Detection of Viruses via Statistical Gene-Expression Analysis

[1]Minhua Chen, [1]David Carlson, [2]Aimee Zaas, [2]Christopher Woods, [2]Geoffrey S. Ginsburg, [2]Joseph Lucas, and [1]Lawrence Carin

[1]Duke University, Electrical and Computer Engineering Department
[2]Institute for Genome Sciences & Policy, Department of Medicine, Duke University

***Abstract*— We develop a new Bayesian construction of the elastic net, with variational Bayesian analysis. This modeling framework is motivated by analysis of gene-expression data for viruses, with a focus on H3N2 and H1N1 influenza, as well as Rhino virus and RSV (respiratory syncytial virus). Our objective is to understand the biological pathways responsible for the host response to such viruses, with the ultimate objective of developing a clinical test to distinguish subjects infected by such viruses from subjects with other symptom causes ($e.g.$, bacteria). In addition to analyzing these new data sets, we provide a detailed analysis of the Bayesian elastic net, and compare it to related models.**

***Index Terms*— Variable selection, Elastic Net, Grouping effect, Bayesian Lasso, Multi-task learning**

I. Introduction

Gene expression measurements have proven to be valuable tools for medical diagnosis and for investigating fundamental biology [1]. A principal motivation of this paper is the investigation of gene expression data after the human body has interacted with a virus. Specifically, as discussed in detail in Section VII, after receiving institutional review board (IRB) approval, we inoculated human volunteers with influenza, and took blood samples periodically over several days, with the goal of monitoring the gene expression values of these subjects as the virus-body interaction evolved. Roughly half of the subjects ultimately became symptomatic, and a goal of this study is to infer which genes are important for distinguishing those who become symptomatic versus from those who do not. We performed two independent such influenza challenge studies of the type discussed above, and therefore model construction is performed with one data set, and independent validation is performed with the second. One of the challenges was performed with the H3N2 virus strain, and the other with H1N1. We also performed challenge studies with Rhino virus and RSV (respiratory syncytial virus), and analyze gene-expression data for these viruses as well.

To analyze such gene-expression data, we consider linear regression of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{X}$ is the $n \times p$ data matrix, each row corresponding to a sample, $\boldsymbol{\beta}$ is the $p \times 1$ regression coefficient vector, $\boldsymbol{y}$ is the $n \times 1$ regression response, and $\boldsymbol{\epsilon}$ is $n \times 1$ additive noise vector (or model error). Here n is the number of samples and p is the feature dimension; for the gene-expression problem there are $p-1$ genes, with an additional coefficient for the model offset. In a simple example the components of $\boldsymbol{y}$ may take on binary values, related to an outcome (we subsequently generalize the model for binary outcomes through use of a probit link function).

The problem of interest is estimation of $\boldsymbol{\beta}$. When performing linear regression, one is interested in prediction accuracy as well as in model physical/biological interpretation [2]. For example, in gene-expression analysis, since there are typically tens of thousands of genes (p on the order of tens of thousand) and most of them are irrelevant with the response, it is desirable to only include within the model those genes associated with the biology (suggesting a sparseness constraint on the components of $\boldsymbol{\beta}$). Lasso [3] has been proposed to address the feature-selection problem, by imposing a sparseness constraint on the regression coefficients. The objective function of Lasso is

$$\hat{\boldsymbol{\beta}}(\text{LASSO}) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{2}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the ℓ_1 norm of $\boldsymbol{\beta}$. The ℓ_2 norm in the first term accounts for prediction accuracy, while the ℓ_1 penalty term yields a sparse solution.

Lasso has been employed successfully in many applications in signal processing and machine learning [4], [5]. However, if some relevant features are highly correlated with each other, Lasso tends to arbitrarily select only a few of these [2], ignoring the rest. There are two disadvantages of this: (i) arbitrarily ignoring correlated and equally important features degrades model interpretability, and (ii) including only one or a few of the correlated features may undermine robustness. In order to achieve sparse and grouped variable selection simultaneously, Zou and Hastie [2] proposed the following Elastic Net criterion:

$$\hat{\boldsymbol{\beta}}(\text{Naive ENet}) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2 \tag{3}$$

The solution is called a naive Elastic Net estimator, and the final estimator is defined as a re-scaled version of the solution of (3). The combination of the ℓ_1 and ℓ_2 penalties in (3) is a compromise between Lasso and ridge regression, and it also makes the optimization problem strictly convex. Theoretical and geometrical reasons are given in [2] for why this optimization criterion can yield sparse and grouped variable selection. In [2], an efficient algorithm (LARS-EN) is also proposed to solve the optimization problem in (3), where λ_1 and λ_2 are two parameters to be tuned via cross validation.

In this paper we are interested in developing a Bayesian construction of the Elastic Net, to complement a recently developed Bayesian construction of Lasso [6]. Authors have earlier attempted to develop a Bayesian Elastic Net construction [7]. However, their method retained the two aforementioned tuning parameters λ_1 and λ_2. Below we demonstrate that a relatively simple extension of the Bayesian Lasso in [6] may be employed to yield a Bayesian Elastic Net, and in so doing one of the tuning parameters may be removed analytically. This simplifies practical model usage significantly, in that setting of tuning parameters is easier.

After developing the Bayesian Elastic Net, it is relatively straightforward to utilize it in typical hierarchical models, that

allow enhanced modeling flexibility. Specifically, for classification problems it is also useful to use the response y in (1) as an input to a logistic or probit link function [8], [9]. In addition, there are many problems in which we are interested in performing multiple distinct but related regressions. For example, one may be interested in developing a probit-regression classifier for gene-expression data, using a Bayesian Elastic Net model to infer a set of important and possibly correlated genes. The data used to develop *multiple* such models may be highly related, and this should be exploited within the analysis.

The problem of jointly learning multiple models, exploiting shared information, has been referred to as multi-task learning [10], [11]. We here extend the Bayesian Elastic Net to a multi-task setting. After validating the model based on published gene-expression data, we apply it to our motivating objective of analyzing expression data associated with virus-body interaction, using new data from our challenge studies.

The remainder of the paper is organized as follows. In Section II we review Bayesian Lasso and an earlier version of the Bayesian Elastic Net, this followed by introduction of our proposed model; we also discuss extension of the model to a probit construction. The multi-task framework is discussed in Section III, and the posterior density function is estimated efficiently via variational Bayesian analysis, as discussed in Section IV. In Section V we provide a discussion of how to set model parameters. Several results are presented in Section VI using previously published data. In Section VII we present results on a new and motivating data sets we have collected, on gene-expression data for influenza, Rhino virus and RSV. Conclusions are discussed in Section VIII.

II. Bayesian Sparse Linear Linear Regression

A. Bayesian View of Lasso and Elastic Net

From a Bayesian viewpoint, the Lasso model in (2) can be interpreted as a maximum *a posterior* (MAP) estimator with Laplace priors placed independently on the components of $\boldsymbol{\beta}$ [6], [12]. In order to make the model inference tractable, the Laplace prior is written in the following hierarchical form [6], [12], [13]:

$$\begin{aligned} p(\boldsymbol{\beta}|\tau,\boldsymbol{\gamma}) &= \prod_{j=1}^{p} \frac{\sqrt{\gamma_j\tau}}{2} \exp\left(-\sqrt{\gamma_j\tau}|\beta_j|\right) \\ &= \prod_{j=1}^{p} \int \mathcal{N}(\beta_j; 0, \tau^{-1}\alpha_j^{-1}) \text{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right) d\alpha_j \end{aligned} \tag{4}$$

Here τ is the precision of the additive noise in (1), α_j is the latent precision parameter for β_j, and InvGa$(\cdot)$ denotes the inverse Gamma distribution. Further, Gamma priors may be imposed on individual Lasso parameters γ_j. The complete Bayesian Lasso model is expressed as [6], [12]

$$\begin{aligned} \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}\boldsymbol{I}) \\ \beta_j &\sim \mathcal{N}(\beta_j; 0, \tau^{-1}\alpha_j^{-1}) \\ \tau &\sim \text{Ga}(\tau; c_0, d_0) \\ (\alpha_j, \gamma_j) &\sim \text{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right) \text{Ga}(\gamma_j; a_0, b_0) \end{aligned} \tag{5}$$

for $j = 1, 2, \cdots, p$. It should be noted that [12] and [6] start with a simpler model where no prior is imposed on γ_j. Instead, a fixed parameter λ replaces γ_j. The model in (5) is proposed as an extension to the model in [12] by using feature-specific hyperparameters. The hierarchical prior on β_j is also called normal-exponential-gamma distribution [14]. Defining $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \cdots, \gamma_p]^\top$ and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_p]^\top$, the full likelihood of Bayesian Lasso becomes

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}\boldsymbol{I}) \text{Ga}(\tau; c_0, d_0) \\ &\times \prod_{j=1}^{p} \mathcal{N}(\beta_j; 0, \tau^{-1}\alpha_j^{-1}) \text{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right) \text{Ga}(\gamma_j; a_0, b_0) \end{aligned} \tag{6}$$

For the hyperparameters we typically choose $a_0 = b_0 = c_0 = d_0 = 10^{-6}$, resulting in noninformative Gamma priors. Both Markov Chain Monte Carlo (MCMC) and variational Bayesian (VB) inference algorithms can be derived for the above model [6], [12].

