EMVS Vignette

Veronika Rockova and Gemma Moran

December 13, 2019

1 Introduction

This vignette describes the algorithm "EMVS", the EM approach to Bayesian Variable Selection (Rockova and George, 2014) and its usage in the R package `EMVS`. This R package implementation of EMVS has two options for prior specification:

1. A "conjugate" or "scale-invariant" prior on the regression coefficients, as detailed in Rockova and George (2014);

2. An "independent" prior where the regression coefficients and error variance are treated as independent *a priori*, which is recommended by Moran, Rockova and George (2018).

This vignette details the EMVS algorithm where the second, independent prior formulation is used, and provides an example of its usage in the package.

2 Model

Consider the classical linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(0, \sigma^2 \mathbf{I}_n) \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is a vector of responses, $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p] \in \mathbb{R}^{n \times p}$ is a fixed regression matrix of p potential predictors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is a vector of unknown regression coefficients and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the noise vector of independent normal random variables with σ^2 as their unknown common variance.

The goal of EMVS is to find which predictors $\mathbf{x}_i$ should be included in the model. In the Bayesian paradigm, this is facilitated by the introduction of binary latent variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$, $\gamma_i \in \{0, 1\}$, where $\gamma_i = 1$ indicates that $\mathbf{x}_i$ is to be included in the model.

The hierarchical model is given by:

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}, v_0, v_1) = N_p(\mathbf{0}, \mathbf{D}_\gamma) \tag{2}$$

where $\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$ with $d_i = (1 - \gamma_i)v_0 + \gamma_i v_1$ for $0 \leq v_0 < v_1$. Following George and McCulloch (1997), Rockova and George (2014) recommend setting the hyperparameters v_0 and v_1 to be small and large fixed values respectively; this yields the canonical "spike" and "slab" of Bayesian variable selection. Note that the above prior does not depend on the error variance σ^2, unlike the original formulation in Rockova and George (2014).

The prior on the error variance is an inverse gamma:

$$\pi(\sigma^2) = \text{IG}(\nu/2, \nu\lambda/2). \tag{3}$$

To incorporate uncertainty regarding which variables should be included in the model, EMVS additionally specifies a prior for the latent indicator variables $\boldsymbol{\gamma}$. This prior is iid Bernoulli:

$$\pi(\boldsymbol{\gamma}|\theta) = \theta^{|\boldsymbol{\gamma}|}(1-\theta)^{p-|\boldsymbol{\gamma}|} \tag{4}$$

where $\theta \in [0,1]$ and $|\boldsymbol{\gamma}| = \sum_{i=1}^{p} \gamma_i$. The proportion of non-zero regression coefficients, θ, is unknown: this parameter is assigned a beta prior:

$$\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}, \quad a, b > 0. \tag{5}$$

3 Algorithm

The EMVS algorithm treats the latent indicators $\boldsymbol{\gamma}$ as "missing data" and indirectly maximizes the posterior $\pi(\boldsymbol{\beta}, \theta, \sigma^2|\mathbf{Y})$ by iteratively maximizing the following objective function:

$$\mathcal{Q}(\boldsymbol{\beta}, \theta, \sigma|\boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{2(k)}) = E_{\boldsymbol{\gamma}|\cdot}[\log \pi(\boldsymbol{\beta}, \theta, \sigma^2, \boldsymbol{\gamma}|\mathbf{Y})|\boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{2(k)}, \mathbf{Y}] \tag{6}$$

where $E_{\boldsymbol{\gamma}|\cdot}(\cdot)$ denotes the conditional expectation $E_{\boldsymbol{\gamma}|\boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{2(k)}, \mathbf{Y}}(\cdot)$. At the kth iteration, given $(\boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{2(k)})$, an E-step is first applied, which computes the expectation of the right side of (6) to obtain $\mathcal{Q}$. This is followed by an M-step, which maximizes $\mathcal{Q}$ over $(\boldsymbol{\beta}, \theta, \sigma^2)$ to yield the values of $(\boldsymbol{\gamma}^{(k+1)}, \theta^{(k+1)}, \sigma^{2(k+1)})$.

