EMSHS package for Bayesian Shrinkage Model using Structural Information in **R**

C. Chang[*] and **P. Suthaharan**[†] and **S. Kundu**[‡] and **Q. Long**[§]

Abstract

The package **EMSHS** implements a scalable, yet adaptive Bayesian shrinkage approach that exploits the prior network information for improved variable selection and prediction. We attempt to simplify the understanding of this package using a hypothetical, high-dimensional data with $n = 25$ observations and $p = 50$ predictors in R.

Keywords: Adaptive bayesian shrinkage, EM algorithm, structured high dimensional variable selection.

EMSHS

1. Introduction

This vignette is written as a supplementary documentation to Chang, Kundu, and Long (2018) in hopes of providing a more visual understanding for the motivation of our work to develop a scalable structured variable selection approach. We use R code from the **E**xpectation **M**aximization estimator for bayesian **SH**rinkage approach with **S**tructural information incorporated (**EMSHS**) package to illustrate this approach - in both the absence and presence of graph information - on high-dimensional data from a hypothetical cancer genomics example.

The rest of the vignette is organized as follows. We introduce a hypothetical cancer genomics example in section 2, the EM estimator for Bayesian Shrinkage approach in the *absence* of graph information in section 2.1, the EM estimator for Bayesian Shrinkage approach in the *presence* of graph information in section 2.2, and concluding remarks about the **EMSHS** package in section 3.

2. Example

Microarray analysis and next generation sequencing in genomics yield increasingly large amounts of data containing more than tens of thousands of variables. In genomics studies, it is common to collect gene expressions from $p \approx 20,000$ genes, which is often considerably larger than the number of subjects (n) in these studies, resulting in a classical small n large

[*]email: changgee@pennmedicine.upenn.edu, Department of Biostatistics, University of Pennsylvania
[†]email: psuthah@emory.edu, Department of Biostatistics, Emory University
[‡]email: suprateek.kundu@emory.edu, Department of Biostatistics, Emory University
[§]email: qlong@pennmedicine.upenn.edu, Department of Biostatistics, University of Pennsylvania

p problem.

Here, we present a hypothetical cancer genomics microarray example to elucidate the explanation of the **EMSHS** package. Our hypothetical example is as follows:

> As a group of cancer research scientists, we are interested in identifying which set of human genes has a significant impact on the risk for developing a specific breast cancer. We peruse through a genomics data set and become dumbfounded by how big the data is - 20,000 human genes in a series of 100 primary breast cancers. We hear about this scalable, adaptive Bayesian shrinkage approach that we can possibly implement to more robustly select the variables (i.e., the human genes) that are most significant. We load in the **EMSHS** package and perform our analysis.

As you can see, we are dealing with a classical small n large p problem, where $n = 100$ primary breast cancers and $p = 20,000$ human genes. For simplicity, we will consider a smaller n and p: $n = 25$ primary breast cancers and $p = 50$ human genes.

We now present our `EMSHS` function which has the following input parameters

```
EMSHS <- function(y,X,mus,nu,E=NULL,
                  a_sigma=1,b_sigma=1,a_omega=2,b_omega=1,
                  w=1,eps=1e-5){}
```

where the input parameter - X - can be initialized based on our aforementioned observations ($n = 25$) and predictors ($p = 50$). We generate y as $X * B + e$, where B represents our sparse true beta matrix and e represents the error term. Additionally, we can initialize our shrinkage and adaptivity parameters, mus and nu, respectively. Although our function accepts a vector of shrinkage parameters, for our example, we will use a single value for mus. Our goal is to produce estimated betas($\hat{B}$) that are close to our true beta (B)

```
set.seed(100)

X <- matrix(rnorm(25*50), ncol = 50) # An n by p design matrix
B <- matrix(c(1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
              0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
              0,0,0,0,0,0,0,0,0,0), ncol = 1) # True beta
e <- matrix(rnorm(25*1), ncol = 1) # error
y <- matrix(X %*% B + e, ncol = 1) # An n by 1 response vector
mus <- 2.3 # The shrinkage parameter
nu <- 0.3 # The adaptivity parameter
```

