Bayesian Methods for Proteomic Biomarker Discovery

Belinda Hernández [1,3], Stephen R Pennington* [3], Andrew C Parnell *[1,2]

[1] School of Mathematical Sciences (Statistics), University College Dublin, Belfield Campus, Dublin 4, Ireland

[2]Insight: The National Centre for Data Analytics, University College Dublin, Belfield Campus, Dublin 4, Ireland

[3] School of Medicine and Medical Science, UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield Campus, Dublin 4, Ireland

*Joint contribution

Corresponding Author

Email: Belinda.Hernandez@ucdconnect.ie

Abstract

The advent of liquid chromatography mass spectrometry has seen a dramatic increase in the amount of data derived from proteomic biomarker discovery. These experiments have seemingly identified many potential candidate biomarkers. Frustratingly, very few of these candidates have progressed to the stage of routine clinical use. It is becoming apparent that the statistical methods used to assess the performance of new candidate biomarkers are a major limitation in their development. Bayesian methods offer some advantages over traditional statistical and machine learning methods. In particular they can incorporate external information into current experiments so as to guide biomarker selection. Further, they can be more robust to over-fitting than other approaches, especially when the number of samples used for discovery is small.

In this review we provide an introduction to Bayesian inference and demonstrate some of the advantages of using a Bayesian framework. We summarise how Bayesian methods have been used previously in proteomics and other areas of bioinformatics. Finally, we describe some popular and emerging Bayesian models from the statistical literature and provide a worked tutorial including code snippets to show the reader how to use these methods for the discovery of proteomic biomarkers in their own experiments.

1 Introduction

Advances in proteomic technology, in particular the widespread use of liquid chromatography mass spectrometry (LC-MS), have meant that vast amounts of information regarding protein and peptide features can now be easily collected from bodily fluids and tissue, making them an ideal target to find biomarkers of disease. A mass spectrum sample can be represented as a series of peaks where the mass to charge ratio (m/z) is depicted on the x-axis and the molecule intensity on the y-axis. In statistical and bioinformatic analysis each m/z ratio is treated as a separate variable where its value is the intensity or abundance of the molecule at the given m/z ratio. Each peak generally corresponds to a protein fragment or peptide and so the objective of most biomarker discovery experiments is to find a subset of peptides that best discriminate between the outcome groups [1]. It is widely accepted that use of individual biomarkers are unlikely to sufficiently capture the complexity and possible heterogeneity of a given disease [2][3][4]. For this reason, most studies focus on finding a panel or signature of differentially expressed protein or peptide features that are both sensitive and specific enough to accurately predict a treatment or disease state.

It has now become clear that the issue of finding a sensitive and specific panel of biomarkers is much more complex than initially anticipated. The area of proteomic biomarker discovery was initially met with high hopes and great enthusiasm; however this fervor has weaned in recent years due to the inability of many studies to validate candidate biomarkers that were initially thought to be highly discriminatory [5][6]. Because of this, few proteomic biomarkers have reached clinical utility despite much government and industry investment [7][8]. Many articles have reflected on the shortcomings of these earlier studies and have laid out guidelines to rectify the oversights of initial experiments [7][9][10].

Bayesian methods have been widely used in many areas of bioinformatics and proteomics mainly due to the fact that they lend themselves nicely to the challenge of analyzing complex, noisy and often incomplete data [11]. Their growing popularity over the last 20 years is mainly attributable to

advances in computational power which make fitting Bayesian models much more attainable for large datasets [12]. This article will review the literature on Bayesian methods in proteomics in general before focusing on how Bayesian methods can be used for the statistical analysis of mass spectra data for proteomic biomarker discovery. We will discuss the benefits of using Bayesian models compared to other traditional and machine learning methods and will also identify some reasons why Bayesian models might attain superior performance in the validation of separate cohorts. We also highlight methods used in other areas of research and other recent developments in Bayesian analysis which could prove to be useful in future applications of proteomic biomarker discovery experiments. Section 5.4 includes a worked proteomic discovery example for the prediction of cardiovascular disease where two of these methods are tested and compared against each other. This section also includes code samples which the reader can use to run these models for their own experiments using freely available software. Readers not interested in following the tutorial may wish to skip Section 5.4 and proceed directly to Section 6.

2 What is Bayesian Inference?

At the heart of all Bayesian methods is Bayes' theorem,

$$p(\theta \mid Y) \propto p(\theta) \times p(Y|\theta)$$

often expressed in words as:

posterior is proportional to prior times likelihood

In the above equation *Y* is the experimental data and θ are the unknown parameters (e.g. peptide importance values). The posterior distribution $p(\theta \mid Y)$ is the joint probability distribution of the unknown parameters given the observed data. Bayes' theorem states that the posterior distribution can be calculated from a combination of a probability distribution on the unknown parameters of interest $p(\theta)$ known as the prior distribution and a conditional probability distribution $p(Y|\theta)$ of the data Y given the parameters θ, known as the likelihood.

Commonly, the prior distribution $p(\theta)$ represents the knowledge about the parameters of interest θ before any data is collected. Its shape represents the degree of certainty or knowledge about θ; for example a distribution with a sharp peak would express high confidence in our knowledge of θ whereas a flat or uninformative prior would express no prior knowledge about the parameters of interest. When data become available after an experiment has been conducted, the information about the data and the parameters of interest are combined through Bayes' theorem to produce $p(\theta \mid Y)$. The main aim of any Bayesian analysis is to identify a credible set of values that the parameters θ can take given the observed data *Y* [12][13], i.e. find the posterior distribution.

Often the full form of the posterior distribution is unavailable due to the calculation of the normalising constant in the proportionality constraint. This problem is neatly sidestepped by using fitting methods such as Markov Chain Monte Carlo (MCMC) which make inference about the posterior distribution by sampling from it rather than computing it explicitly.

3 Motivation for using Bayesian Methods

One of the main advantages of Bayesian methods over non-Bayesian statistical and machine learning techniques is the ability to incorporate external information about the parameters through the prior distribution. In proteomic experiments in particular a great deal is already known about the parameters of interest before an experiment takes place which can be incorporated into the prior distribution. For example, if it was known that certain peptide features tend to have high technical variability and be less reproducible (as is often the case in MS analysis with low abundant features whose intensity is near the limit of detection of the mass spectrometer) a less informative prior could be used on these peptides as opposed to the higher abundance, more reproducible features.

One of the main reasons for the failure of many initial discovery studies to validate according to [14] is the failure to accurately model sources of experimental and biological variability. Many traditional pre-existing techniques have been used to analyse the data resulting from proteomic biomarker experiments such as support vector machines, random forests, lasso regression and various other classification methods [15][16][17]. However, the one disadvantage common to all these methods is that they ignore the uncertainty introduced to the data and assume that the experimental data are the only data available. This uncertainty can however be modelled and incorporated into a Bayesian framework in a consistent manner [18].