The objective function is of the form:

$$\mathcal{Q}(\boldsymbol{\beta}, \theta, \sigma^2|\boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{2(k)}) = C + \mathcal{Q}_1(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{2(k)}) + \mathcal{Q}_2(\theta|\boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{2(k)}) \tag{7}$$

where

$$\mathcal{Q}_1(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{2(k)}) = -\frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 - \frac{n+\nu+2}{2}\log(\sigma^2) - \frac{\nu\lambda}{2\sigma^2} \tag{8}$$

$$- \frac{1}{2}\sum_{i=1}^{p} \beta_i^2 E_{\boldsymbol{\gamma}|\cdot}\left[\frac{1}{v_0(1-\gamma_i) + v_1\gamma_i}\right], \tag{9}$$

and

$$\mathcal{Q}_2(\theta|\boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{2(k)}) = \sum_{i=1}^{p} \log\left(\frac{\theta}{1-\theta}\right) E_{\boldsymbol{\gamma}|\cdot}\gamma_i + (a-1)\log(\theta) + (p+b-1)\log(1-\theta). \tag{10}$$

3.1 E-Step

We have:

$$E_{\boldsymbol{\gamma}|\cdot}\gamma_i = P(\gamma_i = 1|\boldsymbol{\beta}^{(k)}, \theta^{(k)}) = p_i^* \tag{11}$$

where

$$p_i^* = \frac{\theta^{(k)}\phi_{v_1}(\beta_i^{(k)})}{\theta^{(k)}\phi_{v_1}(\beta_i^{(k)}) + (1-\theta^{(k)})\phi_{v_0}(\beta_i^{(k)})} \tag{12}$$

where $\phi_v(x) = \frac{1}{\sqrt{2\pi v}}\exp(-x^2/2v)$ is the normal density with zero mean and variance, v.

To complete the E-step, we have

$$E_{\boldsymbol{\gamma}|\cdot}\left[\frac{1}{v_0(1-\gamma_i)+v_1\gamma_i}\right] = \frac{E_{\boldsymbol{\gamma}|\cdot}(1-\gamma_i)}{v_0} + \frac{E_{\boldsymbol{\gamma}|\cdot}\gamma_i}{v_1} \tag{13}$$

$$= \frac{1-p_i^*}{v_0} + \frac{p_i^*}{v_1} \tag{14}$$

$$\equiv d_i^*. \tag{15}$$

We denote the matrix $\mathbf{D}^* = \text{diag}(d_1^*, \ldots, d_p^*)$.

3.2 M-Step

The objective function (6) yields closed form updates for each of $(\boldsymbol{\beta}^{(k+1)}, \theta^{(k+1)}, \sigma^{2(k+1)})$. These are:

$$\boldsymbol{\beta}^{(k+1)} = [\mathbf{X}^T\mathbf{X} + \sigma^{2(k)}\mathbf{D}^*]^{-1}\mathbf{X}^T\mathbf{Y}, \tag{16}$$

$$\sigma^{2(k+1)} = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(k+1)}\|^2 + \nu\lambda}{n+\nu+2}, \tag{17}$$

$$\theta^{(k+1)} = \frac{\sum_{i=1}^p p_i^* + a - 1}{a+b-p-2}. \tag{18}$$

In problems where $p \gg n$, the calculation cost of (16) is substantially reduced by using the Sherman-Morrison-Woodbury formula to obtain:

$$\boldsymbol{\beta}^{(k+1)} = \sigma^2[\mathbf{D}^{*-1} - \mathbf{D}^{*-1}\mathbf{X}^T(\sigma^2\mathbf{I}_{n\times n} + \mathbf{X}\mathbf{D}^{*-1}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{D}^{*-1}]\mathbf{X}^T\mathbf{Y}. \tag{19}$$

The EMVS algorithm iterates between the above steps until convergence. The default convergence criterion is: $\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\|^2 \leq 10^{-5}$.

3.3 Thresholding the EM Output for Variable Selection

Once we have obtained our MAP estimates $(\widehat{\boldsymbol{\beta}}, \widehat{\theta}, \widehat{\sigma}^2)$, we can find the most probable $\boldsymbol{\gamma}$ given those values. This is obtained by setting

$$\widehat{\gamma}_i = \begin{cases} 1 & \text{if } P(\gamma_i = 1|\widehat{\boldsymbol{\beta}}, \widehat{\theta}, \widehat{\sigma}^2) \geq 0.5 \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

This selection of $\widehat{\boldsymbol{\gamma}}$ is equivalent to thresholding the $\widehat{\beta}_i$ (Rockova and George, 2014). This thresholding occurs at the intersection points $\pm\beta_i^*$ of the $P(\gamma_i = 1|\widehat{\boldsymbol{\beta}}, \widehat{\theta})$ weighted mixture of the spike-and-slab priors, namely,

$$\pm\beta_i^*(v_0, v_1, \widehat{\theta}) = \pm\sqrt{2v_0\log(\omega_i c)c^2/(c^2-1)} \tag{21}$$

where $c^2 = v_1/v_0$ and $\omega_i = (1-\widehat{\theta})/\widehat{\theta}$. Then, the thresholding rule is

$$\widehat{\gamma}_i = \begin{cases} 1 & \text{if } |\widehat{\beta}_i| \geq \beta_i^*(v_0, v_1, \widehat{\theta}). \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

