huopaniemi *et al*., 2009; Steuer, 2006) of metabolomic data. Because of the collinearity that arises from the compounds' concurrent involvement in biological processes, it is reasonable to model individual compounds as members of coherent groups of compounds.

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

We achieve this by introducing another level of hierarchy to the peak-clustering model (Suvitaival *et al.*, 2014). We show that by not only clustering spectral peaks into latent compounds but in addition by clustering these compounds into coherently responding latent groups, we can detect weak covariate effects in the data more accurately.

2 METHODS

The method introduced in this work extends the peak-clustering model (Suvitaival *et al.*, 2014) to reduce the risk of false discoveries in highly collinear metabolomic data. We address the problem of small sample size and low signal-to-noise ratio by introducing stronger structure to the model. This enables us to detect weak covariate effects that are present in multiple compounds.

In this section, we describe the proposed hierarchical Bayesian two-level model. In the equations that follow, we use the indices i, j, m, k and l to denote the samples, the variables, the first-level clusters, the second-level clusters and the covariate level of the sample, respectively, as

$$\begin{aligned} i &= 1, \ldots, N \quad \text{(samples, } i.e. \text{ experimental runs)}, \\ j &= 1, \ldots, P \quad \text{(variables, } i.e. \text{ peaks)}, \\ m &= 1, \ldots, M \quad \text{(first-level clusters, } i.e. \text{ compounds)}, \\ k &= 1, \ldots, K \quad \text{(second-level clusters, } i.e. \text{ groups of compounds)}, \\ l &= 1, \ldots, L_a \quad \text{(covariate level, } i.e. \text{ sample group)} \end{aligned}$$

where N, P, M, K and L_a are their respective total numbers.

The observed data are the spectral peaks, indexed by j, following their identification in the samples, indexed by i. However, the association between the peaks and the compounds, indexed by m, is unknown as is the total number, M, and the identity of the compounds.

The data on the peaks are organized into two arrays: first, the peak height information is arranged into a $P-$by$-N$ matrix $\mathbf{X} \in \mathbb{R}^{P \times N}$, which after the log -transformation and centering based on the control group, $l = 1$, is real valued with missing values where a peak was not detected. Secondly, the peaks' pairwise similarity information is arranged into the array $\mathbf{Q} \in [0, 1]^{N \times P \times P}$. We choose to measure the similarity between two spectral peaks by computing the Pearson correlation between the peaks over a retention time window. This leads to a measure, where peaks j and j', if co-occurring within the retention time window in the sample i, have a positive similarity value $q_{i,j,j'}$. Similarity values for pairs with a missing peak or a negative correlation coefficient are set to zero and thus are effectively considered as missing values in the model. The peak similarity data enable the model to cluster together adduct and isotope peaks, which have a different mass-to-charge ratio but which appear at a coinciding retention time.

2.1 Peak clustering based on chromatographic similarity

We follow the approach detailed by Suvitaival *et al.* (2014) in the peak-clustering stage, presented briefly here for completeness.

In the following equations for inferring the $P-$by$-M$ clustering matrix $\mathbf{V}$, we use the variable $\varepsilon_{j,j'} \equiv \mathbf{v}_{j,\cdot}\mathbf{v}_{j',\cdot}^{\mathrm{T}} \in \{0, 1\}$ to indicate whether the peaks j and j' are in the same or different clusters (1 or 0, respectively). In the notation, the subset operator '$\cdot$' indicates that the entire dimension of the array is included—here, all the M clusters (columns) of the clustering matrix $\mathbf{V}$.

2.1.1 Peak similarity A pair of peaks from one compound can only occur close by in retention time, whereas a pair of peaks from two different compounds does not have such restriction set by the measurement device. This means that the observed similarity between same-compound peaks is expected to be higher than between different-compound peaks. Thus, the similarity value $q_{i,j,j'}$ between the peaks j and j' in a sample i is assumed to have been generated by one of the two components, 'in' or 'out', both of which are a spike-and-slab mixture (Mitchell and Beauchamp, 1988) with a beta distribution as the 'slab'. These two components are parametrized as

$$\begin{aligned} p(q_{i,j,j'} | \varepsilon_{j,j'} = 1) = (1 - p_0^{\mathrm{in}}) \mathrm{Beta}(q_{i,j,j'} | a_{\mathrm{in}}, b_{\mathrm{in}}) \\ + p_0^{\mathrm{in}} \delta(q_{i,j,j'}) \end{aligned} \tag{1}$$

for a pair of peaks in the same cluster and

$$\begin{aligned} p(q_{i,j,j'} | \varepsilon_{j,j'} = 0) = (1 - p_0^{\mathrm{out}}) \mathrm{Beta}(q_{i,j,j'} | a_{\mathrm{out}}, b_{\mathrm{out}}) \\ + p_0^{\mathrm{out}} \delta(q_{i,j,j'}) \end{aligned} \tag{2}$$

for a pair of peaks in different clusters. Both of the beta distributions have parameters a and b, which are set in the way that the same-cluster and different-cluster components favor large and small values of similarity $q_{i,j,j'}$, respectively.

