Published in final edited form as:
Metron. 2014 August 1; 72(2): 217–229. doi:10.1007/s40300-014-0047-y.

Negotiating Multicollinearity with Spike-and-Slab Priors

Veronika Ročková [Postdoctoral Research Associate] and
Department of Statistics, University of Pennsylvania, Philadelphia, PA, 19106,
vrockova@wharton.upenn.edu

Edward I. George [Professor of Statistics]
Department of Statistics, University of Pennsylvania, Philadelphia, PA, 19106,
edgeorge@wharton.upenn.edu

Abstract

In multiple regression under the normal linear model, the presence of multicollinearity is well known to lead to unreliable and unstable maximum likelihood estimates. This can be particularly troublesome for the problem of variable selection where it becomes more difficult to distinguish between subset models. Here we show how adding a spike-and-slab prior mitigates this difficulty by filtering the likelihood surface into a posterior distribution that allocates the relevant likelihood information to each of the subset model modes. For identification of promising high posterior models in this setting, we consider three EM algorithms, the fast closed form EMVS version of Rockova and George (2014) and two new versions designed for variants of the spike-and-slab formulation. For a multimodal posterior under multicollinearity, we compare the regions of convergence of these three algorithms. Deterministic annealing versions of the EMVS algorithm are seen to substantially mitigate this multimodality. A single simple running example is used for illustration throughout.

1 Posterior Resolution of the Likelihood

Suppose we observe data that consists of $\boldsymbol{y}$, an $n \times 1$ response vector, and $\boldsymbol{X} = [x_1,\dots, x_p]$, an $n \times p$ matrix of p potential standardized predictors that are related by a Gaussian linear model

$$f(\boldsymbol{y}|\alpha,\boldsymbol{\beta},\sigma)=\mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\mathrm{I}_n), \quad (1.1)$$

where $\beta = (\beta_1,\dots,\beta_p)$ is a $p \times 1$ vector of unknown regression coefficients, and σ is an unknown positive scalar. (We assume throughout that $\boldsymbol{y}$ has been centered at zero to avoid the need for an intercept). For this setup we shall also suppose that only an unknown subset of the coefficients in β are zero, and that the goal is the identification and estimation of this subset. Of particular interest to us will be addressing this problem in the presence of multicollinearity, where it is well known that the maximum likelihood estimator (also the least squares estimator) is an unreliable and unstable estimator of β.

Correspondence to: Edward I. George.

A fundamental Bayesian approach to this variable selection problem is obtained by introducing a “spike-and-slab” Gaussian mixture prior on β. Conditionally on a random vector of binary latent variables $\gamma = (\gamma_1,\ldots, \gamma_p)'$, $\gamma_i \in \{0, 1\}$, this prior is defined by

$$\pi(\boldsymbol{\beta}|\sigma,\boldsymbol{\gamma})=\mathrm{N}_p(\mathbf{0},\sigma^2\boldsymbol{D}_\gamma^{1/2}\boldsymbol{R}\boldsymbol{D}_\gamma^{1/2}), \quad (1.2)$$

where ***R*** is a preset covariance matrix and

$$\boldsymbol{D}_\gamma=\mathrm{diag}\{[(1-\gamma_1)\upsilon_0+\gamma_1\upsilon_1],\ldots,[(1-\gamma_p)\upsilon_0+\gamma_p\upsilon_1]\} \quad (1.3)$$

for $0 \le \upsilon_0 < \upsilon_1$, George and McCulloch (1993). This prior on β is then combined with suitable priors on α, σ and γ. Typical default choices include the relatively noninuential inverse gamma prior on σ^2, $\pi(\sigma^2) = IG(\nu/2, \nu\lambda/2)$ with $\nu = \lambda = 1$, and the exchangeable beta-binomial prior on γ, obtained by coupling the iid Benoulli form $\pi(\gamma \mid \theta) = \theta^{|\gamma|}(1 - \theta)^{p-|\gamma|}$, $(|\gamma|:=\sum_{i=1}^p\gamma_i)$, with a uniform prior on $\theta \in [0, 1]$. We will restrict attention to these choices throughout.

Under (1.2), the β_i components of β are marginally distributed as

$$\pi(\beta_i|\sigma,\gamma)=(1-\gamma_i)\mathrm{N}(0,\sigma^2\upsilon_0)+\gamma_i\mathrm{N}(0,\sigma^2\upsilon_1), \quad (1.4)$$

a mixture of a “spike distribution” $N(0, \sigma^2\upsilon_0)$ and a “slab distribution” $N(0, \sigma^2\upsilon_1)$. For the purpose of variable selection, the idea is to set υ_0 small and υ_1 large, so that the induced posterior will segregate the β_i coefficients into those that are attributed to the spike distribution and so are inconsequential, and those that are attributed to the slab distribution and so are important. As an alternative to setting υ_1 to be a large fixed value, one may instead add a prior $\pi(\upsilon_1)$ to induce heavy-tailed slab distributions.

We shall be interested here in considering the effect of the two particular choices, $\boldsymbol{R} = I_p$ and $\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$. The choice $\boldsymbol{R} = I_p$, under which the components of β are independently distributed, serves to decrease the posterior correlation between these components. The *g*-prior (Zellner, 1986) related choice $\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, which is proportional to the correlation structure of the likelihood estimates of β, serves to reinforce this structure in the posterior. These two choices have opposite effects on the posterior correlation.

After integrating out α, the induced marginal posterior distribution on β is of the form

$$\pi(\boldsymbol{\beta}|\boldsymbol{y})=\sum_\gamma\pi(\boldsymbol{\beta}|\gamma,\boldsymbol{y})\pi(\gamma|\boldsymbol{y}) \quad (1.5)$$

where the model posterior $\pi(\gamma \mid \boldsymbol{y})$ puts more weight on those models which are more likely to have generated ***y***. The form of each conditional component of (1.5),

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{y}) \propto \int L(\boldsymbol{\beta},\sigma|\boldsymbol{y})\pi(\boldsymbol{\beta}|\sigma,\boldsymbol{\gamma})\pi(\sigma)d\sigma, \quad (1.6)$$

shows how the likelihood $L(\beta, \sigma \mid \boldsymbol{y})$, which does not at all depend on γ, is filtered to determine the distribution of β for each of the subset models determined by γ.

