HHS Public Access

Author manuscript

IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2015 November 01.

Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2014 ; 11(6): 984–994. doi:10.1109/TCBB.2014.2325035.

Latent feature decompositions for integrative analysis of multi-platform genomic data

Karl B. Gregory,
PhD candidate in Department of Statistics, Texas A&M University, College Station, TX, 77843-3143, USA

Amin A. Momin,
Department of Clinical Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, TX, 77230-1402, USA

Kevin R. Coombes, and
Department of Biomedical Informatics, The Ohio State University Wexner Medical Center, USA

Veerabhadran Baladandayuthapani
Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, 77230-1402, USA

Karl B. Gregory: kbgregory@stat.tamu.edu; Veerabhadran Baladandayuthapani: Veera@mdanderson.org

Abstract

Increased availability of multi-platform genomics data on matched samples has sparked research efforts to discover how diverse molecular features interact both within and between platforms. In addition, simultaneous measurements of genetic and epigenetic characteristics illuminate the roles their complex relationships play in disease progression and outcomes. However, integrative methods for diverse genomics data are faced with the challenges of ultra-high dimensionality and the existence of complex interactions both within and between platforms.

We propose a novel modeling framework for integrative analysis based on decompositions of the large number of platform-specific features into a smaller number of latent features. Subsequently we build a predictive model for clinical outcomes accounting for both within- and between-platform interactions based on Bayesian model averaging procedures. Principal components, partial least squares and non-negative matrix factorization as well as sparse counterparts of each are used to define the latent features, and the performance of these decompositions is compared both on real and simulated data. The latent feature interactions are shown to preserve interactions between the original features and not only aid prediction but also allow explicit selection of outcome-related features. The methods are motivated by and applied to, a glioblastoma multiforme dataset from The Cancer Genome Atlas to predict patient survival times integrating gene expression, microRNA, copy number and methylation data.

For the glioblastoma data, we find a high concordance between our selected prognostic genes and genes with known associations with glioblastoma. In addition, our model discovers several relevant cross-platform interactions such as copy number variation associated gene dosing and epigenetic regulation through promoter methylation. On simulated data, we show that our proposed method successfully incorporates interactions within and between genomic platforms to

aid accurate prediction and variable selection. Our methods perform best when principal components are used to define the latent features.

Index Terms

Latent feature; genomic data; high-dimensional; interactions; integrative models; Bayesian model averaging

1 Introduction

Recent advances in emerging technologies as well as their reduction in costs have steadily expanded the range of available genomic data types, providing detailed genome-wide information at the DNA (e.g. copy number variation), transcriptomic (e.g. gene/mRNA expression), epigenomic (e.g. methylation) and proteomic levels [1]. Previous studies, focusing on one platform at a time, have discovered several important oncogenic roles these molecular characteristics play in disease evolution and progression. These studies have identified common targets/genes in which somatic alterations—that is, post-conception mutations, such as methylation and copy number changes, which do not involve rearrangement of the original DNA sequence—are associated with cancer development [2]. Such alterations are highly pronounced in tumor tissue, emphasizing their importance in tumor progression [3]. Now, with multiple molecular platforms being assayed on the same subjects, analytic methods for integrating diverse data types to discover joint oncogenic roles are desired not only to understand the underlying cancer biology but also to aid in clinical prediction using such data [4]. This is one of the fundamental goals in cancer genomics, since this information can be used to determine the evolution of the disease as well as to design personalized therapies based on molecular markers, which is one of the most important problems in cancer research today [5].

However, principled genome-analytic methods for such data are faced with two important challenges: (1) high-dimensionality—a very large number of variables (e.g hundreds/thousands of genes) measured relative to the sample size (e.g. a few hundred patients)—and (2) the presence of complex correlations and interactions both within and between platform-specific features such as {mRNA vs. mRNA}, {mRNA vs. copy number} and {mRNA vs microRNA} interactions. Efforts to address high-dimensionality have seen considerable success in various contexts using a variety of statistical and machine learning approaches. For single-platform analyses, principal components analysis has proven useful for gene expression data, where interactions between genes may be modeled by taking the product of principal component score vectors [6]. Partial least squares has been used to predict a response from a large number of gene expression levels to identify genes of particular importance[7]. Nonparametric Bayesian clustering methods have been proposed to define methylation subgroups for diagnosis and prognosis [8].

Recently, integrative multi-platform analyses have demonstrated improvements in prediction and selection of significant genomic variables related to disease progression. A hierarchical integrative Bayesian analysis, modeling the joint effects of microRNA, mRNA and epigenetic factors on survival times revealed several genes of importance in glioblastoma

[4]. Classification models for clinical decision making have seen marked improvement from incorporating multiple data platforms—the integration of proteomic and gene expression data in a support vector machine yielding better predictive results than the use of a single platform [9]. A sparse version of canonical correlation analysis has been helpful in describing relationships between genetic and epigenetic characteristics when several genomic assays are taken on each subject [10]. However, most of these methods did not take the multi-level interaction effects, within and between platforms, into account when fitting the integrative models—a gap the present article seeks to fill.

In this paper we present a method to integrate diverse groups of high-dimensional genomic features which are believed to interact, and use this knowledge to develop a prediction model for a clinical outcome (e.g. survival time). A schematic of the method is shown in Figure 1. Briefly, the proposed methods (a) use latent feature decompositions such as principal components, partial least squares and non-negative matrix factorization to achieve low-dimensional representations of the data, (b) predict clinical outcomes using functions of these latent features which explicitly allow for interactions, (c) achieve simultaneous dimension reduction and variable selection and (d) are computationally efficient enough to scale to high-dimensional datasets.

The proposed method is applied to a glioblastoma multiforme (GBM) dataset from The Cancer Genome Atlas (TCGA), which provides a comprehensive and coordinated effort to collect multi-platform genomics data from matched tumor samples (see http://cancergenome.nih.gov/). Our motivating data contains the survival times (to death) of patients diagnosed with glioblastoma as well as gene, microRNA expression, copy number variation and methylation measurements from tumor samples. These platforms are believed to be interdependent in general, and it is likely that they interact in glioblastoma [11]. In our analysis, the interactions between the platforms exhibited an important role in patient survival, and we found a number of markers that are influential in GBM progression that have been previously reported in literature. In addition, our model discovers several relevant cross-platform interactions such as copy number variation, associated gene dosing and epigenetic regulation through promoter methylation. The reliability of the predictions and variable selection results of our proposed models are demonstrated on synthetic data.

The Methods section of the paper details the integrative model building process and its rationale, explaining the fitting process and a variable selection algorithm. In Results and Discussion, we present our integrative analysis of the TCGA-based GBM data as well as a simulation study to evaluate the operating characteristics of our methods. Conclusions and future directions of research are drawn in the last section.

Methods

The proposed procedure

—Consider observing the $n \times p_1, \ldots, n \times p_K$ matrices $\mathbf{X}_1, \ldots, \mathbf{X}_K$ along with the $n \times 1$ vector Y, which are, respectively, the values of K groups (platforms/assays in our case) of $p_1, \ldots, p_K$ genomic features and the responses (clinical outcomes) from a random sample of n units.

It is desired to predict the values in Y from the K groups of features and the interactions among them.

A general (conceptual) model incorporating the interactions within and between the groups of features can be written as

$$Y = f_1(\mathbf{X}_1) + f_2(\mathbf{X}_2) + \cdots + f_k(\mathbf{X}_K) \quad (1)$$

$$+ g_{11}(\mathbf{X}_1 * \mathbf{X}_1) + g_{22}(\mathbf{X}_2 * \mathbf{X}_2) + \cdots + g_{KK}(\mathbf{X}_K * \mathbf{X}_K) \quad (2)$$

$$+ g_{12}(\mathbf{X}_1 * \mathbf{X}_2) + g_{13}(\mathbf{X}_1 * \mathbf{X}_3) + \cdots + g_{(K-1)K}(\mathbf{X}_{K-1} * \mathbf{X}_K) + e \quad (3)$$

where "**A***B**" denotes the matrix containing, in the *i*th row, the cartesian product of the entries of the *i*th rows of **A** and **B** (i.e. the values of the interaction terms for observation *i*), and "**A***A**" is a matrix containing, in the *i*th row, all pairwise products of the entries in the *i*th row of **A** (so that there are no second order terms in the model), for $i = 1, \ldots, n$. Let $\{f(\cdot), g(\cdot)\}$ be functionals of a data matrix $\mathbf{X}$, specifications of which are provided below, and e be an $n \times 1$ vector of error terms. The model components have the following interpretations:

- Term (1) represents the *platform-specific effects* modeled as additive components (main effects) for each platform.
- Term (2) represents the *within-platform interaction effects* and consists of interactions among variables from the same platform.
- Term (3) represents the *between-platform interaction effects* and consists of interactions among variables across platforms.