To see more clearly why the above model is the Bayesian version of Lasso, we can integrate out $\boldsymbol{\alpha}$ and the likelihood becomes

$$p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\gamma}) \propto f(\tau, \boldsymbol{\gamma}) \ \exp(-\frac{\tau}{2}(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} 2\sqrt{\gamma_j\tau^{-1}}|\beta_j|)) \tag{7}$$

where $f(\tau, \boldsymbol{\gamma}) = \tau^{\frac{n+p}{2}} \text{Ga}(\tau; c_0, d_0) \prod_{j=1}^{p} \gamma_j^{\frac{1}{2}} \text{Ga}(\gamma_j; a_0, b_0)$. Thus $2\sqrt{\gamma_j\tau^{-1}}$ plays the role of λ in (2). We observe that the log likelihood term is analogous to the optimization criterion of Lasso, except that here adaptive weights are used for penalizing different regression coefficients. This feature-specific shrinkage is also adopted in adaptive Lasso [15].

A natural question arises as to whether a Bayesian Elastic Net model exists, which not only shares the advantages of Bayesian Lasso but also performs grouped variable selection. This problem was first studied in [7], where the Bayesian Elastic Net model was proposed as

$$\begin{aligned} \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}\boldsymbol{I}) \\ \beta_j &\sim \mathcal{N}(\beta_j; 0, \tau^{-1}(\alpha_j + \lambda_2)^{-1}) \\ \tau &\sim \text{Ga}(\tau; c_0, d_0) \\ \alpha_j &\sim \eta\,(\alpha_j/(\alpha_j + \lambda_2))^{\frac{1}{2}} \text{InvGa}\left(\alpha_j; 1, \frac{\gamma}{2}\right) \end{aligned} \tag{8}$$

with $j = 1, 2, \cdots, p$; λ_2 and γ are two parameters to be tuned via cross validation, and η is a normalizing constant. A similar model was also proposed in a recent paper [16]. The full likelihood of this Bayesian Elastic Net model is

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \gamma) \propto \ & \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}\boldsymbol{I}) \text{Ga}(\tau; c_0, d_0) \\ &\times \prod_{j=1}^{p} \mathcal{N}(\beta_j; 0, \frac{\tau^{-1}}{\alpha_j + \lambda_2}) \sqrt{\frac{\alpha_j}{\alpha_j + \lambda_2}} \text{InvGa}\left(\alpha_j; 1, \frac{\gamma}{2}\right) \end{aligned} \tag{9}$$

Again, by integrating out $\boldsymbol{\alpha}$, the likelihood becomes

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\gamma}) &\propto f(\tau, \boldsymbol{\gamma}) \\ \times \ \exp\left(-\tfrac{\tau}{2}\left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + 2\sqrt{\gamma\tau^{-1}}\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\right)\right) & \end{aligned} \tag{10}$$

where $f(\tau,\boldsymbol{\gamma}) = \tau^{\frac{n+p}{2}}\text{Ga}(\tau;c_0,d_0)$. Here $2\sqrt{\gamma\tau^{-1}}$ plays the role of λ_1 in (3). MCMC inference for this model is derived in [7], [16] and promising results are reported for gene selection.

Although theoretically valid, empirical result shows that the Bayesian Elastic Net model in (8) does not yield a solution that is particularly sparse. In this paper, a new Bayesian Elastic Net model is proposed based on the model in (8). Instead of using only one parameter γ for all inverse Gamma distribution, we introduce different γ_j for each precision parameter α_j, and further impose Gamma priors on them, which is analogous to the Bayesian Lasso model in (5). In this way we achieve sparsity and grouped variable selection simultaneously. As a byproduct, we reduce the number of tuning parameters from two to one, since for the new model only λ_2 needs to be tuned (we effectively integrate out λ_1). Another contribution of this paper is that a variational Bayesian solution is derived for the new Bayesian Elastic Net model, which is computationally more efficient than MCMC (and convergence is much easier to diagnose, as discussed further below).

B. Proposed Bayesian Elastic Net

Based on the model in (8), we propose a new Bayesian Elastic Net model:

$$\begin{aligned} \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{y};\boldsymbol{X}\boldsymbol{\beta},\tau^{-1}\boldsymbol{I}) \\ \beta_j &\sim \mathcal{N}(\beta_j;0,\tau^{-1}(\alpha_j+\lambda_2)^{-1}) \\ \tau &\sim \text{Ga}(\tau;c_0,d_0) \\ (\alpha_j,\gamma_j) &\sim \eta\sqrt{\frac{\alpha_j}{\alpha_j+\lambda_2}}\text{InvGa}\left(\alpha_j;1,\frac{\gamma_j}{2}\right)\text{Ga}(\gamma_j;a_0,b_0) \end{aligned} \tag{11}$$

again for $j=1,2,\cdots,p$, and with η a normalizing constant. The difference between the above Bayesian Elastic Net and that proposed in [7], [16] is that we impose a Gamma prior on individual γ_j to avoid tuning, which we have found to yield sparse and grouped solutions. The full likelihood of the model is

$$\begin{aligned} p(\boldsymbol{y},\boldsymbol{\beta},\tau,\boldsymbol{\alpha},\boldsymbol{\gamma}) &\propto \mathcal{N}(\boldsymbol{y};\boldsymbol{X}\boldsymbol{\beta},\tau^{-1}\boldsymbol{I})\text{Ga}(\tau;c_0,d_0) \\ &\times\prod_{j=1}^{p}\mathcal{N}(\beta_j;0,\tau^{-1}(\alpha_j+\lambda_2)^{-1})\left(\alpha_j/(\alpha_j+\lambda_2)\right)^{\frac{1}{2}} \\ &\times\text{InvGa}(\alpha_j;1,\frac{\gamma_j}{2})\text{Ga}(\gamma_j;a_0,b_0) \end{aligned} \tag{12}$$

Here λ_2 is a parameter to be tuned by cross validation. One notes that when λ_2 goes to zero, the above Bayesian Elastic Net model reduces to the Bayesian Lasso model in (5). By integrating out $\boldsymbol{\alpha}$, the full likelihood can be expressed as

$$\begin{aligned} p(\boldsymbol{y},\boldsymbol{\beta},\tau,\boldsymbol{\gamma}) &\propto f(\tau,\boldsymbol{\gamma}) \\ \times\exp\left(-\tfrac{\tau}{2}(\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\textstyle\sum_{j=1}^{p}2\sqrt{\gamma_j\tau^{-1}}|\beta_j|+\lambda_2\|\boldsymbol{\beta}\|_2^2)\right) \end{aligned} \tag{13}$$

and $f(\tau,\boldsymbol{\gamma}) = \tau^{\frac{n+p}{2}}\text{Ga}(\tau;c_0,d_0)\prod_{j=1}^{p}\gamma_j^{\frac{1}{2}}\text{Ga}(\gamma_j;a_0,b_0)$. The expression related to $\boldsymbol{\beta}$ has almost the same form as that in (3), except that we assign different weights for individual $|\beta_j|$, so that each β_j receives a different degree of shrinkage. The idea of adaptive shrinkage for the Elastic Net is also exploited in [17].

