The Bayesian Elastic Net: Classifying Multi-Task Gene-Expression Data

[1]Minhua Chen, [1]David Carlson, [2]Aimee Zaas, [2]Christopher Woods,
[2]Geoffrey S. Ginsburg, [3]Alfred Hero III, [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
[3]University of Michigan, Electrical & Computer Engineering Department

Abstract

Highly correlated relevant features are frequently encountered in variable-selection problems, with gene-expression analysis an important example. It is desirable to select all of these highly correlated features simultaneously as a group, for better model interpretation and robustness. Further, irrelevant features should be excluded, resulting in a sparse solution (of importance for avoiding over-fitting with limited data). We address the problem of sparse and grouped variable selection by introducing a new Bayesian Elastic Net model. One advantage of the proposed model is that by imposing priors on individual parameters in the Laplace distribution, we reduce the number of tuning parameters to one, as compared with two such parameters in the original Elastic Net. In addition, we extend the new Bayesian Elastic Net model to the problem of probit regression, in order to deal with classification problems with a sparse but correlated set of covariates (features). Extension to multi-task learning is also considered, with inference performed using variational Bayesian analysis. The model is validated by first performing experiments on simulated data and on previously published gene-expression data; in these experiments we also perform comparisons to the original Elastic Net and to Bayesian Lasso. Finally, we present and analyze a new gene-expression data set for the time-evolving properties of influenza, measured using blood samples from human subjects in a recent challenge study.

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 the time evolution of gene expression data after the human body has interacted with the influenza virus. Specifically, as discussed in detail in Section VI, after receiving institutional review board (IRB) approval, we inoculated 20 human volunteers with Influenza-A, 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 will become symptomatic versus those who will not (with the ultimate objective of using gene-expression data to make this diagnosis as soon as possible, ideally prior to any symptoms).

Factor analysis is a powerful tool for gene-expression and related analysis [2], [3]. Factor models are motivated by the idea that the sparse factor loadings will capture the *family* of genes responsible for a biological pathway, with the expression values of a given subject represented as a linear combination of factor loadings (and associated pathways). While a factor analysis framework is effective for the aforementioned influenza data near and after the onset of symptoms [4], we have found that factor analysis is not as effective at earlier times. This has motivated study here of a model based on sparse linear regression, with the expectation that at early times (pre-symptomatic) the number of important genes will be smaller than those associated with a fully formed pathway.

While we are motivated by the problem of developing a classifier based upon a sparse group of gene-expression values, our model construction is applicable as well to many other related problems. We therefore present the model in a general setting, with discussion as well on how it is related to our specific problem of interest.

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 (or error) vector. 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 defined by estimating $\boldsymbol{\beta}$. When performing linear regression, one is interested in prediction accuracy as well as in model physical/biological interpretation [5]. 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 [6] 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 [7], [8]. However, if some relevant features are highly correlated with each other, Lasso tends to arbitrarily select only a few of these [5], 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 [5] 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 [5] for why this optimization criterion can yield sparse and grouped variable selection. In [5], 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 [9]. Authors have earlier attempted to develop a Bayesian Elastic Net construction [10]. 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 [9] 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 [11], [12]. 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.

For our motivating problem, we have gene expression data from subjects who will become symptomatic and those who will not. We wish to design a classifier for each time at which we have data, and since the classifiers are expected to be related to one another, we wish to learn the models jointly. The problem of jointly learning multiple models, exploiting shared information, has been referred to as multi-task learning [13], [14]. 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 time-evolving expression data associated with influenza-body interaction, using new data from our challenge study.

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. Several results are presented in Section V using previously published data. In Section VI we present results on a new and motivating data set we have collected, on time-evolving gene-expression data for influenza. Conclusions are discussed in Section VII.

II. Bayesian Sparse Linear Linear Regression

A. *Bayesian View of Lasso and Elastic Net*

From a Bayesian viewpoint, the Lasso estimator in (2) can be interpreted as a maximum *a posterior* (MAP) estimator with Laplace priors placed independently on the components of $\boldsymbol{\beta}$ [9], [15]. In order to make the model inference tractable, the Laplace prior is written in the following hierarchical form [9], [15], [16]:

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

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 [9], [15]

$$\begin{aligned} \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}\boldsymbol{I}) \\ \beta_j &\sim \mathcal{N}(\beta_j; 0, \tau^{-1}\alpha_j^{-1}) \\ \tau &\sim \text{Ga}(\tau; c_0, d_0) \qquad (4) \\ (\alpha_j, \gamma_j) &\sim \text{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right) \text{Ga}(\gamma_j; a_0, b_0) \end{aligned}$$

for $j = 1, 2, \cdots, p$. It should be noted that [15] and [9] start with a simpler model where no prior is imposed on γ_j. Instead, a fixed parameter λ replaces γ_j. The model in (4) is proposed as an extension to the model in [15] by using feature-specific hyperparameters. The hierarchical prior on β_j is also called normal-exponential-gamma distribution [17]. 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

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

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 [9], [15].