2.1. EM Estimator: Absence of Graph Information

Let's first consider the case where we are unaware of any structural graph information among

the $p = 50$ human genes. Therefore, we would initialize the following parameters with the default values

```
E <- NULL # An e by 2 matrix of edges. NULL implies there are no edges

a_sigma <- 1 # The shape parameter of the prior for residual variance
b_sigma <- 1 # The rate parameter of the prior for residual variance
a_omega <- 2 # The shape parameter of the prior for nonzero omega values
b_omega <- 1# The rate parameter of the prior for nonzero omega values

w <- 1 # A weight vector for samples
eps <- 1e-5 # The algorithm stops if relative improvement goes below eps
```

where E is set to NULL, suggesting that there are no prior graph information of connections among predictors (i.e., certain human genes are not linked or correlated with other human genes for measuring the risk of primary breast cancer, in this context), the shape parameters - a_sigma and a_omega - and rate parameters - b_sigma and b_omega - are set to the above default values because we have no specific prior distribution of the connectivity of the predictors (i.e., genes), the weight vector w is is set to 1 because we are considering that all predictors (i.e., human genes) are as significant as the other predictors (i.e., human genes) in our data, and eps is set to $1e{-}5$.

Figure 1 depicts an unstructured graph of $p = 50$ human genes. We now run our `EMSHS` function for unstructured graph information

```
em_no_edge <- EMSHS(y,X,mus,nu,E=NULL,
                    a_sigma=1,b_sigma=1,a_omega=2,b_omega=1,
                    w=1,eps=1e-5)
```

and obtain the following outputs - *niter*, *beta*, *sigma*, *lambda*, and *omega*. As stated previously, we are interested in generating $\hat{B}'s$ that are close to our sparse true B's. Thus, in this documentation, we will focus on the *beta* output. Refer to Chang *et al.* (2018) for information about the other outputs.

We begin to extrapolate the beta output of the **EMSHS** function in hopes of concluding interesting information among the human genes and the risk for primary breast cancer.

```
em_no_edge$beta
```

```
##              [,1]
##  [1,]  0.0000000
##  [2,]  0.0000000
##  [3,]  1.1460205
##  [4,]  1.5952202
##  [5,]  0.9244451
##  [6,]  0.0000000
##  [7,]  0.0000000
```

```
##  [8,]  0.0000000
##  [9,]  0.0000000
## [10,]  0.0000000
## [11,]  0.4731300
## [12,]  0.0000000
## [13,]  0.0000000
## [14,]  0.0000000
## [15,]  0.0000000
## [16,]  0.0000000
## [17,]  0.0000000
## [18,]  0.0000000
## [19,]  0.0000000
## [20,]  0.0000000
## [21,]  0.0000000
## [22,]  0.0000000
## [23,]  0.0000000
## [24,]  0.0000000
## [25,]  0.0000000
## [26,]  0.0000000
## [27,]  0.0000000
## [28,]  0.0000000
## [29,]  0.0000000
## [30,]  0.0000000
## [31,]  0.0000000
## [32,]  0.0000000
## [33,]  0.0000000
## [34,]  0.8651741
## [35,]  0.0000000
## [36,]  0.0000000
## [37,]  0.0000000
## [38,]  0.0000000
## [39,] -0.4058939
## [40,]  0.0000000
## [41,]  0.0000000
## [42,]  0.0000000
## [43,]  0.0000000
## [44,]  0.0000000
## [45,]  0.0000000
## [46,]  0.0000000
## [47,]  0.0000000
## [48,]  0.0000000
## [49,] -0.6306526
## [50,]  0.0000000
```

Upon further scrutiny, we observe that there are some false negatives - gene 1, gene 2 - and false positives - gene 11, gene 34, gene 39, gene 49 - in our $\hat{B}$ matrix. That is, with no prior