We note that a conjugate prior is not available for λ_2, which is why we do not integrate it out, like we did λ_1. However, as we discuss when presenting results, we have found the algorithm to not be overly sensitive to the particular setting of λ_2, and therefore we have performed cross-validation here based on a library of possible λ_2.

C. Probit Regression

Many sparse and grouped variable selection problems arise in the form of classification instead of regression. For classification problems, there is no observable regression response $\boldsymbol{y}$; we only have label information $\boldsymbol{z} = [z_1,z_2,\cdots,z_n]^\top$ with $z_i\in\{-1,1\}$, for the binary case, with the basic ideas discussed below extendable beyond binary labels. We extend the model in (11) to the classification problem by introducing a probit link between the *latent* regression response $\boldsymbol{y}$ and the observed label information $\boldsymbol{z}$, similar to the approaches in [12] and [13]. The resulting Bayesian Elastic Net model for probit regression is

$$\begin{aligned} z_i &\sim 1(z_i=\text{sign}(y_i)) \\ \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{y};\boldsymbol{X}\boldsymbol{\beta},\tau^{-1}\boldsymbol{I}) \\ \beta_j &\sim \mathcal{N}(\beta_j;0,\tau^{-1}(\alpha_j+\lambda_2)^{-1}) \\ \tau &\sim \text{Ga}(\tau;c_0,d_0) \\ (\alpha_j,\gamma_j) &\sim \eta\sqrt{\frac{\alpha_j}{\alpha_j+\lambda_2}}\text{InvGa}\left(\alpha_j;1,\frac{\gamma_j}{2}\right)\text{Ga}(\gamma_j;a_0,b_0) \end{aligned} \tag{14}$$

with $i=1,2,\cdots,n; j=1,2,\cdots,p$, and $1(\cdot)$ is an indicator function, which equals 1 if the argument is satisfied and is 0 otherwise. Thus $z_i\sim 1(z_i=\text{sign}(y_i))$ indicates that $z_i=1$ if $y_i\geq 0$ and $z_i=-1$ otherwise. It is common to fix $\tau=1$ [13], which can be implemented by assigning large values to c_0 and d_0 (*e.g.*, we typically employ 10^6).

III. Multi-Task Learning

The proposed Bayesian Elastic Net may be readily extended to a multi-task setting, in which multiple regression or classification tasks are performed jointly, with variables shared among the tasks. This sharing mechanism allows information borrowing among tasks, enhancing learning. For example, the regression coefficients may differ from task to task, but the sparsity pattern of the regression coefficients can be shared, thereby imposing the belief that irrelevant features are the same or similar among all the tasks and can be pruned jointly [10], [11].

Assume M tasks, each represented as a regression model:

$$\boldsymbol{y}^{(m)} = \boldsymbol{X}^{(m)}\boldsymbol{\beta}^{(m)}+\boldsymbol{\epsilon}^{(m)} \quad ; \quad m=1,2,\cdots,M \tag{15}$$

where $\boldsymbol{X}^{(m)}$ is the $n^{(m)}\times p$ design matrix for task m and each row of $\boldsymbol{X}^{(m)}$ corresponds to a sample; $n^{(m)}$ is the number of samples and p is the feature dimension; $\boldsymbol{\beta}^{(m)}$ is the $p\times 1$ dimensional regression coefficients for task m; $\boldsymbol{\epsilon}^{(m)}$ is the $n^{(m)}\times 1$ dimensional additive noise (or error). For classification problems, $\boldsymbol{y}^{(m)}$ is not directly observed, instead only the label information $\boldsymbol{z}^{(m)} = [z_1^{(m)},z_2^{(m)},\cdots,z_{n^{(m)}}^{(m)}]^\top$ is known ($z_i^{(m)}\in\{-1,1\}$). A Bayesian Elastic Net prior is shared across the M tasks (the sparseness properties are

shared, but not the exact regression weights). The multi-task model may be expressed as

$$\begin{aligned} z_i^{(m)} &\sim 1(z_i^{(m)} = \text{sign}(y_i^{(m)})) \\ y_i^{(m)} &\sim \mathcal{N}(y_i^{(m)}; (\boldsymbol{x}_i^{(m)})^\top \boldsymbol{\beta}^{(m)}, (\tau^{(m)})^{-1}) \\ \beta_j^{(m)} &\sim \mathcal{N}\left(\beta_j^{(m)}; 0, (\tau^{(m)})^{-1}(\alpha_j + \lambda_2^{(m)})^{-1}\right) \\ \tau^{(m)} &\sim \text{Ga}(\tau^{(m)}; c_0, d_0) \\ (\alpha_j, \gamma_j) &\sim \eta(\prod_{m=1}^{M} \sqrt{\frac{\alpha_j}{\alpha_j + \lambda_2^{(m)}}}) \text{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right) \text{Ga}(\gamma_j; a_0, b_0) \end{aligned}$$

for $i = 1, 2, \cdots, n^{(m)}; m = 1, 2, \cdots, M$ and $j = 1, 2, \cdots, p$. The form above is for multi-task *classification*; for regression we simply remove the top-level probit layer. Here η is a normalizing constant, and we allow the tuning parameter $\lambda_2^{(m)}$ in general to be different among different tasks, although in practice we have set it to a constant independent of m.

In the above discussion, the task-dependent $\boldsymbol{\beta}^{(m)}$ share the same Elastic Net prior (the α_j are shared for all M tasks, implying that the multiple tasks share similar non-zero components β_j). This is appropriate for the examples considered in Sections VI and VII, but in general it may not be appropriate that all M learning tasks share the *same* set of α_j. We may alternatively assume $\boldsymbol{\alpha}^{(m)} \sim G$, where $G \sim \text{DP}(\alpha_0 G_0)$; here $\boldsymbol{\alpha}^{(m)}$ represents a *vector* composed of components $\alpha_j^{(m)}$, the α_j associated with task m. The $\text{DP}(\alpha_0 G_0)$ is a Dirichlet Process (DP) [18] prior with precision $\alpha_0 \in \mathbb{R}^+$ and base probability measure G_0. The G_0 may be set as the proposed Bayesian Elastic Net prior. Specifically, we may constitute $G = \sum_{i=1}^{\infty} \pi_i \delta_{\boldsymbol{\alpha}_i^*}$, with the π_i drawn from a stick-breaking construction [19] with parameter α_0, and with the $\boldsymbol{\alpha}_i^*$ drawn from the Bayesian Elastic Net prior. In this setting the M tasks cluster, and *within each cluster* similar components of $\boldsymbol{\beta}^{(m)}$ are important, but not in general across different clusters. While this has not been needed in our examples, it demonstrates the flexibility of hierarchical Bayesian construction, into which the proposed Elastic Net prior may be directly inserted.

IV. Variational Bayesian Inference

A. Basic model

We present a variational Bayesian (VB) inference algorithm for the proposed Bayesian Elastic Net model. Variational Bayesian [20] analysis compromises between inference accuracy and computational efficiency. The full likelihood $p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ for the new Bayesian Elastic Net model is given in (12). In variational Bayes, we seek a distribution $Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ to approximate the exact posterior $p(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}|\boldsymbol{y})$. From Jensen's inequality,

$$\begin{aligned} \log p(\boldsymbol{y}) &= \log \int Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \frac{p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})}{Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})} d\boldsymbol{\beta} d\tau d\boldsymbol{\alpha} d\boldsymbol{\gamma} \\ &\geq \int Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \log \frac{p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})}{Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})} d\boldsymbol{\beta} d\tau d\boldsymbol{\alpha} d\boldsymbol{\gamma} \qquad (17) \\ &= \log p(\boldsymbol{y}) - \text{KL}\left(Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \| p(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}|\boldsymbol{y})\right) \end{aligned}$$

Expression (17) is called the variational lower bound. The KL distance term is nonnegative, and is zero if and only if the two distributions in it are identical. Furthermore, we assume that $Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ factors as

$$Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \doteq Q(\boldsymbol{\beta}) Q(\tau) Q(\boldsymbol{\alpha}) Q(\boldsymbol{\gamma}) \qquad (18)$$

Then the variational lower bound in (17) becomes

$$J = \int Q(\boldsymbol{\beta}) Q(\tau) Q(\boldsymbol{\alpha}) Q(\boldsymbol{\gamma}) \log \frac{p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})}{Q(\boldsymbol{\beta}) Q(\tau) Q(\boldsymbol{\alpha}) Q(\boldsymbol{\gamma})} d\boldsymbol{\beta} d\tau d\boldsymbol{\alpha} d\boldsymbol{\gamma} \qquad (19)$$

By maximizing the lower bound in (19) with respect to $Q(\boldsymbol{\beta})$, $Q(\tau)$, $Q(\boldsymbol{\alpha})$ and $Q(\boldsymbol{\gamma})$, we can effectively minimize the KL distance between the approximated posterior $Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ and the true posterior $p(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}|\boldsymbol{y})$. Following the general update rule of VB inference [20], the update equations for each Q function are derived as follows.