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

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

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 [18].

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 [10], where the Bayesian Elastic Net model was proposed as

$$\begin{aligned} \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{X\beta}, \tau^{-1}\boldsymbol{I}) \\ \beta_j &\sim \mathcal{N}(\beta_j; 0, \tau^{-1}(\alpha_j + \lambda_2)^{-1}) \\ \tau &\sim \mathrm{Ga}(\tau; c_0, d_0) \\ \alpha_j &\sim \eta\left(\alpha_j/(\alpha_j + \lambda_2)\right)^{\frac{1}{2}} \mathrm{InvGa}\left(\alpha_j; 1, \frac{\gamma}{2}\right) \end{aligned} \tag{5}$$

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 [19]. The full likelihood of this Bayesian Elastic Net model is

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

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

$$p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\gamma}) \propto f(\tau, \boldsymbol{\gamma}) \exp\left(-\frac{\tau}{2}\left(\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + 2\sqrt{\gamma\tau^{-1}}\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\right)\right)$$

where $f(\tau, \boldsymbol{\gamma}) = \tau^{\frac{n+p}{2}}\mathrm{Ga}(\tau; c_0, d_0)$. Here $2\sqrt{\gamma\tau^{-1}}$ plays the role of λ_1 in (3). MCMC inference for this model is derived in [10], [19] and promising results are reported for gene selection.

Although theoretically valid, empirical result shows that the Bayesian Elastic Net model in (5) 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 (5). 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 (4). 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.

The proposed Bayesian Elastic Net model is also extended to probit regression, to deal with classification problems with a sparse set of correlated important covariates. Further, we employ the new Bayesian Elastic Net prior on multi-task learning, in which multiple sparse classifiers are learned jointly.

B. Proposed Bayesian Elastic Net

Based on the model in (5), we propose a new Bayesian Elastic Net model:

$$\begin{aligned}
\boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{X\beta}, \tau^{-1}\boldsymbol{I}) \\
\beta_j &\sim \mathcal{N}(\beta_j; 0, \tau^{-1}(\alpha_j + \lambda_2)^{-1}) \\
\tau &\sim \text{Ga}(\tau; c_0, d_0) \qquad (6)\\
(\alpha_j, \gamma_j) &\sim \eta\,(\alpha_j/(\alpha_j + \lambda_2))^{\frac{1}{2}}\,\text{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right)\text{Ga}(\gamma_j; a_0, b_0)
\end{aligned}$$

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 [10], [19] 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

$$\begin{aligned}
p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \propto\, &\mathcal{N}(\boldsymbol{y}; \boldsymbol{X\beta}, \tau^{-1}\boldsymbol{I})\text{Ga}(\tau; c_0, d_0) \\
&\times \prod_{j=1}^{p} \mathcal{N}(\beta_j; 0, \tau^{-1}(\alpha_j + \lambda_2)^{-1})\,(\alpha_j/(\alpha_j + \lambda_2))^{\frac{1}{2}} \qquad (7)\\
&\times \text{InvGa}(\alpha_j; 1, \frac{\gamma_j}{2})\text{Ga}(\gamma_j; a_0, b_0)
\end{aligned}$$

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 (4). By integrating out $\boldsymbol{\alpha}$, the full likelihood can be expressed as

$$p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\gamma}) \propto f(\tau, \boldsymbol{\gamma}) \exp\left(-\frac{\tau}{2}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \sum_{j=1}^{p} 2\sqrt{\gamma_j \tau^{-1}}|\beta_j| + \lambda_2\|\boldsymbol{\beta}\|_2^2)\right)$$

and $f(\tau, \boldsymbol{\gamma}) = \tau^{\frac{n+p}{2}} \mathrm{Ga}(\tau; c_0, d_0) \prod_{j=1}^{p} \gamma_j^{\frac{1}{2}} \mathrm{Ga}(\gamma_j; a_0, b_0)$. 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 [20].

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, it is possible to define a finite, discrete set of λ_2, place a uniform prior on these, and perform Bayesian inference on λ_2. In this manner we may realize a fully Bayesian architecture. 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 opted for cross-validation here (which also allows a "fair" comparison to results computed via the original Elastic Net, for which two-parameter cross-validation is performed).

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 (6) 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 [15] and [16]. The resulting Bayesian Elastic Net model for probit regression is

$$\begin{aligned}
z_i &\sim 1(z_i = \mathrm{sign}(y_i)) \\
\boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}\boldsymbol{I}) \\
\beta_j &\sim \mathcal{N}(\beta_j; 0, \tau^{-1}(\alpha_j + \lambda_2)^{-1}) \\
\tau &\sim \mathrm{Ga}(\tau; c_0, d_0) \qquad (8) \\
(\alpha_j, \gamma_j) &\sim \eta \left(\alpha_j/(\alpha_j + \lambda_2)\right)^{\frac{1}{2}} \mathrm{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right) \mathrm{Ga}(\gamma_j; a_0, b_0)
\end{aligned}$$

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 $\tau = 1$ [16], 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 [13], [14].

