BIOINFORMATICS ORIGINAL PAPER Vol. 30 no. 10 2014, pages 1370–1376
doi:10.1093/bioinformatics/btu064

Genome analysis Advance Access publication January 30, 2014

Bayesian joint analysis of heterogeneous genomics data

Priyadip Ray[1,†], Lingling Zheng[2,*,†], Joseph Lucas[3,‡] and Lawrence Carin[4,‡]

[1]G.S.Sanyal School of Telecommunications, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India, [2]Computational Biology & Bioinformatics, Duke University, Durham, NC 27705, USA, [3]Quintiles, Durham, NC 27703, USA and [4]Electrical & Computer Engineering Department, Duke University, Durham, NC 27705, USA

Associate Editor: John Hancock

ABSTRACT

Summary: A non-parametric Bayesian factor model is proposed for joint analysis of multi-platform genomics data. The approach is based on factorizing the latent space (feature space) into a shared component and a data-specific component with the dimensionality of these components (spaces) inferred via a beta-Bernoulli process. The proposed approach is demonstrated by jointly analyzing gene expression/copy number variations and gene expression/methylation data for ovarian cancer patients, showing that the proposed model can potentially uncover key drivers related to cancer.

Availability and implementation: The source code for this model is written in MATLAB and has been made publicly available at https://sites.google.com/site/jointgenomics/

Contact: catherine.ll.zheng@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 6, 2013; revised on January 9, 2014; accepted on January 26, 2014

1 INTRODUCTION

Gene expression profiles are commonly used in the study of cancer biology (Garber *et al.*, 2001). However, human cancers are highly heterogeneous because of the combined effects of genetic instabilities and epigenetic alterations. Hence, it is notoriously difficult to gain understanding of the mechanisms of cancer progression from gene expression data alone. The root cause of cancer is due to several possible events. For instance, copy number variations (CNVs) under selective pressure change the dosage of key tumor-inducing and tumor-suppressing genes, which thereby affect mRNA transcription and neoplastic cell proliferation (Tang *et al.*, 2012). On the other hand, the mechanism of epigenetic alteration is more complicated. DNA methylation patterns are globally disrupted in tumor cells, characterized by both global hypomethylation and region-specific hypermethylation at CpG islands (Robertson, 2001).

There is a recent explosion in the availability of cancer datasets of multiple heterogeneous data types from sources like the Cancer Genome Atlas project (TCGA; cancergenome.nih.gov). It is hoped that the availability of simultaneous measurements of multiple biological features such as gene expression, DNA copy number change, DNA methylation and others will offer an increased insight into the root causes of cancer phenotypes. These insights have the potential to distinguish genetic/epigenetic changes that promote cancer progression (driver mutation) from those with no selective advantage (passenger mutation) (Akavia *et al.*, 2010), which could revolutionize the treatment of cancer by suggesting novel targets for cancer therapeutics and by developing biomarkers to identify the patients who will most benefit from such therapeutics (Li *et al.*, 2009).

There are numerous publications on combining different types of DNA modifications with gene expression. Perhaps the most natural of these are the brute force methods such as expression quantitative trait loci (eQTL) analysis (Kendziorski *et al.*, 2006). Joint analysis of single nucleotide polymorphism (SNP) data with gene expression data by expression quantitative trait loci involves testing every gene-single nucleotide polymorphism pair for association with a *t*-test, and then correcting for multiple hypothesis testing. CNAmet (Louhimo and Hautaniemi, 2011) defines a similar approach to relate gene expression changes with either copy number change or DNA methylation. Other approaches use well-established models for each of the individual data types, and then combine the results into a statistic that addresses the problem of interest. The approach of Jeong *et al.* (2010) is an example of this for the identification of genes that are regulated by DNA methylation. A shortcoming of all of these approaches is that they do not reduce the dimension of the individual datasets through an accounting of their respective correlation structures.

In Lanckriet *et al.* (2004), the authors used kernel functions predefined for each data type, and mapped to the same vector space, which allows joint analysis in the common range of the kernels. Copy number and expression in cancer (Akavia *et al.*, 2010) has been proposed as a Bayesian scoring function that measures how well a set of candidate gene regulators correlates with the expression of gene modules (groups of genes that are correlated with each other). Another approach (Lucas *et al.*, 2010) uses a sparse factor (SF) model to describe the correlation structure of the gene expression data but uses *post hoc* hypothesis tests to draw connections between gene expression and copy number data. These approaches allow for effective dimension reduction but do not use correlation structure in one dataset to inform the estimations of correlation in the others. Zheng and Lucas (2012) extended this approach by the development of a two-SF model using aneuploidy and gene expression as an example. It assumes some factors are shared by two data where

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
‡The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/30/10/1370/267809 by guest on 18 November 2020

there is a statistical correlation, and some factors are unique to each dataset. The common factors are further linked by sampling gene expression factors centered on the CNV factors with some Gaussian noise, thereby preventing the difference in data size from overwhelming the information available on associations between them. However, it is difficult to check model assumptions in high-dimension by specifying different numbers of factors.

In the recent years, model-based integration approach has become popular for analyzing TCGA data. Xu *et al*. (2012) proposed a three-node Bayesian graphical model to represent the dependence structure of CNVs, methylation and gene expression. This method identifies the mRNA regulatory event separately for each candidate gene but does not consider gene modules as a whole. iBAG (Jennings *et al*., 2012; Wang *et al*., 2013) specifies gene-methylation effect and other regulatory module controlled effect and then uses a two-component regression model to identify gene/biomarkers that are associated with clinical outcomes (e.g. survival time, tumor stage and other demographic variables). In this way, it assumes linear and independent effects between gene expression and methylation as well as patient survival and gene expression. However, given the complexity of epigenetic regulatory mechanisms, more flexible Bayesian non-parametric models may be preferred.