To fit the above model, we must specify the functionals $f(\cdot)$ and $g(\cdot)$. The simplest case is a linear model for which $f_k(\mathbf{X}_k) = \mathbf{X}_k\boldsymbol{\beta}_k$, and $g_{kl}(\mathbf{X}_k * \mathbf{X}_l) = (\mathbf{X}_k * \mathbf{X}_l)\,\boldsymbol{\theta}_{kl}$ where $\boldsymbol{\theta}_{kl}$ is a parameter vector having the same length as $(\mathbf{X}_k * \mathbf{X}_l)$ and $\boldsymbol{\beta}_k$ is $p_k \times 1$, for $k, l \in \{1, \ldots, K\}$. With this specification our model can be written as,

$$Y = \beta_0 + \sum_k \sum_{j=1}^{p_k} \beta_{kj} X_{kj} + \sum_k \sum_{1 \le i < j \le p_k} \gamma_{kji} X_{kj} X_{ki} + \sum_{k>l} \sum_{j=1}^{p_k} \sum_{i=1}^{p_l} \eta_{klji} X_{kj} X_{li} + e \quad (4)$$

where X_{kj} is the jth column of $\mathbf{X}_k$, β_0 is the intercept, β_{kj} are members of $\boldsymbol{\beta}_k$, and γ_{kji} and η_{klji} are members of $\boldsymbol{\theta}_{ij}$ for $k, l \in \{1, \ldots, K\}$. Correspondingly, the first sum is of main effects, the second sum is of interactions between predictors belonging to a common platforms, and the third sum is of the first-order interactions between predictors belonging to different platforms. If $p = p_1 + \cdots + p_k$, then (4) has $p + p(p - 1)/2$ terms, which will often exceed the number of observations n. Considering higher-order interactions will quadratically increase the number of terms, making model fitting unstable or infeasible in some cases. The GBM data set, even after subsetting the predictors according to a pathway, has data on $n = 163$

patients for $p = 1298$ predictors, for which there are a total of (1298)(1297)/2 = 841753 possible two-way interactions!

To overcome this challenge, we consider lower-dimensional projections of the input features that will capture most of the information in the data, which are defined as

$$\mathscr{D}(X_{k1}, \ldots, X_{kp_k}) \equiv \{T_{k1}, \ldots, T_{km_k}\},$$

where $\mathcal{D}$ is a dimension-reduction technique inducing a latent decomposition of the original features (specific choices described below) and $\mathbf{T}_k$ is an $n \times m_k$ matrix of the latent feature scores derived from $\mathbf{X}_k$ such that $m_k < p_k$ for $k = 1, \ldots, K$. The decomposition is done such that, with $m \equiv m_1 + \cdots + m_K$, $m + m(m - 1)/2 < n$ holds. In our construction of $\mathcal{D}$, m_k is of much lower-dimension (tens) than p_k (hundreds/thousands), thus Y may be modeled using the latent features and their interactions such that $f_k(\mathbf{X}_k)$ and $g_{kl}(\mathbf{X}_k * \mathbf{X}_l)$ become, respectively, $\tilde{f}_k(\mathbf{T}_k)$ and $\tilde{g}_{kl}(\mathbf{T}_k * \mathbf{T}_l)$, for $k, l \in \{1, \ldots, K\}$.

With this specification our latent feature-based model can be expressed as

$$Y = \tilde{\beta}_0 + \sum_k \sum_{j=1}^{m_k} \tilde{\beta}_{kj} T_{kj} + \sum_k \sum_{1 \le i < j \le m_k} \tilde{\gamma}_{kji} T_{kj} T_{ki} + \sum_{k>l} \sum_{j=1}^{m_k} \sum_{i=1}^{m_l} \tilde{\eta}_{klji} T_{kj} T_{li} + \tilde{e}, \quad (5)$$

where T_{kj} is the jth column of $\mathbf{T}_k$, $\tilde{\beta}_0$ is the intercept, $\tilde{\beta}_{kj}$ is the main effect of the jth latent feature of the kth variable group, $\tilde{\gamma}_{kji}$ is the interaction effect between the ith and jth latent feature from the kth variable group, and $\tilde{\eta}_{klji}$ is the interaction effect between the jth latent feature from the kth variable group and the ith latent feature for the lth variable group, for $k, l \in \{1, \ldots, K\}$.

Thus, in order to fit model (5), a dimension-reduction technique $\mathcal{D}$ is applied to each matrix $\mathbf{X}_k$ of original features to produce an $n \times m_k$ matrix $\mathbf{T}_k$ such that $m_k \le p_k$ for $k = 1, \ldots, K$ and $m + m(m - 1)/2 < n$. Then $\mathbf{T}_1, \ldots, \mathbf{T}_K$ will contain the realizations of $m_1, \ldots, m_K$ latent "scores" observed on n units. Compared to model (4), in which the GBM data would have 1298 main effects and 841753 interactions across 4 groups of predictors, a latent feature decomposition might choose $m_1 = m_2 = m_3 = m_4 = 4$, for example, and model (5) would have 16 main effects and 16(15)/2 = 120 interaction effects for a total of 136 effects, thus making model fitting scalable, preserving interactions and maintaining predictive accuracy (as we will show).

Model fitting: After this dimension reduction of mapping from the original feature space of X to T-space of latent scores, we conduct a model fitting procedure ($\mathcal{S}$) and a variable selection procedure to estimate the parameters in (5). The specific choices of $\mathcal{D}$ $\{m_k\}_{k=1}^{K}$, and $\mathcal{S}$ are now detailed below.

Choice of latent decompositions

The choices of $\mathcal{D}$ considered in this paper are principal components (PC), partial least squares (PLS), nonnegative matrix factorization (NMF), and their sparse counterparts (SPC, SPLS, SNMF), respectively. There are several other choices available such as kernels, non-linear embeddings etc. Our choice of these six latent decompositions is guided by the fact that each of them produces linear combinations of the original features of the form $T_{kj} = \omega_{kj1}X_{k1} + \cdots + \omega_{kjp_k}X_{kp_k}$ for $j = 1, \ldots, m_k$, which allows a clear interpretation of the dependencies in terms of the original features. The sparse versions are implemented because they encourage most of the ω values to be zero, considerably aiding variable selection. Though these six decompositions are proposed as candidate choices, only one will be selected when analysis is carried out. We now describe the modeling aspects and rationale for each of these decompositions as well as the procedures to determine the number of latent features (i.e the effective rank m_k) to retain for fitting model (5).

Principal components—PC is a common dimension reduction approach in bioinformatics [6] in which a singular value decomposition is applied to $\mathbf{X}_k$ for $k = 1, \ldots, K$ where m_k is the rank of the decomposition. The resulting linear combinations of the columns of $\mathbf{X}_k$ are orthogonal and successively summarize the maximum possible amount of variation in $\mathbf{X}_k$. To determine the number of principal components to retain, a bi-crossvalidation scheme is used in which $r \times s$ submatrices of $\mathbf{X}_k$ are removed and imputed from the remaining rows [12]. A sparse version of principal components (SPC) is implemented using the R package 'PMA' [13], which executes the algorithm found in Witten et al. [14]. The sparsity parameter as well as the rank of the decomposition are chosen via bi-crossvalidation.

Partial least squares—Under PLS, latent features from $\mathbf{X}_k$ are successively defined such that they have a high covariance with $\mathbf{y}$ for $k = 1, \ldots, K$ [15]. Because of the apparent advantage of incorporating response values into the dimension reduction, PLS has gained significant popularity in genomic analyses [16]. The SIMPLS algorithm [17] is used here to carry out the decomposition. For the sparse version (SPLS), the R package 'spls' [18] is used, where the sparsity parameter and number of latent features are chosen using 10-fold crossvalidation [7].

Non-negative matrix factorization—The NMF procedure, which is often used for pattern identification and class discovery in genetic applications [19], decomposes each transposed covariate matrix $\mathbf{X}_k^{'}$, $k = 1, \ldots, K$ such that $\hat{\mathbf{X}}_k^{'} = \mathbf{W}_k\mathbf{H}_k$, where $\mathbf{W}_k$ is $p_k \times m_k$ and $\mathbf{H}_k$ is $m_k \times n$. The columns of $\mathbf{W}_k$ define m_k "metagenes" which are linear combinations of the original p_k variables. The n columns of $\mathbf{H}_k$ contain expression levels of the metagenes defined in $\mathbf{W}_k$. The R package 'nmf' [20] is used to perform the decomposition, where the rank m_k of $\mathbf{W}_k$ is chosen via the same bi-crossvalidation routine as for principal components. For the sparse version (SNMF), the constrained algorithm proposed by Kim and Park [21] is implemented using the same R package, where the regularization parameters are set to $\eta = \max\{|\mathbf{X}_{k(i,j)}|\}^{1/2}$ and $\beta = .1$. The rank is again chosen via bi-crossvalidation.

Predictive modeling fitting via Bayesian model averaging

Having defined our latent features and obtained the latent feature scores, we plug them into model (5) for predictive model fitting. A Bayesian model averaging (BMA) algorithm is proposed for the choice of $\mathcal{S}$ for fitting model (5) on the latent features. The BMA procedure explores the "model space" by fitting a large number of models (all possible models, if computation allows), and assigning to each model (determined by the number of significant variables) a posterior probability. Also, for each variable that is a candidate for entering the model, the BMA algorithm reports a probability that its regression coefficient is nonzero. One can then choose either the highest probability model or a model retaining every variable for which the regression coefficient is nonzero with a probability exceeding some threshold. Here, the highest-probability model is chosen. The values of the regression coefficients deemed nonzero in the highest-probability model are estimated using the averages of their values across all models considered throughout the MCMC-based exploration of the model space. It is well-established that averaging the regression coefficients across all possible models in this way yields more favorable predictive results than fitting a single assumed-to-be-true model. See Raftery et. al. [22] for detailed discussion. We use the companion R package 'BMA' for the implementation.