In Equations (1) and (2), missing values are modeled through the 'spike' δ, which is a Dirac delta function that introduces a point mass at value zero of its argument. Many missing values are expected, as peaks from two different compounds rarely appear at the same time. The prior probability of a missing value is determined by p_0, which receives a higher value in the different-compound component than in the same-compound component. Different retention time, thus, is strong evidence for assigning the peaks into different clusters.

The likelihood,

$$\begin{aligned} \mathcal{L}(\mathbf{Q}|\mathbf{V}) = \prod_{i=1}^{N} \prod_{j=1}^{P-1} \prod_{j'=j+1}^{P} p(q_{i,j,j'} | \varepsilon_{j,j'} = 1)^{\varepsilon_{i,j'}} \\ \times p(q_{i,j,j'} | \varepsilon_{i,j,j'} = 0)^{1-\varepsilon_{j,j'}} \end{aligned} \tag{3}$$

for the data, $\mathbf{Q}$, is then computed through a product over all the samples, all the pairs of peaks, and the same-compound and different-compound terms.

2.1.2 Unknown compounds To accommodate the unknown set of compounds in the data, we set a Dirichlet process prior (Escobar, 1994) for the peak clusters. In this way, we not only can infer the assignments of the peaks into clusters representing the compounds but we can also infer the number of compounds, M, in the data, leading to a $P-$by$-M$ clustering matrix $\mathbf{V}$.

In the Dirichlet process for the cluster assignments of the peaks, the probability of the assignment of a peak j into an existing cluster m,

$$p(v_{j,m} = 1 | \mathbf{Q}, \mathbf{V}_{-j,\cdot}) \propto s_m \mathcal{L}\left(\mathbf{Q} | \mathbf{V}_{-j,\cdot}, v_{j,m} = 1\right) \tag{4}$$

is weighted by the number of other peaks in the cluster, $s_m \equiv \mathbf{v}_{-j,m}^{\mathrm{T}} \mathbf{v}_{-j,m}$. The probability is not dependent on the previous assignment of the peak j, which is left out both from the likelihood term and from the count. This is expressed in the equation by the subset operator '$-j$' that excludes the row j from the clustering matrix $\mathbf{V}$. As an alternative to the existing M clusters, a new cluster is created with the probability

$$p(v_{j,M+1} = 1 | \mathbf{Q}, \mathbf{V}) \propto \alpha_{\mathrm{DP}} \mathcal{L}\left(\mathbf{Q} | \mathbf{V}_{-j,\cdot}, v_{j,M+1} = 1\right) \tag{5}$$

which is affected by the Dirichlet process concentration parameter α_{DP}.

We infer the posterior distribution of the model via Gibbs sampling. Following the sampling, we acquire a point estimate of the distribution of the clustering as the least-squares clustering (Dahl, 2006) relative to the posterior mean clustering. The inferred clustering, $\mathbf{V}$, is then used as a preprocessor for the inference of covariate effects, which is discussed next.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

2.2 Compound clustering for modeling compound correlations

Following the preprocessing, described in Section 2.1, the matrix of observed peak intensities, $\mathbf{X} \in \mathbb{R}^{P\times N}$, for the P peaks and the N samples has been decomposed into M clusters through the P-by-M clustering matrix $\mathbf{V}$. Then, each sample i,

$$\mathbf{x}_{\cdot,i} \sim \mathcal{N}\left(\mathbf{V}\mathbf{x}^{\text{lat}}_{\cdot,i}, \Lambda\right) \tag{6}$$

is represented by a column of the latent variable matrix $\mathbf{X}^{\text{lat}} \in \mathbb{R}^{M\times N}$. Again, the subset operator '·' indicates that the entire dimension of the array is included, for instance, not only the single peak j of the observations matrix $\mathbf{X}$ or the single cluster m of the latent variable $\mathbf{X}^{\text{lat}}$. Following from Equation (6), the peak intensity data are assumed to have arisen from the clusters representing the M unknown compounds through a generative process with additive noise Λ. Further, we assume that the noise is independent by constraining the $P-$by$-P$ noise matrix Λ into a diagonal form with entries

$$\sigma_j^2 \sim \text{Scale} - \text{Inv}-\chi^2\left(n_0, \sigma_0^2\right) \tag{7}$$

that follow the scaled inverse-χ^2 distribution with the scale σ_0^2 and n_0 prior observations.