A common feature of proteomic discovery datasets is that the number of variables p tends to be much larger than the number of samples n, giving rise to problems of 'over-fitting' when traditional methods are used [19]. Over-fitting means that the chosen model fits the current data set too precisely, giving over-optimistic estimates of model performance that would not be repeated on an external validation cohort. The model is thus apportioning signal to random noise rather than identifying a true underlying model. Traditionally over-fitting is discouraged during the model building phase by adding a penalty for the complexity of the model. This is known as regularisation. Bayesian models overcome over-fitting in a similar manner, though the penalty is more explicitly stated via the prior distribution. Further flexibility can be obtained by marginalising over (i.e. removing through integration) or shrinking parameters [20][21] and so when used correctly will have a better chance of validating on a separate cohort. For example Kuschner *et al* used a Bayesian Belief Network (BN) on both simulated and authentic proteomic data to discriminate between patients with sub types of Human T-cell Leukemia Virus type 1, and found that a BN with informative priors far outperformed traditional Linear and Quadratic Discriminant analysis with regards to cross validated and test accuracy [22]. They also found that biomarkers selected by the BN were far more stable over multiple iterations than the other methods tested. Similarly, Vannucci, Shaw and Brown [23] used probit models with Bayesian mixture priors and latent variables to classify women with ovarian cancer from their mass spectrum profiles and found that their method performed accurately and selected biomarker panels which were consistent with the literature.

4 Bayesian models currently used in Proteomics

4.1 Biomarker Discovery

The objective of many biomarker discovery experiments is twofold; first to accurately classify samples into groups and second to select a subset of predictive peptide features or proteins which can further be validated and measured using specific targeted assays. Hence the analysis of

biomarker discovery data is not only a classification prediction problem but could also be viewed as a variable selection problem for high dimensional data where the number of important parameters is small.

With respect to proteomic mass spectrometric biomarker discovery, Bayesian models have not been as widely adopted as in other areas of proteomics. Some examples of Bayesian feature selection techniques have however emerged in the proteomics literature [22][23]. Yu and Chen [24] use an ovarian cancer dataset from the National Cancer Institute to showcase their proposed version of a hierarchical Bayesian neural network. They essentially deal with high dimensionality by filtering variables through the use of a Kolmogorov-Smirnov test which is used to set the hyper-parameters on the prior distribution of a variable being selected to a model. Deng, Geng and Ali [25] also proposed an interesting application of a Bayesian network where they used both microarray and mass spec experiments to choose biomarkers of prostate cancer. In this way their algorithm not only chose biomarkers which reported high predictive accuracy but also those which were supported by multiple sources of biological information. More recently the work of Serang *et al* used a Bayesian goodness of fit approach to detect the true number of differential features in an LC-MS experiment [26]. This method has the advantage that is avoids the need to specify many of the arbitrary cut-off choices common in most proteomic analyses such as the cut off for a "significant" fold change or a "significant" q value. With respect to isobaric labelled mass spectrometry data, Jow, Boys and Wilkinson proposed a hierarchical Bayesian method which was found to perform well in a variety of simulated and real proteomic experiments [27]. This Bayesian methodology in particular has the advantage that it can easily integrate multiple data experiments into a single model. Koh *et al* proposed an alternative model based approach for analysing data resulting from label based proteomic experiments [28]. They note that with labeling approaches many of the ratios used to identify differentially expressed proteins ignore the fact that some proteins are quantified using more peptides and are therefore more reproducible than others. Their proposed method incorporates this knowledge regarding the reproducibility of the protein quantities and models the hierarchical relationship between peptides and proteins directly giving greater importance to proteins which are the most reproducible.

4.2 Other areas of Proteomics

Bayesian methods have infiltrated many other areas of proteomic research apart from biomarker discovery. At a functional and structural level there are many examples of early adoptions to a probabilistic framework in areas such as sequence alignment [29][30] and predicting protein structures [31][32][33]. Bayesian methods have also proven popular for finding protein functions as well as predicting protein-protein interactions [34][35].

Mass spectrometry, and in particular peptide and protein identification from mass spectrum fragmentation data has also seen many contributions from Bayesian models. Serang , MacCross and Stafford Noble [36] proposed a Bayesian framework for protein inference with degenerate peptides to calculate the posterior probability of a given peptide belonging to a protein, which was found to outperform the popular software ProteinProphet on a number of datasets. Li et al [37] also proposed a fully probabilistic approach to protein identification, however their method was found to be too computationally intensive for use on large datasets [36]. ProteinProphet itself uses an empirical Bayes approach in which the prior is estimated from the data. Their method uses a two component mixture model to understand the distribution of peptide search scores observed for all

designated peptides [38][39]. Another software ProFound identifies proteins by searching through existing sequence databases and uses Bayes' theorem to calculate the posterior probability that each protein in the database is the current sample protein being analysed, given the experimental data and other available background information. Proteins are then ranked according to their posterior probability [40].

5 Possible Bayesian applications for biomarker discovery

The use of Bayesian methods for the statistical analysis of proteomic mass spec discovery data is still quite a new and emerging area and to-date has not reached the maturity of other areas of bioinformatics and systems biology. As mentioned previously, mass spectrometry discovery data tends to have far more variables (p) than samples (n) (commonly referred to as small n large p) however it is expected that very few of the variables measured are truly related to the response or outcome. For this reason many studies have focused on finding a small biomarker panel which can accurately predict disease [5][41]. There is a vast literature on Bayesian models used for feature selection on small n large p datasets in other areas of research, which thus could be applied to proteomic biomarker discovery. This section will outline some existing Bayesian models from the statistical literature which could have interesting applications to the area of proteomic biomarker discovery.

5.1 Bayesian Lasso

The Lasso model for linear regression is one of the most widely used methods for variable selection in high dimensional data [42]. The Lasso has also proven popular in various areas of proteomics. Huang et al [43] proposed ProteinLasso which uses the Lasso for protein inference; Friedman, Hastie and Tibshirani also showcased the graphical Lasso to find protein networks in flow cytometry cell signaling data [44]. For other examples see [45][46]. As previously stated, regularisation is a popular form of variable selection where a penalty is applied to the parameters in order to discourage complex models where many variables are chosen. The Lasso is a regularised version of ordinary least squares regression (for a continuous response) which balances model fit and model complexity by adding a penalty parameter which controls the absolute sum of the regression coefficients included in the model. The higher the penalty the more coefficients will have a value of zero and will be effectively eliminated from the model.