Variable selection

Selecting a list of important variables is done as follows. For every $k = 1, \ldots, K$, a dimension-reduction technique $\mathcal{D}$ will create from $\mathbf{X}_k$ a set of m_k latent feature score vectors comprising $\mathbf{T}_k$ which are linear combinations of the column vectors $X_{k1}, \ldots, X_{kp_k}$ such that $T_{kj} = \omega_{kj1}X_{k1} + \cdots + \omega_{kjp_k}X_{kp_k}$ for $j = 1, \ldots, m_k$. It is common to attribute greater importance to variables to which $\mathcal{D}$ assigns greater weights. The model selection procedure $\mathcal{S}$ produces a set of indices $\mathcal{M} \subset \{(k, j) : j = 1, \ldots, m_k, k = 1, \ldots, K\}$ such that the set of latent features $\mathcal{T} \equiv \{T_{kj} : (k, j) \in \mathcal{M}\}$ is retained in the model, where members of $\mathcal{T}$ can appear either as main effects or as part of an interaction. For each $k = 1, \ldots, K$, it is suggested that the variables $X_{k1}, \ldots, X_{kp_k}$ can be sorted in order of importance by considering the maximum of the magnitudes of the weights assigned to each variable across all the latent features from group k retained by $\mathcal{S}$. That is, letting $\mathcal{M}_k = \{(l, j) : l = k, (l, j) \in M\}$ index the latent features from group k retained by $\mathcal{S}$, the $p_k \times 1$ vector

$$(\max\{|\omega_{lj1}|:(l,j) \in \mathcal{M}_k\}, \ldots, \max\{|\omega_{ljp_k}|:(l,j) \in \mathcal{M}_k\}) \quad (6)$$

provides a means of ranking the variables $X_{k1}, \ldots, X_{kp_k}$ by their largest contributions to the retained latent features. Let $\{a^2_{k(1)}, \ldots, a^2_{k(p_k)}\}$ be the vector of maximum loading magnitudes from (6) sorted in descending order. Square each component and divide by the sum of the squared components to get $b_{k(j)} = a^2_{k(j)}/(a^2_{k(1)} + \cdots + a^2_{k(p_k)})$ for $j = 1, \ldots, p_k$. Then, let the variables associated with $a_{k(1)}, \ldots, a_{k(q)}$ be deemed significant where $q = \min\{r : b_{k(1)} + \cdots + b_{k(r)} > \vartheta\}$, for some choice of $\vartheta \in (0, 1)$. If $\vartheta = .8$, the variables selected will have squared maximum loading magnitudes which comprise at least 80% of the sum of squared maximum loading magnitudes for the variables in that group.

Summary of model fitting—In summary, our model fitting and variable selection is done via the following steps:

Step 1: Apply the chosen latent feature decomposition to each platform-specific feature matrix, specifying (or choosing via crossvalidation) the number of latent features m_1, …, m_K to retain from each platform.

Step 2: Fit a regression model for the response Y, conditioning upon all the latent features and their interactions, using Bayesian model averaging procedures.

Step 3: Based on the highest (posterior) probability model from the previous step, we conduct variable/feature selection by analyzing the loadings of the significant latent features retained in the model.

Results and discussion

Integrative analysis of TCGA glioblastoma data

—We apply our proposed procedure for integrative modeling using latent decompositions to a multi-platform glioblastoma multiforme (GBM) dataset collected under the aegis of TCGA [23]. GBM is among the most aggressive of brain cancers, and is known to be driven by diverse genetic and epigenetic mechanisms. In our analysis, the response variable is the uncensored survival time (time from initial diagnosis to death) obtained from $n = 163$ patients. Alongside the survival times, the data contain genomic features from 4 platforms: mRNA (gene expression), microRNA (miRNA), copy number and methylation measurements.

Our goal is to integrate the data from these four genomic platforms—which comprise four distinct feature groups—to predict patient survival times and to identify genes of particular prognostic significance in GBM. To illustrate our methodology, we focus on the signaling pathways that contain genes most frequently altered in GBM: PI3K/AKT, RAS, P53 and RB pathways, obtained from the literature and public pathway databases [24]. Thus our GBM data for downstream modeling consists of 49 genes mapped to these pathways with corresponding mRNA, methylation and copy number variation measurements along with the miRNA profile for each patient, consisting of 534 expression levels, all of which are used as inputs in the model. Thus our original feature matrices for the four platforms have $p_{rna} = 49$, $p_{mirna} = 534$, $p_{meth} = 179$, and $p_{cn} = 536$ features, for a total of $p = 1298$ features.

The proposed method under the six choices of latent feature decomposition is applied to the data to predict the (log-transformed) survival times, and variable selection is performed with a threshold $\vartheta = .8$, i.e. 80% of the "energy" is retained. For a fair comparison with a procedure which does not explicitly incorporate interactions but is known to provide high predictive accuracy and stable variable selection (for correlated variables) we use the elastic net regression (EN) proposed by Zou and Hastie [25]. The EN provides an extension of the LASSO to the $p \gg n$ case and directly chooses a subset/groups of the original predictors to retain while simultaneously estimating the regression coefficients. The tuning parameters are chosen via 10-fold crossvalidation and we use all 1298 ungrouped covariates. The choice for the best model is guided by the simulation results (presented later) as well as by a predictive evaluation of their performances on the test data sets using crossvalidation. We use a 10-fold

crossvalidation procedure, where one-tenth of the rows are omitted from the data at a time and their response values are predicted after fitting the model on the remaining rows. Once predicted values are obtained for all ten splits of the data, the mean squared error of prediction (MSEP) is computed by averaging the mean squared errors over the ten folds.

Boxplots of the crossvalidation MSEP acheived over the ten folds for each of the six decompositions with and without the inclusion of interactions and for the EN appear in Figure 2. It can be seen that overall the inclusion of the interaction terms enables a better fit, lowering the MSEP in all cases except for under the SNMF decomposition (which has exhibited poor performance under simulation as well). The PC, SPC, and NMF decompositions and the EN achieved lower MSEP values than the PLS, SPLS, and SNMF decompositions.

Key results: The Bayesian model averaging procedure did not retain any of the NMF latent features in the model, so the estimated response is a constant, and no variables were chosen. It is striking, then, that the NMF decomposition should have acheived such a low MSEP in the crossvalidation procedure. This seems to suggest that the models fit using the PC and SPC decompositions and the EN are little better than predicting the response with its mean value. However, given the greater reliability of the PC, SPC, and EN procedures in variable selection, the list of variables they identify may be of use. The BMA regressions on the PC and SPC latent features and their interactions and the EN model are displayed in Table 1, along with a list of variables identified as significant using the variable selection procedure with $\vartheta = .8$.

Graphical displays of the variable selections for the SPC model are shown in Figures 3 and 4. Figure 3 displays the sorted loading magnitudes for the 536 copy number measurements and the 49 mRNA gene expression levels. The height of the shaded region represents the cumulative proportion of the sum of squared loading magnitudes. When the height of the shaded area exceeds $\vartheta = .8$, no more variables are selected. Figure 4 displays the loading magnitudes for the copy number measurements and mRNA expression levels plotted in the original orderings.

The PC and SPC decompositions each chose to retain four copy number, three mRNA, four miRNA, and four methylation latent features, so that $m_{rna} = 4$, $m_{mirna} = 3$, $m_{meth} = 4$, and $m_{cn} = 4$. Of all these latent features and their interactions, the Bayesian model averaging scheme selected a single interaction between a copy number and an mRNA latent feature with high posterior probability. The loadings produced by PC and SPC were very similar, such that the same variables were identified as significant under both decompositions: three copy number variation measurements and four mRNA expression levels. The BMA fitted models on the PC and SPC latent features were also identical.

Figure 5 depicts high-probability models fit during the BMA procedure for all six choices of latent feature decomposition. Each panel displays a matrix in which the rows correspond to input variables and the columns to models selected in the BMA search with models in descending order from left to right according to their posterior probabilities. Red (blue) corresponds to negative (positive) associations with the outcomes. Model selection

consistency and convergence is demonstrated by long contiguous horizontal bands, indicating that a variable appears consistently in the posterior MCMC exploration. Under the PC and SPC decompositions, the label CN2xMR2 represents the single copy number × mRNA latent feature interaction included in the highest-probability model; this interaction appears consistently in the selected models, indicating that its final selection is the result of convergence. For the other decompositions, with the exception of NMF, which is shown in the simulation study to be a relatively poor choice, the latent features chosen in the final model are consistently included during the BMA procedure, indicative of convergence.