In this work, we add another level of hierarchy to the model by assuming that the compounds form groups that respond to the experimental covariates in a coherent manner (Fig. 1). We deviate from the formulation of the earlier work and do not let the covariate effects, α, operate directly on the latent variable, $\mathbf{X}^{\text{lat}}$, which represents the coherent variation within a cluster of peaks. Instead, we introduce a higher-level latent variable, $\mathbf{Z} \in \mathbb{R}^{K\times N}$, to represent the coherent variation within a group of compounds. The compound-specific latent variable,

$$\mathbf{x}^{\text{lat}}_{\cdot,i} \sim \mathcal{N}\left(\mathbf{W}\mathbf{z}_{\cdot,i}, \psi^2\mathbf{I}\right) \tag{8}$$

is then generated from the K higher-level clusters, represented by the higher-level latent variable, $\mathbf{Z}$, through the $M-$by$-K$ clustering matrix $\mathbf{W}$.

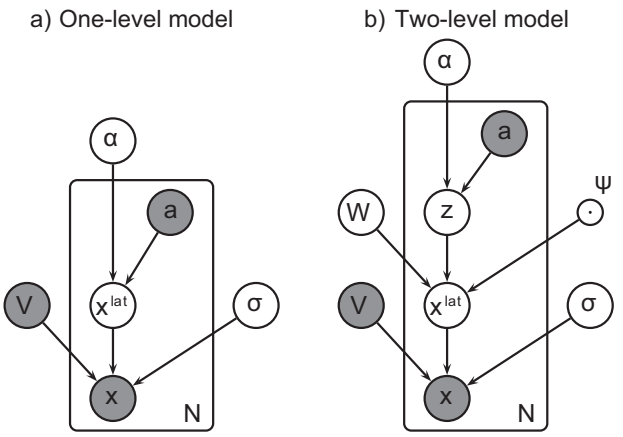


Fig. 1. Plate diagrams of the one-level peak-clustering model (**a**) (Suvitaival *et al.*, 2014) and the two-level compound-clustering model (**b**) (proposed in this work). The two-level model has a second level of hierarchy for modeling coherently responding groups of compounds. The shaded variables are observed: the intensity data $\mathbf{X}$, the covariate vector $\mathbf{a}$ and the peak-clustering matrix $\mathbf{V}$, which is acquired from the peak-clustering stage. White variables are inferred: the compound-specific latent variable $\mathbf{X}^{\text{lat}}$, the peak-specific variance σ^2, the compound-clustering $\mathbf{W}$, the compound group-specific latent variable $\mathbf{z}$ and the covariate effects α. The compound-level variance parameter ψ^2 is selected via cross-validation

The residual variation in the levels of a compound, which is not explained by its group, is controlled by the higher-level variance parameter, $\psi^2 \in \mathbb{R}_+$, which scales the $M-$by$-M$ identity matrix $\mathbf{I}$. In this way, the model can refine the information in the noisy observations first at the compound level and then at the compound group level. To infer the compound clustering, we set a uniform multinomial prior for the K higher-level clusters.

With the second level of hierarchy introduced to the model, the covariate effects, α, no longer operate directly on the compound-specific latent variable, $\mathbf{X}^{\text{lat}}$. Instead, we specify the effects to contribute to the higher-level latent variable, $\mathbf{Z}$. In a one-way experimental design, this means that the higher-level latent variable for the sample i,

$$\mathbf{z}_{\cdot,i} \sim \mathcal{N}\left(\alpha_{\cdot,a_i}, \mathbf{I}\right) \tag{9}$$

is generated from the K covariate effects, $\alpha_{\cdot,a_i}$, that correspond to the covariate level of the sample i, selected by the categorical indicator, a_i. This formulation encourages coherently responding compounds to be clustered together in the model.