It was noticed that the coefficients returned by the original Lasso correspond to the mode of the posterior distribution in a Bayesian setting when a Laplace (double exponential) prior is placed on the parameter vector $p(\beta|\sigma^2)$ where σ^2 refers to the model variance; which led to the inception of the Bayesian Lasso [47]. The Bayesian Lasso allows for the full posterior of the model coefficients to be explored rather than just a point estimate, and so can give more instructive information regarding variable selection. Also, tuning parameters which control how harsh a penalty is placed on the model coefficients can be treated as unknown random variables and so their posterior distribution can also be sampled. This avoids the need for ad-hoc choices of tuning parameters such as those used in the traditional Lasso model. The Bayesian Lasso has been used in various biomedical and bioinformatics studies in recent years and has proven a popular approach for variable selection in data which have a sparse parameter space [48][49][50]. There is a freely available R package to run the Bayesian Lasso called "reglogit" [51] which is showcased in Section

5.4. The Bayesian Lasso can also be run using a package called rJAGS [52] which is shown in Section 5.4.

5.2 Other Priors for Variable Selection

There is a wide array of literature proposing different shrinkage priors other than that of the Laplace prior (used by the Bayesian Lasso above) which have been shown to be optimal in various settings. For example Dunson DB and Lee suggest use of a generalized double Pareto prior [53] and Griffin and Brown suggest a normal-gamma prior on β, which is a generalisation of the double exponential prior [54]. One of the most popular rivals to the Laplace prior of the Bayesian Lasso is the Horseshoe prior [55]. Carvalho, Polson and Scott claim that the main advantages to the use of the Horseshoe prior is that it is robust to large signals and is very effective in shrinking noise variables [55]. The horseshoe distribution is very heavy tailed with an infinitely large spike at zero. This means that coefficients near zero can be shrunk very efficiently but also that coefficients far from zero will not be shrunk as severely, allowing for large signal if it is evident in the data. The horseshoe prior, Bayesian Lasso and ridge regression can be run using package "monomvn" in R [56].

A less severe shrinkage prior, though still widely used for Bayesian variable selection, is Zellner's g prior, where the prior on the parameter vector takes the form: $\beta \sim N(\beta_0, g\sigma^2(X^TX)^{-1})$. Here β_0 is the value around which the regression coefficients are thought to centre (usually taken to be 0); the prior on σ^2 is generally a non-informative prior and g is the hyper-parameter on the model coefficients controlling the degree of shrinkage [57]. This is a popular prior because of its computational simplicity for calculating marginal likelihoods (the likelihood function where some parameters have been removed through integration), and the fact that only g has to be estimated. There have been many suggestions on how to treat g. Some authors suggest placing a prior distribution whereas others suggest using fixed values or estimating the value for g using Empirical Bayes methods; see [58][59] for more information. Use of the g prior with a probit model has previously been proposed in the context of gene microarray studies to classify a number of diseases including colon cancer and lukemia [60] and also in gene selection for expression data [61], as well as in a ridge regression for high dimensional microarray breast cancer data [59]. Regression using the g prior can be run in R using the BMS package [62].

An alternative to shrinkage is to directly model the probability of inclusion of a variable. A popular version of this for regression problems is that of Kuo and Mallick [63]. They introduce a vector of indicator variables which signify inclusion or exclusion of each parameter in the model as shown below [63].

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j I_j x_{ij} + \varepsilon_i$$

Here y_i is the value of response variable y for observation i , β_j is the coefficient for variable j, I_j is the indicator term which is 1 if variable j is to be included in the model and 0 otherwise, x_{ij} is the value of explanatory variable j for observation i and ε_i is random noise associated with observation i. Usually independent priors are placed on the β vector and the indicator vector I. One of the advantages of this method is that it doesn't require much tuning. However it can be slow to fit [64].

5.3 Bayesian non-parametric models

Both the Bayesian Lasso and other shrinkage priors used for regression models assume that the variables are linearly related and that variable interactions are known in the model specification. In this section we discuss examples of non-parametric Bayesian models which could be applied to proteomic biomarker discovery.

5.3.1 Bayesian CART

Classification and Regression decision trees (CART) are a popular method used in many areas of proteomics and in particular biomarker discovery [65][66][67] largely due to the fact they do not assume the covariates are linearly related to the response, often an unreasonably strong assumption in complex biological data. A key aspect of decision trees is that, unlike other regression models, they automatically select and include important variable interactions as part of the model building process and do not require these interactions to be explicitly specified in the model building. Decision trees consist of internal nodes where questions are asked based on a split rule consisting of a variable and a threshold, and terminal nodes which separate the observations into distinct groups. Observations which satisfy the split rule at each internal node are sent to the left hand daughter node and those which do not are sent to the right hand daughter node. Observations are iteratively split into left and right hand daughter nodes as they pass through each internal node in turn until a terminal node is reached. Figure 1 shows an example of a simple decision tree for a binary classification problem. Using Figure 1 as an example model to predict between two groups (control and treatment) it can be seen that all patients with Actin levels $\leq$ 5.3 would be predicted as belonging to the control group; all patients with Actin levels > 5.3 and Keratin levels $\leq$ 10.2 would be predicted as treatment and all patients with Actin levels > 5.3 and Keratin levels > 10.2 would be predicted as control. As can be seen, decision trees are easily interpretable, naturally perform variable selection, allow for complex non-linear interactions and also perform prediction, which are the main reasons for their popularity in biomarker discovery studies [65][66][67].

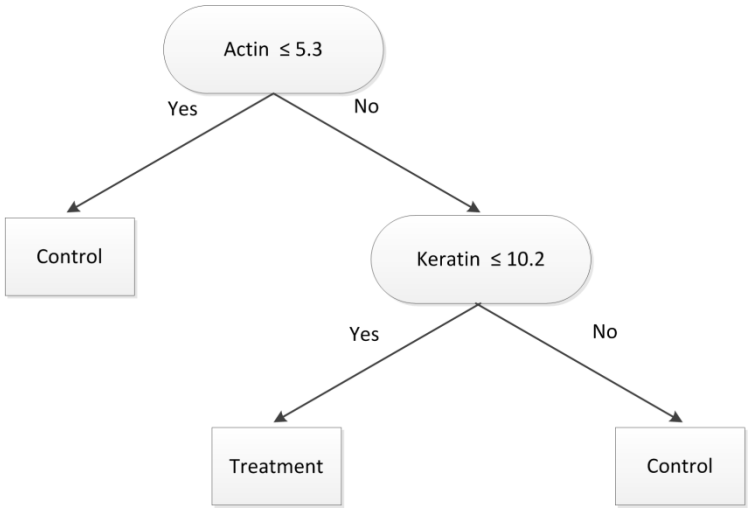

Figure 1 – Example of a binary classification decision tree

The Bayesian version of CART (Bayesian CART) has not, to our knowledge, yet been used in proteomic biomarker discovery but could have very interesting applications in this area as they combine the advantages of Bayesian models with those of traditional decision trees discussed above. For a Bayesian CART model, the data in each terminal node of the tree is assumed to follow a multinomial distribution for classification problems. This allows the probability of belonging to a

given class to be computed which can provide richer information than merely knowing the predicted labels.