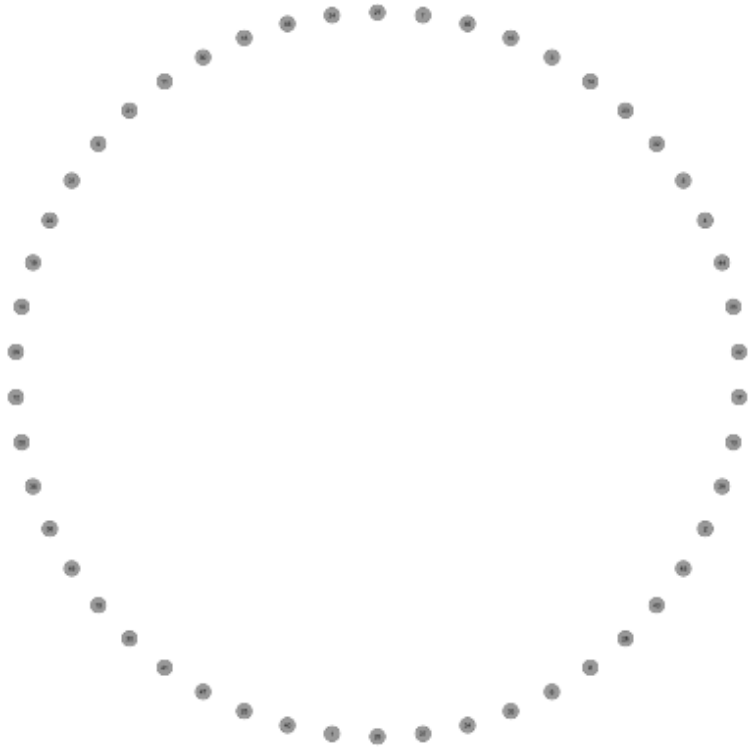


Figure 1: **Unstructured graph information**. This is a graphical visualization of unstructured nodes (i.e., $p = 50$ human genes) with no prior information of connections among the genes.

graph information, the function is incorrectly indicating that particular genes - gene 1, gene 2 - are not influencing the risk for a certain primary breast cancer when, in fact, those genes should have an influence. Also, the function is incorrectly indicating that particulars genes - gene 11, gene 34, gene 39, gene 49 - are influencing the risk for primary breast cancer when, in fact, those genes have have zero influence. For example, in our true beta matrix, genes 1 and 2 were assumed to have an influence (i.e., values were set to 1) on the risk for a certain primary breast cancer (y). However, our estimated beta matrix shows that genes 1 and 2 do not have an influence, suggesting the presence of false negatives. Likewise, genes 10, 34, and 38 were assumed to have no influence (i.e., values were set to 0) on the risk for a certain primary breast cancer (y). However, our estimated beta matrix shows that genes 10, 34, and 38, do indeed, have an influence, suggesting the presence of false positives. To minimize this paradox, we will incorporate graph information and highlight the robustness of the **EMSHS** function.

2.2. EM Estimator: Presence of Graph Information

Until now, we have discussed the case where we were limited in our knowledge of any structure among the genes. Let's now consider the case where we are aware of structural graph information among the $p = 50$ human genes. We observe a structure as illustrated in Figure 2. From this, we construct an E matrix of edges that represent the interactions among the genes.

Now that we have constructed our E matrix, we can run our `EMSHS` function with our structural information.

```
em_edge <- EMSHS(y,X,mus,nu,E,
                 a_sigma=1,b_sigma=1,a_omega=2,b_omega=1,
                 w=1,eps=1e-5)
```

We begin exploring the output given the structural information of the human genes. We see that we have removed the false negatives and minimized the amount of false positives as a result of our graph information, highlighting the robustness of variable selection invoked by the **EMSHS** function.

```
em_edge$beta
```

```
##              [,1]
##  [1,]  0.7769684
##  [2,]  1.0514431
##  [3,]  0.5384613
##  [4,]  0.5099838
##  [5,]  0.7383805
##  [6,]  0.0000000
##  [7,]  0.0000000
##  [8,]  0.0000000
##  [9,]  0.0000000
## [10,]  0.4159533
## [11,]  0.0000000
## [12,]  0.0000000
## [13,]  0.0000000
## [14,]  0.0000000
## [15,]  0.0000000
## [16,]  0.0000000
## [17,]  0.0000000
## [18,]  0.0000000
## [19,]  0.0000000
## [20,]  0.0000000
## [21,]  0.0000000
## [22,]  0.0000000
## [23,]  0.0000000
## [24,]  0.0000000
## [25,]  0.0000000
## [26,] -0.2240287
## [27,]  0.0000000
## [28,]  0.0000000
## [29,]  0.0000000
## [30,]  0.0000000
## [31,]  0.0000000
```

```
## [32,]  0.0000000
## [33,]  0.0000000
## [34,]  0.0000000
## [35,]  0.0000000
## [36,]  0.0000000
## [37,]  0.0000000
## [38,]  0.0000000
## [39,]  0.0000000
## [40,]  0.0000000
## [41,]  0.0000000
## [42,]  0.0000000
## [43,]  0.3092875
## [44,]  0.0000000
## [45,]  0.0000000
## [46,]  0.0000000
## [47,]  0.0000000
## [48,]  0.0000000
## [49,]  0.0000000
## [50,]  0.0000000
```