The EN chose only a single main effect out of the 1298, the miRNA expression level hsa-miR-383. No two-way interactions were included in the EN model, since considering all of them would mean including 1298(1297)/2 = 841753 additional regressors, which is computationally impractical. Including the main effects and interactions for just the copy number and mRNA platforms would still involve 536 + 49 + 585(584)/2 = 171405 regressors. This highlights the gains of our approach, as these numerous interactions are implicitly brought into the model through the latent feature interactions.

The fitted models and lists of selected features from the proposed procedure under all six decompositions can be found in Additional file 1.

Biological ramifications

Lists of genes (for probes associated with expression, copy number and methylation) along with mature miRNAs were prepared for the PC, SPC, PLS, SPLS, and SNMF individual models (under the NMF decomposition, no latent features were retained). Ingenuity pathway analysis (IPA build 192063) was performed to identify their bio-functional roles using the core analysis pipeline. The various models contain both common and unique latent features. To determine their significance in cancer, we performed functional analysis on the genes (probes associated with expression, copy number variation and methylation) and miRNAs using IPA. A majority of the selected features in the models are of known assocation with cancer processes and GBM; see Additional file 2 for detailed descriptions of the genes selected by each model and their associations with cancer.

Briefly, we found many genes from the models play critical roles in cell cycle, cell division, proliferation and other mechanisms associated with cancer progression and metastasis. This is highlighted by the fact that *CCND1*, *CCND2*, *CDKN2C* and *MDM2*, the common genes between PC, SPC, PLS and SPLS models, have been implicated in GBM and other tumors[26], [27]. While *CCND1*, *CCND2*, and *CDKN2C* directly regulate cell cycle and cell division, *MDM2* is a regulator of p53 through ubiqitination [28]. Additionally, genes such as *ERBB3*, *PDGFRA*, *PIK3CA* and *EGFR* within the models also play important roles in cancer signaling pathways.

The models also contain a number of interacting features from different platforms. Partially, this redundancy can be attributed to the influence of copy number variations on gene expression through gene dosing [29] and epigenetic regulation of gene expression through promoter methylation [30]. The presence of both gene expression and methylation probes for the *HRAS*, *CCND1*, *CCND2*, *ERBB3* and *PDGFRA* genes in the PLS model concurs with

previous studies which have shown them to be epigenetically regulated [30], [31], [32]. Also, the presence of copy number variation and gene expression features within the SNMF model indicates the influence of gene dosing, which is consistent with previous observations of higher *EGFR* expression in tumors with gene amplifications [33].

Another mechanism that may impact the model is the modulation of gene expression by small non-coding RNA (miRNA). To identify the putative regulatory impact of miRNAs on mRNA, we constructed interaction networks using the IPA Build-Connect algorithm for miRNA and genes selected by the PLS model. Only experimentally validated relations were included in the network construction. The network is depicted in Figure 6, which shows only molecules that are connected (the layout was modified using Path Designer), and reveals important regulatory relationships between miRNA (blue nodes) and mRNA (red nodes) that may impact GBM survival.

Simulation study

We further evaluated the operating characteristics of proposed procedure using simulations. Of particular interest is the the effectiveness of incorporating interactions for prediction and variable selection. We benchmarked our simulations to be close to real data in terms of the sample size and number of features. We used all the six choices of the latent decomposition but chose $K = 2$ groups/platforms of $p_1 = 300$ and $p_2 = 600$ features, in order to not specify all possible interactions across four groups. To evaluate the gains in predictive power by including latent feature interactions in the model, the simulation is run with and without these interactions, and the mean squared error of prediction on test data sets over 50 simulation runs is recorded.

In generating the original features, the belief that significant genes often act upon the response in concert, or as part of a group of genes, is imposed [34]. In each variable group, 95% of the variables are spurious—not having any effect upon the response—and are uncorrelated. The remaining variables are given nonzero slope coefficients and are correlated. Let $s_k = .95*p_k$ for $k = 1, 2$ be the number of spurious variables in each covariate group. Then

$$X_{kj} \sim \begin{cases} Z_k + e_{kj} & j = 1, \ldots, p_k - s_k \\ f_{kj} & j = p_k - s_k + 1, \ldots, p_k \end{cases}$$

where Z_1, Z_2, e_{1j}, e_{2j}, $j = 1, \ldots, p_k - s_k$ are independent and follow a gamma distribution with mean 1 and variance 1/2, and f_{kj}, $j = 1, \ldots, p_k$ are independent and follow a gamma distribution with mean 2 and variance 1 for $k = 1, 2$. Thus each X_{kj}, $j = 1, \ldots, p_k$, $k = 1, 2$, has mean 2 and variance 1, and the first $p_k - s_k$ variables, which shall be given nonzero slope coefficients, are correlated. After centering the predictors and setting

$$\beta_{kj} = \begin{cases} 3 & j = 1, \ldots, p_k - s_k, k = 1, 2 \\ 0 & o.w. \end{cases}$$
$$\eta_{klji} = \begin{cases} 5 & 1 \leq j, i \leq 3, k = 2, l = 1 \\ 0 & o.w. \end{cases}$$

the response Y is built from model (4), with $e \sim N(0, \sigma^2)$ with $\sigma^2 = 4$.

The simulation is run 50 times with training sets of size $n = 160$, and each time the mean squared error of prediction (MSEP) is computed on a test data set with $n = 400$ observations. Variable selection is done with $\vartheta = .8$.

The performance of the proposed method under the six choices of latent feature decomposition is again compared to that of the EN. Figure 7 shows boxplots of the MSEP values achieved on the testing datasets for the 50 simulation runs under each decomposition with and without the inclusion of latent feature interactions in the latent feature model. The models which included latent feature interactions acheived lower prediction errors across the test datasets for all six decompositions, indicating that the latent feature interactions bring additional predictive power when interactions are present among the original features. The SPLS decomposition with interactions acheived the lowest prediction error, followed by the SPC, PC, and SNMF decompositions. In addition, the median MSEP for all the models with interactions is lower than that of the EN, indicating the benefit of incorporating interactions in predictive modeling.

To compare the variable selection capabilities of the model under the six decompositions and the EN, true positive (selection of an original feature with a nonzero slope coefficient) and true negative (avoiding selection of an original feature with a slope coefficient of zero) rates are reported in Table 2 with standard errors and are depicted in Figure 8. The EN and the proposed procedure under the SPLS decomposition achieved the highest true positive rates, but also had the lowest true negative rates. The NMF and SNMF procedures appear overly conservative for variable selection, with very low true positive rates and high true negative rates—more so the NMF, never selecting a single spurious variable (and selecting very few significant variables) during all 50 simulation runs. The PC, SPC, and PLS procedures are the more balanced, and of these the PLS had a lower true positive rate and inferior predictive power.

To further compare the variable selection capabilities of the proposed procedure under the six decompositions and the EN, the conditional probability that a variable is significant or spurious, given that the model has deemed it so, are computed. In the simulation, 5% of the original features in each group had nonzero coefficients. Given that 100u% of the original features have nonzero coefficients, the probabilities P(significant | selected) and P(spurious | not selected) for each method are more readily interpretable than TP and TN rates. For a given proportion u of significant features, these probabilities can be computed as

$$P(\text{significant}|\text{selected}) = uP(TP)\{uP(TP)+(1-u)[1-P(TN)]\}^{-1}$$
$$P(\text{spurious}|\text{not selected}) = (1-u)P(TN)\{(1-u)P(TN)+u[1-P(TP)]\}^{-1} \quad (7)$$

These functions are plotted in separate panels in Figure 9 against u and $1 - u$ for $u \in (0, .1)$. The dashed vertical line through each plot indicates the value of u at which the simulation was run. If true positive and true negative rates remain constant across u, then the conditional probabilities in (7) and (7) will follow the plotted lines. The EN and the proposed procedure under SPLS are correct with the least probability when they *do* select a variable, and correct with the greatest probability when they do *not* select a variable. The probabilities that the NMF or SNMF are correct in failing to select a variable are very low. The proposed procedure under the PC and SPC decompositions provides reliable selection of variables, as well as middle-of-the-road reliability in avoiding the selection of spurious variables, and so these are again the best-performing methods.

To investigate the consistency/stability of our procedure under small perturbations in the data, we conducted a bootstrap analysis. We applied our model fitting and variable selection procedure using the PC and SPC decompositions as well as the EN to 500 random bootstrap samples of the data. For each bootstrap sample, we recorded the variables selected by each method and measured stability by evaluating the variable selection consistency across the bootstrap samples, where values close to 1 indicate stable performance. Figure 10 displays for the PC, SPC, and EN procedures, the frequency at which each of the 1298 variables (in the GBM data) was selected over 500 bootstrap samples of the data. For each procedure, the variables are sorted from most to least frequently selected. The dashed line traces the bootstrap variable selection frequencies and the vertical arrows mark variables included in the original model fit. The proposed procedure under the PC decomposition achieved the greatest stability, choosing a small proportion of the variables with high frequency (close to 1). The SPC decomposition chose fewer variables on average, though it still demonstrated a tendency to select the same variables across bootstrap samples. The EN procedure tended to select only a single variable, but only in a small proportion of the bootstrap samples. This further suggests that PC-based latent decompositions are best suited for such high-dimensional data.