For jointly modeling heterogeneous data, Bayesian and semi-Bayesian latent variable models have been developed to factorize the latent space into a shared and data-specific part (Archambeau and Bach, 2008; Klami and Kaski, 2008). However, in these approaches the number of latent factors is chosen a priori. Alternatively, one may consider multiple factor models, each with a different number of factors, and perform model selection based on information criteria such as Akaike information criterion (AIC) (Akaike, 1987) or Bayesian information criterion (BIC) (Schwarz, 1978). However, as stated earlier, it is often challenging to check modeling assumptions in high-dimensions. Hence, a non-parametric or semiparametric model is desirable.

In this article, we propose a non-parametric Bayesian factor analysis approach for integrating heterogeneous genomics data, with the number of factors inferred from the available data. Our proposed approach is based on factoring the latent space into shared and data-specific components, using a beta-Bernoulli process (Griffiths and Ghahramani, 2005; Paisley and Carin, 2009; Thibaux and Jordan, 2007) to infer the dimension of these latent spaces. We demonstrate the proposed approach on the joint analysis of genomics data for ovarian cancer patients, including three different data types: gene expression, CNV and DNA methylation data. We demonstrate that the joint analysis of gene expression/CNV and gene expression/DNA methylation levels can potentially identify genomic and epigenomic regulators influencing cancer pathophysiology outcomes.

The remainder of the article is organized as follows. In Section 2, we present the proposed hierarchical Bayesian model for jointly analyzing heterogeneous genomics data. In Section 3, we discuss priors on factor loadings and factor scores. Section 4 outlines an MCMC inference algorithm. Section 5 provides experimental results for the proposed model on the analysis of multiple genomics data. Finally, we provide concluding remarks in Section 6.

2 JOINT BAYESIAN FACTOR ANALYSIS

Let $\{\boldsymbol{X}^{(r)}\}_{r=1,R}$ represent data from R different modalities, where $\boldsymbol{X}^{(r)} = (\boldsymbol{x}_1^{(r)}, \ldots, \boldsymbol{x}_M^{(r)}) \in \mathbb{R}^{N_r \times M}$. Here N_r represents the dimensionality of the observation vectors of modality r (e.g. the total number of genes) and M represents the total number of samples, e.g. patients. In SF modeling, learning a single shared matrix of factor loadings for different signal classes has been proposed in Mairal *et al*. (2008). However, for heterogeneous data such as that considered here, learning a shared set of factor loadings is more difficult.

The joint factor model may be represented as

$$\boldsymbol{X}^{(r)} = \boldsymbol{D}^{(r)}\left(\boldsymbol{W}^{(c)} + \boldsymbol{W}^{(r)}\right) + \boldsymbol{E}^{(r)} \tag{1}$$

The matrix $\boldsymbol{D}^{(r)} = (\boldsymbol{d}_1^{(r)}, \ldots, \boldsymbol{d}_K^{(r)}) \in \mathbb{R}^{N_r \times K}$ consists of the factor loadings specific to data modality r, factor scores $\boldsymbol{W}^{(r)} = (\boldsymbol{w}_1^{(r)}, \ldots, \boldsymbol{w}_M^{(r)}) \in \mathbb{R}^{K \times M}$ are specific to data from modality r, $\boldsymbol{W}^{(c)} = (\boldsymbol{w}_1^{(c)}, \ldots, \boldsymbol{w}_M^{(c)}) \in \mathbb{R}^{K \times M}$ consists of the factor scores common among all modalities and $\boldsymbol{E}^{(r)} = (\epsilon_1^{(r)}, \ldots, \epsilon_M^{(r)}) \in \mathbb{R}^{N_r \times M}$ consists of the noise/residual specific to data of modality r.

Note that one may alternatively consider

$$\boldsymbol{X}^{(r)} = \boldsymbol{D}^{(rc)}\boldsymbol{W}^{(c)} + \boldsymbol{D}^{(r)}\boldsymbol{W}^{(r)} + \boldsymbol{E}^{(r)} \tag{2}$$

where $\boldsymbol{D}^{(rc)}$ corresponds to factor loadings associated with common factors, with $\boldsymbol{D}^{(rc)}$ reflective of how these common factors are viewed by modality r; $\boldsymbol{D}^{(r)}$ are factor loadings associated with factors specific only to modality r. The framework in (1), which we use throughout, allows the ability to share factor loadings between the common and modality-specific factors, and the manner in which $\boldsymbol{W}^{(c)}$ and $\boldsymbol{W}^{(r)}$ are modeled allows sufficient flexibility to yield (2) if the data so warrant.