All the covariate effects, $\alpha \in \mathbb{R}^{K\times A}$, for the K higher-level clusters and A levels of the covariate, $\mathbf{a} \in \{1, \ldots, A\}^N$, are independent and identically distributed,

$$\alpha_{\cdot,l} \sim \begin{cases} \delta\left(\alpha_{\cdot,l}\right), & l = 1 \\ \mathcal{N}(\mathbf{0}, \mathbf{I}), & l = 2, \ldots, L_a \end{cases} \tag{10}$$

except for the baseline level, $l = 1$, which represents the control group and for which the effect is by definition fixed to zero.

In a two-way experimental design, there is a second covariate, $\mathbf{b} \in \{1, \ldots, B\}^N$, with B distinct levels and the corresponding effects, $\beta \in \mathbb{R}^{K\times B}$, analogous to $\mathbf{a}$, A and α, respectively. Additionally, in a two-way design, there is an interaction effect, $(\alpha\beta) \in \mathbb{R}^{K\times A\times B}$, between the two covariates. Together, these three covariate effects influence the higher-level latent variable,

$$\mathbf{z}_{\cdot,i} \sim \mathcal{N}\left(\alpha_{\cdot,a_i} + \beta_{\cdot,b_i} + (\alpha\beta)_{\cdot,a_i,b_i}, \mathbf{I}\right) \tag{11}$$

additively. Again, the covariate effects are independent and identically distributed,

$$\alpha_{\cdot,c}, \beta_{\cdot,d}, (\alpha\beta)_{\cdot,c,d} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{12}$$

at all the levels, $c \in \{2, \ldots, A\}$ and $d \in \{2, \ldots, B\}$, of the covariates $\mathbf{a}$ and $\mathbf{b}$, respectively, except for the baseline levels, $c = 1$ or $d = 1$, where the covariate effects are by definition fixed to zero.

2.3 Model selection

The lower-level clustering of peaks into compounds follows the Dirichlet process (Escobar, 1994), leading to a non-parametric determination of the complexity for the lower-level latent variable.

At the higher clustering level, the model is subject to complexity selection with respect to the number of higher-level clusters, K, and the higher-level variance parameter, ψ^2. For these parameters, we make the selection jointly based on cross-validation.

The variance parameters, ψ^2 and σ^2, control the flow of information from the observed data, $\mathbf{X}$, up the hierarchy of the model toward the inferred covariate effects, α. Small values of the variance parameters allow the information to more readily propagate toward the covariate effects, enabling the detection of weaker covariate effects, while large values protect from excessive false-positive effects.

For data with a simple experimental design, the number of higher-level clusters, K, can remain low while still capturing the responses to the covariates. In a more complex experimental design with multiple covariates and their levels, the number of higher-level clusters may need to be larger to capture the richness of the association between the observed data and the experimental covariates. However, the number of higher-level clusters, K, is most crucially restricted by the availability of

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

Table 1. Design in the simulated experiments 1, 2 and 3 (columns in the table; Sections 3.1.1, 3.1.2 and 3.1.3, respectively)

Experiment	1	2	3
Number of samples ('case' + 'control'), N	10 + 10	10 + 10	10 + 10
Number of observed variables per a lower-level cluster (peaks), P	7	2	2
Number of lower-level clusters per a higher-level cluster (compounds), M	7	7	1, 3, . . . , 19
Number of higher-level clusters (groups of similarly responding compounds), K	7	1	1
Covariate effects of the higher-level clusters, $\alpha_{\cdot,2}$	[+2, −1, +0.5, 0, 0, 0, 0]	0, 0.2, . . . , 2.0	0.2
Validation range of the number higher-level clusters, K	1, . . . , 7	-	-
Validation range of the higher-level variance parameter, ψ^2	0.05, 0.1, . . . , 0.5	0.1, 0.2, . . . , 0.5	0.1, 0.2, . . . , 0.5

replicates: the required number of samples increases exponentially with the increasing complexity of the experimental design.

3 RESULTS

Next, we show that we can benefit from the inherent structure of the mass spectral data in two ways: first, we can make the inference of compound-specific covariate effects more accurate by using multiple peaks from the compound. Second, in the regime of low signal-to-noise ratio, we can further improve the accuracy by imposing stronger structure on the model, and infer the covariate effects on groups of coherently responding compounds.

We present experimental results on simulated data and metabolomic benchmark data (Franceschi *et al*., 2012) with known changes in the concentrations of chemical compounds.

In the experiments, we compare three approaches for the inference of covariate effects:

0. *Single peak*. The standard ANOVA model, where the covariate effects are computed as the average difference of the intensity of a single peak between the sample groups.

1. *1-level*. A Bayesian approach for inferring the covariate effects using data from multiple spectral peaks (Suvitaival *et al*., 2014).