Assume M tasks, each represented as a regression model:

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

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

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

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 V and VI, 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) [21] 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 [22] 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 [23] compromises between inference accuracy and computational efficiency. In VB inference, the full posterior of the model parameters are approximated as a product of marginal posteriors. These marginal posteriors are inferred by minimizing the KL distance between the approximated posterior and the true joint posterior, yielding an analytical solution under particular circumstances. The main advantage of VB is its computationally efficiency and fast convergence, which can be monitored via the variational lower bound.

The full likelihood $p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ for the new Bayesian Elastic Net model is given in (7). In variational Bayes, we seek a distribution $Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ to approximate the exact posterior $p(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}|\boldsymbol{y})$. According to Jensen's inequality,

$$\begin{aligned} \log p(\boldsymbol{y}) &= \log \int Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \frac{p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})}{Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})} d\boldsymbol{\beta} d\tau d\boldsymbol{\alpha} d\boldsymbol{\gamma} \\ &\geq \int Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \log \frac{p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})}{Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})} d\boldsymbol{\beta} d\tau d\boldsymbol{\alpha} d\boldsymbol{\gamma} \\ &= \log p(\boldsymbol{y}) - \mathrm{KL}\left(Q(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \| p(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}|\boldsymbol{y})\right) \end{aligned} \tag{9}$$

Expression (9) 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

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

Then the variational lower bound in (9) becomes

$$J = \int Q(\boldsymbol{\beta}) Q(\tau) Q(\boldsymbol{\alpha}) Q(\boldsymbol{\gamma}) \log \frac{p(\boldsymbol{y}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})}{Q(\boldsymbol{\beta}) Q(\tau) Q(\boldsymbol{\alpha}) Q(\boldsymbol{\gamma})} d\boldsymbol{\beta} d\tau d\boldsymbol{\alpha} d\boldsymbol{\gamma} \tag{10}$$

By maximizing the lower bound in (10) 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 [23], the update equations for each Q function are derived as follows.

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

$$Q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{11}$$

with

$$\boldsymbol{\Sigma} = \left(\langle\tau\rangle \boldsymbol{X}^\top \boldsymbol{X} + \langle\tau\rangle(\text{diag}(\langle\boldsymbol{\alpha}\rangle) + \lambda_2 \boldsymbol{I})\right)^{-1} \tag{12}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\langle\tau\rangle \boldsymbol{X}^\top \boldsymbol{y}\right) \tag{13}$$

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 $\boldsymbol{\alpha}$:

$$Q(\boldsymbol{\alpha}) = \prod_{j=1}^{p} Q(\alpha_j) = \prod_{j=1}^{p} \text{InvGaussian}\left(\alpha_j; g_j, h_j\right) \tag{14}$$

with

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

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:

$$\text{InvGaussian}\left(\alpha_j; g_j, h_j\right) = \left(\frac{h_j}{2\pi\alpha_j^3}\right)^{\frac{1}{2}} \exp\left(-\frac{h_j(\alpha_j - g_j)^2}{2g_j^2\alpha_j}\right) \quad (\alpha_j > 0)$$

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

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

$$Q(\boldsymbol{\gamma}) = \prod_{j=1}^{p} Q(\gamma_j) = \prod_{j=1}^{p} \text{Ga}(\gamma_j; a_j, b_j) \tag{15}$$

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:

$$Q(\tau) = \text{Ga}(\tau; c, d) \tag{16}$$

with

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

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 (10) is feasible but omitted here.

Calculation of $\boldsymbol{\Sigma}$ in (12) 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 [24], 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 [5], 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

$$\hat{\boldsymbol{\beta}} = \xi \cdot \tilde{\boldsymbol{\beta}}$$

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 (13). Here ξ is a scaling constant, and it is recommended in [5] 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:

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

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

It is important to emphasize that the need for rescaling is *removed* for the case of classification when employing probit regression, as discussed below (this extension has not been considered previously with the Elastic Net, to the authors' knowledge, although related methods have been employed for an ℓ_p norm [25]).