We wish to impose the condition that any $\boldsymbol{x}_i^{(r)}$ is a sparse linear combination of the factor loadings. Hence, the factor scores are represented as,

$$\boldsymbol{w}_i^{(r)} = \boldsymbol{s}_i^{(r)} \odot \boldsymbol{b}_i^{(r)} \quad \text{and} \quad \boldsymbol{w}_i^{(c)} = \boldsymbol{s}_i^{(c)} \odot \boldsymbol{b}_i^{(c)} \tag{3}$$

where $\boldsymbol{s}_i^{(r)} \in \mathbb{R}^K$, $\boldsymbol{s}_i^{(c)} \in \mathbb{R}^K$, $\boldsymbol{b}_i^{(r)} \in \{0,1\}^K$, $\boldsymbol{b}_i^{(c)} \in \{0,1\}^K$ and $\odot$ represents the Hadamard product (element-wise vector product). The sparse binary vectors $\boldsymbol{b}_i^{(r)}$ are drawn from the following beta-Bernoulli process (Griffiths and Ghahramani, 2005; Paisley and Carin, 2009; Thibaux and Jordan, 2007)

$$\boldsymbol{b}_i^{(r)} \sim \prod_{k=1}^{K} \text{Bernoulli}(\pi_k),\ \pi \sim \prod_{k=1}^{K} \text{Beta}(c\alpha, c(1-\alpha)) \tag{4}$$

with π_k representing the k^{th} component of π and $\alpha \in (0,1)$. In practice K is finite, and the above equation represents a finite approximation to the beta-Bernoulli process, where the number of non-zero components of each $\boldsymbol{b}_i^{(r)}$ is a random variable drawn from Binomial(K, α). If α is set to $\frac{\rho}{K}$, in the limit $K \to \infty$ this reduces to the number of non-zero components in $\boldsymbol{b}_i^{(r)}$ being drawn from Poisson(ρ); this corresponds to the Indian buffet process (IBP) (Griffiths and Ghahramani, 2005; Paisley and Carin, 2009; Thibaux and Jordan, 2007). Therefore, we may explicitly impose a prior belief on the number of non-zero components in $\boldsymbol{w}_i^{(r)}$. The shared binary vectors $\boldsymbol{b}_i^{(c)}$ are modeled similarly as $\boldsymbol{b}_i^{(r)}$.

Downloaded from https://academic.oup.com/bioinformatics/article/30/10/1370/267809 by guest on 18 November 2020

The properties of the beta process induce sparsity in the feature space by encouraging sharing of features. An example of a binary feature matrix drawn from a beta-Bernoulli process is shown in Figure 1. If K is set to a relatively large value, the finite beta-Bernoulli process allows inference of the number of factors needed to represent the data. Conceptually, in a finite beta-Bernoulli process, we start with a large number of parameters (instead of an infinite number of parameters, as done in an IBP), and then allow the data to infer the appropriate (low-dimensional) model order during posterior inference. In principle, one might desire to actually treat the model as infinite within inference; however, Paisley and Carin (2009) show that the finite approximation of the beta-Bernoulli process, for large K, yields a very accurate approximation to the IBP, and hence we anticipate little, if any, change in results for an infinite model from those obtained in this article.

The noise or residual in (1) is modeled as

$$\epsilon_i^{(r)} \sim \mathcal{N}(0, \gamma_\epsilon^{(r)-1} \boldsymbol{I}_{N_r}), \quad \gamma_\epsilon^{(r)} \sim \text{Gamma}(a_0, b_0) \tag{5}$$

where $\boldsymbol{I}_{N_r}$ represents the $N_r \times N_r$ identity matrix.

The construction in (1) imposes the belief that there are underlying (low-dimensional) features represented by the factor scores that may be shared across modalities, via $\boldsymbol{W}^{(c)}$; however, each modality has a unique mapping from these low-dimensional factor scores to the high-dimensional data, reflected by $\boldsymbol{D}^{(r)}$. Further, each modality may also have idiosyncratic low-dimensional features, characterized by $\boldsymbol{W}^{(r)}$. The common and idiosyncratic features are learned jointly, via the simultaneous analysis of all modalities. A unique feature of the above construction is that it allows complete sharing of some low-dimensional features across different data modalities as well as partial sharing, i.e. a shared feature may be slightly perturbed via $\boldsymbol{W}^{(r)}$ and shared across different modalities.

3 MODEL CONSTRUCTION

3.1 Priors for factor loadings and factor scores

In the absence of covariates, the factor loadings may be drawn i.i.d. (independent and identically distributed) from a Gaussian distribution (for ease of notation, we henceforth drop the modality index r, unless referring to multiple data modalities simultaneously),

$$\boldsymbol{d}_k \sim \mathcal{N}(0, \gamma_s^{-1} \boldsymbol{I}_N), \quad \gamma_s \sim \text{Gamma}(a_5, b_5) \tag{6}$$

One may impose covariate-dependent factor loadings [e.g. factor loadings drawn from a Gaussian process or a mixture of Gaussian processes (Gramacy and Lee, 2007; Meeds and Osindero, 2006; Rasmussen and Ghahramani, 2002; Ray and Carin, 2011; Tresp, 2001)] or an additional Markov structure (Fox *et al.*, 2008) on the factor loadings to impose the correlation. However, for the genomics data considered in this article, it was deemed unnecessary to impose such structures.

The factor scores may be drawn i.i.d. (independent and identically distributed) from a Gaussian distribution as,

$$\boldsymbol{s}_i^{(r)} \sim \mathcal{N}(0, \gamma_r^{-1} \boldsymbol{I}_K) \qquad \boldsymbol{s}_i^{(c)} \sim \mathcal{N}(0, \gamma_c^{-1} \boldsymbol{I}_K) \tag{7}$$