3. Conclusion

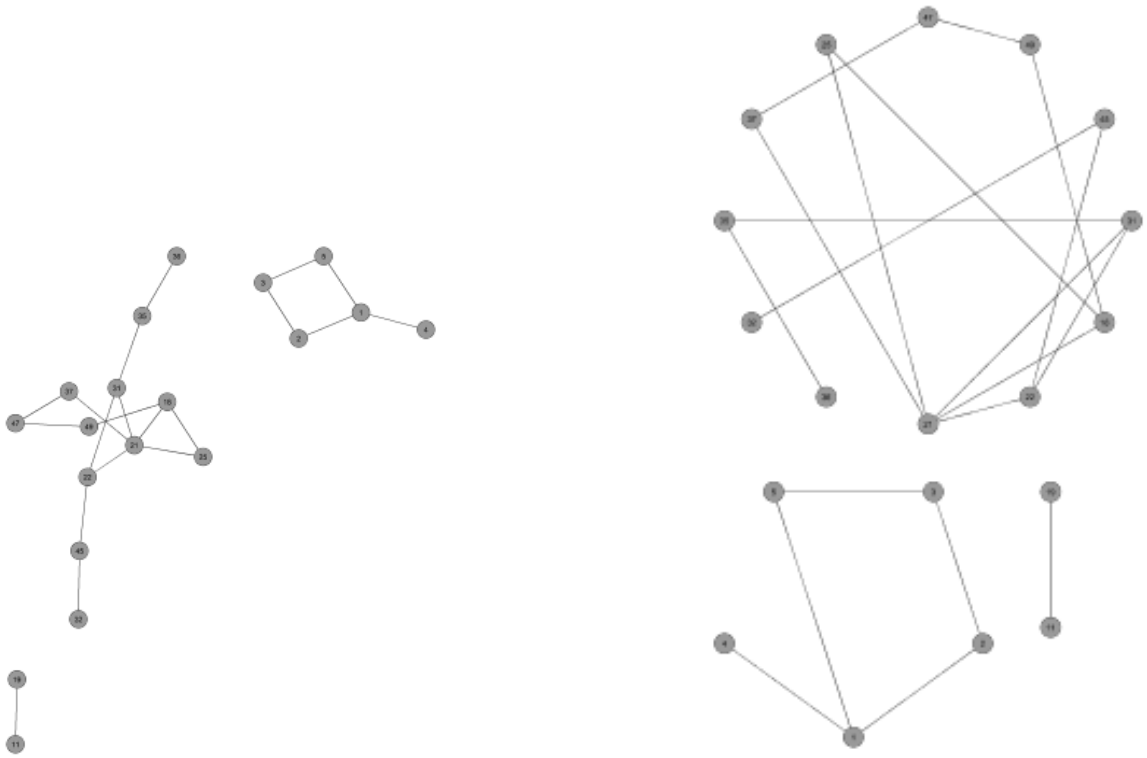
This vignette is designed to help the user of the **EMSHS** package more readily execute the code to achieve good variable selection, prediction and computational scalability within high-dimensional settings. For a more technical review of this scalable, adaptive Bayesian shrinkage approach, refer to Chang *et al.* (2018).

References

Chang C, Kundu S, Long Q (2018). "Scalable Bayesian Variable Selection for Structured High-dimensional Data." *Biometrics*, pp. 1–27. URL https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12882.

Affiliation:

Changgee Chang
Department of Statistics
Perelman School of Medicine
University of Pennsylvania
E-mail: changgee@pennmedicine.upenn.edu
URL: https://dbe.med.upenn.edu/biostat-research/ChanggeeChang



(a) Structural Layout. (b) Circular Layout

Figure 2: **Structured graph information**. This is a graphical visualization - represented in a structural (a) and circular (b) layout - of nodes (i.e., significant genes) with prior information of edges (i.e., significant connections). Here, we can leverage the association structure of the genes and produce biologically meaningful outcomes, and lead to improvements in prediction and variable selection.

Praveen Suthaharan
Department of Biostatistics and Bioinformatics
Graduate Student at Rollins School of Public Health
Emory University
E-mail: psuthah@emory.edu

Suprateek Kundu
Department of Biostatistics and Bioinformatics
Assistant Professor at Rollins School of Public Health
Emory University
E-mail: suprateek.kundu@emory.edu
URL: https://sites.google.com/view/suprateek/

Qi Long
Department of Biostatistics, Epidemiology and Informatics
Professor at Perelman School of Medicine
University of Pennsylvania
E-mail: qlong@pennmedicine.upenn.edu
URL: https://dbe.med.upenn.edu/biostat-research/QiLong