Integrating Multi-Platform Genomic Data Using Hierarchical Bayesian Relevance Vector Machines

Sanvesh Srivastava
Department of Statistics
Purdue University
srivasta@purdue.edu

Wenyi Wang*, Pascal O. Zinn†, Rivka R. Colen¶, Veerabhadran Baladandayuthapani‡,
Departments of Bioinformatics and Computational Biology*, Radiology †, Biostatistics‡,
The University of Texas M.D. Anderson Cancer Center
Department of Neurosurgery, Baylor College of Medicine¶
wwang7@mdanderson.org, zinn@bcm.edu, rrcolen@gmail.com, veera@mdanderson.org

Abstract—We present a statistical framework, hierarchical relevance vector machine (H-RVM), for improved prediction of scalar outcomes using interacting high-dimensional input covariates from different sources. We illustrate our methodology for integrating genomic data from multiple platforms to predict observed clinical phenotypes. H-RVM is a hierarchical Bayesian generalization of the relevance vector machine and its learning algorithm is a special case of the computationally efficient variational method of hierarchic kernel learning framework. We apply H-RVM to data from the Cancer Genome Atlas based Glioblastoma study to predict imaging-based tumor volume by integrating gene and miRNA expression data and show that H-RVM performs much better in prediction as compared to competing methods.

Keywords—Bayesian modeling; multiple kernel learning; genomics; high-dimensional data analysis; prediction

I. INTRODUCTION

Recent biotechnological advances, like microarrays and next-generation sequencing, have enabled measurements of genomic activities at a very detailed resolution (e.g., base pair, SNP). Typically, genomic data are collected from these technologies for a limited number of individuals to relate phenotypes (e.g., patient outcomes) with genomic measurements for future therapeutic studies. Further, genomic data are now available from multiple platforms and resolutions for the same individual. For example, data have been collected in The Cancer Genome Atlas (TCGA) project, from multiple platforms that measure genomic activities such as gene expression, DNA copy number, miRNA expression from multiple tumor types (see http://cancergenome.nih.gov for more details). Magnetic resonance imaging (MRI) is a technology for measuring imaging-based cancer phenotypes, like tumor volume, and radiogenomics is a relatively new field that relates genomic activities/regions with such MRI-based phenotypes (for details see [1] and references therein). Under the assumption that these multiple sources of genomic activities interact and affect the observed phenotype, it is appropriate to model not only their marginal effects but also their interactions for prediction. In this work we present H-RVM framework for predicting scalar outcomes (phenotypes) using multiple sources of interacting high-dimensional covariates. We illustrate the use of H-RVM for integrating genomic data from multiple platforms and for predicting the observed phenotypes, by using a TCGA-based Glioblastoma (GBM) radiogenomics study [1] (herein GBM-MRI data) and by modeling the interaction between the gene and miRNA expression to predict the MRI-based tumor volumes.

One of the main challenges in modeling the statistical dependence in such high-throughput studies is that a large number of covariates (usually in tens/hundreds) are available for a relatively small number (usually in tens/hundreds) of samples (e.g., patients); therefore classical statistical approaches, such as linear models and hierarchical clustering, are prone to over-fitting [2], [3]. In these situations, [3] recommends accounting for high-dimensionality by using approaches that borrow information across covariates to compensate for the limited information available across samples, which leads to better and more reliable inference. Current statistical approaches in radiogenomics use linear parametric models and hierarchical clustering for inferring the relation between MRI phenotypes and genomic features [4]. Although these approaches are computationally efficient, interpretable, and simple, they are unreliable for two main reasons. First, due to the parametric, independence (i.e. non-interacting covariates), and linear assumptions, the sampling distributions of the test statistics, may not be justified in high-dimensional settings [3]. Second, the patterns observed through classical clustering approaches could overfit for these data [2] due to the presence of these higher order interactions [3]. In addition, integrating multiple sources of diverse genomic data leads to genomic discoveries that are better predictive of MRI phenotypes [1].

The H-RVM framework presented here: i. models the relation between a scalar outcome (tumor volume) and high-dimensional covariates (gene and miRNA expression) through a data-adaptive and flexible nonparametric approach, ii. borrows information among covariates through a hierarchical Bayesian framework, iii. acknowledges and models the interaction between covariates through nonlinear functionals, iv. has parameters that have direct interpretability in terms of which sets of covariates (genes, miRNAs, and their interactions) have more profound effect on the
outcome, and uses a computationally efficient variational Bayes approach [5], and thus easily scales to large datasets.