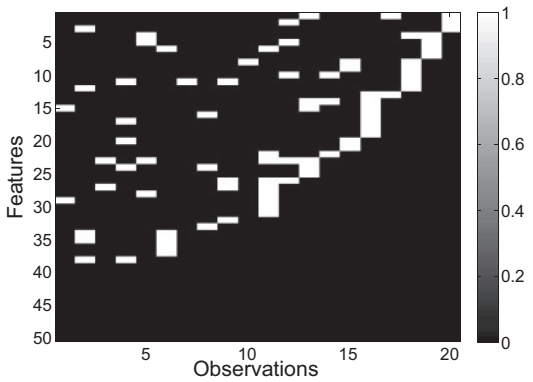

Fig. 1. A feature selection matrix drawn from a beta-Bernoulli prior with $\alpha = 0.08$ and $c = 100$. For ease of illustration, the rows of the binary matrix have been reordered from top to bottom by the magnitude of the binary number expressed by that row, taking the rightmost column as the most significant bit. It can be observed that the prior favors sharing of features, with some dominant features shared by many observations and a few features that are possessed by only a few observations

We impose broad gamma prior on γ_r and γ_c: $\gamma_r \sim \text{Gamma}(a_2, b_2)$ and $\gamma_c \sim \text{Gamma}(a_3, b_3)$.

3.2 Imposing sparsity

In many biological applications, such as those considered in this article, it is desirable that the factor loading matrix is sparse (Carvalho *et al.*, 2008). To impose sparsity on the factor loadings, we use a Student-t sparseness-promoting prior (Tipping, 2001). In this construction, d_{jk}, the j$^{\text{th}}$ component of $\boldsymbol{d}_k$, is drawn

$$d_{jk} \sim \mathcal{N}(0, \tau_{jk}^{-1}) \tag{8}$$

$$\tau_{jk} \sim \text{Gamma}(a_1, b_1) \tag{9}$$

However, there are multiple ways by which one may desire to impose sparsity, such as using the spike-slab prior (Carvalho *et al.*, 2008; Chen *et al.*, 2011; Ishwaran and Rao, 2005). This consists of a discrete-continuous mixture of a point mass at zero, referred to as the 'spike', and any other distribution, such as the Gaussian distribution, known as the 'slab'. A hierarchical beta-Bernoulli construction of the spike-slab prior for imposing sparsity on the factor loadings is provided in Chen *et al.* (2011). We found that the spike-slab prior works as well as the model presented above; however, for the sake of brevity, we include only the results for the Student-t sparseness prior in this article. Note that for modeling the heterogeneous genomic data considered in this article, we have this imposed sparsity both on the factor loadings (via a student-t sparseness-promoting prior) and on the factor scores (via a beta-Bernoulli process).

4 MCMC INFERENCE

The conditional posterior distribution of all the model parameters for the joint factor model may be derived analytically. We use a Gibbs sampler to draw samples from the posterior distribution of the model parameters. The number of Gibbs burn-in samples is set to 3000 and the number of collection samples is set

Downloaded from https://academic.oup.com/bioinformatics/article/30/10/1370/267809 by guest on 18 November 2020

to 1000. Broad gamma hyper priors are chosen for the variance terms with $a_0 = b_0 = a_2 = b_2 = a_3 = b_3 = 10^{-5}$. The results are relatively insensitive to these settings, and various other settings such as $a_0 = b_0 = a_2 = b_2 = a_3 = b_3 = 10^{-3}$ or $a_0 = b_0 = a_2 = b_2 = a_3 = b_3 = 10^{-6}$ yielded very similar results. The shrinkage parameters on the factor loadings are set at $a_1 = 10^{-3}$ and $b_1 = 10^{-6}$ (for gene-copy number analysis) and $a_1 = 1$ and $b_1 = 10^{-2}$ (for gene-methylation analysis). The shape parameters on the beta-Bernoulli process are set as follows: (i) for gene-copy number analysis, the common and data-specific factors have the same hyperparameters, $c = 400,\ \alpha = 2.5\text{e} - 9$; (ii) for gene-methylation analysis, the hyper priors on the common factors are $c = 700,\ \alpha = 1.5\text{e} - 4$, on the gene-expression-specific factors are $c = 500,\ \alpha = 2\text{e} - 4$ and on the methylation-specific factors are $c = 10,\ \alpha = 0.01$.

The mixing of the MCMC sampler was also carefully examined. The sampler was run extensively for different number of burn-in and collection samples. It was also run multiple times in parallel with different initial values. The results of these experiments were found to be consistent and repeatable across such runs. Details of the Gibbs update equations are provided in the Supplementary Material.

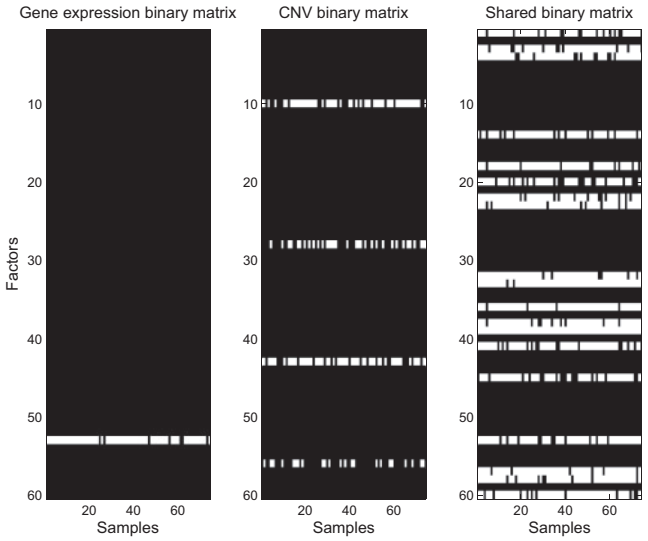

Fig. 2. The inferred feature selection matrices that are unique to data of modality r ($\boldsymbol{B}^{(r)}$) and common to both data modalities ($\boldsymbol{B}^{(c)}$). From left to right, the figures are binary matrices unique to gene expressions, CNVs and shared between gene expressions and CNVs, respectively. The y-axis is the indicator of each factor, and x-axis represents the 74 subjects. The inferred factors and samples selected by the model are assigned as 1 (white) and 0 (black) otherwise. Results are shown for the maximum-likelihood collection sample (for illustration purposes)