Insight into the effect of the spike-and-slab prior (1.2) on the posterior components through (1.6) is clearest when $\boldsymbol{R} = \mathrm{I}_p$. In this case, $L(\beta, \sigma \mid \boldsymbol{y})$ is multiplied by

$$\pi(\boldsymbol{\beta}|\sigma,\boldsymbol{\gamma})=\left[\prod_{\gamma_i=0}\phi_{\sigma^2\upsilon_0}(\beta_i)\right]\left[\prod_{\gamma_i=1}\phi_{\sigma^2\upsilon_1}(\beta_i)\right], \quad (1.7)$$

where $\varphi_{\sigma^2\upsilon}(\cdot)$ is the $N(0, \sigma^2\upsilon)$ density function. Thus, $L(\beta, \sigma \mid \boldsymbol{y})$ is downweighted when both β_i is large and $\gamma_i = 0$. In particular, when $\upsilon_0 = 0$, $L(\beta, \sigma \mid \boldsymbol{y})$ is multiplied by 0 when both $\beta_i \neq 0$ and $\gamma_i = 0$, effectively conditioning on $\beta_i = 0$. In contrast, $L(\beta, \sigma \mid \boldsymbol{y})$ is relatively unaffected by those β_i for which $\gamma_i = 1$, as long as they are not too extreme as determined by υ_1.

The translation of the likelihood information into distinct components via (1.5) facilitates the identification of those submodels best supported by the data. The enhanced clarity provided by the posterior is especially pronounced in the presence of multicollinearity, as illustrated by the simple simulated example described in the next section.

2 Motivating Example with Multicollinearity

We constructed $n = 100$ observations on $p = 2$ predictors according to $\mathrm{N}_p(0, \Sigma)$ with $\Sigma=(\rho_{ij})_{i,j=1}^p$ and $\rho_{ij} = 0.9^{|i-j|}$. Setting $\beta = (1, 0)'$, we then generated responses from $\mathrm{N}_n(\boldsymbol{X}\beta, \sigma^2\mathrm{I}_n)$ with $\sigma^2 = 3$. The resulting maximum likelihood estimate (MLE) $\hat{\beta}_{MLE} = (0.55, 0.38)'$ is at the center of the likelihood surface depicted in Figure 1(a). The instability of the MLE due to the collinearity between the predictors has led it to misallocate the signal across the coordinates.

Now consider what happens when we add the spike-and-slab prior with $\upsilon_0 = 0$, $\upsilon_1 = 1000$ and $\boldsymbol{R} = \mathrm{I}_2$. The posterior $\pi(\beta \mid \boldsymbol{y})$ has translated the likelihood information through (1.5) to the $\pi(\gamma \mid \boldsymbol{y})$ weighted sum of the four $\pi(\beta \mid \gamma, \boldsymbol{y})$ components corresponding to $\gamma = (0, 0)', (1, 0)', (0, 1)', (1, 1)'$, respectively. By using $\upsilon_0 = 0$, which yields a point mass at zero for the spike distribution, these four components support β values of the form $(0, 0)'$, $(\beta_1, 0)'$, $(0, \beta_2)'$ and $(\beta_1, \beta_2)'$, respectively.

These components can be identified as the four regions of posterior accumulation in Figure 1(b), which depicts the four posterior modes and associated posterior student-t contours (and 95% HPD intervals for the 1-dimensional regions). The probability mass is distributed among these components according to the posterior model probabilities $\pi[(0, 0)'|\boldsymbol{y}] < 0.001$, $\pi[(0, 1)'|\boldsymbol{y}] = 0.187$, $\pi[(1, 0)' \mid \boldsymbol{y}] = 0.764$ and $\pi[(1, 1)' \mid \boldsymbol{y}] = 0.049$. The global mode $\hat{\beta}_{MAP} = (0.91, 0.00)'$, a much better estimate than the MLE, is sitting atop the $(\beta_1, 0)$ component, which has been most heavily weighted by $\pi(\gamma \mid \boldsymbol{y})$ for this data. We note in passing that replacing $\boldsymbol{R} = \mathrm{I}_2$ by $\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$ here, yields virtually the same posterior as in Figure 1(b).

It will also be of interest to get some insight into the effect of using $\upsilon_0 > 0$, so let us also consider what happens when we change $\upsilon_0 = 0$ to $\upsilon_0 = 0.005$ in the above. As before, the posterior $\pi(\beta \mid \boldsymbol{y})$ is the weighted sum of the four $\pi(\beta \mid \gamma, \boldsymbol{y})$ components corresponding to $\gamma = (0, 0)', (1, 0)', (0, 1)', (1, 1)'$, respectively. However because $\upsilon_0 > 0$ employs a continuous spike distribution, none of these posterior components rule out any β values. Instead, the first three posterior components are distinguished only by more heavily weighting β values close to $(0, 0)'$, $(\beta_1, 0)'$, $(0, \beta_2)'$.

Under the independence prior choice $\boldsymbol{R} = I_2$, these components can still be identified as the four regions of posterior accumulation in Figure 2(a). However, in contrast to Figure 1(b), they are now all full dimensional and not quite as cleanly separated. Nonetheless, the global mode $\hat{\beta}_{MAP} = (0.88, 0.03)'$, is still a much better estimate than the MLE, sitting atop the $\gamma = (1, 0)'$ component, which again has been most heavily weighted by $\pi(\gamma \mid \boldsymbol{y})$ for this data. This posterior has still been effective at mitigating the effect of multicollinearity.

When we instead consider the g-prior choice $\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, it is interesting to see how the posterior $\pi(\beta \mid \boldsymbol{y})$, now depicted in Figure 2(b), has changed. The four components now appear as an intermediate change between the $\upsilon_0 = 0$ posterior in Figure 1(b) and the $\upsilon_0 > 0$ posterior in Figure 2(a). The g-prior structure has served to maintain the multicollinear structure in the components corresponding to $\gamma = (0, 0)', (1, 0)', (0, 1)'$, keeping them more similar to their lower dimensional counterparts in Figure 1(b). With this posterior, the global mode $\hat{\beta}_{MAP} = (0.901, 0.001)'$ is now even closer to the true $\beta = (1, 0)'$.

3 EM Algorithms for Posterior Mode Identification

As a practical matter, identification of the posterior mode under the spike-and-slab prior formulation can be challenging when the number of predictors is large. For this purpose, EM algorithms can provide a fast deterministic search alternative to stochastic search approaches such as SSVS George and McCulloch (1993, 1997). In this section, we describe three such EM algorithms which are designed for different choices of υ_0 and $\boldsymbol{R}$. The first of these was proposed by Rockova and George (2014) (hereafter RG14) for ($\upsilon_0 > 0$, $\boldsymbol{R} = I_p$), involving closed form E-step and M-step updates. The other two are new algorithms designed for the cases ($\upsilon_0 > 0$, $\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$) and ($\upsilon_0 = 0$, $\boldsymbol{R} = I_p$), respectively, and exploit mean-field approximations in the E-step.