1) Update equation for $\boldsymbol{\beta}$:

$$Q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad (20)$$

with

$$\begin{aligned} \boldsymbol{\Sigma} &= \left(\langle\tau\rangle \boldsymbol{X}^\top \boldsymbol{X} + \langle\tau\rangle(\text{diag}(\langle\boldsymbol{\alpha}\rangle) + \lambda_2 \boldsymbol{I})\right)^{-1} \qquad (21) \\ \boldsymbol{\mu} &= \boldsymbol{\Sigma}\left(\langle\tau\rangle \boldsymbol{X}^\top \boldsymbol{y}\right) \qquad (22) \end{aligned}$$

Here $\text{diag}(\langle\boldsymbol{\alpha}\rangle)$ corresponds to the Lasso (ℓ_1) shrinkage and $\lambda_2 \boldsymbol{I}$ corresponds to the ridge (ℓ_2) shrinkage. We also have $\langle\boldsymbol{\beta}\rangle = \boldsymbol{\mu}$, $\langle\boldsymbol{\beta}\boldsymbol{\beta}^\top\rangle = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$ and $\langle\beta_j^2\rangle = \langle\boldsymbol{\beta}\boldsymbol{\beta}^\top\rangle_{jj}$.

2) Update equation for $\boldsymbol{\alpha}$:

$$Q(\boldsymbol{\alpha}) = \prod_{j=1}^{p} Q(\alpha_j) = \prod_{j=1}^{p} \text{InvGaussian}\left(\alpha_j; g_j, h_j\right) \qquad (23)$$

with

$$g_j = \sqrt{\frac{\langle\gamma_j\rangle}{\langle\tau\rangle\langle\beta_j^2\rangle}}, \qquad h_j = \langle\gamma_j\rangle \qquad (24)$$

Here $\text{InvGaussian}\left(\alpha_j; g_j, h_j\right)$ denotes the inverse Gaussian distribution with mean g_j and shape parameter h_j:

$$\begin{aligned} \text{InvGaussian}\left(\alpha_j; g_j, h_j\right) &= \left(\frac{h_j}{2\pi\alpha_j^3}\right)^{\frac{1}{2}} \\ &\times \exp\left(-\frac{h_j(\alpha_j - g_j)^2}{2g_j^2\alpha_j}\right) \quad (\alpha_j > 0) \qquad (25) \end{aligned}$$

with $\langle\alpha_j\rangle = g_j$ and $\langle\alpha_j^{-1}\rangle = g_j^{-1} + h_j^{-1}$.

3) Update equation for $\boldsymbol{\gamma}$:

$$Q(\boldsymbol{\gamma}) = \prod_{j=1}^{p} Q(\gamma_j) = \prod_{j=1}^{p} \text{Ga}(\gamma_j; a_j, b_j) \qquad (26)$$

with $a_j = a_0 + 1, b_j = b_0 + \frac{1}{2}\langle\alpha_j^{-1}\rangle$ and $\langle\gamma_j\rangle = a_j/b_j$.

4) Update equation for τ:

$$Q(\tau) = \text{Ga}(\tau; c, d) \qquad (27)$$

with

$$c = c_0 + \frac{n+p}{2} \quad (28)$$
$$d = d_0 + \frac{1}{2}(\langle \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \rangle + \sum_{j=1}^{p} \langle \beta_j^2 \rangle (\langle \alpha_j \rangle + \lambda_2))$$

and $\langle \tau \rangle = c/d$.

In all equations listed above, $\langle \cdot \rangle$ means expectation with respect to the $Q(\cdot)$ functions. Computation of the variational lower bound in (19) is feasible but omitted here.

Calculation of $\boldsymbol{\Sigma}$ in (21) involves inversion of a $p \times p$ matrix. In many applications, the feature dimension p is much larger than the number of samples n, and inversion of a $p \times p$ matrix is very expensive. By applying the matrix inversion lemma [21], we only need invert a small matrix with dimension $n \times n$. In this way, computation and *storage* of the matrix $\boldsymbol{\Sigma}$ is avoided.

As indicated in [2], the Elastic Net model has the problem of double shrinkage, since ridge and Lasso shrinkage are in play at the same time. In order to correct this double shrinkage effect, a post-processing step is performed, which is a *rescaling* operation

$$\hat{\boldsymbol{\beta}} = \xi \cdot \tilde{\boldsymbol{\beta}} \quad (29)$$

where $\tilde{\boldsymbol{\beta}}$ is the naive elastic net solution and $\hat{\boldsymbol{\beta}}$ is the final solution. In the VB solution, $\tilde{\boldsymbol{\beta}} = \boldsymbol{\mu}$ as expressed in (22). Here ξ is a scaling constant, and it is recommended in [2] that $\xi = 1 + \lambda_2$. However, through experiments we found that this ξ value does not work well for our VB solution. In practice we used the following criterion to find an appropriate ξ:

$$\xi = \arg\min_{\xi} \|\boldsymbol{y} - \xi \cdot \boldsymbol{X}\boldsymbol{\mu}\|^2 = \frac{(\boldsymbol{X}\boldsymbol{\mu})^\top \boldsymbol{y}}{(\boldsymbol{X}\boldsymbol{\mu})^\top (\boldsymbol{X}\boldsymbol{\mu})} \quad (30)$$

Thus, by using rescaling as a post-processing step, the final expression for the Bayesian Elastic Net estimator is $\hat{\boldsymbol{\beta}} = \xi\boldsymbol{\mu} = \frac{(\boldsymbol{X}\boldsymbol{\mu})^\top \boldsymbol{y}}{(\boldsymbol{X}\boldsymbol{\mu})^\top (\boldsymbol{X}\boldsymbol{\mu})}\boldsymbol{\mu}$.

B. Inference with probit

The inference for the model in (14) is similar to that in (11), as given from (20) to (27), except that now we need to introduce another $Q(\cdot)$ function for the latent regression response $\boldsymbol{y}$:

$$Q(\boldsymbol{y}) = \prod_{i=1}^{n} Q(y_i) \propto \prod_{i=1}^{n} 1(z_i = \text{sign}(y_i)) \cdot \mathcal{N}\left(y_i; \theta_i, \sigma^2\right) \quad (31)$$

where $\theta_i = \boldsymbol{x}_i^\top \langle \boldsymbol{\beta} \rangle$ and $\sigma^2 = \langle \tau \rangle^{-1}$. This is a truncated normal distribution, with truncation region determined by z_i. If $z_i = 1$ then $Q(y_i) \propto 1(y_i \geq 0) \cdot \mathcal{N}\left(y_i; \theta_i, \sigma^2\right)$, otherwise $Q(y_i) \propto 1(y_i < 0) \cdot \mathcal{N}\left(y_i; \theta_i, \sigma^2\right)$. Statistics of the truncated normal distribution $Q(y_i)$ can be computed as follows:

$$\langle y_i \rangle = \theta_i + z_i \sigma \frac{\phi\left(\theta_i/\sigma\right)}{\Phi\left(z_i\theta_i/\sigma\right)}$$
$$\langle y_i^2 \rangle = \sigma^2 + \langle y_i \rangle \cdot \theta_i$$
$$-\langle \log Q(y_i) \rangle = \frac{\log(2\pi\sigma^2)}{2} + \frac{\langle (y_i - \theta_i)^2 \rangle}{2\sigma^2} + \log \Phi(\frac{z_i\theta_i}{\sigma}) \quad (32)$$

Here $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative density function for standard normal distribution, respectively. Consequently, $\boldsymbol{y}$ in (22) should be replaced with its expectation $\langle \boldsymbol{y} \rangle$, and we should also treat $\boldsymbol{y}$ as a random variable for the expectation $\langle \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \rangle$ in (27). In addition, the variational lower bound for this model should include the term $-\sum_{i=1}^{n} \langle \log Q(y_i) \rangle$, which is the entropy of $Q(\boldsymbol{y})$. Since only the sign of y_i is related with the output label z_i, no extra re-scaling step is required in this classification.