5 JOINT ANALYSIS OF HETEROGENEOUS GENOMICS DATA

5.1 Data description

The data in this study include ovarian cancer gene expression, CNVs and methylation data collected from TCGA project. We aim to integrate gene expression/CNVs and gene expression/methylation from 74 ovarian cancer patients. For computing purposes, we downsized the original massive data into smaller sets. Independent gene-by-gene filtering (based on criteria such as overall mean and overall variance) is typically used to reduce data dimension as well as increase the number of discoveries in high-throughput experiments (Bourgon *et al.*, 2010; Gentleman *et al.*, 2005; Talloen *et al.*, 2007). In our analysis, a filtering criterion was established for the gene expression data (Affymetrix HT_HGU133A) to eliminate probes with sample mean <6 or SD <0.4, which resulted in a gene expression dataset downsized from 22 277 to 5976. Comparative genomic hybridization (CGH) data were filtered to remove Agilent Human Genome CGH 244A probes containing missing values. This set was further filtered by keeping only one in 50 probes, leaving 4443 probes. Methylation data (Illumina Infinium human methylation 27K bead assay) was filtered to retain only higher variance samples (resulting in 4722 probes) and was inverse-probit transformed to lie on the real line.

5.2 Analysis of gene expression and CNVs data

We applied the joint Bayesian factor model to gene expression and CGH to identify the factors that are representative of correlated changes in gene expression and DNA CNVs. We set the upper bound on the number of factors as $K = 60$ and obtained 1 specific to gene expression, 4 unique to CNVs and 19 shared between both modalities (Fig. 2). The results are relatively insensitive to the choice of K, and various other choices yielded similar results. Figure 3 shows the correlation structure of the probe sets (gene expression) and CGH clones (CNVs) that are included in joint factor number 41 (the factor numbering is arbitrary, and changes between collection samples, with these results illustrative; in these and related results we depict the maximum likelihood collection sample). As expected, correlation between the factor genes for those patients who were included in this factor is higher than for those not included.

It is well known that some variations in cancer gene expression are caused by gene dosage changes because of CNVs. In addition, because of the mechanism by which CNV occurs, it tends to happen in contiguous regions. Of the 19 CNV factors identified, one is a nearly perfect representation of batch effects in the data and the remaining 18 display copy number amplification/deletion in specific chromosomal regions. Most of these show similar gene expression changes in the same region. We demonstrate this behavior in Figure 4, which shows that the largest factor loadings from both CNV and gene expression for factor 18 are clustered around the same region of chromosome 8.

We identified highly associated CNVs in the chromosomal arm 8q12.3–8q24.13 (factor 18), which is a known region for frequent high-level amplification associated with disease progression in human cancers (Frank *et al.*, 2007; Pils *et al.*, 2005). The rediscovery of genes in this region also validates our approach. For example, E2F5 (8q21.2, Unique ID: 1875), an important gene in the regulation of cell cycle, is known to be overexpressed in ovarian epithelial cancer (Kothandaraman *et al.*, 2010). Overexpressed genes, MTDH (8q22.1, Unique ID: 92140) and EBAG9 (8q23, Unique ID: 9166), have been recognized in a variety of cancers including ovarian and breast cancers

Downloaded from https://academic.oup.com/bioinformatics/article/30/10/1370/267809 by guest on 18 November 2020

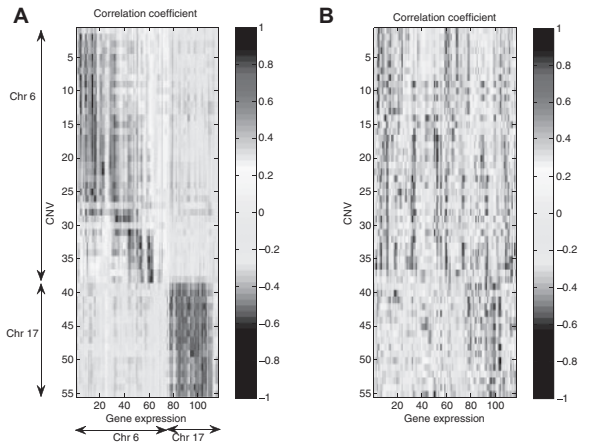

Fig. 3. Correlation structure between gene expression and CNVs of top loaded genes from factor 41. The figure displays correlation coefficients between the two data. Panels (**A**) and (**B**) show the correlation results from patients selected and dropped out by the model, respectively. It is observed that CNVs from chromosome 6 and genes from chromosome 17 have a reverse correlation pattern