B. Inference with probit

The inference for the model in (8) is similar to that in (6), as given from (11) to (16), except that now we need to introduce another $Q(\cdot)$ function for the latent regression response $\boldsymbol{y}$:

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

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 $Q(y_i) \propto 1(y_i \geq 0) \cdot \mathcal{N}(y_i; \theta_i, \sigma^2)$, otherwise $Q(y_i) \propto 1(y_i < 0) \cdot \mathcal{N}(y_i; \theta_i, \sigma^2)$. Statistics of the truncated normal distribution $Q(y_i)$ can be computed as follows:

$$\begin{aligned} \langle y_i \rangle &= \theta_i + z_i \sigma \frac{\phi(\theta_i/\sigma)}{\Phi(z_i\theta_i/\sigma)} \\ \langle y_i^2 \rangle &= \sigma^2 + \langle y_i \rangle \cdot \theta_i \\ -\langle \log Q(y_i) \rangle &= \frac{\log(2\pi\sigma^2)}{2} + \frac{\langle (y_i - \theta_i)^2 \rangle}{2\sigma^2} + \log \Phi(\frac{z_i \theta_i}{\sigma}) \end{aligned} \tag{17}$$

Here $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative density function for standard normal distribution, respectively. Consequently, $\boldsymbol{y}$ in (13) should be replaced with its expectation $\langle \boldsymbol{y} \rangle$, and we should also treat $\boldsymbol{y}$ as a random variable for the expectation $\langle \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 \rangle$ in (16). 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

$$p(z_\star = 1 | \tau, \boldsymbol{x}_\star) = \Phi\left(\frac{\boldsymbol{x}_\star^\top \boldsymbol{\mu}}{(\tau^{-1} + \boldsymbol{x}_\star^\top \boldsymbol{\Sigma} \boldsymbol{x}_\star)^{\frac{1}{2}}}\right) \tag{18}$$

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.

C. Multi-task inference

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 tasks as expressed below, instead of a inverse Gaussian distribution in (14).

$$Q(\boldsymbol{\alpha}) = \prod_{j=1}^{p} Q(\alpha_j) = \prod_{j=1}^{p} \mathrm{GIG}\left(\alpha_j; q_j, g_j, h_j\right)$$

with

$$q_j = \frac{M}{2} - 1, \quad g_j = \sum_{m=1}^{M} \langle \tau^{(m)} \rangle \langle (\beta_j^{(m)})^2 \rangle, \quad h_j = \langle \gamma_j \rangle$$

Here $\mathrm{GIG}\left(\alpha_j; q_j, g_j, h_j\right)$ denotes the Generalized Inverse Gaussian distribution with parameter q_j, g_j and h_j:

$$\mathrm{GIG}\left(\alpha_j; q_j, g_j, h_j\right) = \frac{(g_j/h_j)^{q_j/2}}{2K_{q_j}(\sqrt{g_j h_j})} \alpha_j^{q_j - 1} \exp\left(-\frac{1}{2}\left(g_j \alpha_j + h_j \alpha_j^{-1}\right)\right) \quad (\alpha_j > 0)$$

where $K_{q_j}(\cdot)$ is the modified Bessel function of the third kind, and

$$\langle \alpha_j \rangle = \frac{K_{q_j+1}(\sqrt{g_j h_j})}{\sqrt{\frac{g_j}{h_j}} K_{q_j}(\sqrt{g_j h_j})}$$

$$\langle \alpha_j^{-1} \rangle = \frac{\sqrt{\frac{g_j}{h_j}} K_{q_j-1}(\sqrt{g_j h_j})}{K_{q_j}(\sqrt{g_j h_j})} = \left(g_j \langle \alpha_j \rangle - 2q_j\right)/h_j$$

Recursive equation $K_{q_j+1}(t) = \frac{2q_j}{t} K_{q_j}(t) + K_{q_j-1}(t)$ is used in the above calculation. Specifically for $M = 1$, $q_j = \frac{M}{2} - 1 = -\frac{1}{2}$, the above GIG distribution reduces to inverse Gaussian distribution, which agrees with the single task update equation in (14).

V. Experiments on Published Data

A. Simulation Data

Our first example is from the original Elastic Net paper [5], in which two tuning parameters are employed, as compared to the one parameter associated with the proposed model. We simulate data from the following model:

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

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:

$$\begin{aligned}
&\boldsymbol{X}_i = \boldsymbol{Z}_1 + \boldsymbol{\omega}_i \quad \boldsymbol{Z}_1 \sim \mathcal{N}(\boldsymbol{Z}_1; \boldsymbol{0}, \boldsymbol{I}) \quad \boldsymbol{\omega}_i \sim \mathcal{N}(\boldsymbol{\omega}_i; 0, 0.01\boldsymbol{I}) \quad i = 1, 2, ..., 5 \\
&\boldsymbol{X}_j = \boldsymbol{Z}_2 + \boldsymbol{\omega}_j \quad \boldsymbol{Z}_2 \sim \mathcal{N}(\boldsymbol{Z}_2; \boldsymbol{0}, \boldsymbol{I}) \quad \boldsymbol{\omega}_j \sim \mathcal{N}(\boldsymbol{\omega}_j; 0, 0.01\boldsymbol{I}) \quad j = 6, 7, ..., 10 \\
&\boldsymbol{X}_k = \boldsymbol{Z}_3 + \boldsymbol{\omega}_k \quad \boldsymbol{Z}_3 \sim \mathcal{N}(\boldsymbol{Z}_3; \boldsymbol{0}, \boldsymbol{I}) \quad \boldsymbol{\omega}_k \sim \mathcal{N}(\boldsymbol{\omega}_k; 0, 0.01\boldsymbol{I}) \quad k = 11, 12, ..., 15 \\
&\boldsymbol{X}_m \sim \mathcal{N}(\boldsymbol{X}_m; \boldsymbol{0}, \boldsymbol{I}) \quad m = 16, 17, ..., 40
\end{aligned}$$