We derive H-RVM as a generalization of the relevance vector machine (RVM) [6] to accommodate interactions between the input covariates and induce dependence and show that its learning/estimation algorithm is a special case of hierarchic kernel learning framework (HKL) [7] in Section II, apply H-RVM to the GBM-MRI data in Section III, and propose extensions of H-RVM framework in Section IV.

II. Hierarchical Relevance Vector Machine

A. Notation and Terminology

Data for H-RVM can potentially be from multiple (>3) sources, but for ease of exposition we assume that the data are triplet: a scalar outcome – tumor volume as measured by MRI, gene and miRNA expression measurements for N patients. We will denote column vectors (matrices) in bold lowercase (uppercase) alphabets. The training data are t with N rows for tumor volume, X and Y for column-centered and standardized gene and miRNA expression each with N rows and G and M columns, respectively; rows correspond to patients and columns represent the genes and miRNAs, respectively. Centering and standardization remove any systematic mean or scaling effects due to the different data sources. We denote the gene and miRNA expression for i th patient as row vectors (of length G and M, respectively) \( x_i^T = (x_{i1}, \ldots, x_{iG}) \) and \( y_i^T = (y_{i1}, \ldots, y_{iM}) \). The interaction of genes and miRNAs is modeled on the \( \mathcal{X} \times \mathcal{Y} \) domain (\( \mathcal{X} \) and \( \mathcal{Y} \) are the space of gene and miRNA expression) as a vector of length GM. Specifically, the first-order interaction term for i th patient is \( (xy)^T = (x_{i1}y_{i1}, \ldots, x_{i1}y_{iM}, \ldots, x_{iG}y_{i1}, \ldots, x_{iG}y_{iM}) \).

Higher-order interaction terms could be specified analogously, but we do not pursue it here.

B. Hierarchical Bayesian Model For H-RVM

Kernel learning (KL) is an approach for nonparametric classification and regression [5]. Assuming that the relation between t and X has the form

\[
t_i = \alpha_0 + \sum_{j=1}^{N} \alpha_j K(x_j, x_i | \sigma^2),
\]

where \( K(\cdot) \) is a kernel matrix that maps X to its feature space with kernel parameter \( \sigma^2 \) and \( \alpha \)'s are the corresponding weights. Therefore, learning (1) is equivalent to learning the parameters \( \{ \alpha_i \}_{i=1}^{N} \). Multiple Kernel Learning (MKL) extends (1) by replacing K with a weighted average of L kernel matrices \( \{ K_l \}_{l=1}^{L} \),

\[
t_i = \alpha_0 + \sum_{j=1}^{N} \alpha_j \sum_{l=1}^{L} \beta_l K_l(x_j, x_i | \sigma_l^2).
\]

MKL improves the flexibility of KL by using L kernel parameters \( \{ \sigma_l^2 \}_{l=1}^{L} \) and weights \( \{ \beta_l \}_{l=1}^{L} \). A variety of approaches exist to learn the kernel parameters, weights, and \( \{ \alpha_i \}_{i=0}^{N} \) for MKL (for details see [7], [5], [8]).

H-RVM extends the KL framework (1) by positing different kernels for each input matrix (gene, miRNA, and their interaction) using different weighting schemes of MKL (2). Although our methodology is independent of the choice of kernels, in this work we use gaussian radial basis function (RBF) kernels [5] for gene (K_1), miRNA (K_2), and their interaction (K_3), defined in the following manner for the \( (i,j) \) th features,

\[
(K_s)_{ij} = e^{-\frac{||x_i - x_j||^2}{2\sigma^2}} = K(\bullet_i, \bullet_j),
\]

\( * = 1, 2, 3 \) and \( \bullet = x, y, (x,y) \), respectively.

Thus \( K_0, K_1, K_2 \) and \( K_3 \) are the predictors that correspond to gene, miRNA, and their interaction, where \( \alpha^T = (\alpha_0, \ldots, \alpha_N) \) are the corresponding regression coefficients.