The EM algorithm has been previously considered in the context of Bayesian shrinkage estimation under sparsity priors (Figueiredo (2003)), Kiiveri (2003), Griffin and Brown (2012, 2005). Literature on similar computational procedures for spike-and-slab models is far more sparse. EM-like algorithms for point mass variable selection priors were considered by Hayashi and Iwata (2010) and Bar et al. (2010), which approximate the E-step by neglecting the correlation among the selection indicators. The EM algorithm for the point mass prior that we describe below uses a mean-field approximated E-step, which takes into account the fact that the selection indicators can be dependent. A similar dependence also occurs in the case ($\upsilon_0 > 0$ and $\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$). There we again take advantage of the mean field approximation, which induces "dependent thresholding" for variable selection rather than univariate thresholding along individual coordinate axes.

3.1 The EMVS Algorithm for $\upsilon_0 > 0$ and $R = I_p$

For the problem of model identification, RG14 proposed EMVS, an approach based on a fast closed form EM algorithm that quickly identifies posterior modes of $\pi(\beta, \theta, \sigma^2 \mid \boldsymbol{y})$ under a spike-and-slab prior with $\upsilon_0 > 0$ and $\boldsymbol{R} = I_p$. The modal β values are then thresholded to identify nearby high posterior γ models under the posterior for which $\upsilon_0 = 0$. We refer to this EM algorithm as the EMVS algorithm.

As described in detail in RG14, the EMVS algorithm proceeds by iteratively maximizing the objective function

$$
\begin{aligned}
&Q\left(\boldsymbol{\beta},\theta,\sigma|\psi^{(k)}\right)\\
&=-\frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2}{2\sigma^2}\\
&-\log\sigma^{n-1+p+\nu}-\frac{\nu\lambda}{2\sigma^2}-\frac{1}{2\sigma^2}\sum_{i=1}^{p}\beta_i^2\mathsf{E}_{\gamma|\cdot}\left[\frac{1}{\upsilon_0(1-\gamma_i)+\upsilon_1\gamma_i}\right]+\sum_{i=1}^{p}\log\left(\frac{\theta}{1-\theta}\right)\mathsf{E}_{\gamma|\cdot}[\gamma_i]+p\log(1-\theta),
\end{aligned}
\quad (3.1)
$$

where $\psi^{(k)} = (\beta^{(k)}, \theta^{(k)}, \sigma^{(k)})$ and $\mathsf{E}_{\gamma|\cdot}[\cdot]$ denotes expectation conditionally on $[\psi^{(k)}, \boldsymbol{y}]$. At the *k*th iteration, an E-step is first applied, computing the expectations in (3.1), followed by an M-step that maximizes over (β, θ, σ) to yield the values of $\psi^{(k+1)} = (\beta^{(k+1)}, \theta^{(k+1)}, \sigma^{(k+1)})$.

The E-step expectations are obtained quickly from the closed form expressions

$$
\mathsf{E}_{\gamma|\cdot}[\gamma_i]=\frac{\pi(\beta_i^{(k)}|\sigma^{(k)},\gamma_i{=}1)\theta^{(k)}}{\pi(\beta_i^{(k)}|\sigma^{(k)},\gamma_i{=}1)\theta^{(k)}+\pi(\beta_i^{(k)}|\sigma^{(k)},\gamma_i{=}0)(1-\theta^{(k)})}\equiv p_i^{\star} \quad (3.2)
$$

and

$$
\mathsf{E}_{\gamma|\cdot}\left[\frac{1}{\upsilon_0(1-\gamma_i)+\upsilon_1\gamma_i}\right]=\frac{1-p_i^{\star}}{\upsilon_0}+\frac{p_i^{\star}}{\upsilon_1}\equiv d_i^{\star}. \quad (3.3)
$$

Note that these closed form expressions are available when $\upsilon_0 > 0$ but not when $\upsilon_0 = 0$.

For the M-step maximization, the $\beta^{(k+1)}$ value that globally maximizes *Q* is obtained by the generalized ridge estimator

$$
\boldsymbol{\beta}^{(k+1)}=(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{D}^{\star})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad (3.4)
$$

where $\boldsymbol{D}^{\star}=\mathrm{diag}\{d_i^{\star}\}_{i=1}^{p}$ is the *p* × *p* diagonal matrix with entries $d_i^{\star}>0$, the well-known solution to the ridge regression problem

$$
\boldsymbol{\beta}^{(k+1)}=\mathrm{argmin}_{\beta\in\mathbb{R}^p}\{\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2+\|\boldsymbol{D}^{\star 1/2}\boldsymbol{\beta}\|^2\}. \quad (3.5)
$$

In problems where *p* >> *n*, the calculation of (3.4) can be enormously reduced by using the Sherman-Morrison-Woodbury formula to obtain an expression which requires a *n* × *n* matrix inversion rather than a *p* × *p* matrix inversion. Alternatively, as described in George et al. (2013), the solution of (3.5) can be obtained even faster with the stochastic dual coordinate ascent algorithm of Shalev-Shwartz and Zhang (2013).

The maximization of *Q* is then completed with the simple updates

$$\sigma^{(k+1)}=\sqrt{\frac{\left\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(k+1)}\right\|^2+\left\|\boldsymbol{D}^{\star 1/2}\boldsymbol{\beta}^{(k+1)}\right\|^2+\nu\lambda}{n+p+\nu}} \quad (3.6)$$

and

$$\theta^{(k+1)}=\frac{1}{p}\sum_{i=1}^{p}p_i^{*}. \quad (3.7)$$

3.2 An Approximate EM Algorithm when $\upsilon_0 > 0$ and $\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$

When $\upsilon_0 > 0$ and $\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, an EM algorithm with a fast closed form E-step is no longer available. However, as we now show, an EM algorithm for variable selection becomes feasible with deterministic mean field approximations.

The expectation of the complete data log-likelihood here requires computation of both the first and second moments of the vector of latent γ indicators. This is better seen from the following expression for the objective function

$$\begin{aligned} &Q\left(\boldsymbol{\beta},\theta,\sigma|\psi^{(k)}\right)\\ &=-\frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\\ &-\log\sigma^{n-1+p+\nu}-\frac{\nu\lambda}{2\sigma^2}-\frac{\mathsf{E}_{\gamma|\cdot}[\boldsymbol{\beta}'\boldsymbol{D}_{\gamma}^{-1/2}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D}_{\gamma}^{-1/2}\boldsymbol{\beta}]}{2\sigma^2} \quad (3.8)\\ &+\sum_{i=1}^{p}\log\left(\frac{\theta}{1-\theta}\right)\mathsf{E}_{\gamma|\cdot}[\gamma_i]+p\log(1-\theta). \end{aligned}$$