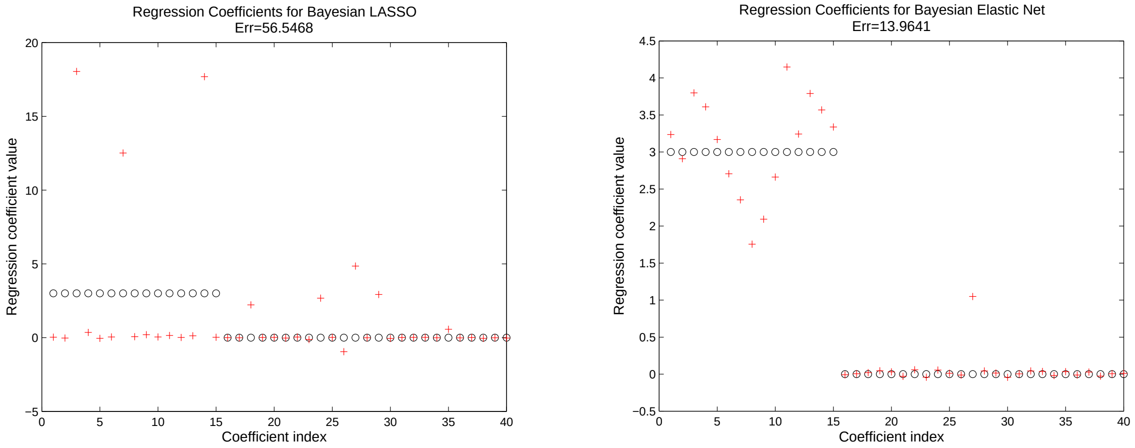
Here $\boldsymbol{X}_i$ denotes the i^{th} column of $\boldsymbol{X}$, with dimension $n \times 1$. The ground truth of the regression coefficient $\boldsymbol{\beta}$ is set to be $\boldsymbol{\beta} = [3, 3, \cdots, 3, 0, 0, \cdots, 0]^\top$ with first 15 elements equal to 3 and the rest 0. Therefore 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 ($m = 16, 17, ..., 40$). After cross validation, we choose $\lambda_2 = 80$ for this example.

The performance measure is the residue in the testing set:

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

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 the mismatch between the estimated model ($\hat{\boldsymbol{\beta}}$) and the true model ($\boldsymbol{\beta}$). We generate 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 practice, 100 such runs are performed. The mean of these 100 median Err is 29.58, which is slightly better than the error reported in the original paper [5] (34.5). The Bayesian Lasso gives median of Err around 85, much larger than Bayesian Elastic Net. Typical estimation results are plotted in Figure 1.

From the figure we observe that Bayesian Lasso does not perform group selection, since for each group only one or two features are selected. In contrast, Bayesian Elastic Net can select all 15 important features, thus the performance in testing set is more robust. Besides Elastic Net, another algorithm named group Lasso [26] can also do grouped variable selection. However, we do not compare with group Lasso here, since we do not assume that the grouping information is known *a priori*.



(a)

(b)

Fig. 1. Typical estimation results of Bayesian elastic net ($\lambda_2 = 80$) and Bayesian Lasso ($\lambda_2 = 0$). Black circles are the ground truth. (a) Bayesian Lasso, (b) Bayesian Elastic Net

B. Single-task gene analysis

We consider the leukaemia data in [27], 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 leukaemia (ALL) and the other is type-2 leukaemia (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. 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 choose $\lambda_2 = 10$ to achieve a balance between sparsity and grouped variable selection in the training set; we also tried $\lambda_2 = 15$ and $\lambda_2 = 20$, and the results are quite similar. Twenty genes are selected by the Bayesian Elastic

Net, with 0/38 training errors and 1/34 testing error. The selected genes are: **X95735**, **U50136**, **Y12670**, **M23197**, D49950, **X85116**, **M55150**, **M16038**, **X17042**, **M80254**, **L08246**, **U82759**, M22960, **M84526**, **U46751**, **M27891**, **M83652**, **Y00787**, M81933, Y00339. Genes in boldface are also reported in [27].