Note that we slightly abuse the notation \( \{ K_l \}_{i=1}^{N} \) here, which have an extra row of 1’s for the intercept appended to the kernels in (3); therefore \( \{ K_l \}_{i=1}^{N} \) here have dimensions \( (N+1) \times N \). Using (2), we combine \( \{ K_l \}_{i=1}^{N} \) through the weight vector \( \beta^T = (\beta_1, \beta_2, \beta_3) \) to predict t with the constraint that \( \sum_{m=1}^{3} \beta_m = 1 \), and

\[
t = (\beta_1 K_0^T + \beta_2 K_1^T + \beta_3 K_2^T) \alpha = K^T \alpha.
\]

The convexity constraint \( \sum_{m=1}^{3} \beta_m = 1 \) ensures the joint kernel is positive definite and \( \beta_i \) denotes the influence of \( i \) th source in predicting tumor volume. Although similar to (2), (4) differs in two important ways. First, (4) obtains kernels using (1) for different data sources, namely gene and miRNA expression, instead (2) uses kernels with different kernel parameters from the same data source. Second, we allow for dependence between data sources via the interaction kernel (K_3), but MKL does not; instead it uses a convex combination of the different kernels using the weights (\( \beta \)) to aid prediction. These two modifications allow H-RVM to generalize (1) and combine the learning framework of (2) to accommodate interactions.

H-RVM learns parameters \( \alpha \) and \( \beta \) in (4) from t, X, Y, but instead of using MKL methods, we will reformulate (4) as a hierarchical Bayesian model. This reformulation facilitates i. the interpretation of H-RVM as a hierarchical Bayesian extension of RVM [6], which is a special case of Bayesian KL and ii. the use of variational learning algorithm of HKL [7] for H-RVM, since we condition on known kernels. The Bayesian model for H-RVM can be summarized as,

\[
t \mid \alpha, \beta, \gamma, X, Y \sim N(t \mid K_0^T \alpha, \gamma^{-1} I),
\]

\[
\gamma \sim \text{Gamma}(\gamma | c_\gamma, d_\gamma),
\]

\[
\alpha \mid \phi \sim N(\alpha \mid 0, \text{diag}(\phi^{-1})),
\]

\[
\phi \equiv \{ \phi_i \}_{i=0}^{N} \sim \prod_{i=0}^{N} \text{Gamma}(\phi_i | c_\phi, d_\phi),
\]

\[
\beta \sim \text{Dirichlet}(\beta \mid a_1, a_2, a_3),
\]
RVM assumes that the expectations, shape and rate parameters (a priori) predictive power of $i$th covariates (i.e., gene and miRNA measurements) for tumor volume. Small $\alpha_i$ and large $\phi_i$ indicate low predictive power. The covariates with high $\alpha_i$'s a posteriori are the relevance vectors: covariates that are most predictive of tumor volume. $\gamma$ is the precision of the error distribution.

The interpretation of the weights, $\beta^T = (\beta_1, \beta_2, \beta_3)$, is the following: if we fix $\beta_1 = 1$, then (5) – (8) correspond to a RVM that predicts $t_i$ (tumor volume) with $x_i$ (gene expression) as covariate. Similarly, if we separately fix $\beta_2 = 1$ and $\beta_3 = 1$, then (5) – (8) correspond to RVMs that predict $s_i$ (tumor volume) using $y_i$ (miRNA expression) and $(x,y)_i$ (gene and miRNA interaction) as covariates, respectively (for further details see [6]). H-RVM introduces another hierarchy and combines these three RVMs as a weighted average with the weights generated from a Dirichlet distribution (9). Because of this extra level of hierarchy, H-RVM extends and is more flexible than RVM for predicting tumor volume using gene and miRNA expression and their interactions (see Section III).