In summary, across all the choices of latent feature decomposition, the models in which interactions were included exhibited the greatest predictive power. The sparse decompositions, as expected, achieved better variable selection results, especially in the case of SPLS over PLS and SNMF over NMF, the PC and SPC producing similar TP and TN rates in the simulation. Except for under the NMF and SNMF decompositions, the proposed procedure was competitive with the EN. The PC and SPC decompositions appeared to acheive the best results between predictive performance and variable selection and we recommend either of these to be the default option for data analysis.

Conclusions

We propose a modeling framework for integrative analysis of diverse genomic data based on latent decompositions of the original feature space. Our models explicitly allow interactions

both within and between platform-specific components to predict a clinical outcome. We demonstrate that in the presence of high-dimensionality, decomposing the feature groups into smaller groups of latent features and allowing these latent features to interact is an effective way to make predictions as well as to identify important variables among the original sets of features. Among the latent feature decompositions investigated, the latent features derived via principal components and its sparse version demonstrated the best performance for prediction and variable selection under simulated settings, and also provided the best fit to our motivating glioblastoma dataset under crossvalidation; these are the decompositions ultimately recommended to the user. An R package implementing the procedure shall be forthcoming.

Additional files

Additional file 1 — selectedFXtables.pdf

A pdf file containing the fitted models for the glioblastoma data for all six decomposition methods with crossvalidation variable selection results. This file is available at: http://odin.mdacc.tmc.edu/~vbaladan/Veera_Home_Page/Software_files/SelectedFXtables.pdf

Additional file 2 — BioFunctions_IPA_020713.xls

An excel file containing separate sheets for the PC, SPC, PLS, SPLS, and SNMF models, each describing the genes selected by the model with respect to their known associations with cancer and a summary sheet describing the concordance of the model-selected genes with genes known to be associated with cancer.. This file is available at: http://odin.mdacc.tmc.edu/~vbaladan/Veera_Home_Page/Software_files/BioFunctions_IPA_020713.xls

Acknowledgments

VB research is partially supported by NIH grant R01 CA160736 and the Cancer Center Support Grant (CCSG) (P30 CA016672). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. AAM was partially supported by the Lorraine Dell Program in Bioinformatics for Personalization of Cancer Medicine.

References

1. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CMT, Beyene J. Data integration in genetics and genomics: methods and challenges. Human Genomics and Proteomics. 2009
2. Weir B, Zhao X, Meyerson M. Somatic alterations in the human cancer genome. Cancer Cell. Nov. 2004 6:433–438. [PubMed: 15542426]
3. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nature Reviews, Genetics. Jun.2002 3:415–428.
4. Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, Do KA. ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. Bioinformatics. 2013; 29(2): 149–159. [PubMed: 23142963]
5. Chin L, Gray JW. Translating insights from the cancer genome into clinical practice. Nature. Apr. 2008 452:553–563. [PubMed: 18385729]
6. Ma S, Dai Y. Principal component analysis based methods in bioinformatics studies. Breifings in Bioinformatics. Jan; 2011 12(6):714–722.

7. Chun H, Keleandscedil S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society Series B. 2010; 72(1):3–25. [Online]. Available: http://ideas.repec.org/a/bla/jorssb/v72y2010i1p3-25.html.
8. Zhang L, Meng J, Liu H, Huang Y. A nonparametric bayesian approach for clustering bisulfate-based dna methylation profiles. BMC Genomics. 2012; 13(Suppl 6)
9. Daemen A, Gevaert O, Ojeda F, Debucquoy A, Suykens J, Sempoux C, Machiels JP, Haustermans K, De Moor B. A kernel-based integration of genome-wide data for clinical decision support. Genome Medicine. 2009; 1(4):39. [Online]. Available: http://www.genomemedicine.com/content/1/4/39. [PubMed: 19356222]
10. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical Applications in Genetics and Molecular Biology. 2009; 8(28)
11. Etcheverry A, Aubrey M, de Tayrac M, Vauleon E, Boniface R, Guenot F, Saikali S, Hamlat A, Riffaud L, Menei P, Quillien V, Mosser J. Dna methylation in glioblastoma: impact on gene expression and clinical outcome. BMC Genomics. 2010; 11(701)
12. Owen AB, Perry PO. Bi-crossvalidation of the SVD and the nonnegative matrix factorization. The Annals of Applied Statisics. 2009; 3(2):564–594.
13. Witten, D.; Tibshirani, R.; Gross, S.; Narasimhan, B. PMA:Penalized Multivariate Analysis. 2011. [Online]. Available: http://CRAN.R-project.org/package=PMA
14. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009; 10(3):515–534. [PubMed: 19377034]
15. Wold, H. Estimation of Principal Components and Related Models by Iterative Least squares. New York: Academic Press; 1966. p. 391-420.
16. Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics. 2006; 8(1):32–44. [PubMed: 16772269]
17. Mevik B-H, Wehrens R. The pls package: principal component and partial least squares regression in r. Journal of Statistical Software. 2007; 18
18. Chung, D.; Chun, H. Sparse Partial Least Squares (SPLS) Regression and Classification. 2012. [Online]. Available: http://www.stat.wisc.edu/_chungdon/spls/
19. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences of the United States of America. Mar; 2004 101(12):4164–4169. [PubMed: 15016911]
20. Gaujoux, R. Algorithms and framework for Nonnegative Matrix Factorization (NMF). 2010. [Online]. Available: http://nmf.r-forge.r-project.org
21. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics. Jun; 2007 23(12):1495–1502. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btm134. [PubMed: 17483501]
22. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. Journal of the American Statistical Association. Mar; 1997 92(437):179–191.
23. 2013 Feb. [Online]. Available: https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp
24. 2013 Jan. [Online]. Available: http://cbio.mskcc.org/cancergenomics/gbm/pathways
25. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B. 2005; 67:301–320.
26. Buschges R, Weber R, Actor B, Lichter P, Collins VP, Reifenberger G. Amplification and expression of cyclin d genes (ccnd1, ccnd2 and ccnd3) in human malignant gliomas. Brain Pathology. Jul; 1999 9(3):435–442. [PubMed: 10416984]
27. Furnari F, Fenton T, Bachoo R, Mukasa A, Stommel J, Stegh A, Hahn W, Ligon K, Louis D, Brennan C, Chin L, DePinho R, Cavenee W. Malignant astrocytic glioma: genetics, biology, and paths to treatment. Genes Dev. 2007; 21(21):2683–710. [PubMed: 17974913]
28. Moll U, Petrenko O. The mdm2-p53 interaction. Mol Cancer Res. 2003; 1(14):1001–8. [PubMed: 14707283]
29. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME,

Dermitzakis ET. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. Feb 9; 2007 315(5813):848–53. automatic medline import. [PubMed: 17289997]

30. Baylin S. Dna methylation and gene silencing in cancer. Nat Clin Pract Oncol. Dec; 2005 2(Suppl 1):S4–11. [PubMed: 16341240]
31. Hervouet E, Vallette FM, Cartron PF. Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation. Epigenetics. Oct; 2009 4(7):487–499. [Online]. Available: http://www.landesbioscience.com/journals/7/article/9883/. [PubMed: 19786833]
32. Castro, M.; Grau, L.; Puerta, P.; Gimenez, L.; Venditti, J.; Quadrelli, S.; Sànchez-Carbayo, M. Multiplexed methylation profiles of tumor suppressor genes and clinical outcome in lung cancer. 2010. p. 86[Online]. Available: http://www.translational-medicine.com/content/8/1/86
33. Lopez-Gines C, Gil-Benso R, Ferrer-Luna R, Benito R, Serna E, Gonzalez-Darder J, Quilis V, Monleon D, Celda B, Cerdá-Nicolas M. New pattern of egfr amplification in glioblastoma and the relationship of gene copy number with gene expression profile. Mod Pathol. 2010
34. Segal MR, Dahlquist KD, Conklin BR. Regression approaches for microarray data analysis. Journal of Computational Biology. 2003; 10:961–980. [PubMed: 14980020]

Biographies

Karl B. Gregory Karl Gregory received a BS degree in economics and statistics from Central Michigan University in 2009 and completed a MS in statistics at Texas A&M University in 2011. He is currently working toward a PhD degree in statistics at Texas A&M University. His research interests are in high-dimensional inference with time series and genomic applications as well as in bootstrap methodology. He spent the summer of 2012 at The University of Texas MD Anderson Cancer Center as a doctoral research intern

Amin A. Momin received his MS in Bioinformatics in 2004 and PhD in Biology in 2010 from Georgia Institute of Technology, Atlanta GA. He currently works as a Platfrom research investigator in the Department of Clinical Cancer Prevention for the Proteomics group. His current research includes identification of plasma cancer markers, novel therapeutic targets and protein modifications.

Kevin R. Coombes is currently a Professor of Biomedical Informatics at The Ohio State University Wexner Medical Center. He received his Ph.D. in pure mathematics from the University of Chicago in 1982. Performing research in algebraic K-theory and arithmetic algebraic geometry, he rose through the academic ranks at MIT, the University of Michigan, and the University of Maryland at College Park. In the late 1990's, he became interested in bioinformatics, which he worked on at the University of Texas M.D. Anderson Cancer Center from 1999–2013. His current research focuses on statistical, mathematical, and computational methods to process, analyze, and understand highly multivariate biological data arising from high throughput technologies. He is particularly interested in (1) methods that incorporate existing biological knowledge early in the analytical process and (2) methods that integrate diverse types of biological data with a view toward predicting clinically relevant patient outcomes. He also specializes in forensic bioinformatics, a subfield that he helped found.