4 Dynamic Posterior Exploration

The speed of the EM algorithm allows for posterior modes to be found for multiple values of the spike parameter v_0. This facilitates a *dynamic posterior exploration* strategy where the slab parameter v_1 is held fixed and v_0 is gradually decreased to approximate the ideal point mass spike prior. This is akin to an annealing strategy; when $v_0 = v_1$, the posterior is convex (in this case, it is equivalent to ridge regression) and as v_0 is decreased, it becomes multimodal. By starting with large v_0 and using the resultant solution as a "warm start" for the next (smaller) value of v_0, the procedure can more easily find modes of the posterior. Decreasing v_0 also serves to shrink smaller coefficients to zero and find a sparse solution to the regression problem. The original paper (Rockova and George, 2014) illustrates a "forward" strategy for EMVS by starting with small v_0 and gradually increasing it. However, our recommendation is to use a "backward" strategy wherein we start with large v_0 and gradually *decrease* it. In this "backward" strategy, the algorithm also stabilizes at a solution for small v_0, in many cases eliminating the need to choose a particular "best" spike parameter.

The `EMVS` function in the R package has three options for dynamic posterior exploration: `direction = c("backward", "foward", "null")`. These are described below.

- `direction = "backward"` *(default, recommended)*: this option starts with the largest value of `v0` and finds a solution to the EMVS algorithm. This solution is used as the inital value for the algorithm with the second largest value of `v0`. This process is repeated until the smallest value of `v0`.
- `direction = "forward"`: this option is similar to the `backward` option, but starts with the smallest value of `v0` and then uses the resultant solution as the initial value for the next largest value of `v0`, repeating the process until the largest value of `v0`.
- `direction = "null"`: this option uses `beta_init` as the initial value for each value of `v0` and as such is not a dynamic posterior exploration strategy.

5 Prior Specification

As mentioned in the introduction, the `EMVS` function in the R package has two options for the prior specification: `independent = c(TRUE, FALSE)`. These are described below.

- `independent = TRUE` *(default, recommended)*: this option implements EMVS with the "independent" prior detailed in Section 2 of this vignette.
- `independent = FALSE`: this option implements EMVS with the "conjugate" or "scale-invariant" prior on the regression coefficients detailed in the original paper (Rockova and George, 2014).

The reason we recommend the "independent" prior for EMVS is that it yields better error variance estimates (Moran et al., 2018). Intuitively, the "conjugate" prior adds p "pseudo-observations" of the error variance σ^2, which can result in severe underestimation of the error variance. For more details, see Moran et al. (2018).

6 Example

In this section, we demonstrate the basic usage of `EMVS` for both the independent and conjugate prior implementations. We conclude with a comparison of the error variance estimates from the two, highlighing the benefit of the independent prior formulation.

We begin by loading the package:

```
library(EMVS)
```

We create a toy dataset with $n = 100$ and $p = 1000$:

```
set.seed(12022018)
n = 100
p = 1000
X = matrix(rnorm(n * p), n, p)
beta = c(1.5, 2, 2.5, rep(0,p-3))
Y = X[,1] * beta[1] + X[,2] * beta[2] + X[,3] * beta[3] + rnorm(n)
```

6.1 Independent Prior

We set the parameters for EMVS: the "ladder" of values of v_0, the slab parameter v_1, the initial $\boldsymbol{\beta}$ and the hyperparameters of the beta distribution, a, b.

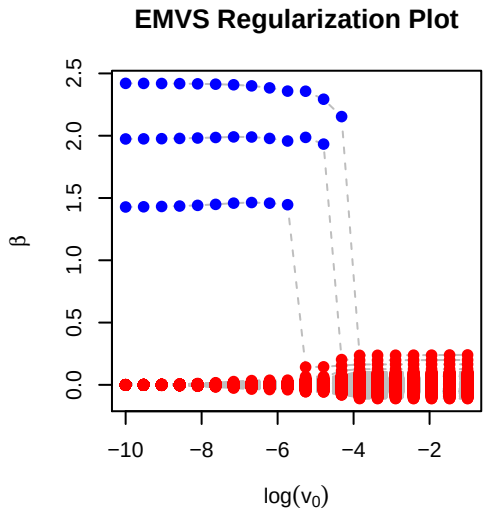
```
# independent prior on regression coefficients and variance
v0 = exp(seq(-10, -1, length.out = 20))
v1 = 1
beta_init = rep(1, p)
sigma_init = 1
a = b = 1
```