Chipman, George and McCulloch proposed a version of Bayesian CART which samples from the posterior distribution of the trees using MCMC [68]. This essentially means that many trees are sampled from the posterior distribution by creating a chain of k trees $T_0, T_1 \dots T_k$. At each iteration k, a new tree is proposed and either accepted or rejected according to how well it matches the observed data and the prior distribution. The new trees are proposed by either growing or pruning (making the most recently accepted tree bigger or smaller) or changing or swapping internal nodes of the most recently accepted tree. This algorithm continues iteratively sampling trees until the model parameter estimates become stable. At this stage it is usually assumed that the algorithm has converged and we have samples from the posterior distribution of trees.

The tree size and shape is determined by a prior probability of a terminal node splitting $P_{split} = \alpha\,(1 + \mathrm{d_i})^{-\beta}$ where d refers to the depth of the current terminal node i and α and β are penalties on the tree size and shape respectively. As discussed, trees are sampled by growing, pruning, changing or swapping the current tree in the algorithm. Grow and prune moves are synonymous with node birth and death where a random terminal node is either converted into an internal node by further splitting it into two daughter nodes (grow) or a random internal node with two terminal daughter nodes is collapsed into a terminal node (prune). Growing or pruning a tree alters the size of the tree. Tree structure is altered by changing or swapping nodes where a new split rule is chosen for a given internal node in a tree or an internal parent-daughter pair of nodes is swapped around respectively.

5.3.2 Other Bayesian Tree Models

There have been many other variations on the Bayesian CART model of [68]. One application which was developed independently and at the same time is [69]. The Bayesian CART model of [69] differs from [68] in that their tree prior only requires one parameter determining the number of terminal nodes in the tree and does not include further parameters determining the tree shape and size. They also use a similar MCMC approach to sample from the posterior distribution of tree models [68]. More recently [70] proposed an alternative prior on the tree $\pi(T)$ as shown below:

$$\pi(T) = \alpha\big(m_0(T)\big) \prod_{u \in a(T)} \beta\big(m_{l(u)}(T) \big|\, m_u(T)\big)$$

Here $m_0(T)$ refers to the number of terminal nodes in tree T, α and β determine the tree size and shape respectively, m_u refers to the number of terminal nodes in the sub-tree below node u and $m_{l(u)}$ is the number of terminal left daughter nodes in the sub-tree below node u. In addition to the four proposal steps of [68] and [69] described above, [70] suggest an additional proposal move to sample new trees called the "restructure proposal", this move searches for alternative trees that would result in the same terminal nodes as the current accepted tree. They claim that this overcomes the problem of slow mixing found in previous Bayesian CART models, as radically different trees can be proposed rather than just local changes of previously accepted trees.

5.3.3 Bayesian Additive Regression Trees (BART)

Bayesian Additive Regression Trees (BART) [71] is a fundamental extension of Bayesian CART and is a non-parametric tree-based ensemble method which brings with it all the advantages of a fully probabilistic model. The BART model uses a sum of multiple Bayesian CART trees as proposed in [68] in its formulation. Posterior predictions are then constructed by adding the MCMC samples over all trees. Point and credible intervals for each predicted data point can also be obtained by taking the appropriate quantiles from the distribution of the MCMC samples for each predicted value of the response variable.

The idea of CART ensemble methods is not new, in fact the most widely used of these methods, the random forest, was invented in 2001 [72] and has been a popular method in the area of proteomic biomarker discovery [41][73][74]. BART, to the authors' knowledge, has not yet been applied to a proteomic setting, however a multiclass version has been used to classify satellite images [75] and a very recent application to gene regulation data using informative priors for biomarker selection has also been implemented [76]. Another extension of the BART method has also been proposed for use on high dimensional survival data and was successfully showcased on a number of gene expression datasets [77].

The recent addition of a more efficient parallelised software package in R called bartMachine [78] means that this method can now be easily implemented for high dimensional datasets such as those commonly found in MRM and smaller proteomic biomarker discovery experiments. Due to memory constraints and computational complexity of the model however, some preprocessing of the data may be needed for very high dimensional data. Code snippets to implement BART in R have been included in the next section.

Bayesian CART and BART models offer many advantages over traditional tree and tree-ensemble methods for biomarker discovery. Their main benefit is the fact that credible intervals can be constructed around the point estimate of the probability of belonging to a given class. This could have large implications for the quality of clinical decisions made based on the output of such methods. Hence Bayesian tree models provide the user with a much richer output on which to make decisions compared to traditional CART models.

A summary of the main advantages of the Bayesian Lasso, Bayesian Decision Tree methods and BART discussed in this section can be seen in Table 1.

Table 1: *Comparison of Bayesian Lasso model with Bayesian decision tree based models*

	Bayesian Lasso	**Bayesian Decision Tree**	**BART**
Will work for small n large p data	✓	✓	✓
Automatically includes high order variable interactions		✓	✓
Eliminates non-predictive variables	✓		
Provides a variable importance score			✓

Finds linear relationships	✓		
Finds non-linear relationships		✓	✓

5.4 Worked Example: Implementation of Bayesian Inference for biomarker discovery in R

This section will provide a worked example using the Bayesian Lasso and Bayesian additive regression trees (BART) models described in the previous section using freely available software in the statistical programming language R. Here we use an illustrative example of LC-MS data which was collected for 500 patients, 150 of which had a cardiovascular disease and 350 of whom were healthy. A total of 37 proteins were measured by MRM for each patient. The objective of this study was two-fold as with any biomarker discovery experiment:

1. To build a classifier which can accurately predict between the two groups and
2. To find a subset of proteins/peptide features which are important in discriminating between the groups

The dataset described is provided in the supplementary material where all identifying information has been removed. Therefore peptide features in the following tutorial will be referred to using their column number in the dataset provided rather than their sequence. This data set is shown for illustrative purposes only, however the analyses and code used here are equivalent for higher dimensional datasets where p>>n as is common in shotgun discovery experiments. To compare fairly across all analyses the data set was split into a training and test set where a random sample of 400 patients were chosen to build the model and the remaining 100 were used to test the model. This ensures that none of the models over-fit to random artifacts in the data. In the following code snippets the response variable for the training data will be referred to as "`training_response`" and the response variable for the test data as will be referred to as "`test_response`". Similarly, the explanatory variables for the 400 training observations will be referred to "`training_data`" and the explanatory variables for the 100 test observations will be referred to as "`test_data`".

It should be noted that with shotgun proteomic experiments where the number of peptides measured can reach tens of thousands much consideration should be given to the appropriate sample size to use. Although sample size calculation is beyond the scope of this article the interested reader may wish to refer to previous work where we show how varying sample sizes can affect the overall classification performance of a model [79].