### C. Variational Inference For H-RVM

We use HKL [7] for learning the posterior distributions of parameters for H-RVM. Following [7], we have two sets of parameters $\Theta = \{\alpha, \beta, \phi, \gamma\}$ and $\Psi = \{c_\phi, d_\phi, c_\gamma, d_\gamma, a_1, a_2, a_3\}$ where $\Theta$ are the model parameters and $\Psi$ are the hyperparameters. Contrary to HKL, H-RVM assumes that $a_1, a_2,$ and $a_3$ are hyperparameters. We use the form of variational posterior as

$$p(\Theta, \Psi, t, X, Y) \approx Q(\Theta) = Q(\alpha)Q(\beta)Q(\phi)Q(\gamma).$$

We have removed the conditioning on $\Psi, t, \text{ and } \{K_i\}_{i=1}^3$ in $Q$s and will denote $E_x[f]$ as $\langle f \rangle_x$ for simplicity. Following [7], the variational posterior distributions are available for $\alpha, \phi,$ and $\gamma$, as $Q(\alpha) = \mathcal{N}(\mu_\alpha, \Sigma_\alpha)$ where

$$\mu_\alpha = \langle \gamma \rangle \gamma N_\beta K_\beta^T K_\beta^{-1} \phi_0^T \phi_0 + \text{diag}(\langle \phi \phi \rangle)$$

$$\Sigma_\alpha = \langle \gamma \rangle \gamma ^2 \Sigma_\alpha K_\beta^T K_\beta \phi_0$$

and for $i = 0, \ldots, N$ and $j = 1, \ldots, N$,

$$Q(\phi_i) = \text{Gamma} \left( \frac{1}{2} + c_\phi, \frac{1}{2}(\alpha_i^2 + \phi_0^T) \right)$$

$$Q(\gamma) = \text{Gamma} \left( \frac{N}{2} + c_\gamma, \frac{1}{2}(\|e\|^2 + \gamma_d) \right)$$

$$\|e\|^2 = \|t - K_\beta^T \alpha\|^2.$$  (13)

The expectations, $\langle \alpha \rangle, \langle \alpha \alpha^T \rangle, \langle \phi \rangle,$ and $\langle \gamma \rangle,$ are available from (11) – (13) as in [7]. Instead of $Q(\beta)$, its non-normalized version $Q^*(\beta)$ is available from [7] as

$$Q^*(\beta) = \prod_{m=1}^3 \beta_m^{a_m - 1} \exp \left( - \frac{\Omega_\beta - 2 \beta^T b_m}{2} \right).$$

We use the importance sampling method of [7] to calculate $\langle \beta \rangle, \langle \log \beta \rangle,$ and $\langle \beta^T \beta \rangle$. Further, we estimate $\Psi$ by the type II maximum likelihood procedure as recommended in [7]. The kernel parameters $\{\sigma_i^2\}_{i=1}^3$ are learned from three RVMs for each of the three sources through crossvalidation recommended by [6].

### III. DATA ANALYSIS

In the GBM-MRI data [1] we have information from 75 patients on their gene expressions (22267 probes), miRNA expressions (1510 probes), age, and tumor volume. The tumor volume is measured by quantitative MRI volumetrics [9]. Although we can use all the probes, we preselect most important 1000 gene and 200 miRNA probes via univariate selection to estimate their effects on tumor volume. Because our interest is in integrating genomic data from multiple platforms, herein we remove the age effect after univariate regression and use H-RVM to model the residual log$_2$ tumor volume ($t$) using column-centered and column-standardized gene and miRNA expression matrices (II-A).

H-RVM learns $\alpha$ and $\beta$ using HKL (II-C), and $\beta$ shows that gene expression, miRNA expression, and their interaction contribute 20%, 36%, and 44% in predicting $t$, respectively — thus indicating that the interaction term is more predictive than the individual components. Similar to [6], we also obtain the relevance vectors for H-RVM that are the covariates for the individuals (II-B) with nonzero $\alpha_i$ and are influential in predicting $t$. The $\alpha_i$'s for 44th, 50th, and 53rd individual are outliers based on the boxplot of $\alpha$, indicating that they have significant effect on $t$ compared to other individuals.
We further compare the predictions and residual errors of H-RVM with RVMs using following models: i. gene-only RVM (Gene-RVM), ii. miRNA-only RVM (MiRNA-RVM), iii. interaction-only RVM (Interaction-RVM), and with iv. the lasso [2] with gene and miRNA expressions as linear predictors. Figure 1 shows the prediction results for these five models. We find that H-RVM’s predictions closely agree with true tumor volumes $t$ since the linear fit between them has slope and intercept close to 1 and 0, respectively. This observation is further confirmed by H-RVM’s mean squared error, 0.004, that is lowest among five models. The next highest slope is for Gene-RVM and its value is 0.73, which shows that it overall under-predicts $t$. Similar observation holds for the other three models. Figure 2 shows the boxplots of residual errors for the five models. H-RVM’s residual error distribution has noticeably smaller spread and range compared with other models, which is due to H-RVM’s good predictive performance. Gene-RVM’s and Interaction-RVM’s residual errors have smaller spread and a better fit compared with that for lasso or MiRNA-RVM. This analysis shows that H-RVM has a better predictive performance than the lasso or RVMs using single platform data for GBM-MRI data. Further, because we model log$_2$ tumor volume, the gain for tumor volume predictions are exponentially higher for H-RVM compared with its competitors. Therefore, integrating multi-platform genomic data to predict $t$ through the H-RVM framework is recommended for the GBM-MRI data.