In contrast, by using Bayesian Lasso ($\lambda_2 = 0$), only one gene is selected (**X95735**), with 0/38 training error and 3/34 testing errors. The original Elastic Net paper [5] reported 3/38 training errors and 0/34 testing error. In [5] the authors coded the type of leukaemia as a $0-1$ response $\boldsymbol{y}$, and rescaling was required, with this not needed for the Bayesian Elastic Net 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.

C. Multi-task gene analysis

We apply the multi-task Bayesian Elastic Net model introduced in Section III to gene-expression data for small round blue cell tumors [28]. These data were previously used in [14]. 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 (18).

Following the pre-processing procedure in [14], 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).

TABLE I

CLASSIFICATION ERRORS FOR MULTI-TASK (MT) BAYESIAN LASSO AND MULTI-TASK (MT) BAYESIAN ELASTIC NET. '$3-2$' MEANS 3 ERRORS IN THE TRAINING SET AND 2 ERRORS IN THE TESTING SET.

	Task 1	Task 2	Task 3	Task 4	Combined
MT Lasso	$2-3$	$0-0$	$0-0$	$2-2$	$3-2$
MT ENet	$0-1$	$0-0$	$0-0$	$0-0$	$0-0$

For this experiment, we choose $\lambda_2^{(m)} = 10$ for all tasks to achieve a balance between sparsity and grouped variable selection (again, the results were relatively insensitive to the choice of λ_2). Comparisons are made to multi-task Bayesian Lasso, for which $\lambda_2^{(m)} = 0$.

For multi-task Bayesian Lasso, only four genes (**770394**, **236282**, **812105**, **784224**) are selected and other correlated genes are suppressed. The training and testing errors using these four biomarkers are 3/63 and 2/20.

For multi-task Bayesian Elastic Net, twelve genes are selected: **770394**, **1435862**, **377461**, 814260, 183337, **236282**, **812105**, **325182**, **383188**, **784224**, **796258**, **207274**. Genes in boldface are also reported in [14]. These twelve genes cluster naturally into four groups (see Figure 2), corresponding to four tumor categories, and genes within each group are strongly correlated. The training and testing errors using these twelve biomarkers are 0/63 and 0/20. The testing result is the same as that reported in [14]. The ability of Bayesian Elastic Net to do grouped variable selection not only improves interpretation but also improves the prediction performance in the testing set (compared to MT Lasso; see Table 1).

VI. ANALYSIS OF TIME-EVOLVING INFLUENZA EXPRESSION DATA

A. Data collection

A healthy volunteer intranasal challenge with influenza A/Wisconsin/67/2005 (H3N2) was performed at Retroscreen Virology, LTD (Brentwood, UK), using 17 pre-screened volunteers who provided informed consent. On day of inoculation, a dose of 106 TCID50 Influenza A manufactured and processed under current good manufacturing practices (cGMP) by Bayer Life Sciences, Vienna, Austria) was inoculated intranasally per standard methods at a varying