5.4.1 Package reglogit

An MCMC implementation of the Bayesian Lasso which is equivalent to a logistic regression with double exponential priors can be run using the R package reglogit [51][80]. The Bayesian Lasso model is run by default as follows:

```
1. set.seed(100)
2. #set number of iterations
3. T = 1000
4. reg_logit_model = reglogit(T, training_response,training_data, normalize=FALSE)
```

Simple prediction of the class of each sample can also be calculated via

```
reg_logit_preds = predict(reg_logit_model, XX=test_data))
```

The area under the ROC curve can be calculated as follows:

```
library(ROCR)
prediction=prediction(reg_logit_preds$mp,test_response)
performance=performance(prediction,"auc")
reg_logit_auc=performance@y.values[[1]][1]
```

and the classification rate can be calculated using the following code:

```
class_rate=sum(reg_logit_preds$c == test_response)
```

Analysis of the output for the cardiovascular disease data showed that the Bayesian Lasso model performed quite well in this instance and identified 74% of the test samples correctly with an area under the ROC curve of 0.65 (see Figure 2).

Important variables can be chosen by looking at the posterior distribution of the variable parameters. The 2.5%, 50% and 97.5% quantiles of the posterior distribution of the variable parameters can be viewed using the following code:

```
burnin = (1:(T/10))
quants = t(apply(reg_logit_model$beta[-
burnin,],2,function(x)quantile(x,probs=c(0.25,0.5,0.975))))
```

Choosing the variables whose 95% credible intervals did not include zero, we found that the peptides 3, 12 and 13 were given non-zero coefficients and hence were important in distinguishing between those patients who had experienced cardiovascular disease or not.

5.4.2 JAGS

The Bayesian Lasso can also be run using a package called rjags [52] which requires installation of the JAGS software. The use of rjags does require some programming skills as all models are written by the user. However this allows the user full access to all information regarding the model and the ability to easily change priors and other model assumptions. An example of how a Bayesian Lasso might be run in JAGS is included in the following code:

```
require(rjags)
data=list(y=response,X=data,n=nrow(data),p=ncol(data))
#initialise parameter values
init=list(lambda=1,alpha=0,beta=rep(0,p))
modelstring="
  model {
    for (i in 1:n) {
      #write a logistic regression
      logit(theta[i])<-alpha+inprod(X[i,],beta)
      y[i]~dbern(theta[i])
    }
    for (j in 1:p) {
      #set a double exponential prior on the beta parameters (The Bayesian LASSO is
    equivalent to placing a double exponential prior on beta)
      beta[j]~ddexp(0,lambda)
    }
```

```
    #set a vague prior on the intercept term. JAGS uses precision instead of variance so this is equivalent to alpha~N(0,100)
    alpha~dnorm(0,0.01)
    #set a weakly informative prior on hyper-parameter lambda
    lambda~dgamma(0.1,0.1)
  }
"
```

Here non-informative priors have been set on the beta and lambda parameters, however these could be easily changed if previous experiments or expert opinion deemed some peptides to be better candidate biomarkers than others. The data used in this model includes all 500 samples where the response variable for the test samples has been set to NA. rjags will automatically give predicted values for all samples whose response is flagged as NA.

Once the model has been run, the model parameter output can be viewed using the following code:

```
#use the coda.samples command to view parameter output and diagnostics
Output=coda.samples(model=model,variable.names=c("alpha","beta","lambda"),n.iter=1000,thin=10)
```

Convergence of the model parameters and the quantiles of the model parameters can also be checked:

```
#check convergence
gelman.diag(output)
gelman.plot(output)
```

The 95% credible interval for the parameter estimates can be viewed as follows:

```
#get quantiles of the model parameters
quantiles=summary(window(output,burnin=2000))[[2]]
```

Choosing those variable parameters whose 2.5% and 97.5% quantiles do not include 0 found that variables 3, 12 and 27 had non-zero coefficients according to this model, which corresponds to peptides numbered 3, 12 and 27 as being important predictors of cardiovascular disease. In this case the JAGS model predicted 67% of the test cases correctly and gave an area under the ROC curve of 0.66 (see Figure 2).

As mentioned earlier, the posterior distribution of the model parameters can be explored in a Bayesian setting. This means that credible intervals for the probability of having a cardiovascular event can be constructed with these models rather than just having access to a point prediction. The quantiles of the predicted probabilities for each of the test samples can be viewed as follows:

```
#see quantiles for theta
output_theta=coda.samples(model=model,variable.names=c("theta"),
                          n.iter=1000,thin=10)
#plot 95% CI for the probability of having cardiovascular disease
theta_quants=summary(output_theta[,test_samples])[[2]]
```

To illustrate the usefulness of this additional information we shall take two patients numbered 60 and 367. Patient 60 has a median predicted probability of 25.61% of having cardiovascular disease

and patient 367 has a median predicted probability of 4.25%. In the absence of further information (as with machine learning methods) the clinician would give both patients the all clear. However if we look at the 2.5% and 97.5% quantiles for patient 60 we see they range between 7.46% and 62.79% respectively whereas those for patient 367 range from 1.01% to 13.29%. This additional information means that the model is very sure that patient 367 does not have cardiovascular disease; however it is not at all sure as to the class of patient 60. If a clinician was merely basing the prognosis on the median estimate, they would most likely quite confidently diagnose both patients as healthy. However knowing that the estimate for patient 60 could vary anywhere between 7.46% and 62.79% might change their opinion and hence the medical advice offered to this patient. For example patient 60 may be sent for further diagnostic tests. Alternatively, knowing the 95% credible interval for patient 367 ranges from 1.01% to 13.29%, might give the clinician added confidence as to the true diagnosis of this patient. In reality patient 60 did have cardiovascular disease and patient 367 was healthy, so basing predictions on a point estimate would have led to the misdiagnosis of patient 60 in this case. Bayesian decision theory [81], not discussed here, provides a more nuanced approach to making such decisions.

5.4.3 bartMachine

BART can be run using the package bartMachine in R [82]. The bartMachine function creates a Java virtual machine which requires the user to set the amount of memory to be used by the function. This can be set using the options() command in R as shown in the following code. Here we have set 5Gb of RAM as the limit for the virtual machine. The serialize=TRUE command allows the BART model to be saved and loaded at a later date otherwise all information regarding the model will be lost once R is closed. The bartMachine model can be run on the training data as follows:

```
options(java.parameters = "-Xmx5000m")
library(bartMachine)
set.seed(100)
bm=bartMachine(data.frame(training_data),training_response,num_trees=50,num_iterations_after_burn_in = 1000,mem_cache_for_speed = FALSE,serialize = TRUE)
```

Once the model is run on the training data using the code above, predictions and the full posterior of the test samples can be viewed as follows:

```
#get predicted values for test samples
preds=bart_predict_for_test_data(bm,data.frame(scale(cvd_data[test_samps,])),response[test_samps])
#look at the posterior samples for test data
test_samps_posterior=bart_machine_get_posterior(bm,data.frame(scale(cvd_data[test_samps,])))
#get the 2.5%,50% and 97,5% quantiles of the posterior test samples
posterior_quantiles=t(apply(test_samps_posterior[[3]],1,function(x)quantile(x,probs=c(0.025,0.5,0.975))))
```

In this case BART marginally outperformed both implementations of the Lasso for this data in terms of the classification rate of 71% for the test samples. Figure 2 shows the ROC curve for all three methods compared in this section. As can be seen BART also outperformed both implementations of the Lasso in terms of the area under the ROC curve giving an AUC of 0.68.