We then run `EMVS` using the independent prior formulation described in Section 2 of this vignette (`independent = TRUE`). The output is stored in `result_ind`.

```
# independent prior on regression coefficients and variance
result_ind = EMVS(Y = Y, X = X, v0 = v0, v1 = v1, type = "betabinomial",
                  independent = TRUE, beta_init = beta_init,
                  sigma_init = sigma_init,
                  a = a, b = b, log_v0 = TRUE)
```

The function `EMVSplot` plots the estimates of the regression coefficients, $\widehat{\boldsymbol{\beta}}$, over all the values of `v0` (i.e. the regularization plot). Here the v_0 are plotted on the log scale - this is because we set the option `log_v0` = `TRUE` when running `EMVS`.

```
EMVSplot(result_ind, "both", FALSE)
```



Note that the default option `direction = "backward"` results in the EMVS algorithm stabilizing for small v_0. This eliminates the need to choose a v_0; we take the coefficients at the smallest v_0 as our solution.

6.2 Conjugate Prior

We now run `EMVS`, using the conjugate prior formulation as outlined in the original paper (`independent = FALSE`). The output is stored in `result_conj`. We re-initialize both the slab parameter `v1` and the ladder of spike values `v0` as the scale of the variance is different for the independent and conjugate formulations.

```
v0 = seq(0.1, 2, length.out = 20)
v1 = 1000
result_conj = EMVS(Y, X, v0 = v0, v1 = v1, type = "betabinomial",
                   independent = FALSE, beta_init = beta_init,
                   sigma_init = sigma_init, a = a, b = b)
```

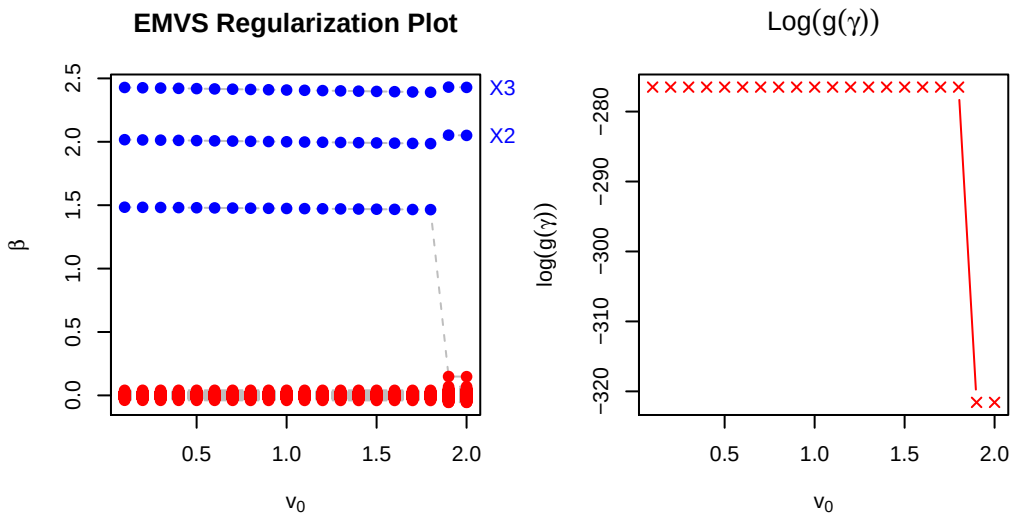
The function `EMVSbest` shows the maximum value of the `log_g` function over all the `v0` values (Rockova and George, 2014) and the non-zero indices of the model with the highest value of `log_g` (the marginal posterior).

```
EMVSbest(result_conj)
```

```
## [1] "Best Model Found"
## $log_g_function
## [1] -276.5027
##
## $indices
## [1] 1 2 3
```

We plot both the regularization path and the values of the `log_g` function.

```
EMVSplot(result_conj, "both", FALSE)
```