2. *2-level*. The Bayesian approach proposed in this work for inferring the covariate effects through the two-level clustering of peaks and coherently responding compounds.

We evaluate each approach by its accuracy at inferring the covariate effects. The accuracy is measured in terms of the mean squared error (MSE).

3.1 Simulated data

We demonstrate the new approach through simulated experiments in regimes, which imitate real metabolomic experiments by their sample size, number of peaks associated with a compound and the general level of noise. We present three experiments where we studied three different aspects of the inference task: (i) the presence of multiple unchanged compounds, (ii) the strength of the change, and (iii) the number of coherently changing compounds. These three experiments are detailed in their respective subsections that follow next. The experiments are summarized in Table 1.

Table 2. The two-level model was more accurate at small effect sizes of the covariate on simulated data

True covariate effect	RMSE			Corrected P-value of difference of 2-level	
	Single	1-level	2-level	to Single	to 1-level
0	1.16	0.53	**0.51**	$1.1 \times 10^{-190**}$	$2.9 \times 10^{-2*}$
+0.5	1.42	0.68	**0.63**	$3.9 \times 10^{-44**}$	$4.0 \times 10^{-3**}$
–1.0	1.03	**0.56**	0.58	$1.7 \times 10^{-27**}$	4.4×10^{-1}
+2.0	1.22	**0.90**	1.13	$5.0 \times 10^{-3**}$	$2.6 \times 10^{-24**}$

Note: The two-level and one-level models, and the single-peak approach ('2-level', '1-level' and 'Single', respectively), were compared by their MSE between the inferred and the true covariate effect. The smallest MSE for each true effect is highlighted in bold. The significance of the difference between the two-level model and the two comparison approaches was tested with the two-sided paired t-test with the Benjamini–Hochberg control (Benjamini and Hochberg, 1995) for the false discovery rate. The result is from the first simulated experiment (Section 3.1.1). */** Significant difference at confidence level 95/99%.

3.1.1 Presence of multiple unchanged compounds In the first simulated experiment, we studied the simultaneous inference of multiple zero and non-zero covariate effects. We generated data with seven similarly responding clusters of compounds, each compound producing seven peaks.

When comparing the inferred covariate effects with the true effects, the two peak-clustering models always had a lower error compared with the single-peak approach (Table 2). Most importantly, the added structure of the two-level model prevented the model from overfitting to the noisy data and improved the accuracy at small and diminishing covariate effects, leading to a decrease in false discoveries.

The number of the higher-level clusters, K, and the variance parameter, ψ^2, were selected jointly via stratified nested 5-fold cross-validation. The procedure was repeated with five independent datasets.

3.1.2 Strength of the change In the second simulated experiment, we studied how the signal-to-noise ratio of the true covariate effect influences the accuracy of inference. We generated

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

independent datasets with seven compounds and progressively increased the generated covariate effect.

The covariate effect was inferred the most accurately by the two-level model throughout the experiment (Fig. 2a). The accuracy of the model-based approaches decreased slightly when the strength of the true effect increased. This followed from the prior assumption that prevents the model from overfitting to unexpectedly strong covariate effects. The prior for the covariate effects places most of the probability mass around zero and thus effectively sets a bias for the inferred effects toward zero (Fig. 2b). The data-based single-peak approach does not have this bias, but its confidence intervals were considerably wider, which lead to a larger error in the inference task. The effects in any real metabolomic dataset most probably are in the weak regime, where the bias is overshadowed by the noise in the data.

3.1.3 Number of coherently changing compounds In the third simulated experiment, we studied how the number of coherently responding compounds influences the accuracy of the two-level model. We generated data with a weak covariate effect and gradually increased the number of compounds.

We discovered that the accuracy of the two-level model increased as the number of coherently responding compounds increased (Fig. 3). The experiment empirically confirmed the expected connection between the two Bayesian models: when there was only one responding compound, the two-level model effectively reduced to the one-level model in terms of the error. As expected, the performance of the single-peak approach and the one-level model remained constant throughout the experiment.

3.2 Benchmark data with known changes in concentrations

Next, we applied the method on real ultra performance liquid chromatography-mass spectrometry data (Franceschi *et al.*, 2012). The recently published benchmark dataset of apple samples includes a set of annotated spike-in compounds with a known increase in the concentration. The samples have been measured in both the positive and negative ion modes.