The expectation of the quadratic form in the fourth summand can be written as

$$\mathsf{E}_{\gamma|\cdot}[\boldsymbol{\beta}'\boldsymbol{D}_{\gamma}^{-1/2}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D}_{\gamma}^{-1/2}\boldsymbol{\beta}]=\left(\frac{1}{\sqrt{\upsilon_1}}-\frac{1}{\sqrt{\upsilon_0}}\right)^2\mathrm{tr}\{\mathrm{diag}\{\boldsymbol{\beta}\}\boldsymbol{X}'\boldsymbol{X}\mathrm{diag}\{\boldsymbol{\beta}\}\mathsf{E}_{\gamma|\cdot}\boldsymbol{\gamma}\boldsymbol{\gamma}'\} \quad (3.9)$$

$$+\frac{2}{\sqrt{\upsilon_0}}\left(\frac{1}{\sqrt{\upsilon_1}}-\frac{1}{\sqrt{\upsilon_0}}\right)\mathsf{E}_{\gamma|\cdot}\boldsymbol{\gamma}'\mathrm{diag}\{\beta\}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}+\frac{1}{\upsilon_0}\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}. \quad (3.10)$$

The first and second conditional moments of γ above are with respect to the following conditional distribution which, based on (1.2), resembles a Markov Random Field (MRF) distribution on a completely connected graph with self-loops, i.e.

$$\pi(\boldsymbol{\gamma}|\boldsymbol{\psi}^{(k)})\propto\exp\{\boldsymbol{\gamma}'\boldsymbol{a}+\boldsymbol{\gamma}'B\boldsymbol{\gamma}\}, \quad (3.11)$$

where

$$\boldsymbol{a}=-\frac{1}{\sigma^{(k)2}}\left(\frac{\sqrt{\upsilon_0}-\sqrt{\upsilon_1}}{\sqrt{\upsilon_1}\upsilon_0}\right)\mathrm{diag}\{\boldsymbol{\beta}^{(k)}\}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}^{(k)}+\left[\frac{1}{2}\log\left(\frac{\upsilon_0}{\upsilon_1}\right)+\log\left(\frac{\theta^{(k)}}{1-\theta^{(k)}}\right)\right]\mathbf{1} \tag{3.12}$$

and

$$\boldsymbol{B}=-\frac{1}{2\sigma^{(k)2}}\left(\frac{1}{\sqrt{\upsilon_1}}-\frac{1}{\sqrt{\upsilon_0}}\right)^2\mathrm{diag}\{\boldsymbol{\beta}^{(k)}\}\boldsymbol{X}'\boldsymbol{X}\mathrm{diag}\{\boldsymbol{\beta}^{(k)}\}. \tag{3.13}$$

Note that the distribution (3.11) deviates slightly from a traditional MRF distribution, which assumes that the matrix ***B*** has zeroes on the diagonal. Nevertheless, the exact computation of $\mathsf{E}_{\gamma|\cdot}\gamma$ and $\mathsf{E}_{\gamma|\cdot}\gamma\gamma'$ is still feasible in small problems. However, for applications involving moderate to large numbers of predictors, approximate computation will be needed.

It is interesting to point out the effect of the *g*-prior on simultaneously selecting variables that are related. For standardized predictors, ***X′X*** is proportional to the sample correlation matrix. A pair (*i*, *j*) of highly collinear predictors will have a large entry $(\boldsymbol{X'X})_{(i,j)}$ in absolute value, which can be potentially magnified by the current parameter estimates $\beta_i^{(k)}$ and $\beta_j^{(k)}$. For example, for large positive $(\boldsymbol{X'X})_{(i,j)}$, large $\beta_i^{(k)}$ and $\beta_j^{(k)}$ of the same sign will lead to a large negative $\boldsymbol{B}_{(i,j)}$ entry, which will in turn lower their probability of co-occurence at the *k*th iteration.

3.2.1 Mean Field Approximated E-step—For the E-step calculations, we deploy a variant of a mean field approximation which outputs approximations to $\mathsf{E}_{\gamma|\cdot}\gamma_i$ for $i = 1,\ldots, p$ as a solution to a series of nonlinear equations. Proceeding as in RG14, the approximations $\hat{\mathsf{E}}_{\gamma|\cdot}\gamma_i = \mu_i$ can be obtained by solving

$$\mu_i=\frac{\exp(a_i+\sum_{j=1}^{p}B_{ij}\mu_j)}{1+\exp(a_i+\sum_{j=1}^{p}B_{ij}\mu_j)},\mathrm{for} i=1,\ldots,p.$$

Because the vector ***a*** and matrix ***B*** can involve rather large numbers, it will be useful in practice to perform the approximation with reparametrized values $\boldsymbol{a}\star = \boldsymbol{a}/C$ and $\boldsymbol{B}\star = \boldsymbol{B}/C^2$ to obtain $\mu_i^{\star}$. The solution can be transformed back by noting $\mathrm{logit}(\mu_i)=C\mathrm{logit}(\mu_i^{\star})$. The constant *C* can be for instance chosen as the maximum of the absolute value of the entries in the matrix ***B*** and the vector ***a***. The matrix of second posterior moments can then be approximated by $\hat{\mathsf{E}}_{\gamma|\cdot}[\gamma\gamma'] = \hat{\mathsf{E}}_{\gamma|\cdot}[\gamma]\hat{\mathsf{E}}_{\gamma|\cdot}[\gamma']$, because the approximating distribution assumes a completely disconnected graph.

3.2.2 The M-step—The M-step proceeds by jointly updating (β,σ,θ), which is equivalent to updating (β,σ) and θ individually. Only the updates for the regression parameters and residual variance require slight modification:

$$\boldsymbol{\beta}^{(k+1)}=\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{D}^{\star 1/2}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D}^{\star 1/2}\right]^{-1}\boldsymbol{X}'\boldsymbol{y},$$

and

$$\sigma^{(k+1)} = \sqrt{\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(k+1)}\|^2 + \|\boldsymbol{\beta}^{(k+1)\prime}\boldsymbol{D}^{\star 1/2}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D}^{\star 1/2}\boldsymbol{\beta}^{(k+1)}\|^2 + \nu\lambda}{n - 1 + p + \nu}}.$$

3.3 An Approximate EM Algorithm when υ_0 = 0 and $\boldsymbol{R}$ = I_p

We now describe a variant of the EM algorithm for the point-mass prior (υ_0 = 0), assuming that the covariates are a priori independent ($\boldsymbol{R}$ = I_p). Here, the joint prior distribution (1.2) is degenerate for all models of dimension smaller than *p*.