Veerabhadran Baladandayuthapani is currently an Associate Professor of Biostatistics at UT MD Anderson Cancer Center. He received his Ph.D. in Statistics from Texas A&M University and Bachelors (honors) degree in Mathematics from the Indian Institute of Technology, Kharagpur, India. His research interests are in Big data analytics and particularly in developing new statistical frameworks and software for analyzing datasets characterized by high dimensionality and complex structures such as high-throughput genomics, proteomics and imaging. These frameworks include hierarchical and spatial functional data analysis, semi-/non-parametric modeling, non-linear methods for classification and prediction, graphical models and machine learning approaches. His work has been published in top statistical, biostatistical, bioinformatics and biomedical journals. He has also co-authored a book on Bayesian analysis of gene expression data. He also holds several grants from NIH and NSF as PI and co-investigator.

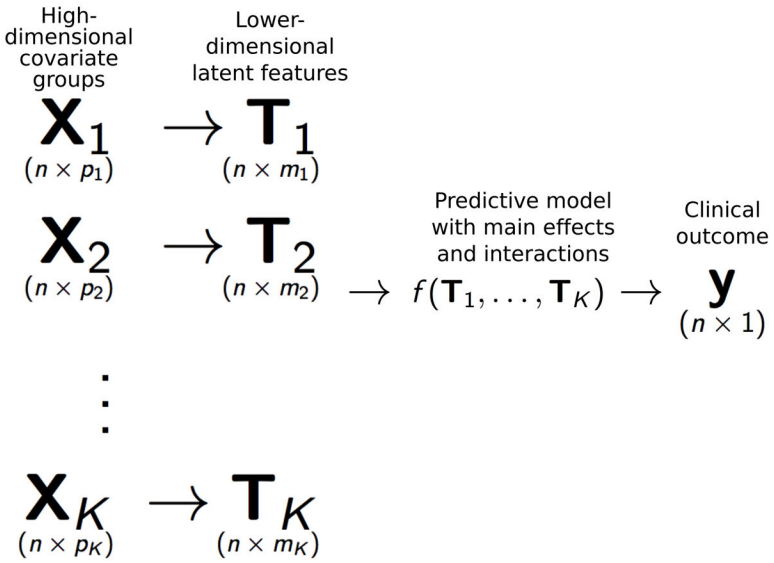

Fig. 1.

High-dimensional measurements from different genomic platforms such as copy number, methylation, and gene expression are decomposed into respective groups of low-dimensional latent features. These latent features and their interactions are used to build a predictive model for the observed clinical outcome such as survival time, dose response, or tumor type.

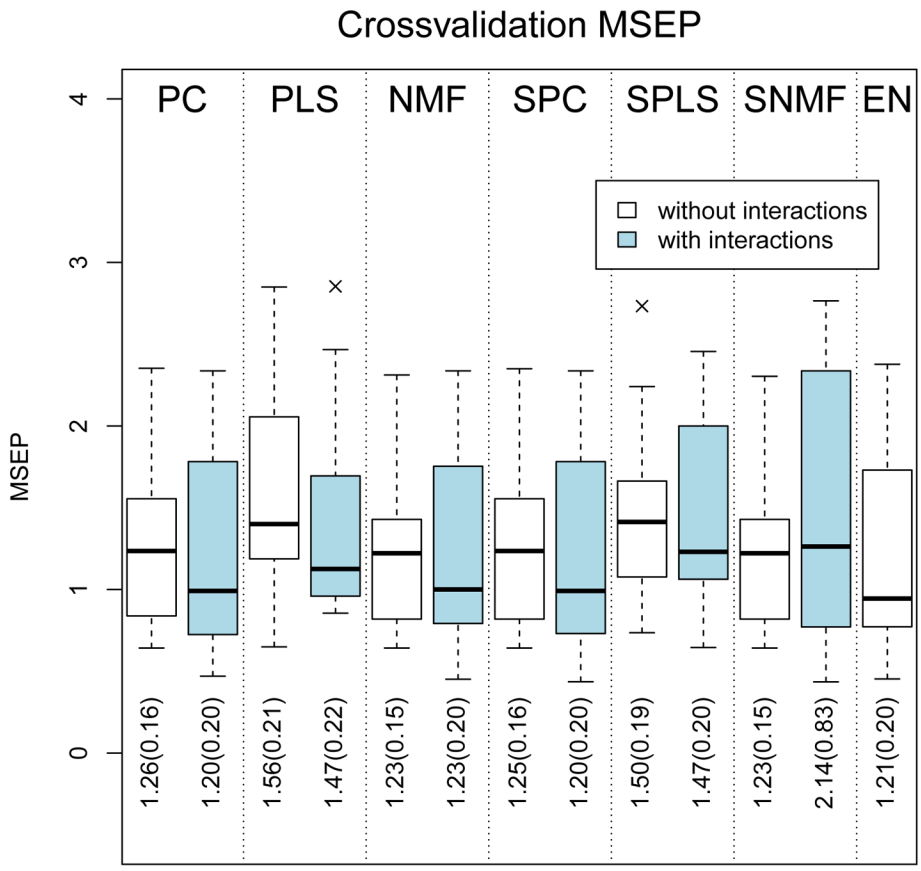

Fig. 2.

Boxplots of mean squared errors of prediction over ten crossvalidation folds for proposed procedure under the six latent feature decompositions with and without including latent feature interactions and the elastic net. Means with standard errors displayed in vertical text beneath boxplots.

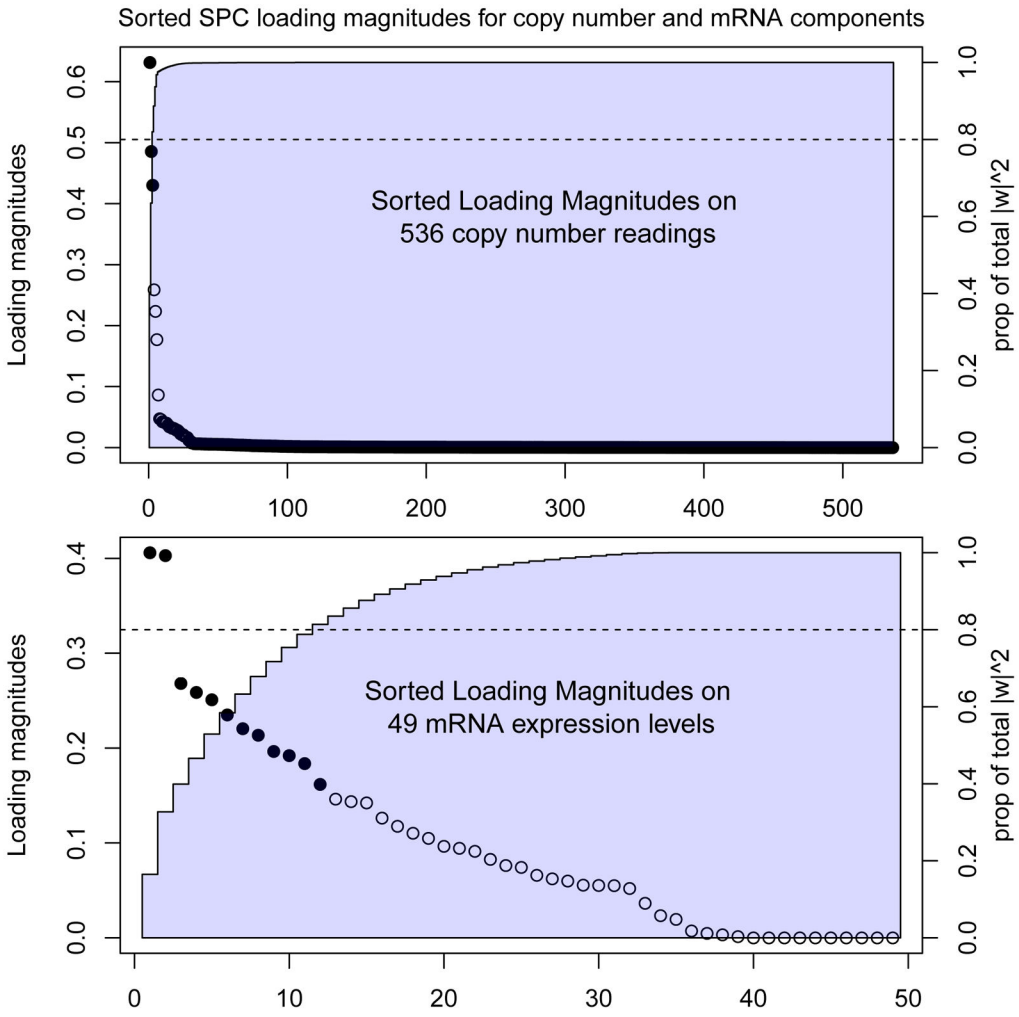

Fig. 3.