The predictive distribution for testing data $\boldsymbol{x}_\star$ can be expressed as

$$p(z_\star = 1 | \tau, \boldsymbol{x}_\star) = \Phi\left(\frac{\boldsymbol{x}_\star^\top \boldsymbol{\mu}}{(\tau^{-1} + \boldsymbol{x}_\star^\top \boldsymbol{\Sigma} \boldsymbol{x}_\star)^{\frac{1}{2}}}\right) \quad (33)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the posterior mean and covariance of $\boldsymbol{\beta}$ derived in the variational Bayesian solution; τ could be replaced with its expectation $\langle \tau \rangle$. For 'hard' decision, we have $z_\star = 1$ if $\boldsymbol{x}_\star^\top \boldsymbol{\mu} > 0$ and $z_\star = -1$ otherwise. Concerning the aforementioned lack of rescaling for the case of classification, note that the accuracy of the *probability* in (33) is difficult to test; therefore, it is possible that rescaling may be necessary to make this probability accurate (but the assessment of such accuracy is difficult). As indicated, in practice we often simply threshold the label decision at a probability of 0.5, and this was done with success in all examples presented below.

The VB solution to the multi-task case is a relatively direct extension of that for the single-task model discussed above. A main difference is that the posterior for α_j now is a generalized inverse Gaussian (GIG) distribution with statistics collected from all M, instead of a inverse Gaussian distribution in (23). These details are omitted here for brevity.

V. Setting Parameters

This paper is motivated by the original Elastic Net (ENet) [2], which builds regression or classification models based on a sparse set of *correlated* features. This should be contrasted with Lasso [3], which is also sparse but it discards multiple correlated features, and therefore Lasso provides less biological interpretation. The ENet involves two parameters, and may be solved by employing the LARS algorithm [2], [22], in which ENet parameter λ_2 is fixed, and one sweeps all ranges of λ_1, and as λ_1 increases the model weights become less sparse. Similarly, Lasso may be solved via LARS, and in this case one sweeps through the single Lasso parameter, again covering a range of sparseness from a single non-zero weight to all weights being non-zero. If $\lambda_2 = 0$, the ENet model becomes Lasso. We note this background because the proposed Bayesian ENet has a similar relationship to Bayesian Lasso [6], with the latter manifested if the Bayesian ENet parameter is set as $\lambda_2 = 0$. The Bayesian Lasso is typically very sparse, and therefore the setting of λ_2 in the Bayesian ENet controls the degree of model sparseness, with $\lambda_2 = 0$ yielding very sparse solutions, and increasing λ_2 yielding less sparse represenstations. Therefore, analogous to the ENet, as we sweep through increasing λ_2 in the Bayesian ENet, we may test models of decreasing levels of sparseness. Thus, the setting of λ_2 in the Bayesian ENet dictates a desired level of

TABLE I
SUMMARY OF RESULTS FROM THE SYNTHESIZED DATA, BASED UPON 100 RANDOM PARTITIONS OF THE DATA.

Method	Average Err	# Features Used on Average
Bayesian ENet	26.9	15
Bayesian Lasso	85.0	8
ENet	34.5	15

model sparseness, much as within the ENet the stopping point of the LARS algorithm specifies a degree of sparseness.

Additionally, as for the original ENet, one may use cross validation to set λ_2 in the Bayesian ENet. In the context of such cross validation, the Bayesian ENet has the advantage of only requiring the setting of a single parameter. In the experiments presented below, we employ cross validation to set λ_2 in the Bayesian ENet. The range of λ_2 considered in this cross validation is selected as to test a range of different levels of sparseness, with a similar consideration required in the original ENet. In practice multiple λ_2 may yield comparable cross-validation performance, and therefore the selection of the λ_2 used is dictated by the desired level of model sparseness, which is linked to the interpretability of the model in terms of underlying biological pathways.

VI. EXPERIMENTS ON PUBLISHED DATA

A. Simulation data

Our first example is from the original Elastic Net paper [2]. We simulate data from the following model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\epsilon \tag{34}$$

where $\boldsymbol{X}$ is a matrix of dimension $n \times p$ with $n = 500, p = 40$, $\epsilon \sim \mathcal{N}(\epsilon; 0, 1)$ and $\sigma = 15$. The first 50 samples are used for training, the second 50 samples are used for validation when tuning the parameter λ_2, and the remaining 400 samples are used for testing. The predictors (design matrix) are simulated as follows:

$$\begin{aligned}
&\boldsymbol{X}_i = \boldsymbol{Z}_1 + \boldsymbol{\omega}_i \quad \boldsymbol{Z}_1 \sim \mathcal{N}(\boldsymbol{Z}_1; \boldsymbol{0}, \boldsymbol{I}) \quad \boldsymbol{\omega}_i \sim \mathcal{N}(\boldsymbol{\omega}_i; 0, 0.01\boldsymbol{I}) \\
&\boldsymbol{X}_j = \boldsymbol{Z}_2 + \boldsymbol{\omega}_j \quad \boldsymbol{Z}_2 \sim \mathcal{N}(\boldsymbol{Z}_2; \boldsymbol{0}, \boldsymbol{I}) \quad \boldsymbol{\omega}_j \sim \mathcal{N}(\boldsymbol{\omega}_j; 0, 0.01\boldsymbol{I}) \\
&\boldsymbol{X}_k = \boldsymbol{Z}_3 + \boldsymbol{\omega}_k \quad \boldsymbol{Z}_3 \sim \mathcal{N}(\boldsymbol{Z}_3; \boldsymbol{0}, \boldsymbol{I}) \quad \boldsymbol{\omega}_k \sim \mathcal{N}(\boldsymbol{\omega}_k; 0, 0.01\boldsymbol{I}) \\
&\boldsymbol{X}_m \sim \mathcal{N}(\boldsymbol{X}_m; \boldsymbol{0}, \boldsymbol{I})
\end{aligned} \tag{35}$$

where $i = 1, 2, ..., 5$; $j = 6, 7, ..., 10$; $k = 11, 12, ..., 15$ and $m = 16, 17, ..., 40$. Here $\boldsymbol{X}_i$ denotes the i^{th} column of $\boldsymbol{X}$, with dimension $n \times 1$. The ground truth of the regression coefficient $\boldsymbol{\beta}$ is set to be $\boldsymbol{\beta} = [3, 3, \cdots, 3, 0, 0, \cdots, 0]^\top$ with first 15 elements equal to 3 and the rest 0. Therefore there are three equally important feature groups ($1 \sim 5, 6 \sim 10, 11 \sim 15$), and within each group there are five highly correlated members. The last 25 features are pure noise ($m = 16, 17, ..., 40$). In the cross-validation, we use a candidate list $[0, 10, 50, 100]$ to choose λ_2 from in the validation set. For different runs of the experiment, the choice of λ_2 may be different, but $\lambda_2 = 0$ is never chosen. The performance measure is the residue in the testing set:

$$Err = \frac{1}{n'}\|\boldsymbol{y}' - \boldsymbol{X}'\hat{\boldsymbol{\beta}}\|^2 - \sigma^2 \tag{36}$$

where $\boldsymbol{X}'$ and $\boldsymbol{y}'$ are design matrix and response in the testing set, with $n' = 400$. The parameter σ^2 is subtracted because the first term contains residue caused by the model itself (the additive noise). Thus Err accounts for the mismatch between the estimated model ($\hat{\boldsymbol{\beta}}$) and the true model ($\boldsymbol{\beta}$). We generated 50 data sets independently, and for each data set we can calculate one Err. We then employ the median of these 50 Err as the performance measure. In this experiment, 100 such runs are performed, with results summarized in Table I. It is observed that both ENet and Bayesian ENet are effective at extracting the relevant features, with the Bayesian ENet yielding slightly better fitting accuracy. Each of the 15 features inferred by the ENet and Bayesian ENet are consistent with truth, while only three of the 8 from Bayesian Lasso are.