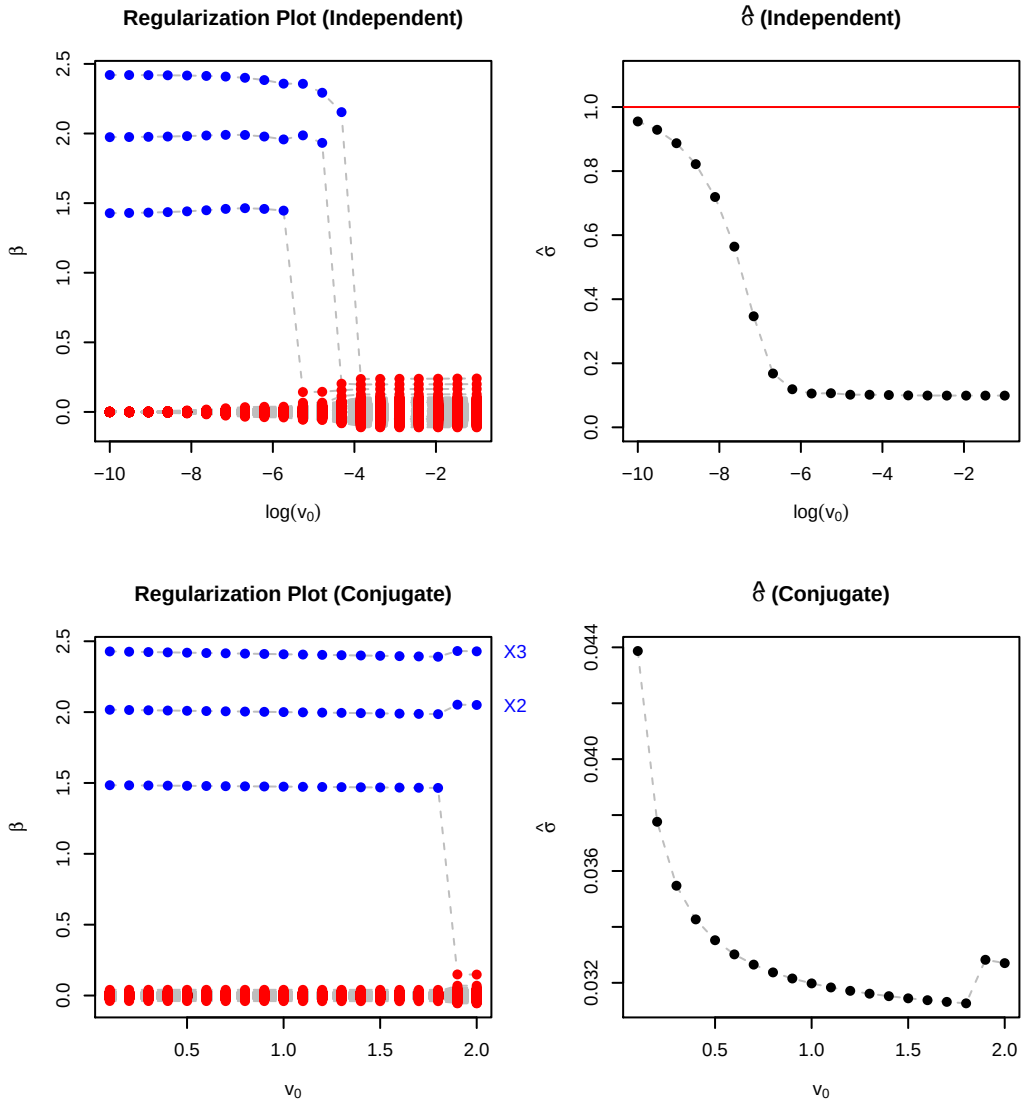


Note that the `log_g` function is unavailable for the `independent = TRUE` implementation as the independent prior formulation does not yield a closed form for the marginal posterior. An approximation will be implemented in a future update to the `EMVS` package. As discussed earlier, however, the "backward" strategy stabilizes for small values v_0 and so we recommend taking the coefficients found at the smallest value of v_0 as the solution. In many cases this eliminates the need for the `log_g` function as a criterion.

6.3 Variance Estimation

We now compare the error variance estimates from the independent and conjugate prior implementations of EMVS.

The below plot shows the estimates of the standard deviation of the error ($\widehat{\sigma}$) for each value of v_0 for both the independent and conjugate prior formulations, as well as the regularization plots for the coefficients, $\boldsymbol{\beta}$.



We can see that the conjugate prior formulation severely underestimates the true error variance ($\sigma = 1$) with the estimate $\widehat{\sigma}_{conj} = 0.0439$. In contrast, the estimate of σ at `v0` = `exp(-10)` for the independent case is $\widehat{\sigma}_{ind} = 0.955$, much closer to the true value.

7 Conclusion

In this vignette, we described the EMVS algorithm of Rockova and George (2014) with an independent prior formulation. We demonstrated how this algorithm can be applied using the `EMVS` R package.

8 References

George, E. I. and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection", *Statistica Sinica*, 7, 339-373

Moran, G. E., Rockova, V. and George, E. I., (2018) "On variance estimation for Bayesian variable selection", [arXiv:1801.03019]

Rockova, V. and George, E. I. (2014), "EMVS: The EM approach to Bayesian variable selection," *Journal of the American Statistical Association*, 109, 828-846