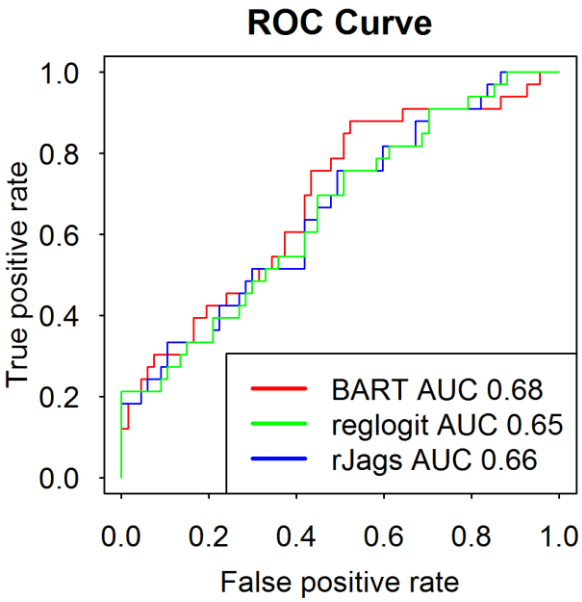

Figure 2 *- ROC curve comparison of Lasso using package reglogit and rJags and BART using package bartMachine*

bartMachine unlike the Lasso does not eliminate variables from the model and as such all variables a re given an importance score which is based on the number of times each variable was selected for e ach tree in the model. The importance score for each variable can be seen by using the following co mmand:

```
1.  #get variable importance scores
2.  get_var_props_over_chain(bm)
```

In this case the bartMachine model found that variables 13, 2, 14, 3, 34, 23 and 12 were given the hi ghest importance. It is interesting to note that all but one of the variables identified by both Lasso im plementations were also identified as the most important according to BART. A vignette for the bart Machine package can be found at [78].

6 Discussion

In this paper we have identified a number of potential areas from the Bayesian statistical literature which could be applied to proteomic discovery data. We have also guided the reader towards using these models for their own experiments by supplying tutorial style example code using freely available software and have showcased these methods using a real proteomic experiment to predict cardiovascular disease.

Bayesian modeling for proteomic biomarker discovery data is a relatively new and emerging field with exciting future opportunities. Many articles have discussed reasons for the failure of biomarker panels from numerous biomarker discovery experiments to validate on separate cohorts. Some of the issues cited in the literature could be alleviated by the incorporation of a probabilistic model. One such reason suggested is that the original model over-fits the data. This is due to the fact that biomarker discovery datasets (especially those emerging from mass spectrometry) tend to be small n large p. Bayesian models when used with sensible priors tend to be more robust to over-fitting than non-Bayesian models as the full posterior distribution of the model parameters is given. Another possible reason for the failure of many experiments to validate is the failure to incorporate

industry and scientific knowledge about underlying processes to the model. With careful modeling Bayesian methods can easily include information about variability introduced to the data through the various experimental and preprocessing stages before data are subjected to statistical analysis rather than ignore this variability.

The area of Bayesian variable selection for high dimensional data is a new and rapidly developing field. Decision tree models and ensemble methods in particular have traditionally been very popular for biomarker discovery as they are easily interpreted, are non parametric and automatically include important interactions without prior specification by the user. There have been many recent advances in Bayesian tree and ensemble methods which could have very interesting applications for proteomic biomarker discovery.

Acknowledgements

We would like to thank Dr Chris Watson and Dr John Baugh for kindly providing the cardiovascular dataset used in this manuscript. We would also like to acknowledge that this work was supported by the Irish Research Council. Protein biomarker discovery work in the Pennington Biomedical Proteomics Group is supported by grants from Science Foundation Ireland (for mass spectrometry instrumentation), the Irish Cancer Society (PCI11WAT), St Luke's Institute for Cancer Research, the Health Research Board (HRA_POR/2011/125), Movember GAP and the EU FP7 (MIAMI). The UCD Conway Institute is supported by the Program for Research in Third Level Institutions as administered by the Higher Education Authority of Ireland.

References

[1] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:2507–17. doi:10.1093/bioinformatics/btm344.

[2] Chambers AG, Percy AJ, Yang J, Camenzind AG, Borchers CH. Multiplexed quantitation of endogenous proteins in dried blood spots by multiple reaction monitoring-mass spectrometry. Mol Cell Proteomics 2013;12:781–91. doi:10.1074/mcp.M112.022442.

[3] Mikolajczyk SD, Song Y, Wong JR, Matson RS, Rittenhouse HG. Are multiple markers the future of prostate cancer diagnostics? Clin Biochem 2004;37:519–28. doi:10.1016/j.clinbiochem.2004.05.016.

[4] Ky B, French B, Levy WC, Sweitzer NK, Fang JC, Wu AHB, et al. Multiple biomarkers for risk prediction in chronic heart failure. Circ Heart Fail 2012;5:183–90. doi:10.1161/CIRCHEARTFAILURE.111.965020.

[5] Surinova S, Schiess R, Hüttenhain R, Cerciello F, Wollscheid B, Aebersold R. On the development of plasma protein biomarkers. J Proteome Res 2011;10:5–16. doi:10.1021/pr1008515.

[6] Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation : the long and uncertain path to clinical utility 2006;24:971–83. doi:10.1038/nbt1235.

[7] Diamandis EP. The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? BMC Med 2012;10:87. doi:10.1186/1741-7015-10-87.

[8] Oon S, Pennington S, Fitzpatrick J, Watson RWG. Biomarker research in prostate cancer-towards utility, not futility. Nat Rev Urol 2011;8:131–8.

[9] Dakna M, Harris K, Kalousis A, Carpentier S, Kolch W, Schanstra JP, et al. Addressing the challenge of defining valid proteomic biomarkers and classifiers. BMC Bioinformatics 2010;11:594. doi:10.1186/1471-2105-11-594.

[10] Alaiya A, Al-Mohanna M, Linder S. Clinical cancer proteomics: promises and pitfalls. J Proteome Res 2005;4:1213–22. doi:10.1021/pr050149f.

[11] Wilkinson DJ. Bayesian methods in bioinformatics and computational systems biology. Brief Bioinform 2007;8:109–16. doi:10.1093/bib/bbm007.

[12] Beaumont M a, Rannala B. The Bayesian revolution in genetics. Nat Rev Genet 2004;5:251–61. doi:10.1038/nrg1318.