Sorted loading magnitudes across copy number and mRNA latent features generated via sparse principal components. Energy plotted with threshold at $\vartheta = .8$ shown. Filled circles correspond to selected variables.

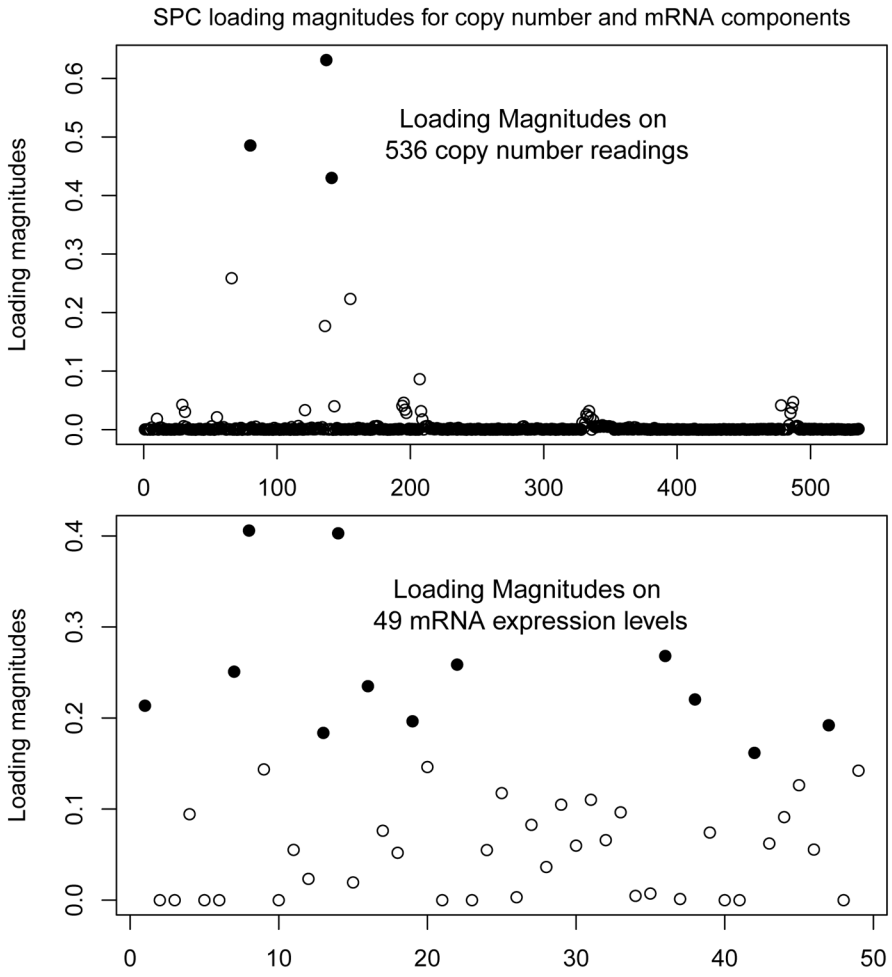

Fig. 4.

Unsorted loading magnitudes across copy number and mRNA latent features generated via sparse principal components. Filled circles correspond to selected variables.

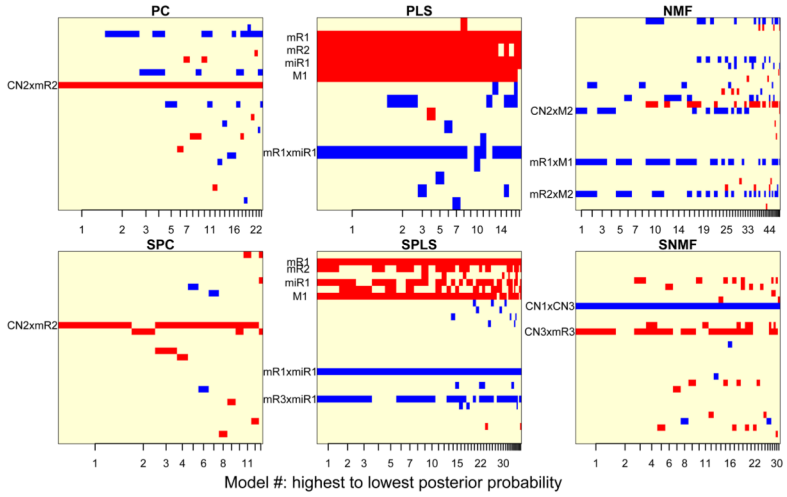

Fig. 5.

Models selected on latent feature main effects and interactions during Bayesian model averaging with model posterior probabilities decreasing from left to right. Each row corresponds to a latent feature and each column to a model. Numerals following platform names (CN = copy number, M = methylation, mR = mRNA, and miR = miRNA) index the latent features, e.g. principal components, derived from that platform. Filled cells indicate inclusion, red and blue signifying positive and negative coefficient estimates, respectively.

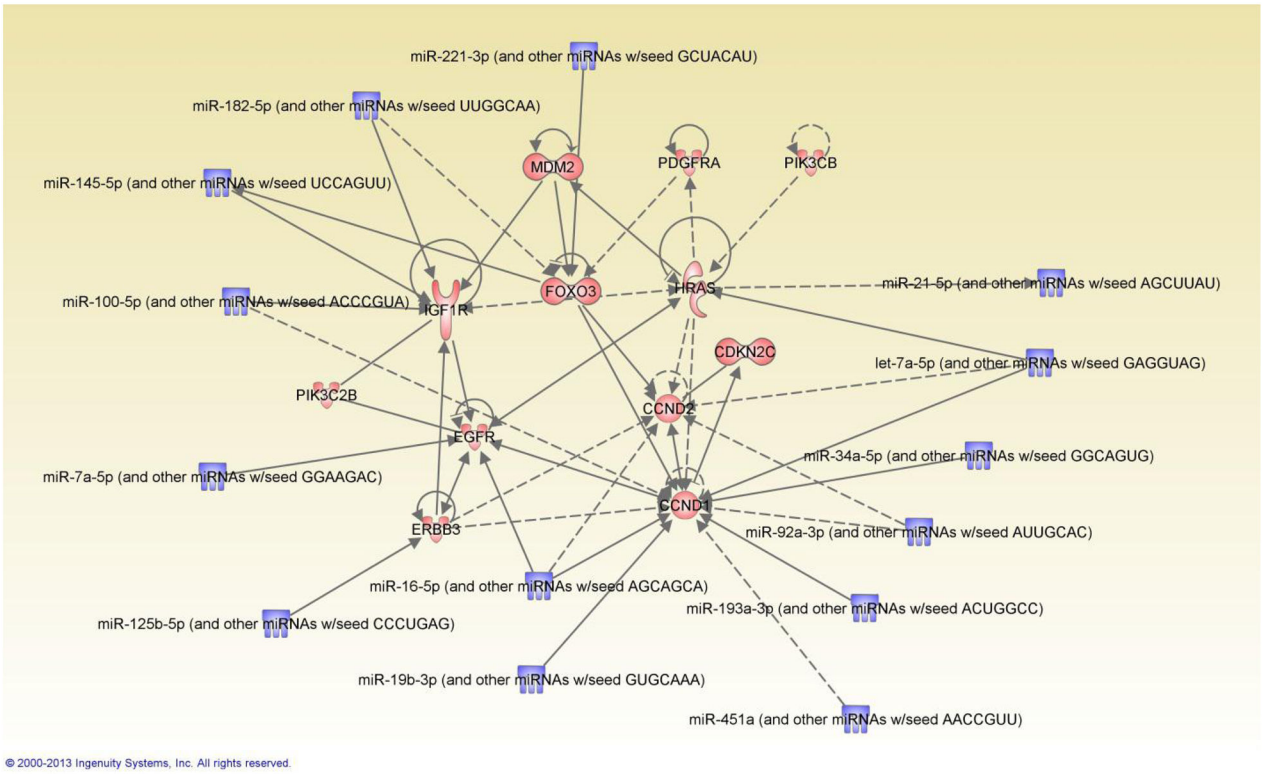

Fig. 6.

The interaction network between genes and miRNAs from the PLS model depicts direct and indirect relationships between genes (red nodes) and miRNAs (blue nodes). The relationships were constructed using IPA (build 192063). The dashed lines indicate indirect relationships while the solid lines depict direct relationships. The miRNA targets are based on experimental validations provided by IPA knowledgebase.