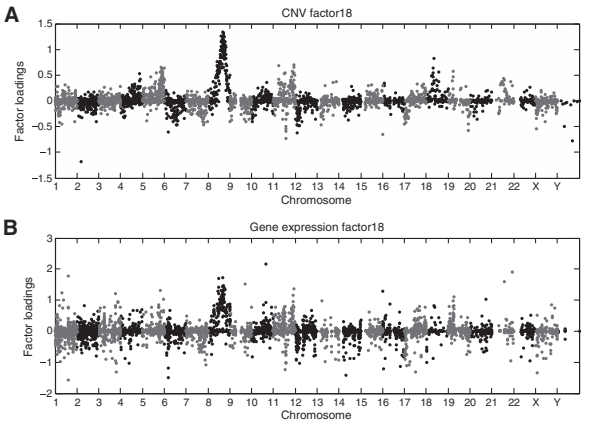

Fig. 4. Factor analytic relationship between CNV and gene expression. The figures show the factor loadings from the 18th factor of the joint factor model fit to CNV (**A**) and gene expression data (**B**), respectively. These results are for the maximum-likelihood collection sample, for ease of interpretation, although a full approximate posterior distribution is inferred. The gray denotes odd-numbered chromosomes and the black denotes even-numbered chromosomes

(Akahira *et al*., 2004; Emdad *et al*., 2007; Rennstam *et al*., 2003). Another gene in this region whose expression level is known to be important in tumor biology is WWP1 (8q21, Unique ID: 11059). This recapitulation of some of the well-known features of aneuploidy in cancer suggests that our joint model is appropriately capturing correlation structure between gene expression and CGH data.

As described above, many factors we obtained are associated with individual chromosomal locations, as demonstrated in Figure 4. However, there is also a subset of factors that are representative of multiple regions. Figure 5 shows that the largest factor loadings in CNV/gene expression for factor 41 come from both chromosomes 6 and 17. This is the explanation of the

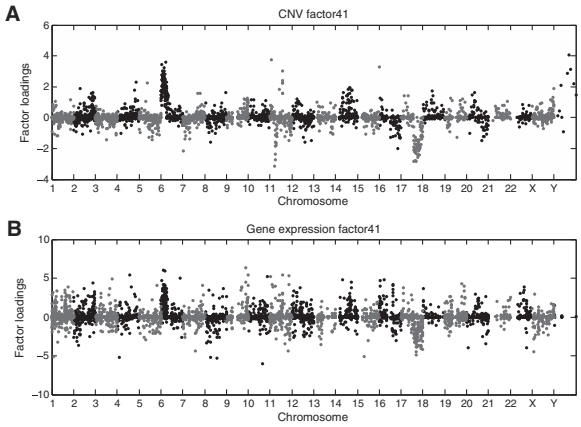

Fig. 5. Dual peaks shown in the loadings of factor 41 of the joint factor model fit to CNV (**A**) and gene expression (**B**) data. The gray denotes odd-numbered chromosomes and the black denotes even-numbered chromosomes

checkerboard regions of positive and negative correlation in Figure 3 as well. The CNVs from the top-ranked CGH probes in the two locations are highly negatively correlated, with copy number gain in chromosome 6 and loss in the other. There are a number of possible mechanistic explanations for this feature. For example, it is possible that wholesale duplication of one region is lethal to the cells without shutting down the apoptosis pathway. Such a shut down might be accomplished by deletion of other regions. Previous approaches to the joint analysis of gene expression and CNV through the use of factor models, such as Lucas *et al*. (2010), have failed to find these relationships.

The proposed joint factor model provides the flexibility of discovering factors that are relevant only for a subset of the subjects. It is interesting to note that a similar model, which enforces that all subjects are included in the inferred factors, performed poorly compared with the proposed model and discovered fewer factors that captured correlated changes in gene expression and CNVs.

5.3 Analysis of gene expression and DNA methylation data

For computational purpose, we selected probes with highest variances across samples and obtained top 1000 probes for both gene expression and methylation. Eighteen common factors were thus obtained. Unlike CNVs, methylation does not typically occur in contiguous regions; therefore, it is not surprising that no regional peaks were detected. Methylation acts as an epigenetic regulator and silences tumor suppressor genes by changing chromosomal structures. We detected a gene, SPON1 (11p15.2, Unique ID: 10418), which appears to be predominantly regulated by methylation of its CpG site (Fig. 6). Elevated expression of this gene relative to normal tissue is a known hallmark of ovarian cancer (Pyle-Chenault *et al*., 2005); however, the mechanism of this overexpression was previously unknown. SPON1 encodes VSGP/F-spondin protein promoting proliferation in vascular smooth cell during ovarian folliculogenesis, which has been identified as a potential diagnostic marker or therapeutic target for

Downloaded from https://academic.oup.com/bioinformatics/article/30/10/1370/267809 by guest on 18 November 2020

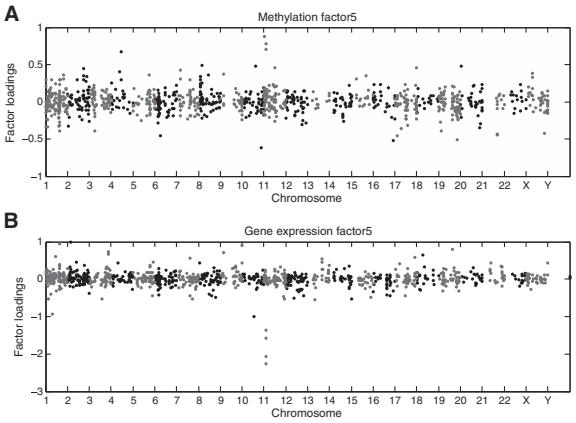

Fig. 6. SPON1 gene identified in the loadings peak from factor 5 of the joint factor model fit to DNA methylation (**A**) and gene expression (**B**) data. The gray denotes odd-numbered chromosomes and the black denotes even-numbered chromosomes