Our derivation proceeds by first rewriting the likelihood for every model γ as

$$f(\boldsymbol{y}|\boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma) = \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{\beta}, \sigma^2\mathrm{I}_n), \quad (3.14)$$

where $\boldsymbol{\Gamma} = \mathrm{diag}\{\gamma_i\}_{i=1}^p$. The objective function for the EM algorithm then becomes

$$\begin{aligned} Q&\left(\boldsymbol{\beta}, \theta, \sigma|\boldsymbol{\psi}^{(k)}\right) \\ &= -\frac{\mathsf{E}_{\gamma|\cdot}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{\beta}\|^2}{2\sigma^2} \\ &\quad - \log\sigma^{n-1+\sum \mathsf{E}_{\gamma|\cdot}[\gamma_i]+\nu} \\ &\quad - \frac{\nu\lambda}{2\sigma^2} - \frac{\mathsf{E}_{\gamma|\cdot}[\boldsymbol{\beta}'\mathrm{diag}\{\gamma_i\}\boldsymbol{\beta}]}{2\sigma^2\upsilon_1} \\ &\quad + \sum_{i=1}^p \log\left(\frac{\theta}{1-\theta}\right)\mathsf{E}_{\gamma|\cdot}[\gamma_i] + p\log(1-\theta), \end{aligned} \quad (3.15)$$

The expectation of the quadratic form in the first summand can be written as

$$\mathsf{E}_{\gamma|\cdot}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{\beta}\|^2 = -\frac{1}{2\sigma^2}\mathrm{tr}\{\mathrm{diag}\{\boldsymbol{\beta}\}\boldsymbol{X}'\boldsymbol{X}\mathrm{diag}\{\boldsymbol{\beta}\}\mathsf{E}_{\gamma|\cdot}\boldsymbol{\gamma}\boldsymbol{\gamma}'\} + \frac{1}{\sigma^2}\mathsf{E}_{\gamma|\cdot}\boldsymbol{\gamma}'\mathrm{diag}\{\boldsymbol{\beta}\}\boldsymbol{X}'\boldsymbol{Y} - \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{y}.$$

As in the EM algorithm for the *g*-prior, we compute the first and second posterior moments of the latent inclusion indicators. The E-step can again be obtained with a mean field approximation, this time using a slightly different MRF distribution

$$\pi(\boldsymbol{\gamma}|\boldsymbol{\psi}^{(k)}) \propto \exp\{\boldsymbol{\gamma}'\boldsymbol{a} + \boldsymbol{\gamma}'B\boldsymbol{\gamma}\}, \quad (3.16)$$

where

$$\boldsymbol{a} = \frac{1}{\sigma^{(k)2}}\mathrm{diag}\{\boldsymbol{\beta}^{(k)}\}\boldsymbol{X}'\boldsymbol{Y} - \frac{1}{2\sigma^{(k)2}\upsilon_1}\mathrm{diag}\{\boldsymbol{\beta}^{(k)}\}\boldsymbol{\beta}^{(k)} + \left[\frac{1}{2}\log\left(\frac{1}{\upsilon_1}\right) + \log\left(\frac{\theta^{(k)}}{1-\theta^{(k)}}\right)\right]\mathbf{1} \quad (3.17)$$

and

$$\boldsymbol{B} = -\frac{1}{2\sigma^{(k)2}}\mathrm{diag}\{\boldsymbol{\beta}^{(k)}\}\boldsymbol{X}'\boldsymbol{X}\mathrm{diag}\{\boldsymbol{\beta}^{(k)}\}. \quad (3.18)$$

Focusing on the vector of sparsity parameters a, negative parameters $\beta_i^{(k)}$ for covariates, which correlate positively with the outcome, induce smaller baseline inclusion probabilities $\exp(a_i)/[1+\exp(a_i)]$. This behavior will become more evident from the plots of convergence regions in the next section.

The M-step update of the joint vector of regression coefficients

$$\boldsymbol{\beta}^{(k+1)} = (\mathsf{E}_{\gamma|\cdot}\boldsymbol{\Gamma}\boldsymbol{X}'\boldsymbol{X}\mathsf{E}_{\gamma|\cdot}\boldsymbol{\Gamma} + \boldsymbol{D}^{\star})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

can be unstable, when entries in the approximated $\mathsf{E}_{\gamma|\cdot}[\Gamma]$ approach zero. In such instances, it will be useful to set $\beta_i^{(k)}$ directly to zero, when the conditional inclusion probability is smaller than a pre-specified threshold. This sparsification step induces more stability, while rendering the exclusion of each covariate throughout the iterations reversible.

The residual variance is then updated according to

$$\sigma^{(k+1)} = \sqrt{\frac{\|\boldsymbol{y} - \boldsymbol{X}\mathsf{E}_{\gamma|\cdot}\boldsymbol{\Gamma}\boldsymbol{\beta}^{(k+1)}\|^2 + \|\boldsymbol{\beta}^{(k+1)\prime}\boldsymbol{D}^{\star}\boldsymbol{\beta}^{(k+1)}\|^2 + \nu\lambda}{n - 1 + \sum \mathsf{E}_{\gamma|\cdot}\gamma_i + \nu}}.$$

4 Geometry of EM Convergence Regions

Despite attractive features such as rapid convergence, economy of storage and computational speed, our EM algorithms may converge to local modes rather than the global mode of main interest. Overcoming this tendency can be especially challenging in multimodal posterior landscapes, such as those induced by our spike-and-slab priors, where performance becomes heavily dependent on the choice of a starting value. To shed some light on the extent to which this occurs with our EM algorithms, we studied this aspect of their performance on the simple example with two correlated predictors from Section 2. Note that the posterior multimodality there has been exacerbated by the strong collinearity between the predictors.

For a regular grid of values on [−0.5, 1.5] × [−0.5, 1.5], we ran each of our three EM algorithms, starting at each point on the grid and recording the mode to which the algorithm converged. These results are displayed in Figure 3(a) for $\upsilon_0 = 0.005$ and $\upsilon_1 = 1000$ with $\boldsymbol{R} = I_2$, in Figure 3(b) for $\upsilon_0 = 0.005$ and $\upsilon_1 = 1000$ with $\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$ and in Figure 3(c) for $\upsilon_0 = 0$ and $\upsilon_1 = 1000$ with $\boldsymbol{R} = I_2$. Numbering the modes 1, 2, 3 and 4, each starting value has been assigned same number as its EM destination mode. This results in a partition of the grid into four regions, delineating the regions of attraction for each mode.

These three figures confirm the susceptibility of the EM algorithms to starting values. The convergence towards the global mode (here mode 3) is not guaranteed unless the starting value belongs to the small region of attraction around it. The geometry of the convergence regions differs depending on the variant of our EM algorithm. The independence covariance matrix $\boldsymbol{R} = I_2$ yields nearly rectangular regions corresponding to thresholding univariate directions. The point-mass prior $\upsilon_0 = 0$ penalizes the directions of sign inconsistency between the starting value and the sample correlation between the variable and the response. The g-prior $\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$ performs dependent thresholding along multivariate directions, rather than coordinate axes.