We started with the raw spectral data [The raw spike-in dataset by Franceschi *et al.* (2012) is available online at http://cri.fmach.eu/Research/Computational-Biology/Biostatistics-and-Data-Management/download/data/Spiked-Apple-Data (June 11, 2013, date last accessed)] to acquire the shapes of the peaks in addition to their heights. The data were preprocessed using MZmine 2 (Pluskal *et al.*, 2010) with default settings.

We evaluated the approaches by the MSE between the inferred and the true covariate effects. If a cluster contained multiple annotated peaks, the error was computed for each of the annotated peaks. Clusters with no annotated peaks were assumed to have a 0% true effect. The effect of the single-peak approach was computed as the average change of the strongest peak of the cluster.

The number of higher-level clusters and the variance parameter were selected jointly through a stratified nested 5-fold cross-validation from the sets $K \in \{1, 2, \ldots, 10\}$ and $\psi^2 \in \{0.25, 0.5, \ldots, 1.5\}$, respectively. Model selection and validation was done independently for the positive and negative ion mode datasets.

The analyses for the data from both the ion modes lead to an outcome, where the Bayesian models were more accurate at

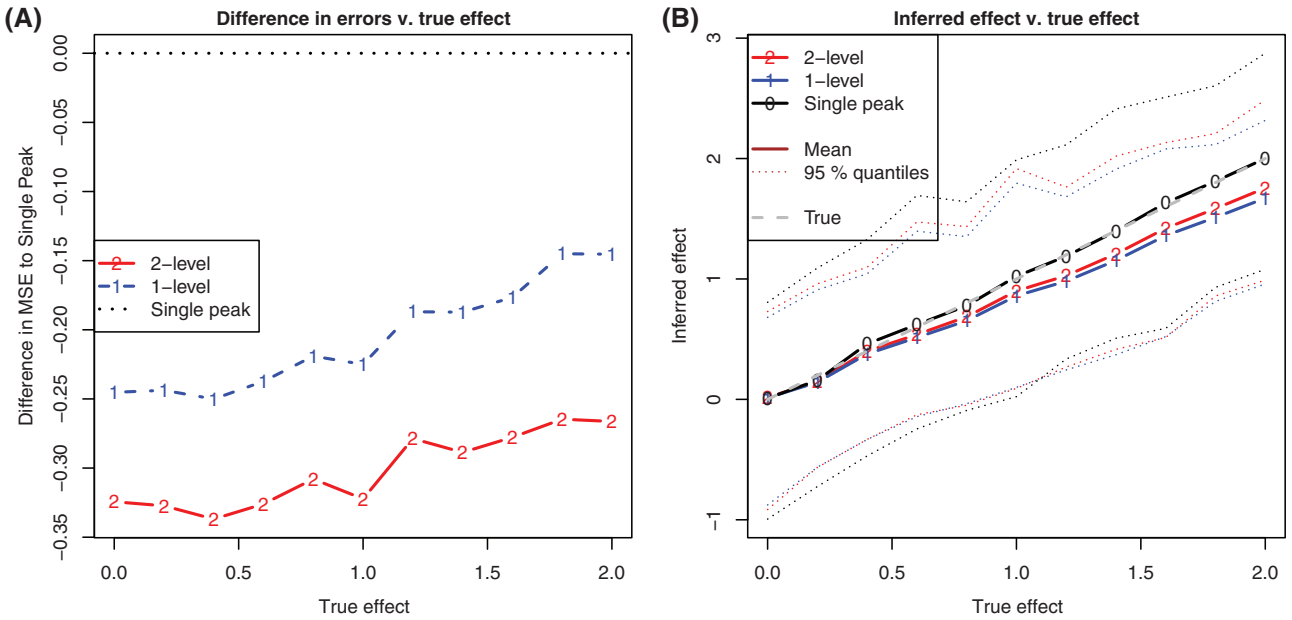


Fig. 2. The peak-clustering and the compound-clustering models ('1-level' and '2-level') reduced uncertainty around the covariate effect compared with the single-peak approach. The hierarchical models have a bias toward zero, which follows from the model assumption incorporated to the prior of the covariate effect. The prior-induced bias lead to a slight increase in the error of the peak-clustering models as the true effect increased but acted to prevent the models from overfitting and thus from false findings at normal effect sizes. (**a**) Pairwise difference in the error between the single-peak approach and each of the two clustering models shown as a function of the magnitude of the true effect. (**b**) Inferred effect as a function of the magnitude of the true effect. Result from the second simulated experiment (Section 3.1.2), where the true covariate effect was varied from 0 to 2