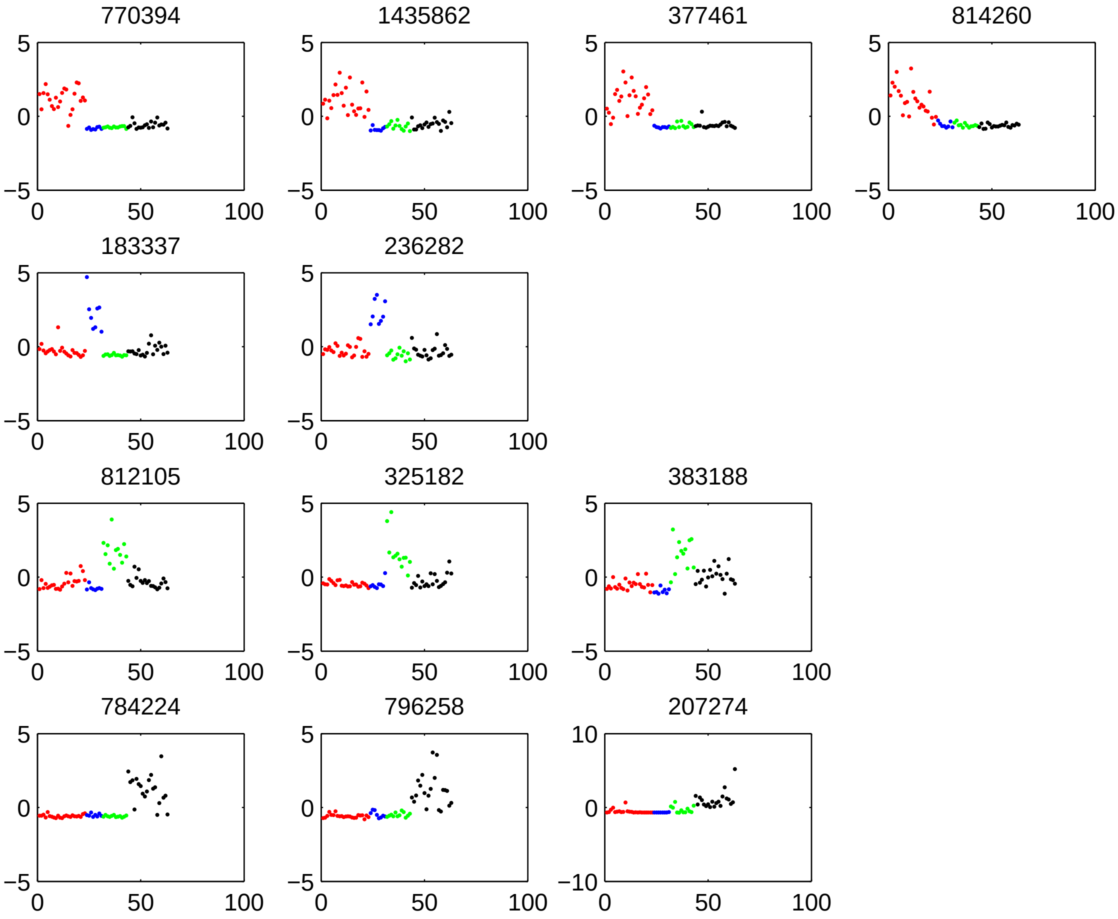


Fig. 2. Gene expression values for the genes selected by the multi-task Bayesian Elastic Net model in the training set. Different colors represent different kind of cancers: red - EWS, blue - BL, green - NB, black - RMS. Note that these genes do a relatively good job of classifying the cancers, as they manifest cancer-dependent expression values (vertical axes).

dose (1:10, 1:100, 1:1000, 1:10000) with four to five subjects receiving each dose. 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.

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 [4].

B. MTL Elastic Net analysis

The question we are trying to answer is whether we can build classifiers to distinguish those individuals who will become symptomatic from those who will not, based on classifiers designed for each of the 14 post-inoculation time points. We use a leave-one-time-out technique to test the models. Specifically, multi-task learning is applied using all data from 13 time points, holding out all data at the left-out time. The learned classifier from the nearest adjoining time is then used to classify all of the data from the held-out set (using the earlier time if there are two such adjoining times). This leave-one-time-out procedure is done for each time point, and the testing errors are plotted in Figure 3. We used the top 500 genes in this analysis, after performing a Fisher-score analysis, and we chose $\lambda_2^{(m)} = 5$ for all tasks.

Throughout the experiments, around nine genes are selected every time (for a given hold-out case) and the specific set of selected genes was relatively stable. As can be seen in Figure 4, all selected genes have some discriminative power across all time points. Multi-task Bayesian Lasso model gives similar classification errors, but only selects two genes at each time. The two genes are within the nine genes of the multi-task Elastic Net model, but change for the different time points (change when a different time point is held out).

More details on these data, with a particular focus on biological interpretation, are provided in [4] and will be provided in additional forthcoming papers.

VII. Conclusion

The Elastic Net model developed in [5] has been extended to a Bayesian version, with the number of model parameters reduced from two to one. Efficient inference is performed using

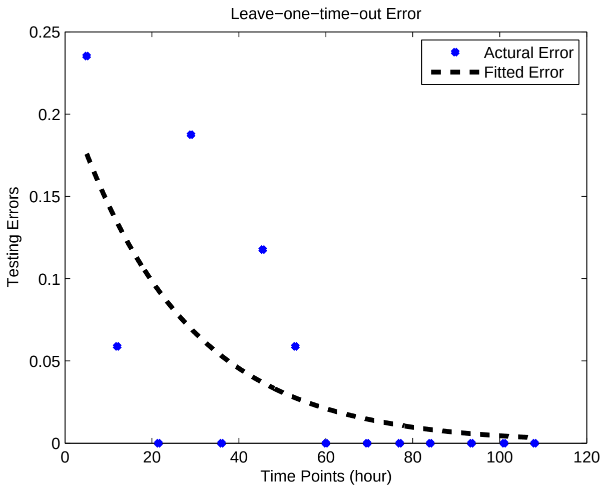


Fig. 3. Leave-one-time-out errors for each of the 14 time points. The actual errors are shown in dots, and an exponential fit is also plotted to smooth the raw classification errors (to aid visualization).

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 model has been examined on several widely examined data sets available in the literature, with comparison to the original Elastic Net and to Bayesian Lasso [9]. The model consistently performed comparably to the original Elastic Net, with the advantage of only requiring one parameter. The Bayesian Lasso (and original non-Bayesian Lasso) typically manifested inferior classification performance, and particularly was less interpretable (it only selected a small subset of correlated features/genes, this motivating the original Elastic Net).

The original Elastic Net was not employed for multi-task classification problems, this constituting a unique feature of the model and results presented here. Using one multi-task example based on published gene-expression data, the proposed model yielded very encouraging classification results while inferring a relatively large set of meaningful genes.