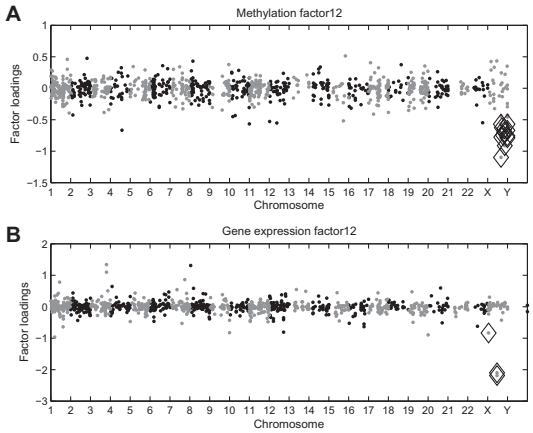

Fig. 7. Loadings from factor 12 with strong correlations between methylation (**A**) and gene expression (**B**) at many different loci. The gray denotes odd-numbered chromosomes and the black denotes even-numbered chromosomes. Diamond markers denote loci where the absolute value of loadings are >0.5, and gene expressions are significantly associated with adjacent methylation sites, detailed in Supplementary Table S4

ovarian carcinoma (Miyamoto *et al.*, 2001; Pyle-Chenault *et al.*, 2005).

In contrast to the almost single gene precision of factor 5, factor 12 shows strong correlation between methylation and gene expression in many different loci on X chromosome (Fig. 7, Supplementary Table S4). The list of CpG sites heavily loaded on this factor is displayed in Supplementary Table S4. Pathway analysis on these candidate genes (Fig. 7A) reveals that many are involved in the regulation of transcription. They may regulate the expression of XIST (Fig. 7B), the dysregulation of which has been found in several cancers (Huang *et al.*, 2002; Sirchia *et al.*, 2009). The correlation of methylation levels at all of these sites combined with their correlated gene expression levels suggests that they might all be the targets of a single methylation program; however, the existence of coordinated methylation enzymes that target these locations is unconfirmed.

We implemented the joint factor model for analysis of multiple genomics data in non-optimized Matlab on a quad core PC with 2.2 GHz CPU and 4 GB ram. The average time per iteration of the Gibbs sampler for the results in Section 5.2 is 72 s and for the results in Section 5.3 is 55 s.

6 CONCLUSION

A non-parametric joint factor analysis method is introduced for modeling multiple disparate but statistically related data. The proposed approach was demonstrated on the joint analysis of heterogeneous genomics data related to ovarian cancer. The proposed model uncovered key drivers of cancer, some of which have been previously reported in literature as well as some new genomic causes of cancer (potentially).

In this article, we have focused on integrating multiple heterogeneous but statistically correlated datasets, via a joint factor analysis approach, where the latent space is factorized into a shared component and data-specific components. The joint factor analysis model allows enough flexibility to incorporate additional information into the model. For example, we could alternatively use a Gaussian process prior or a Sticky-HMM prior (Du *et al.*, 2010) on the loadings of CNVs to incorporate spacial-dependence structure in the model. Hence, more refined regional peaks may be recovered. We plan to investigate such priors in our future work. Besides, we assume data-specific linear mappings between the latent space and the observation space. However, the assumption that the data lie in or close to a low-dimensional subspace is restrictive and a better assumption is that the data lie on a manifold. In the future, we wish to relax the linearity assumption of our joint factor model via a mixture of factor analyzers approach.

ACKNOWLEDGEMENTS

The authors thank two anonymous reviewers and associate editor for their careful reading and constructive comments. The content is solely the contributions of the authors and does not represent the official views of the sponsor.

Funding: This work was supported in part by Biomedical Advanced Research and Development Authority (BARDA) and the Office of Naval Research.

Conflict of Interest: none declared.

REFERENCES

Akahira,J.-I. *et al.* (2004) Expression of EBAG9/RCAS1 is associated with advanced disease in human epithelial ovarian cancer. *Br. J. Cancer*, **90**, 2197–2202.

Akaike,H. (1987) Factor analysis and AIC. *Psychometrika*, **52**, 317–332.

Akavia,U.D. *et al.* (2010) An integrated approach to uncover drivers of cancer. *Cell*, **143**, 1005–1017.

Archambeau,C. and Bach,F. (2008) Sparse probabilistic projections. In: *Proceedings of Neural Information Processing Systems*. pp. 73–80.

Bourgon,R. *et al.* (2010) Independent filtering increases detection power for high-throughput experiments. In: *Proceedings of the National Academy of Sciences*. pp. 9546–9551.

Carvalho,C. *et al.* (2008) High-dimensional sparse factor modelling: applications in gene expression genomics. *J. Am. Stat. Assoc.*, **103**, 1438–1456.

Downloaded from https://academic.oup.com/bioinformatics/article/30/10/1370/267809 by guest on 18 November 2020