5 Mitigating Multimodality with Deterministic Annealing

In order to increase the chances of converging to the global mode, Ueda and Nakano (1998) propose a deterministic annealing EM variant (DAEM) based on the principle of maximum entropy and an analogy with statistical mechanics. In our context, the DAEM algorithm aims at finding a maximum of the negative of the free energy function

$H_t(\boldsymbol{\beta},\boldsymbol{\theta},\sigma)=\frac{1}{t}\log\sum_{\gamma}\pi(\boldsymbol{\beta},\boldsymbol{\theta},\sigma,\boldsymbol{\gamma}|\boldsymbol{y})^t$ where $0 < t < 1$. The problem of optimizing the logarithm of the incomplete posterior distribution is embedded as a special case for $t = 1$.

The parameter $1/t$ corresponds to a temperature parameter and determines the degree of separation between the multiple modes of $F_t(\cdot)$. Large enough values smooth the function to have only one minimum. As the temperature decreases, multiple modes begin to appear and the function gradually resembles the true log incomplete posterior. The influence of poorly chosen starting values can be weakened by keeping the temperature high at the early stage of computation, gradually decreasing it during the iteration process. Alternatively, the free energy function can be optimized for a decreasing sequence of temperature levels $1/t_1 > 1/t_2 > \cdots > 1/t_k$, where the solution at $1/t_i$ serves as the starting point for the computation at $1/t_{i+1}$. Provided that the new global maximum is close to the previous one, this strategy can increase the chances of finding the true global maximum. We will investigate how the tempering affects the ability to converge towards the region of attraction of the global mode.

While the M-step of the DAEM algorithm remains unchanged, the E-step requires the computation of the expected complete log posterior density with respect to a distribution which is proportional to the current estimate of the conditional complete posterior given the observed data raised to the power t. This distribution is particularly easy to derive for mixtures (Ueda and Nakano, 1998). We begin by describing this distribution for the EMVS algorithm with $\upsilon_0 > 0$ and $\boldsymbol{R} = I_p$, in which case this corresponds to a Bernoulli distribution with inclusion probabilities

$$p_i^{t\star}=\frac{\pi(\beta_i|\sigma^{(k)},\gamma_i=1)^t\,\mathsf{P}(\gamma_i=1|\boldsymbol{\theta}^{(k)})^t}{\pi(\beta_i|\sigma^{(k)},\gamma_i=1)^t\,\mathsf{P}(\gamma_i=1|\boldsymbol{\theta}^{(k)})^t+\pi(\beta_i|\sigma^{(k)},\gamma_i=0)^t\,\mathsf{P}(\gamma_i=0|\boldsymbol{\theta}^{(k)})^t}. \tag{5.1}$$

The EMVS algorithm with annealing then proceeds by substituting (5.1) for (3.2). At high temperatures (t close to zero) the probabilities (5.1) become nearly uniform distribution. This leads to a nearly equal penalty on all the coefficients regardless of their magnitude (the diagonal elements of the ridge matrix equal $\frac{\upsilon_0 p_i^{t\star}+\upsilon_1(1-p_i^{t\star})}{\upsilon_1\upsilon_0} \approx \frac{\upsilon_0+\upsilon_1}{2\upsilon_0\upsilon 1}$, inducing a unimodal posterior.

To gauge the effectiveness of deterministic annealing on the EMVS algorithm for $\upsilon_0 = 0.005$ and $\upsilon_1 = 1000$ with $\boldsymbol{R} = \mathrm{I}_2$, we recomputed the domains of attraction for $t = 0.2$ in Figure 4(a) and for $t = 0.1$ in Figure 4(b) for our simple example. Comparison with from Figure 3(a) shows how annealing has dramatically improved the geometry of the convergence regions. By substantially increasing the region of attraction towards the best mode (number 3), annealing has here increased the chances of converging to the global solution.

Deterministic annealing for the case of the *g*-prior ($\upsilon_0 > 0$, $\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$) and the point-mass prior ($\upsilon_0 = 0$, $\upsilon_1 = 1000$, $\boldsymbol{R} = \mathrm{I}_2$) proceeds by finding the marginals of a MRF distribution, raised to the power of the inverse temperature parameter

$$\pi^t(\boldsymbol{\gamma}|\boldsymbol{\psi}^{(k)}) \propto \exp\{t\boldsymbol{\gamma}'\boldsymbol{a}+t\boldsymbol{\gamma}'B\boldsymbol{\gamma}\}. \quad (5.2)$$

The mean field approximation can be used just as before (Section 3.2.1), but with the sparsity and connectivity matrix multiplied by *t*. However, because the entries in both ***a*** and ***B*** can be prohibitively large, the tempering here will only be effective for very small *t*. This is particularly true for the continuous *g*-prior, where both ***a*** and ***B*** involve quantities depending on $1/\upsilon_0$, which can be very large.

The effectiveness of deterministic annealing for these two EM algorithms is displayed in Figure 5. Applied to our simple example, deterministic annealing has not been very effective in increasing the region of attraction toward the global mode.

6 Discussion

We have illustrated how the destabilizing inuence of multicollinearity in variable selection problems can be mitigated by introducing a spike-and-slab prior. Such priors induce posteriors which filter the likelihood information into a weighted sum of cleanly separated posterior modes corresponding to subset models. In order to locate the highest posterior mode, we presented three variants of EM algorithms for variable selection and considered the geometry of their convergence regions in multimodal landscapes. Whereas multicollinearity induces diffculties when using point-mass priors or dependent prior covariances, the EMVS algorithm for a continuous spike-and-slab prior with an independent prior covariance matrix fares superbly when coupled with deterministic annealing.

Acknowledgments

The authors would like to thank the reviewers for very helpful suggestions.

References

Bar H, Booth J, Wells M. An Empirical Bayes Approach to Variable Selection and QTL Analysis. In the Proceedings of the 25th International Workshop on Statistical Modelling, Glasgow, Scotland. 2010:63–68.

Figueiredo MA. Adaptive Sparseness for Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2003; 25:1150–1159.

George EI, McCulloch RE. Variable Selection Via Gibbs Sampling. Journal of the American Statistical Association. 1993; 88:881–889.

George EI, McCulloch RE. Approaches for Bayesian Variable Selection. Statistica Sinica. 1997; 7:339–373.

George E, Rockova V, Lesaffre E. Faster spike-and-slab variable selection with dual coordinate ascent EM. Proceedings of the 28thWorkshop on Statistical Modelling. 2013; 1:165–170.

Griffin, J.; Brown, P. Alternative Prior Distributions for Variable Selection with Very Many More Variables Than Observations. Technical report, University of Warwick, University of Kent; 2005.

Griffin JE, Brown PJ. Bayesian Hyper-LASSOS with Non-convex Penalization. Australian & New Zealand Journal of Statistics. 2012; 53:423–442.