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

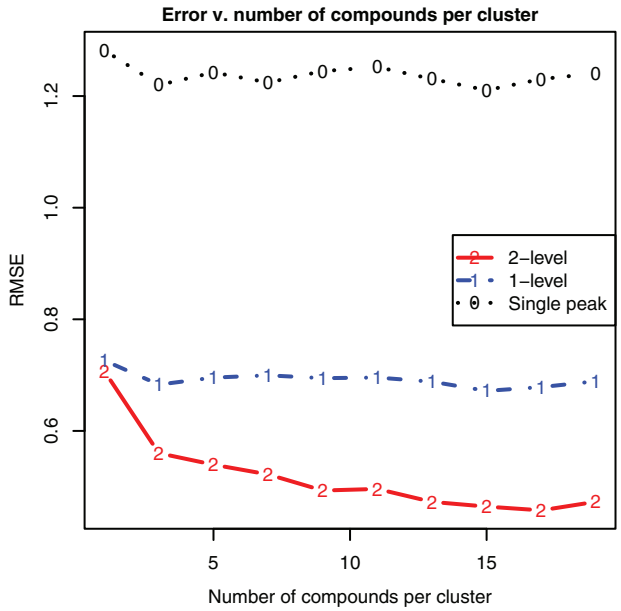


Fig. 3. The error in the covariate effect inferred by the compound-clustering model ('2-level') decreased when more coherently responding compounds were observed. The accuracy of the peak-clustering model and the data-based single-peak approach ('1-level' and 'Single peak', respectively) remained constant. Root mean squared error (RMSE) is shown as a function of the number of compounds (i.e. lower-level clusters) per higher-level cluster. Result from the third simulated experiment (Section 3.1.3), where a weak covariate effect of 0.2 was generated and the number of compounds was gradually increased from 1 to 19

inferring the covariate effects than the single-peak approach by a significant margin (Table 3). The two-level model improved the accuracy compared with the one-level model at weak effect sizes. It is worth noting that the results were strongly in line with the simulated experiments (Section 3.1.1).

4 CONCLUSION

Additional spectral peaks produced by the mass spectrometer as a result of the ionization process have been shown to be useful for the inference of covariate effects when multiple peaks can be confidently associated with one compound. However, even with multiple peaks supporting the inference, small covariate effects may be hidden under the between-sample variation. We addressed this problem by introducing stronger structure to the model of the covariate effects, thereby regularising the covariate effects and making them less dependent on the variation of individual compounds.

We achieved an improvement in the accuracy of the inferred covariate effects by assuming a structure of coherently responding compounds in the data. We proposed a structured model for inferring covariate effects for groups of compounds through two layers of probabilistic clustering. Metabolomic data are known to have collinear structure for similar compounds. This phenomenon is argued to arise from the biological processes that the compounds are involved in. However, the method proposed in this work does not restrict the groups of compounds by their

Table 3. The two-level model is most accurate at small levels of covariate effects and both the Bayesian models are more accurate than the single-peak approach on the metabolomic benchmark data (Section 3.2)

True covariate effect (%)	RMSE			Corrected *P*-value of difference of 2-level	
	Single	1-level	2-level	to Single	to 1-level
(a) Positive ion mode					
+0	0.42	0.31	**0.09**	$<\varepsilon^{**}$	$<\varepsilon^{**}$
+20	0.41	0.22	**0.19**	$3.4 \times 10^{-34**}$	$2.1 \times 10^{-13**}$
+40	0.44	**0.28**	0.33	$3.8 \times 10^{-11**}$	$2.7 \times 10^{-4**}$
+100	1.06	**0.91**	0.92	$1.3 \times 10^{-2*}$	$1.3 \times 10^{-5**}$
(b) Negative ion mode					
+0	0.42	0.31	**0.11**	$<\varepsilon^{**}$	$<\varepsilon^{**}$
+20	0.54	0.26	**0.20**	$2.5 \times 10^{-47**}$	$4.4 \times 10^{-43**}$
+40	0.45	**0.34**	0.35	$6.1 \times 10^{-30**}$	$1.1 \times 10^{-8**}$
+100	0.82	**0.74**	0.88	2.5×10^{-1}	$4.2 \times 10^{-83**}$

Note: The two-level and one-level models, and the single-peak approach ('2-level', '1-level' and 'Single', respectively), are compared by their MSE between the inferred and the true covariate effect. The smallest MSE for each true effect is highlighted in bold. The significance of the difference between the two-level model and the two comparison approaches is tested with the two-sided paired *t*-test with the Benjamini–Hochberg correction (Benjamini and Hochberg, 1995) for the *P*-values. A near-zero value below the machine accuracy is denoted by 'ε'. */** Significant difference at confidence level 95/99%.

chemical or biological similarity but infers the groups only based on their responses to the covariates.