Chen,M. *et al.* (2011) Predicting viral infection from high-dimensional biomarker trajectories. *J. Am. Stat. Assoc.*, **106**, 1–21.
Du,L. *et al.* (2010) Sticky hidden markov modeling of comparative genomic hybridization. *Trans. Signal Process.*, **58**, 5353–5368.
Emdad,L. *et al.* (2007) Astrocyte elevated gene-1: recent insights into a novel gene involved in tumor progression, metastasis and neurodegeneration. *Pharmacol. Ther.*, **114**, 155–170.
Fox,E. *et al.* (2008) An HDP-HMM for systems with state persistence. In: *Proceedings of the 25th International Conference on Machine Learning*. pp. 312–319.
Frank,B. *et al.* (2007) Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. *Carcinogenesis*, **28**, 1442–1445.
Garber,M.E. *et al.* (2001) Diversity of gene expression in adenocarcinoma of the lung. In: *Proceedings of the National Academy of Sciences*. pp. 13784–13789.
Gentleman,R. *et al.* (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for biology and health)*. Springer-Verlag New York, Secaucus, NJ, USA.
Gramacy,R. and Lee,H. (2007) Bayesian treed Gaussian process models with an application to computer modeling. *J. Am. Stat. Assoc.*, **103**, 1119–1130.
Griffiths,T. and Ghahramani,Z. (2005) Infinite latent feature models and the Indian buffet process. In: *Proceedings of Neural Information Processing Systems*. pp. 475–482.
Huang,K.-C. *et al.* (2002) Relationship of XIST expression and responses of ovarian cancer to chemotherapy. *Mol. Cancer Ther.*, **1**, 769–776.
Ishwaran,H. and Rao,J.S. (2005) Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Stat.*, **33**, 730–773.
Jennings,E. *et al.* (2012) Hierarchical Bayesian methods for integration of various types of genomics data. In: *Proceedings of the 2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)*. pp. 5–8.
Jeong,J. *et al.* (2010) An empirical Bayes model for gene expression and methylation profiles in antiestrogen resistant breast cancer. *BMC Med. Genomics*, **3**, 55.
Kendziorski,C.M. *et al.* (2006) Statistical methods for Expression Quantitative Trait Loci (eQTL) mapping. *Biometrics*, **62**, 19–27.
Klami,A. and Kaski,S. (2008) Probabilistic approach to detecting dependencies between datasets. *Neurocomputing*, **72**, 39–46.
Kothandaraman,N. *et al.* (2010) E2F5 status significantly improves malignancy diagnosis of epithelial ovarian cancer. *BMC Cancer*, **10**, 64.
Lanckriet,G.G. *et al.* (2004) A statistical framework for genomic data fusion. *Bioinformatics*, **20**, 2626–2635.
Li,M. *et al.* (2009) Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer. *BMC Med. Genomics*, **2**, 1–13.
Louhimo,R. and Hautaniemi,S. (2011) CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics*, **27**, 887–888.
Lucas,J.E. *et al.* (2010) Latent factor analysis to discover pathway-associated putative segmental aneuploidies in human cancers. *PloS Comput. Biol.*, **6**, e1000920.
Mairal,J. *et al.* (2008) Supervised dictionary learning. In: *Proceedings of Neural Information Processing Systems*. pp. 1033–1040.
Meeds,E. and Osindero,S. (2006) An alternative infinite mixture of Gaussian process experts. In: *Proceedings of Neural Information Processing Systems*. pp. 883–890.
Miyamoto,K. *et al.* (2001) Isolation and characterization of vascular smooth muscle cell growth promoting factor from bovine ovarian follicular fluid and its cDNA cloning from bovine and human ovary. *Arch. Biochem. Biophys.*, **390**, 93–100.
Paisley,J. and Carin,L. (2009) Nonparametric factor analysis with beta process priors. In: *Proceedings of the 26th International Conference on Machine Learning*. pp. 777–784.
Pils,D. *et al.* (2005) Five genes from chromosomal band 8p22 are significantly down-regulated in ovarian carcinoma. *Cancer*, **104**, 2417–2429.
Pyle-Chenault,R. *et al.* (2005) VSGP/F-spondin: a new ovarian cancer marker. *Tumor Biol.*, **26**, 245–257.
Rasmussen,C. and Ghahramani,Z. (2002) Infinite mixtures of Gaussian process experts. In: *Proceedings of Neural Information Processing Systems*. pp. 881–888.
Ray,P. and Carin,L. (2011) Non-parametric Bayesian modeling and fusion of spatio-temporal information sources. In: *2011 Proceedings of the 14th International Conference on Information Fusion (FUSION)*. pp. 1–7.
Rennstam,K. *et al.* (2003) Patterns of chromosomal imbalances defines subgroups of breast cancer with distinct clinical features and prognosis. a study of 305 tumors by comparative genomic hybridization. *Cancer Res.*, **63**, 8861–8868.
Robertson,K. (2001) DNA methylation, methyltransferases, and cancer. *Oncogene*, **20**, 3139–3155.
Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.
Sirchia,S. *et al.* (2009) Misbehaviour of XIST RNA in breast cancer cells. *PloS One*, **4**, e5559.
Talloen,W. *et al.* (2007) I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics*, **23**, 2897–2902.
Tang,X. *et al.* (2012) Functional interaction between responses to lactic acidosis and hypoxia regulates genomic transcriptional outputs. *Cancer Res.*, **72**, 491–502.
Thibaux,R. and Jordan,M.I. (2007) Hierarchical beta processes and the Indian buffet process. In: *Proceedings of the 11th Conference on Artificial Intelligence and Statistic*. pp. 564–571.
Tipping,M.E. (2001) Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, **1**, 211–244.
Tresp,V. (2001) Mixtures of Gaussian processes. In: *Proceedings of Neural Information Processing Systems*. pp. 654–660.
Wang,W. *et al.* (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, **29**, 149–159.
Xu,Y. *et al.* (2012) A Bayesian graphical model for integrative analysis of TCGA data. In: *Proceedings of 2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)*. pp. 135–138.
Zheng,L. and Lucas,J. (2012) *Aneuploidy in Health and Disease*. Uncover cancer genomics by jointly analysing aneuploidy and gene expression, InTech, pp. 22–41.

Downloaded from https://academic.oup.com/bioinformatics/article/30/10/1370/267809 by guest on 18 November 2020