Hayashi T, Iwata H. EM Algorithm for Bayesian Estimation of Genomic Breeding Values. BMC Genetics. 2010; 11:1–9. [PubMed: 20051104]

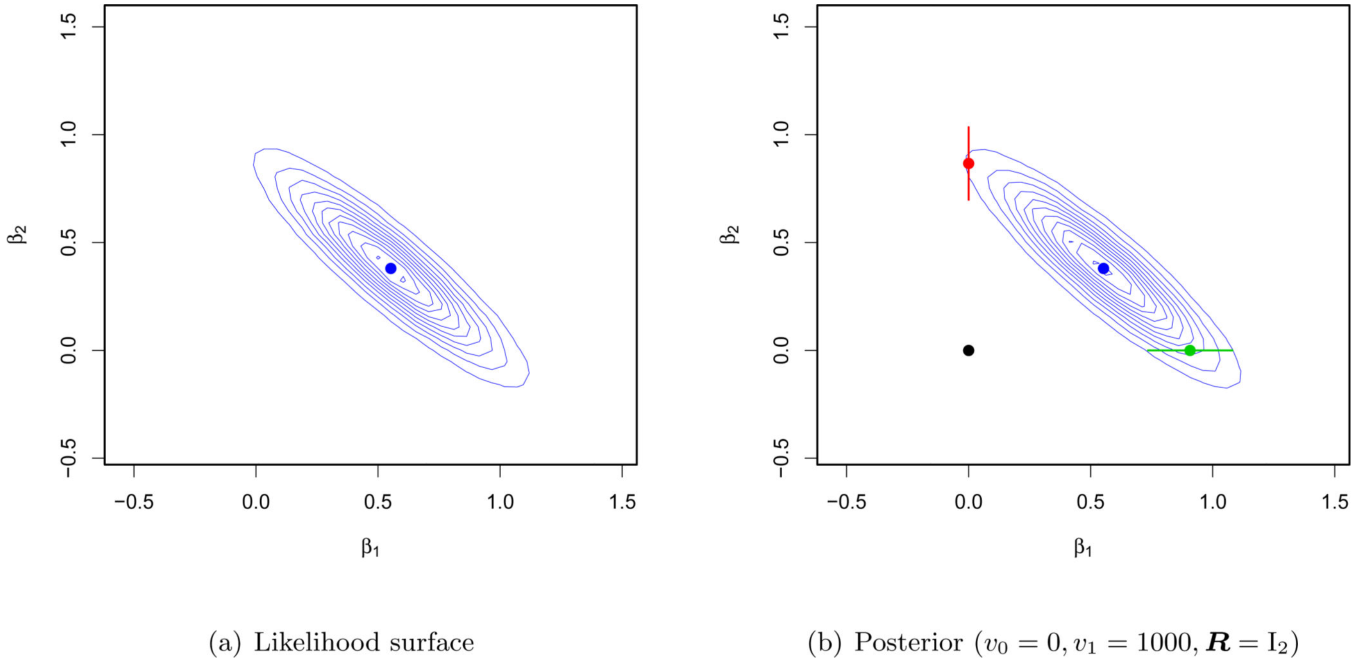
Kiiveri H. A Bayesian Approach to Variable Selection When the Number of Variables is Very Large. Institute of Mathematical Statistics Lecture Notes-Monograph Series. 2003; 40:127–143.

Rockova V, George E. EMVS: The EM approach to Bayesian Variable Selection. Journal of the American Statistical Association. 2014

Shalev-Shwartz S, Zhang T. Stochastic dual coordinate ascent methods for regularized loss minimization. Journal of Machine Learning Research. 2013; 14:567–599.

Ueda N, Nakano R. Deterministic Annealing EM Algorithm. Neural Networks. 1998; 11:271–282. [PubMed: 12662837]

Zellner A. On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. Studies in Bayesian Econometrics and Statistics. 1986

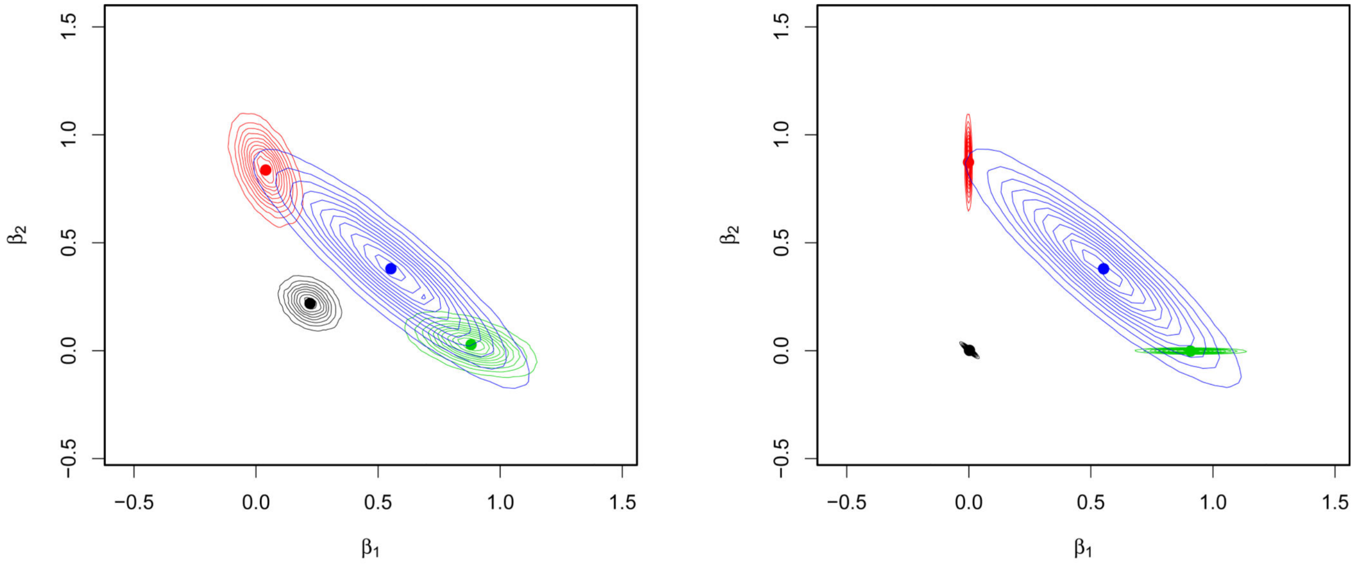


(a) Likelihood surface

(b) Posterior $(v_0 = 0, v_1 = 1000, \boldsymbol{R} = \mathrm{I}_2)$

Figure 1.