[13] Box GEP, Tiao GC. Bayesian Inference in Statistical Analysis. Wiley Clas. Addison-Wesley Pub. Co.; 1973.

[14] Harris K, Girolami M, Mischak H. Definition of Valid Proteomic Biomarkers : A Bayesian Solution. In: Kadirkamanathan V, Sanguinetti G, Girolami M, Niranjan M, Noirel J, editors. Pattern Recognit. Bioinformatics, Lect. Notes Comput. Sci., Springer Berlin Heidelberg; 2009, p. 137–49.

608 [15] Sampson DL, Parker TJ, Upton Z, Hurst CP. A Comparison of Methods for Classifying Clinical
609 Samples Based on Proteomics Data: A Case Study for Statistical and Machine Learning
610 Approaches. PLoS One 2011;6:e24973. doi:10.1371/journal.pone.0024973.

611 [16] Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, et al. Comparison of statistical
612 methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics
613 2003;19:1636–43. doi:10.1093/bioinformatics/btg210.

614 [17] Caruana R, Karampatziakis N, Yessenalina A. An empirical evaluation of supervised learning in
615 high dimensions. Proc. 25th Int. Conf. Mach. Learn. (ICML '08), 2008, p. 96–103.

616 [18] Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data:
617 regularized t -test and statistical inferences of gene changes. Bioinformatics 2001;17:509–19.

618 [19] Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning Data
619 Mining,Inference,and Prediction. 2nd ed. Springer Series in Statistics; 2009.

620 [20] Waterhouse S, MacKay D, Robinson T. Advances in Neural Information Processing Systems 8.
621 In: Touretzky DS, Mozer MC, Hasselmo ME, editors. Bayesian Methods Mix. Expert. 8th ed.,
622 MIT Press; 1995, p. 351–7.

623 [21] Bishop CM, Tipping ME. Bayesian Regression and Classification. NATO Sci Ser IIIComputer Syst
624 Sci n.d.;190:267–88.

625 [22] Kuschner KW, Malyarenko DI, Cooke WE, Cazares LH, Semmes OJ, Tracy ER. A Bayesian
626 network approach to feature selection in mass spectrometry data. BMC Bioinformatics
627 2010;11:177. doi:10.1186/1471-2105-11-177.

628 [23] Vannucci M, Sha N, Brown PJ. NIR and mass spectra classification: Bayesian methods for
629 wavelet-based feature selection. Chemom Intell Lab Syst 2005;77:139–48.
630 doi:10.1016/j.chemolab.2004.10.009.

631 [24] Yu J, Chen X-W. Bayesian neural network approaches to ovarian cancer identification from
632 high-resolution mass spectrometry data. Bioinformatics 2005;21 Suppl 1:i487–94.
633 doi:10.1093/bioinformatics/bti1030.

634 [25] Deng X, Geng H, Ali HH. Cross-platform analysis of cancer biomarkers: a Bayesian network
635 approach to incorporating mass spectrometry and microarray data. Cancer Inform
636 2007;3:183–202.

637 [26] Serang O, States U, Steen J a. Nonparametric Bayesian Evaluation of Di ff erential Protein
638 Quantification. J Proteome Res 2013;12:4556–65.

639 [27] Jow H, Boys RJ, Wilkinson DJ. Bayesian identification of protein differential expression in
640 multi-group isobaric labelled mass spectrometry data. Stat Appl Genet Mol Biol 2014;13:531–
641 51. doi:10.1515/sagmb-2012-0066.

642 [28] Koh HW., Swa HLF, Damian F, Ler SG, Gunaratne J, Choi H. EBprot: Statistical Analysis of
643 Labeling-based Quantitative Proteomics Data. Proteomics 2015.
644 doi:10.1002/pmic.201400620.

[29] Liu JS, Neuwald, Andrew F. Lawrence CE. Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies. J Am Stat Assoc 1995;90:1156–70. doi:10.1080/01621459.1995.10476622.

[30] Webb B-JM, Liu JS, Lawrence CE. BALSA: Bayesian algorithm for local sequence alignment. Nucleic Acids Res 2002;30:1268–77.

[31] Schmidler SC, Liu JS, Brutlag DL. Bayesian Segmentation of Protein Secondary Structure. J Comput Biol 2000;7:233–48. doi:10.1089/10665270050081496.

[32] Brevern AG De, Etchebest C, Hazout S. Bayesian Probabilistic Approach for Predicting Backbone Structures in Terms of Protein Blocks. ProteinsStricture,Function Genet 2000;3:271–87.

[33] Aydin Z, Singh A, Bilmes J, Noble WS. Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure. BMC Bioinformatics 2011;12:154. doi:10.1186/1471-2105-12-154.

[34] Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of Protein Function Using Protein–Protein Interaction Data. J Comput Biol 2003;10:947–60. doi:10.1089/106652703322756168.

[35] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 2003;302:449–53. doi:10.1126/science.1087361.

[36] Serang O, MacCross MJ, Stafford Noble W. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. J Proteome Res 2010;9:5346–57. doi:10.1021/pr100594k.Efficient.

[37] Li YF, Arnold RJ, Li Y, Radivojac P, Sheng Q, Tang H. A bayesian approach to protein inference problem in shotgun proteomics. J Comput Biol 2009;16:1183–93. doi:10.1089/cmb.2009.0018.

[38] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry abilities that proteins are present in a sample on the basis. Anal Chem 2003;75:4646–58.

[39] Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 2007;4:787–97. doi:10.1038/NMETH1088.

[40] Zhang W, Chait BT. ProFound : An Expert System for Protein Identification Using Mass Spectrometric Peptide Mapping Information. Anal Chem 2000;72:2482–9.

[41] Cima I, Schiess R, Wild P, Kaelin M, Schuffler P, Lange V, et al. Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. Proc Natl Acad Sci 2011:1–6. doi:10.1073/pnas.1013699108.

[42] Tibshirani R. lasso.pdf. J R Stat Soc Ser B (Statistical Methodol 1996;58:267–88.

[43] Huang T, Gong H, Yang C, He Z. ProteinLasso: A Lasso regression approach to protein inference problem in shotgun proteomics. Comput Biol Chem 2013;43:46–54. doi:10.1016/j.compbiolchem.2012.12.008.

[44] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 2008;9:432–41. doi:10.1093/biostatistics/kxm045.

[45] Qin T, Tsoi LC, Sims KJ, Lu X, Zheng WJ. Signaling network prediction by the Ontology Fingerprint enhanced Bayesian network. BMC Syst Biol 2012;6 (Suppl S. doi:10.1186/1752-0509-6-S3-S3.