In the experiments, we showed that the two-level model proposed in this work decreases the error of inferred covariate effects in a typical setting, where the true effects are small or diminishing. Through three simulated experiments, testing the approaches with multiple zero-effect clusters, varying effect size and varying number of similarly responding compounds, we demonstrated that the two-level model is more accurate at inferring weak covariate effects from noisy multipeak data when compared with the two comparison approaches. The outcome was repeated on a metabolomic benchmark dataset with known changes in the compound concentrations. Following the reduction in the error for the weak covariate effects, the two-level model is argued to reduce false findings.

To further improve the consistency of the inferred covariate effects, we suggest the following avenues of research: (i) prior knowledge about the similarity of the compounds can be incorporated into the prior of the higher-level clustering, either in terms of the biological processes in which the compounds are involved or in terms of the chemical similarity of the compounds. (ii) The lower-level clustering can be improved to detect even the weakest peaks by incorporating prior knowledge about the relative positions of the peaks associated with one compound. This is possible thanks to the fact that the expected positions of many adduct and isotope peaks can be calculated based on the ionization process and the chemical formula of the compound, respectively. (iii) The covariate effects in the isotope peaks are argued to be highly preserved because the isotope peaks do not arise from variation in the ionization process. Additionally, the expected relative heights of these peaks can be calculated if the identity

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

of the compound is known. Incorporating these properties into the model may be even more useful for the inference of covariate effects than the two aforementioned points.

Funding: Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170; Computational Modeling of the Biological Effects of Chemicals, 140057). Computational resources were provided by the Aalto University School of Science 'Science-IT' project.

Conflicts of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.*, **57**, 289–300.

Broadhurst,D.I. and Kell,D.B. (2006) Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, **2**, 171–196.

Dahl,D.B. (2006) Model-based clustering for expression data via a Dirichlet process mixture model. In: Do,K.A. and Müller,P. (eds) *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, Cambridge, pp. 201–218.

De Iorio,M. *et al.* (2004) An ANOVA model for dependent random measures. *J. Am. Stat. Assoc.*, **99**, 205–215.

Escobar,M.D. (1994) Estimating normal means with a Dirichlet process prior. *J. Am. Stat. Assoc.*, **89**, 268–277.

Franceschi,P. *et al.* (2012) A benchmark spike-in data set for biomarker identification in metabolomics. *J. Chemom.*, **26**, 16–24.

Huopaniemi,I. *et al.* (2009) Two-way analysis of high-dimensional collinear data. *Data Min. Knowl. Discov.*, **19**, 261–276.

Jansen,J.J. *et al.* (2012) Individual differences in metabolomics: individualised responses and between-metabolite relationships. *Metabolomics*, **8**, 94–104.

Katajamaa,M. and Orešič,M. (2007) Data processing for mass spectrometry-based metabolomics. *J. Chromatogr. A*, **1158**, 318–328.

Kau,A.L. *et al.* (2011) Human nutrition, the gut microbiome and the immune system. *Nature*, **474**, 327–336.

Kaufman,C.G. and Sain,S.R. (2010) Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Anal.*, **5**, 123–149.

Kuhl,C. *et al.* (2012) CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.*, **84**, 283–289.

Mitchell,T.J. and Beauchamp,J.J. (1988) Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.*, **83**, 1023–1032.

Orešič,M. (2009) Metabolomics, a novel tool for studies of nutrition, metabolism and lipid dysfunction. *Nutr. Metab. Cardiovasc. Dis.*, **19**, 816–824.

Plumb,R. *et al.* (2004) Ultra-performance liquid chromatography coupled to quadrupole-orthogonal time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, **18**, 2331–2337.

Pluskal,T. *et al.* (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, **11**, 395.

Rogers,S. *et al.* (2009) Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, **25**, 512–518.

Saccenti,E. *et al.* (2014) Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, **10**, 361–374.

Smilde,A.K. *et al.* (2005) ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, **21**, 3043–3048.

Steuer,R. (2006) Review: on the analysis and interpretation of correlations in metabolomic data. *Brief. Bioinform.*, **7**, 151–158.

Suvitaival,T. *et al.* (2014) Stronger findings from mass spectral data through multipeak modeling. *BMC Bioinformatics*, **15**, 208.

Tikunov,Y. *et al.* (2012) MSClust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. *Metabolomics*, **8**, 714–718.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014