TABLE II
SINGLE-TASK PERFORMANCE ON LEUKEMIA DATA OF DIFFERENT MODELS.

Method	Validation Error	Testing Error	# Genes Used
Bayesian ENet	$1/38$	$1/34$	19
Bayesian Lasso	$1/38$	$3/34$	1
ENet	$3/38$	$0/34$	45
SVM-RFE	$2/38$	$1/34$	31

B. Single-task gene analysis

We consider the leukemia data in [23], which consists of 7129 genes and 72 samples. There are 38 samples in the training set and 34 samples in the testing set. The samples consist of two groups, one is type-1 leukemia (ALL) and the other is type-2 leukemia (AML). F-score pre-screening is performed in the training set to select the most important 1000 genes, which is used for *probit* regression modeling (we also show results below for which the Bayesian Elastic Net is directly applied to a data set of 12,023 genes, with no pre-pruning). The design matrix $\boldsymbol{X}$ is composed of the normalized gene-expression values, with the first column being all '1's to account for the bias term, and $\boldsymbol{\beta}$ contains the weights on the bias and gene features. We use a candidate list $[0, 10, 50, 100]$ to choose λ_2. For each candidate, a ten-fold cross-validation is applied and an averaged validation error on the ten-fold is obtained. Then λ_2 is chosen as the candidate that gives minimum validation error. If multiple candidates give the same validation error, then the candidate that results in more selected genes will be used for better model interpretation and robustness. After this cross-validation, $\lambda_2 = 10$ is chosen. The validation and testing results are given in Table II, for the proposed Bayesian Elastic Net (ENet), for Bayesian Lasso [6], for the original ENet [2], and for the support vector machine with feature selection (SVM-RFE) [24].

From Table II we note that the overly sparse Bayesian Lasso gives inferior results, while the Bayesian ENet, ENet

and SVM-RFE yield comparable performance. The SVM-RFE is an iterative algorithm, and the stopping point is often difficult to ascertain. Additionally, in [2] the authors coded the type of leukemia as a $0-1$ response $\boldsymbol{y}$, and rescaling was required, with this not needed for the Bayesian ENet with probit regression. We argue that the probit link adopted here (see Section II-C) is a more principled way to deal with classification problems, and this paper is (to our knowledge) the first to combine probit regression with the Elastic Net model.

The nineteen genes selected by the Bayesian ENet are: **X95735**, **U50136**, **Y12670**, **M23197**, D49950, **X85116**, **M55150**, **M16038**, **X17042**, **M80254**, **L08246**, **U82759**, M22960, **M84526**, **U46751**, **M27891**, **M83652**, **Y00787**, M81933. Genes in boldface are also reported in [23].

C. Multi-task gene analysis

We apply the multi-task Bayesian ENet model introduced in Section III to gene-expression data for small round blue cell tumors [25]. These data were previously used in [11]. There are four classes of samples: Ewing family of tumors (EWS), neuroblastoma (NB), rhabdomyosarcoma (RMS) and Burkitt lymphoma (BL) which is a subset of non-Hodgkin lymphoma. There are 63 training samples and 20 testing samples, each sample containing 2308 gene-expression values. The goal of the analysis is to identify a small set of genes that can classify different types of tumors. This multi-category classification problem can be reformulated as a multi-task learning problem, each task learning a one-versus-all binary classifier. The classification results in all tasks are combined to give the final multi-category classification result using the predictive distribution in (33). Note that these data are only employed to *demonstrate* an application of multi-task learning. In practice one may also desire development of a multi-class classifier directly (rather than multiple binary classifiers).

Following the pre-processing procedure in [11], the 2308 genes are first reduced to 500 based on the marginal correlation (a measure similar to F-score). The design matrix for each task is a 63×501 matrix, with the first column being all '1' s to account for the bias term, and the rest of the 500 columns the normalized gene features. For Task 1, EWS samples are assigned label 1 and all the rest -1. Similar label assignment applies for Tasks 2 (NB), 3 (RMS) and 4 (BL). In order to choose $\lambda_2^{(m)}$, 4-fold cross-validation is performed on the training set using candidate list $[0, 10, 50, 100]$. Using the same criterion explained in the previous section, $\lambda_2^{(m)} = 100$ is chosen. The validation and testing results are given in Table III (the SMALR results are from [11]). Similar to the ENet, there are two parameters that must be set for SMALR, a λ and h_0 a bandwidth parameter.

TABLE III
MULTI-TASK CANCER DATA

Method	Validation Error	Testing Error	# Genes Used
Bayesian ENet	$0/63$	$0/20$	42
Bayesian Lasso	$4/63$	$2/20$	4
SMALR	$0/63$	$0/20$	20

VII. Analysis of Virus Expression Data

A. Data collection

Two *independent* data collections have been executed for influenza, as follows. In each case a healthy volunteer intranasal challenge with influenza was performed at Retroscreen Virology, LTD (Brentwood, UK), using pre-screened volunteers who provided informed consent; one of these challenges was performed with the H3N2 virus and the other with H1N1. On the day of inoculation, a dose of influenza manufactured and processed under current good manufacturing practices (cGMP) by Bayer Life Sciences, Vienna, Austria) was inoculated intranasally per standard methods. Subjects were not released from quarantine until after the 216th hour. Blood and nasal lavage collection continued throughout the duration of the quarantine. All subjects received oral oseltamivir (Roche Pharmaceuticals) 75 mg by mouth twice daily prophylaxis at day 6 following inoculation. All patients were negative by rapid antigen detection (BinaxNow Rapid Influenza Antigen; Inverness Medical Innovations, Inc) at time of discharge. For the first study 17 individuals participated, and in the second 19 individuals participated.

Subjects had the following samples taken 24 hours prior to inoculation with virus (baseline), immediately prior to inoculation (pre-challenge) and at set intervals following challenge: peripheral blood for serum, peripheral blood for PAXgene$^{\text{TM}}$, nasal wash for viral culture/PCR, urine, and exhaled breath condensate. Peripheral blood was taken at baseline, then at 8 hour intervals for the initial 120 hours and then 24 hours for the remaining 2 days of the study. All results presented here are based on gene-expression data from blood samples. Further details of this study, as well as additional related studies, are discussed in [26]. Specifically, in [26] separate data were reported for Rhino virus and for RSV; the data from these two viruses will also be considered below.

B. Single-task classification analysis

We first consider the accuracy of the classification model, by training on H3N2 data and testing on independent H1N1 data. This analysis was performed using all 12,023 genes from the gene-expression data. In the analysis presented here, we focus on data collected at 93.5, 101, and 108 hours after inoculation, this corresponding to the time period when symptoms are first observed clinically, for those who become symptomatic. We use this time period because the different subjects manifested initial symptoms at different times, and in the above time window all subjects here became symptomatic had manifested symptoms. The objective of the classifier was to distinguish those individuals that who were symptomatic with influenza with those who were not. While we are interested in these classification results, our principal interest is in understanding the biological pathways responsible for distinguishing symptomatic and asymptomatic individuals. Hence, while classification performance is important, understanding the important

TABLE IV

PERFORMANCE ON INFLUENZA DATA GENERATED IN THIS STUDY, NEAR TIME OF SYMPTOM ONSET. TRAIN ON H3N2 VIRUS DATA, TEST ON H1N1 VIRUS DATA.

Method	Training Error	Testing Error	# Genes Used
Bayesian ENet	$0/51$	$4/57$	102
Bayesian Lasso	$0/51$	$5/57$	1
ENet	$0/51$	$4/57$	100
Lasso	$0/51$	$5/57$	20
RVM	$0/51$	$5/57$	1
SVM-RFE	$0/51$	$4/57$	16

genes inferred as useful via the classifier is our principal concern in this study. This has motivated development of the Bayesian ENet (and also was the motivation for the original ENet [2]). The practical application of this work is the goal of diagnosing people as being symptomatic with a virus, as opposed to, for example, a bacteria; the clinical symptoms for these two case are often similar. Our goal is to develop genomic tests that exploit the known host pathways of influenza and related viruses, to improve diagnosis (this is discussed further in the Conclusions).