IV. CONCLUSIONS AND FUTURE WORK

In this work we have presented the H-RVM framework that generalizes the MKL framework for integrating high-dimensional data from multiple sources to predict scalar response. We applied H-RVM to a high-dimensional TCGA-based GBM study to predict tumor volume (phenotype) using two data sources: gene and miRNA expressions, and found that H-RVM’s predictive performance is better than KL methods and the lasso that ignore interactions. We hypothesize that H-RVM gains its flexibility and power by modeling the interaction between gene and miRNA expressions. We believe H-RVM might be a useful tool for clinicians, since increased accuracy of noninvasive MRI screens using genomic events is an important problem in GBM.

Although we have presented the application of H-RVM in the context of radiogenomics, the framework is general and can be extended and adapted to data from multiple platforms and with different distributional assumptions. We used the computationally efficient variational method of HKL, which is extremely useful for handling large data from projects such as TCGA. In addition, [8] presents more scalable versions of HKL and MKL that can be adapted to our framework. We have a freely available code for fitting H-RVM at: http://odin.mdacc.tmc.edu/~vbaladan.

While we use an approximate Bayesian learning method, the accuracy of inference for H-RVM can be improved by i. using MCMC approaches that obtain the posterior distribution of the unknown parameters and account for the uncertainty in model parameters and specification, ii. modeling the uncertainty in the scale parameter of the kernels, and iii. exploring the sensitivity of the results to the choice of kernels. Our future work will concentrate on extending the H-RVM framework using Bayesian spike and slab priors to select variables (genes and miRNAs) from the interacting covariates before embedding the data in the space of kernels. This will aid our model interpretation as to which genes might be important players in predicting phenotypes.

ACKNOWLEDGMENT

Research supported by NIH R01 CA160736 and NSF IIS-0914861 (to VB) and CCSG P30 CA016672 (to WW).

REFERENCES

[1] P. O. Zinn, B. Mahajan, ..., and R. R. Colen, “Radiogenomic mapping of edema/cellular invasion mri-phenotypes in glioblastoma multiforme,” PLoS One, vol. 6, no. 10, p. e25451, 2008.
[2] D. Witten and R. Tibshirani, “A framework for feature selection in clustering,” Journal of the American Statistical Association, vol. 105, no. 490, pp. 713–726, 2010.
[3] B. Efron, Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge Univ Pr, 2010.
[4] M. Diehn, C. Nardini, ..., and M. Kao, “Identification of noninvasive imaging surrogates for brain tumor gene-expression modules,” Proceedings of the National Academy of Sciences, vol. 105, no. 13, p. 5213, 2008.
[5] C. M. Bishop, Pattern recognition and machine learning. Springer New York, 2006, vol. 4.
[6] M. Tipping, “Sparse bayesian learning and the relevance vector machine,” The Journal of Machine Learning Research, vol. 1, pp. 211–244, 2001.
[7] M. Girolami and S. Rogers, “Hierarchic bayesian models for kernel learning,” in Proceedings of the 22nd International Conference on Machine Learning, 2005.
[8] M. Gönül, “Bayesian efficient multiple kernel learning,” in Proceedings of the 29th International Conference on Machine Learning, 2012.
[9] P. O. Zinn, P. Sathyan, ..., and R. R. Colen, “A novel volume-age-kps (vak) glioblastoma classification identifies a prognostic cognate microrna-gene signature,” PLoS ONE, vol. 7, no. 8, p. e41522, 08 2012.