The proposed methodology has been motivated by a new influenza challenge study [4] we have undertaken, in which gene-expression data are employed to try to infer which individuals, inoculated with the virus, will ultimately become symptomatic. Gene-expression data are avail-

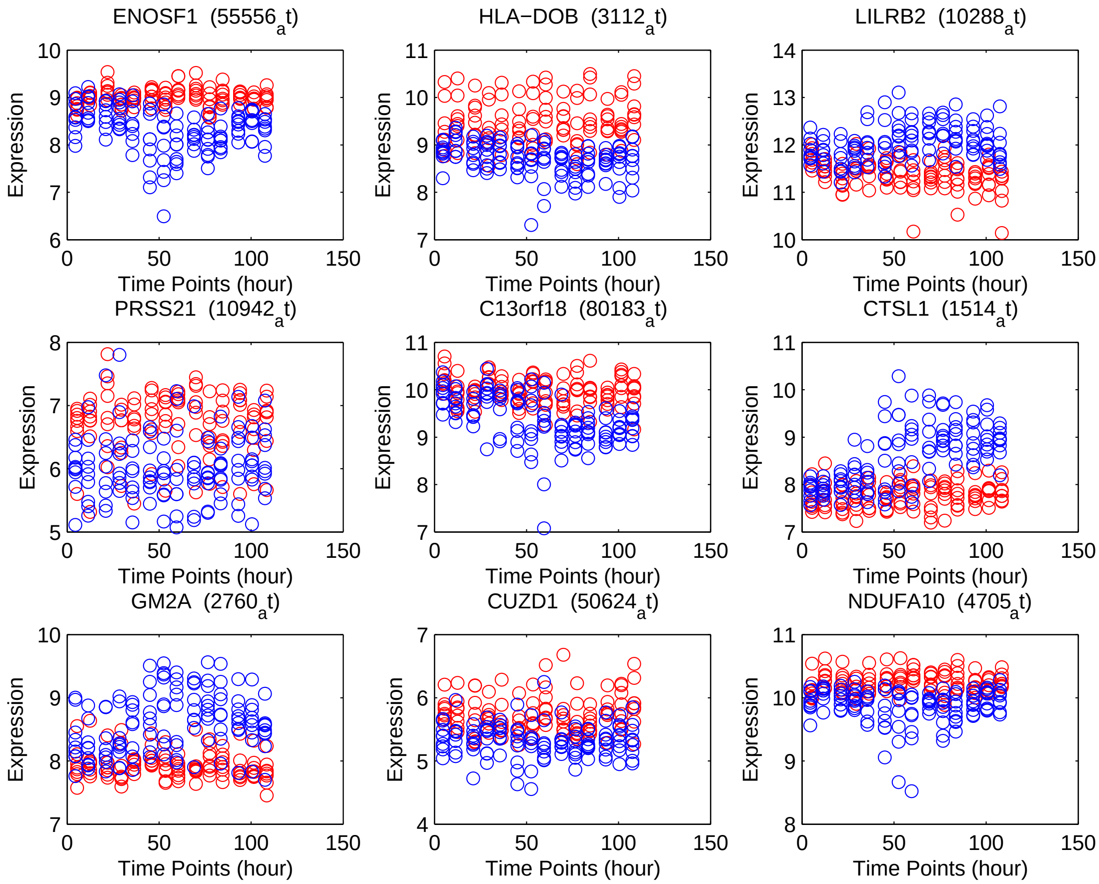


Fig. 4. Nine genes (atop each sub-figure) selected by the multi-task Bayesian Elastic Net model applied to the influenza data (the horizontal axis represents time in hours, and the vertical axis denotes gene-expression value). Red circles represent healthy samples and blue circles represent infected samples.

able at multiple time points after inoculation, and a Bayesian Elastic Net classifier is designed at each time point. We have presented multi-task results on classification performance and have also inferred genes that are relevant for this classification task. The results from this study appear encouraging, in that the inferred genes do indeed appear to be discriminative, and the Elastic Net framework selects multiple correlated genes. The data from this study will be made available to other researchers, such that other models may be examined in the future.

REFERENCES

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

[2] C. Carvalho, J. Chang, J. Lucas, J.R Nevins, Q. Wang, and M. West. High-dimensional sparse factor modelling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103:1438–1456, 2008.

[3] D. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10:515–534, 2009.

[4] 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*, 2009 (to appear).

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

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

[7] 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.

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

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

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

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

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

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

[14] 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.

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

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

[17] 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.

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

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

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

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

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

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

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

[25] Zhenqiu Liu, Feng Jiang, Guoliang Tian, Suna Wang, Fumiaki Sato, Stephen J. Meltzer, and Ming Tan. Sparse logistic regression with lp penalty for biomarker identification. *Statistical Applications in Genetics and Molecular Biology*, 6, 2007.

[26] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society Series B*, 68:49–67, 2006.

[27] 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.

[28] 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.