In Table IV we show classification results using the proposed Bayesian ENet, Bayesian Lasso [6], the original ENet model [2], Lasso [3], the relevance vector machine (RVM) [27], and the SVM with feature extraction (SVM-RFE) [24]. In addition to the testing accuracy, we emphasize the number of genes inferred as important via the model. Specifically, note that the RVM and Bayesian Lasso only infer a single gene, and therefore these are of very little value for biological interpretation. For the ENet we set $\lambda_2 = 10$, and we set λ_1 such that 100 genes were extracted (the ENet analysis was performed using LARS [2], and therefore we may set λ_1 as to infer a desired number of genes, here 100). Similarly, the Lasso results were also computed using LARS, and again we may stop it such that a desired number of genes are employed. However, importantly, the genes selected by Lasso are *not* correlated [2], and therefore these are not useful for biological interpretation (one will not readily find the expected correlated genes associated with a biological pathway). Therefore, while one may (here) select 20 genes with Lasso, there are of little additional interpretative value than the single gene associated with Bayesian Lasso and with the RVM, and this limitation of Lasso was the motivation for the original ENet. We also note that the number of genes selected for SVM-RFE is also arbitrary, due to the selected stop criterion in the iterative algorithm.

Therefore, we note ENet, Lasso, and SVM-RFE each require one to stop the model with a specified number of important genes. The proposed Bayesian ENet also has a related issue in the selection of λ_2, with as mentioned $\lambda_2 = 0$ corresponding to Bayesian Lasso (see Section V). In Table IV we report results for $\lambda_2 = 10$, but a very similar set of important genes was found for $\lambda_2 \in [10, 20, 30]$.

Concerning biological significance, the 50 genes with largest model weights, as inferred via the Bayesian ENet are: OAS1, IFI44, DDX60, IFI44L, RTP4, IFIT1, GBP1, LY6E, RSAD2, SCO2, TRIM22, IFITM1, IFIT3, ISG15, MS4A4A, IFI6, PSME2, XAF1, IFIT5, IFITM3, HERC5, IFI27, SERPING1, PLSCR1, OAS3, STAT1, ZCCHC2, HUMISGF3A, OASL, C1QB, SIGLEC1, AIM2, OAS2, MX1, LOC26010, SMPDL3A, C13orf18, LAP3, TDRD7, PARP12, PSME1, VAMP5, IRF7, SAT1, SMAD1, BLVRA, IDH2, C1QA, MT2A and TAP1. The ENet results inferred a similar set of genes, for the settings associated with Table IV.

A particular challenge of microarray-based gene expression

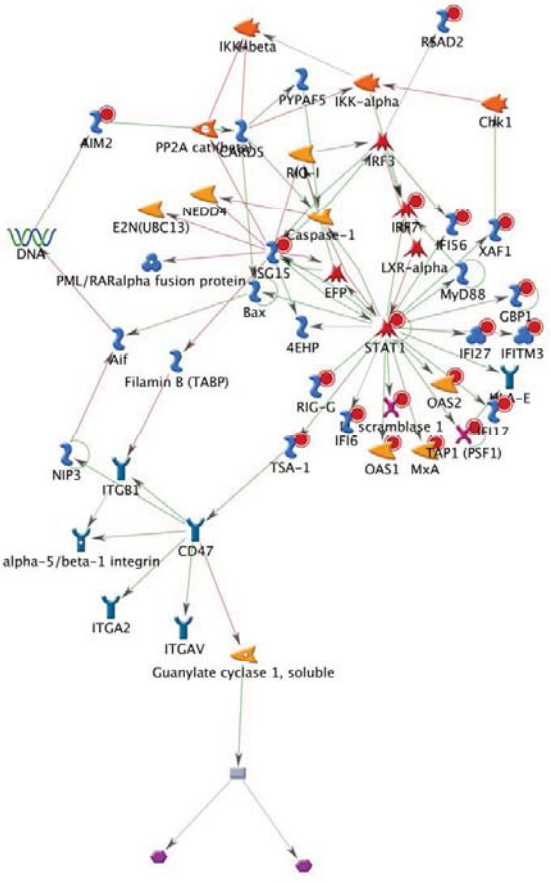

Fig. 1. Multiple components of canonical viral immunity pathways are represented in the top 50 differentially expressed genes identified by the Bayesian elastic net framework. Pathway analysis (www.genego.com) illustrates the STAT-1 network, with overrepresentation of the genes identified by the elastic net framework. Shown is the top network related to genes in the elastic net, the STAT-1 (signal transducer and activator of transcription 1). STAT1, a transcription factor, coordinates cellular responses to type II interferons, a major component of antiviral immunity and is a key mediator of cytokine-induced gene expression as it is activated by many cytokines including type I and type II IFNs, interleukin (IL)-6 and IL-10. In this figure, transcription factors are indicated by red 6-pointed stars (example STAT-1), enzymes by orange arrowheads (example MXA), proteases by orange indented arrowheads (example caspase 1), and binding proteins by blue "S" shapes (example GBP-1). Elements present in the gene list generated by the elastic net are indicated with red targets. A full key to all symbols is shown in Supplemental Figure 1.

signature experiments is to relate the selected genes to the relevant disease state. Thus, to examine the biological significance of these genes, and the inferred biological pathway, we used GATHER^{TM} (www.gather.genome.duke.edu) and Genego^{TM} (www.genego.com) to evaluate the relationship of the aforementioned gene list to various biologic processes. These programs utilize curated available scientific literature regarding known gene relationships and disease pathways to infer the relative likelihood that the genes in a given set are related, and with which known pathways or biologic processes the genes are associated. Gene sets derived from the Bayesian ENet, and the original ENet were analyzed using the above described methods, with representative pathways and relationships shown in Figures 2 for the Bayesian ENet. For the purposes of inferring biologic plausibility, a geneset that is less sparse often provides superior results ($i.e.$, strength of association with the categories of "response to virus" and "immune system process" is greatest with the Bayesian ENet). However, should the goal be to maximize sparseness without loss of inference to biologic plausibility, the Lasso process gene set remains biologically related to host response to viral infection with inclusion of key genes such as OAS1, GBP1 and RTP4.

Specifically, GATHER analysis shows that 23 of the top 50 genes in the Bayesian ENet results are classified in the represented clustered into the gene ontology (GO) category 0009607 (response to biotic stimulus). When using the original ENet and the Lasso models, the number of genes in this GO category are 14 and 6, respectively, indicating a loss of specificity when sparseness is increased. Pathway analysis further implicates specific antiviral immunity pathways as major components of the factors selected by the Bayesian ENet and the original ENet models. Interferon response pathways (shown in Figure 2 as triggered by STAT-1, a transcription factor) are well established as primary elements of host response to viral infection [28]. Sensing of viral infection involves a complex positive feedback loop, with sensing of viral dsRNA by the RNA helicases RIG-1 and MDA-5 and subsequent triggering of interferon production. This results in activation of STAT-1 and other transcription factors, and induction of antiviral immune response genes such as MxA and OAS to result inhibition of viral replication [29]. In contrast, the top pathway identified when using Lasso, the DNAJB6 pathway, is most closely associated with cellular response to stress.

C. MTL analysis and pan-viral pathway

The following example is the principal motivation for developing the multi-task Bayesian ENet. Specifically, we consider the gene expression data from H3N2 and H1N1, as discussed above, while also considering Rhino virus and RSV gene expression data from similar challenge studies, as summarized in [26]. For the latter two data we also consider time samples at points for which all symptomatic samples are manifesting symptoms. For Rhino this corresponds to samples at times 48, 72 and 96 hours, and for RSV 132, 141.5 and 165.5 hours. The total number of samples for H3N2, H1N1, Rhino and RSV are 51, 57, 65 and 36, respectively. Our objective is to infer a "pan-viral" [26] set of genes that are consistently important across these four viruses. To do this, we build four classifiers in a multi-task setting, with the goal of distinguishing symptomatic and asymptomatic subjects for each viruses, with all four classifiers learned jointly. While we may perform various forms of model testing ($e.g.$, leave-one analysis), and all such analyses we have considered yield classification performance analogous to the results summarized in Table IV, here our principal objective is in inferring and interpreting the common biological pathways between these four viruses. Again, the limitations of Lasso, Bayesian Lasso and the RVM for this task (they are each too sparse, and hence not biologically interpretable) has motivated the ENet and proposed Bayesian ENet (but the original ENet cannot be employed for the multi-task analysis). Note that the multi-task analysis assumes that the genes associated with relevant pathways may be shared between the four viruses, but the specific classifier weights on the genes are virus specific, for each of the four viruses considered.