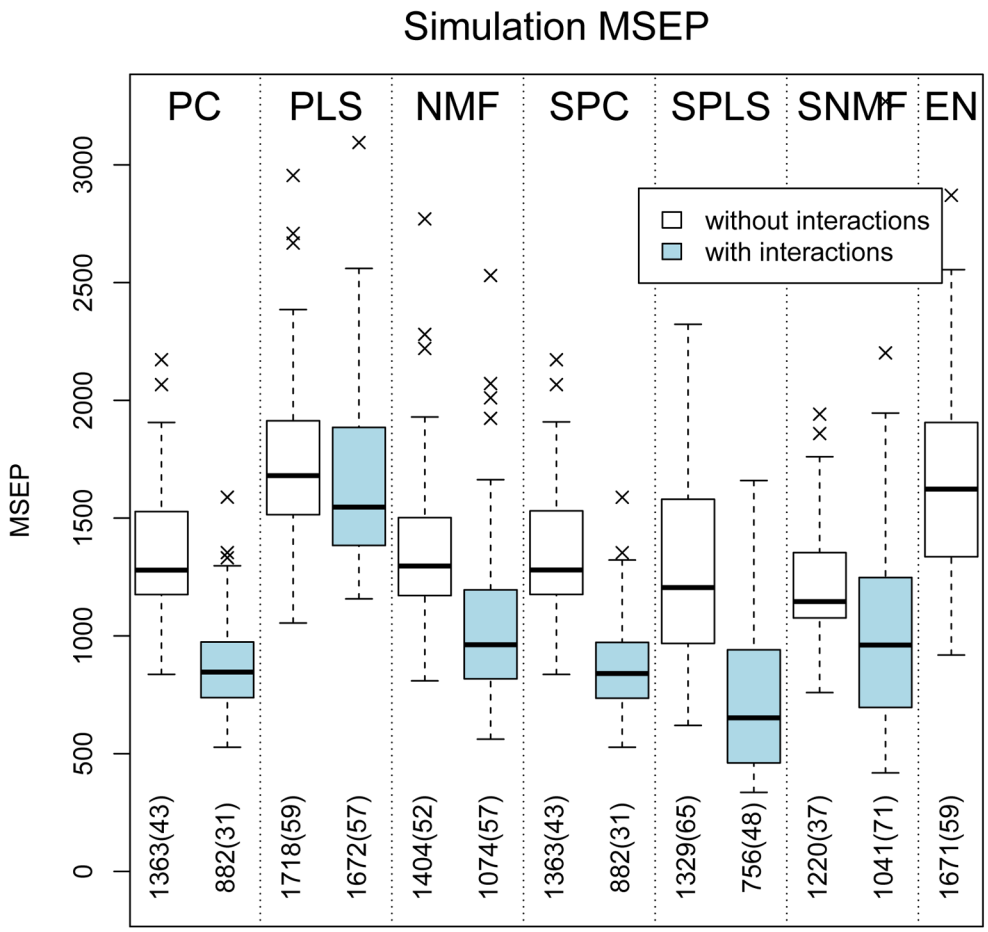

Fig. 7.

Boxplots of mean squared errors of prediction over 50 simulation runs for proposed procedure under the six latent feature decompositions with and without including latent feature interactions and the elastic net. Means with standard errors displayed in vertical text beneath boxplots.

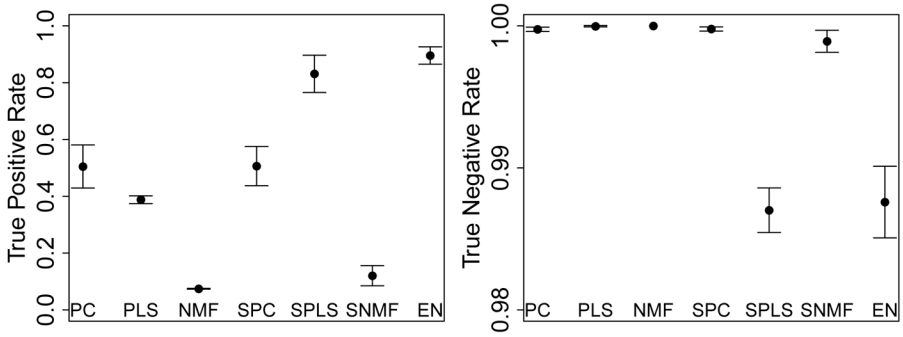

Fig. 8.

True positive and true negative rates with ±1:96*SE* under six latent feature decompositions and the elastic net.

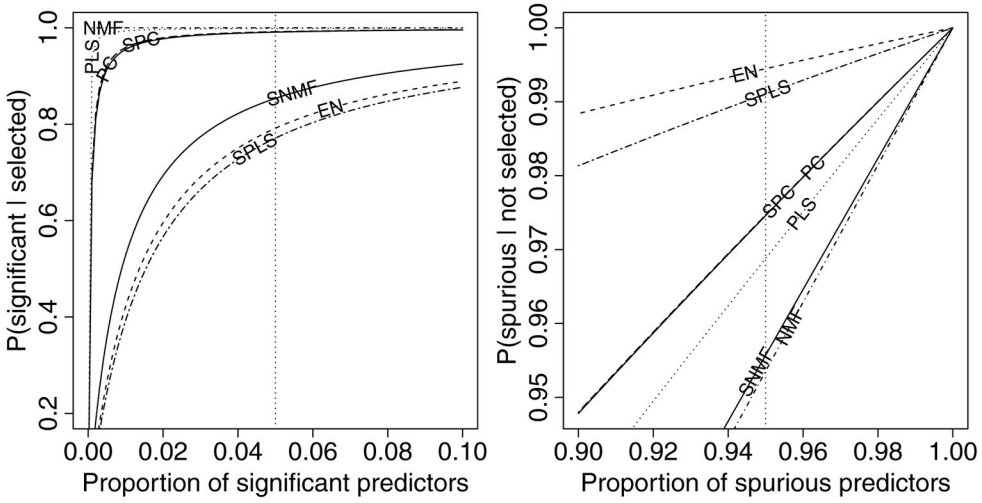

Fig. 9.

Conditional probabilities of significance and spuriousness of a given feature, given that it has been deemed so, across proportions of truly significant features ranging from 0 to .1 for the six decompositions and the elastic net. The vertical line marks simulation conditions.

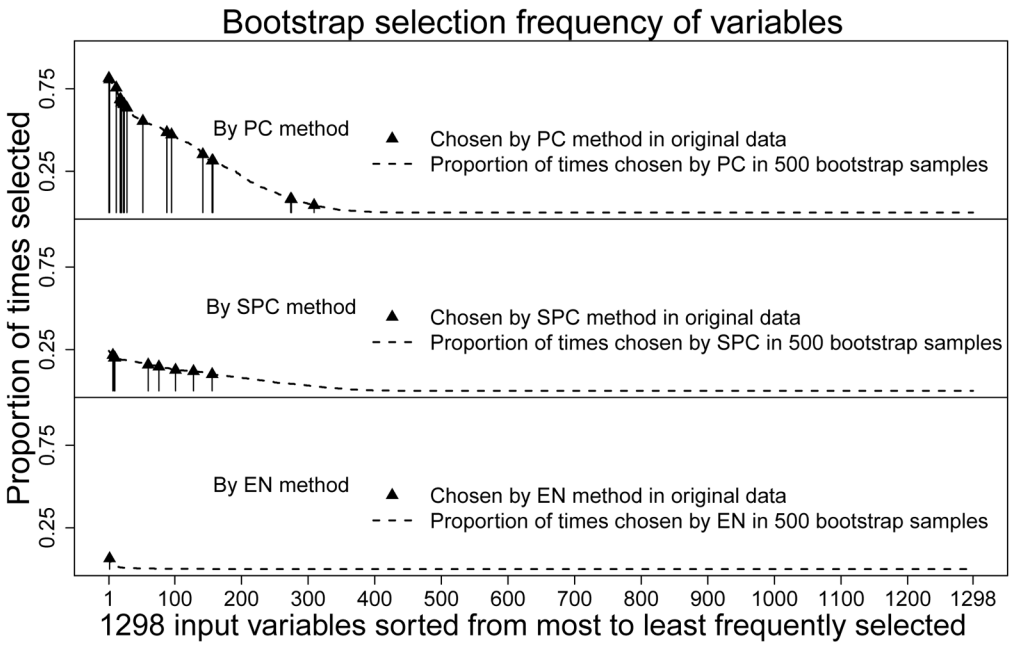

Fig. 10.

Proportion of times each of the 1298 variables was selected across 500 bootstrap samples by the proposed procedure under the PC and SPC decompositions as well as by the EN. Variables are sorted from most to least frequently selected for each method, and vertical arrows mark variables chosen by each method on the original data.

TABLE 1

Models selected by principal components (PC), sparse principal components (SPC), and elastic net models, with selected variables: genes, methylation and copy number probes.

PC & SPC:	$\hat{y} = 5.63 + 0.32 CN_2 mRNA_2$
selected:	*A.14.P132739 A.14.P118388 A.14.P133404 CCND2 EGFR PIK3C2B GAB1 CCND1 PIK3CA AKT1 ERBB3 FOXO1A SPRY2 CDKN2C PIK3R1 CDK4 TP53 FOXO3A*
Elastic Net:	$\hat{y} = 5.63 - 0.004$ hsa-miR-383

selected: hsa-miR-383

TABLE 2

Simulated true positive and true negative rates for identification of significant variables. Standard errors in parentheses.

	PC	PLS	NMF	SPC	SPLS	SNMF	EN
TP	0.5048	0.3880	0.0743	0.5063	0.8310	0.1203	0.8957
se(TP)	(0.0387)	(0.0070)	(0.0008)	(0.0353)	(0.0335)	(0.0181)	(0.0155)
TN	0.9998	1.0000	1.0000	0.9998	0.9870	0.9989	0.9876
se(TN)	(0.0001)	(0.0000)	(0.0000)	(0.0001)	(0.0008)	(0.0004)	(0.0013)