[46] Soltys SG, Le Q, Shi G, Tibshirani R, Giaccia AJ, Koong AC. The Use of Plasma Surface-Enhanced Laser Desorption / Ionization Time-of-Flight Mass Spectrometry Proteomic Patterns for Detection of Head and Neck Squamous Cell Cancers The Use of Plasma Surface-Enhanced Laser Desorption / Ionization Time-of-Flight Mass. Clin Cancer Res 2004;10:4806–12.

[47] Park T, Casella G. The Bayesian Lasso. J Am Stat Assoc 2008;103:681–6. doi:10.1198/016214508000000337.

[48] Li J, Das K, Fu G, Li R, Wu R. The Bayesian lasso for genome-wide association studies. Bioinformatics 2011;27:516–23. doi:10.1093/bioinformatics/btq688.

[49] De los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 2009;182:375–85. doi:10.1534/genetics.109.101501.

[50] Cai X, Huang A, Xu S. Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. BMC Bioinformatics 2011;12:211. doi:10.1186/1471-2105-12-211.

[51] Gramacy RB, Polson NG. Simulation-based regularized logistic regression 2014. doi:10.1214/12-BA719.

[52] Stukalov A, Plummer MM. Package " rjags " 2015. http://cran.r-project.org/web/packages/rjags/rjags.pdf (accessed February 12, 2015).

[53] Armagan A, Dunson DB, Lee J. Generalized Double Pareto Shrinkage. Stat Sin 2013;23:119–43.

[54] Griffin JE, Brown PJ. Inference with normal-gamma prior distributions in regression problems. Bayesian Anal 2010;5:171–88. doi:10.1214/10-BA507.

[55] Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. Biometrika 2010;97:465–80. doi:10.1093/biomet/asq017.

[56] Gramacy RB. Package monomvn 2014. http://cran.r-project.org/web/packages/monomvn/monomvn.pdf (accessed September 12, 2014).

[57] Zellner A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: P.K G, Zellner A, editors. Bayesian Inference Decis. Tech. Essays Honor Bruno Finetti, North-Holland/Elsevier; 1986, p. 233–43.

[58] Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g -priors for Bayesian Variable Selection. J Am Stat Assoc 2008;103:410–23. doi:10.1198/016214507000001337.

[59] Baragatti M, Pommeret D. A study of variable selection using -prior distribution with ridge parameter. Comput Stat Data Anal 2012;56:1920–34. doi:10.1016/j.csda.2011.11.017.

[60] Yang A-J, Song X-Y. Bayesian variable selection for disease classification using gene expression data. Bioinformatics 2010;26:215–22. doi:10.1093/bioinformatics/btp638.

[61] Zhou X, Wang X, R. Dougherty E. A Bayesian approach to nonlinear probit gene selection and classification. J Franklin Inst 2004;341:137–56. doi:10.1016/j.jfranklin.2003.12.010.

[62] Feldkircher, Martin Zeugner S. Package BMS 2013. http://cran.r-project.org/web/packages/BMS/BMS.pdf (accessed September 12, 2014).

[63] Kuo L, Mallick B. Variable Selection for Regression Models. Indian J Stat Spec Issue Bayesian Anal 1998;60:65–81.

[64] O'Hara RB, Sillanpää MJ. A review of Bayesian variable selection methods: what, how and which. Bayesian Anal 2009;4:85–118. doi:10.1214/09-BA403.

[65] Vlahou A, Schorge JO, Gregory BW, Coleman RL. Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data. J Biomed Biotechnol 2003;2003:308–14. doi:10.1155/S1110724303210032.

[66] Yu Y, Chen S, Wang L-S, Chen W-L, Guo W-J, Yan H, et al. Prediction of pancreatic cancer by serum biomarkers using surface-enhanced laser desorption/ionization-based decision tree classification. Oncology 2005;68:79–86. doi:10.1159/000084824.

[67] Markey MK, Tourassi GD, Floyd CE. Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer. Proteomics 2003;3:1678–9. doi:10.1002/pmic.200300521.

[68] Chipman H a., George EI, McCulloch RE. Bayesian CART Model Search. J Am Stat Assoc 1998;93:935–60.

[69] Denison BYDGT, Mallick BANK, Smith AFM. A Bayesian CART algorithm. Biometrika 1998;85:363–77.

[70] Wu Y, Tjelmeland H, West M. Bayesian CART – Prior Specification and Posterior Simulation –. J Comput Graph Stat 2006;16:44–66.

[71] Chipman H a., George EI, McCulloch RE. BART: Bayesian additive regression trees. Ann Appl Stat 2010;4:266–98. doi:10.1214/09-AOAS285.

[72] Breiman L. Random forests. Mach Learn 2001;45:5–32. doi:10.1186/1478-7954-9-29.

[73] Fan Y, Murphy TB, Byrne JC, Brennan L, Fitzpatrick JM, Watson RWG. Applying random forests to identify biomarker panels in serum 2D-DIGE data for the detection and staging of prostate cancer. J Proteome Res 2011;10:1361–73. doi:10.1021/pr1011069.

[74] Baek S, Tsai C-A, Chen JJ. Development of biomarker classifiers from high-dimensional data. Brief Bioinform 2009;10:537–46. doi:10.1093/bib/bbp016.

[75] Agarwal R, Ranjan P, Chipman H. A new Bayesian ensemble of trees approach for land cover classification of satellite imagery. Can J Remote Sens 2013;39:507–20. doi:10.5589/m14-003.

[76] Bleich J, Kapelner A, George EI, Jensen ST. Variable Selection for BART : An Application to Gene Regulation. Ann Appl Stat 2014;8:1750–81.

[77] Bonato V, Baladandayuthapani V, Broom BM, Sulman EP, Aldape KD, Do K-A. Bayesian ensemble methods for survival prediction in gene expression data. Bioinformatics 2011;27:359–67. doi:10.1093/bioinformatics/btq660.

[78] Kapelner A, Bleich J. bartMachine: Machine Learning With Bayesian Additive Regression Trees 2013. http://cran.r-project.org/web/packages/bartMachine/vignettes/bartMachine.pdf (accessed March 2, 2015).

[79] Hernández B, Parnell A, Pennington SR. Why have so few proteomic biomarkers “survived” validation? (Sample size and independent validation considerations). Proteomics 2014;14:1587–92. doi:10.1002/pmic.201300377.

[80] Gramacy RB, Polson NG. Simulation-based regularized logistic regression. Bayesian Anal 2012;7:567–90. doi:10.1214/12-BA719.

[81] Berger JO. Statistical decision theory and Bayesian analysis. 2nd ed. New York: Springer; 1985.

[82] Kapelner A, Bleich J. Package bartMachine 2014. http://cran.r-project.org/web/packages/bartMachine/bartMachine.pdf.

Figure Legends:

Figure 1 – Example of a binary classification decision tree

Figure 2 – ROC curve comparison of Lasso using package reglogit and rJags and BART using package bartMachine

Table Captions:

Table 1 – Comparison of Bayesian Lasso model with Bayesian decision tree based models