For a setting of $\lambda_2 = 10$ (with similar results manifested with other values of λ_2, as discussed above), the following were inferred as the 50 most important genes for the "pan-viral" pathway: RSAD2, OAS1, IFI44L, RTP4, IFIT3, IFITM1, IFI44, PLSCR1, LY6E, ISG15, P2RX5, IFI27, GBP1, KIAA0125, APOBEC3A, EPB41L3, IFIT1, XAF1, PSMB9, TRIM22, SERPING1, HERC5, OASL, SCO2, IFI6, DDX60, BLK, MS4A4A, TNFRSF9, BLVRA, LOC26010, MX1, C1QA, OAS3, IRF7, VAMP5, IFIT5, SMPDL3A, FER1L3, UBE2L6, SIGLEC1, C13orf18, PSME2, IFI35, C1QB, BST2, OAS2, PNOC, RRAS and SRBD1.

In Figure 2 we relate these genes to an inferred pathway, in the manner discussed above. This inferred pathway is deemed to be of high accuracy as the strength of association of the multi-task gene list with this pathway is quite robust (z-score [an indication of how many genes in the gene list are represented in a particular network] 76.83). The top represented pathway, the ISG15 pathway in Figure 2, is highly involved in viral immunity as it is activated by initial viral sensing and subsequent interferon production. ISG15 is known to target the influenza A protein NS1 and result in limitation of viral replication [30]. Clear overlap is seen with the top pathway identified from analysis of the influenza specific data (Figure 1).

D. Computation time

The VB inference employed for the Bayesian ENet is relatively efficient. All computations have been run in non-optimized Matlab on a desktop PC. To give a sense of the level of required computations, for the multi-task influenza experiment above we, we used all 12,023 genes. For a fixed λ_2, the algorithm required approximately 40 minutes, for 5000 VB iterations. The number of VB iterations was set to be large enough such that convergence was assured (more than absolutely necessary to get good results).

VIII. Conclusion

The Elastic Net model developed in [2] has been extended to a Bayesian version, with the number of model parameters

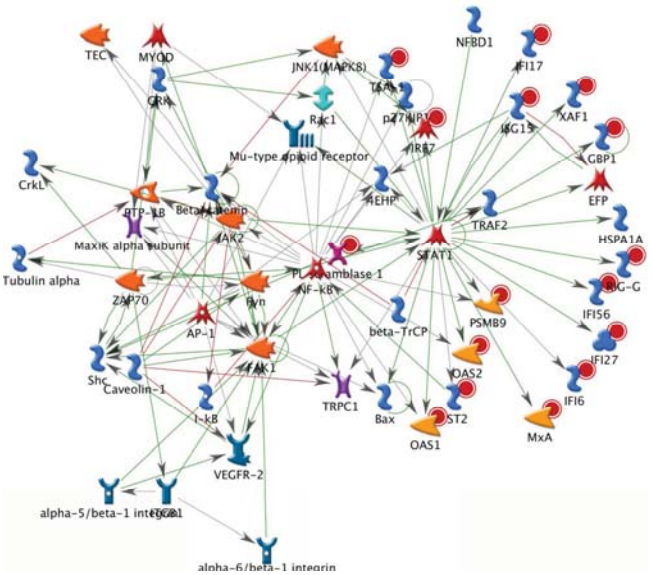

Fig. 2. Multi-task elastic net processes identify genes from a key anti-viral immune pathway. Pathway analysis (www.genego.com) illustrates the ISG15 pathway, with over-representation of genes identified by the multi-task elastic net. ISG15 is a ubiquitin-like modifier that is induced by interferon to restrict viral replication [31]. Downstream elements of ISG15 activation include activation of STAT-1, similar to what is shown in Figure 1. A full key to all symbols is shown in Supplemental Figure 1.

reduced from two to one. Efficient inference is performed using a variational Bayesian analysis. Having developed a Bayesian framework, the model is readily inserted into hierarchical Bayesian settings. For classification problems we have employed a probit link function, and when simultaneously building multiple related models, a multi-task Bayesian Elastic Net has been developed.

The analysis of the virus data discussed in Section VII was the principal motivation for this research. Our goal is to infer a biological pathway(s) responsible for the host response to virus, with the goal of developing a diagnostic tool for distinguishing those with symptoms from virus with those with symptoms from other causes ($e.g.$, bacteria). Hence, the ability to interpret and gain insights from the inferred pathways is critical. Using the top 40 genes inferred from the Bayesian ENet analysis of the "pan viral" pathway, we have developed clinically viable polymerase chain reaction (PCR) chips (the use of 40 genes on such chips has motivated our desire for sparse models, but not as sparse as Lasso or Bayesian Lasso). Our early clinical tests with this PCR chip is highly encouraging, and will be reported upon in detail elsewhere.

REFERENCES

[1] G.J. McLachlan, K.-A. Do, and C. Ambroise. *Analyzing Microarray Gene Expression Data*. Wiley, 2004.

[2] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society Series B*, 67:301–320, 2005.

[3] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society Series B*, 58:267–288, 1996.

[4] M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1: 586–597, 2007.

[5] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.

[6] T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686, 2008.

[7] L. Bornn, A. Doucet, and R. Gottardo. The bayesian elastic net. In *CMS-MITACS Joint Conference*, 2007.

[8] J. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679, 1993.

[9] J. Ashford and R. Sowden. Multi-variate probit analysis. *Biometrics*, 26:535–546, 1970.

[10] S. Ji, D. Dunson, and L. Carin. Multitask compressive sensing. *IEEE Transaction on Signal Processing*, 57:92–106, 2009.

[11] H. Liu, J. Lafferty, and L. Wasserman. Nonparametric regression and classification with joint sparsity constraints. In *Advances in Neural Information Processing Systems (NIPS)*, volume 21, 2008.

[12] A. Kaban. On bayesian classification with laplace priors. *Pattern Recognition Letters*, 28:1271–1282, 2007.

[13] M. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 25:1150–1159, 2003.

[14] J. Griffin and P. Brown. Bayesian adaptive lassos with non-convex penalization. Technical Report 07-2, Centre for Research in Statistical Methodology, University of Warwick, 2007.

[15] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

[16] M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors, and bayesian lassos. 2009 (submitted).

[17] H. Zou and H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 2008.

[18] T. Ferguson. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.

[19] J. Sethuraman. A constructive definition of the dirichlet prior. *Statistica Sinica*, 2:639–650, 1994.

[20] M. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[21] M. Woodbury. Inverting modified matrices. Technical Report 42, Statistical Research Group, Princeton University, 1950.

[22] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression annals of statistics. *Annals of Statistics*, 32, 2004.

[23] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[24] A. Rakotomamonjy. Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, 3, 2003.

[25] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7:673–679, 2001.

[26] A.K. Zaas, M. Chen, J. Lucas, T. Veldman, A.O. Hero, J. Varkey, R. Turner, C. Oien, S. Kingsmore, L. Carin, C.W. Woods, and G.S. Ginsburg. Peripheral blood gene expression signatures characterize symptomatic respiratory viral infection. *Cell Host & Microbe*, 6:207–217, 2009.

[27] M.E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1, 2001.

[28] S.D. Der, A. Zhou, B.R. G. Williams, and R.H Silverman. Identification of genes differentially regulated by interferon α , β, or γ using oligonucleotide arrays. *Proc. Nat. Ac. Sciences*, 95:15623–15628, 1998.

[29] A. Garcia-Sastre. Antiviral response in pandemic influenza viruses. *Emerging Infectious Diseases*, 12:43–47, 2006.

[30] C. Zhao, G.Q. Sun, S. Li, M.-F. Lang, S. Yang, W. Li, , and Y. Shi. MicroRNA let-7b regulates neural stem cell proliferation and differentiation by targeting nuclear receptor TLX signaling. *Proc. Nat. Ac. Sciences*, 107:1876–1881, 2010.

[31] G.A. Versteeg, B.G. Hale, S. van Boheemen, T. Wolff, D.J. Lenschow, and A. Garca-Sastre. Species-specific antagonism of host ISGylation by the influenza B virus NS1 protein. *J. Virology*, 10:5423–5430, 2010.