Likelihood surface and multimodal posterior landscapes under the point-mass spike-and-slab prior

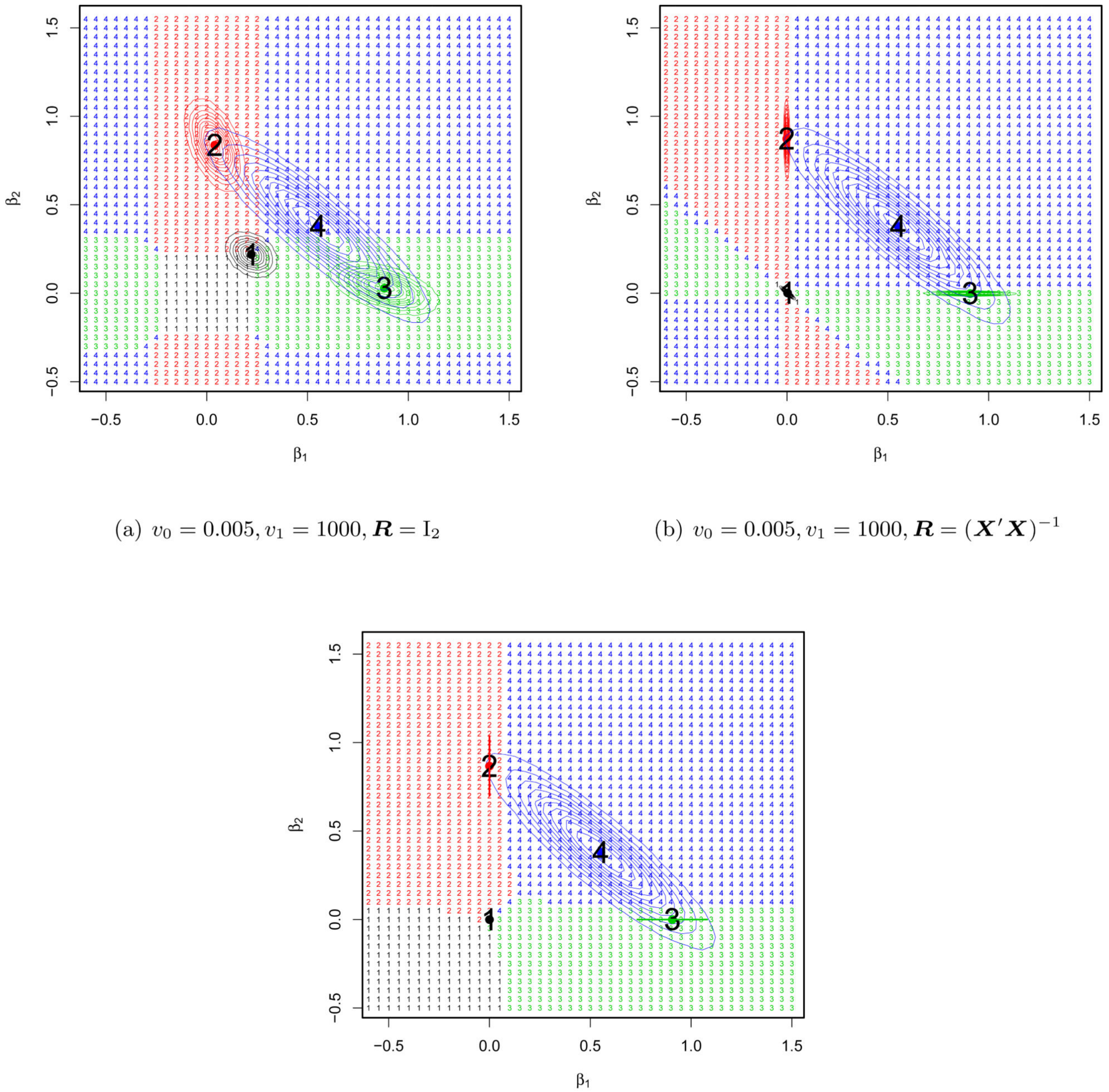


(a) Posterior $(v_0 > 0, v_1 = 1000, \boldsymbol{R} = \mathrm{I}_2)$

(b) Posterior $(v_0 > 0, v_1 = 1000, R = (\boldsymbol{X}'\boldsymbol{X})^{-1})$

Figure 2.

Posterior landscapes under continuous spike and slab priors



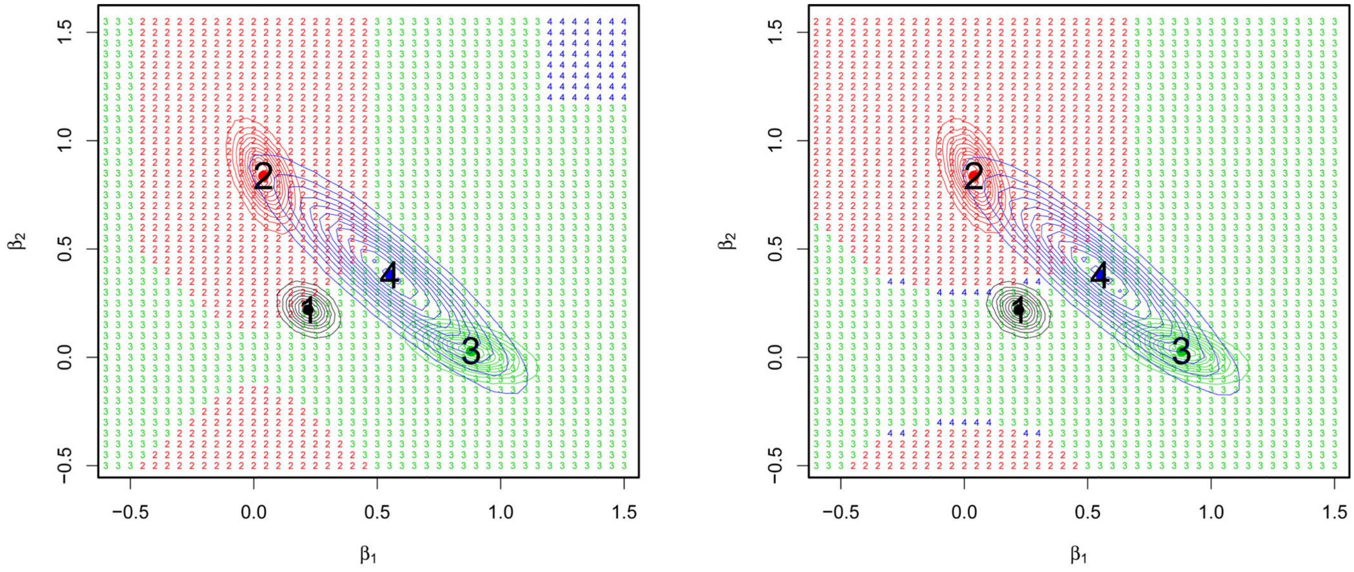
(a) $v_0 = 0.005, v_1 = 1000, \boldsymbol{R} = \mathrm{I}_2$

(b) $v_0 = 0.005, v_1 = 1000, \boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$

(c) $v_0 = 0, v_1 = 1000, \boldsymbol{R} = \mathrm{I}_2$

Figure 3.

Geometry of EM convergence regions

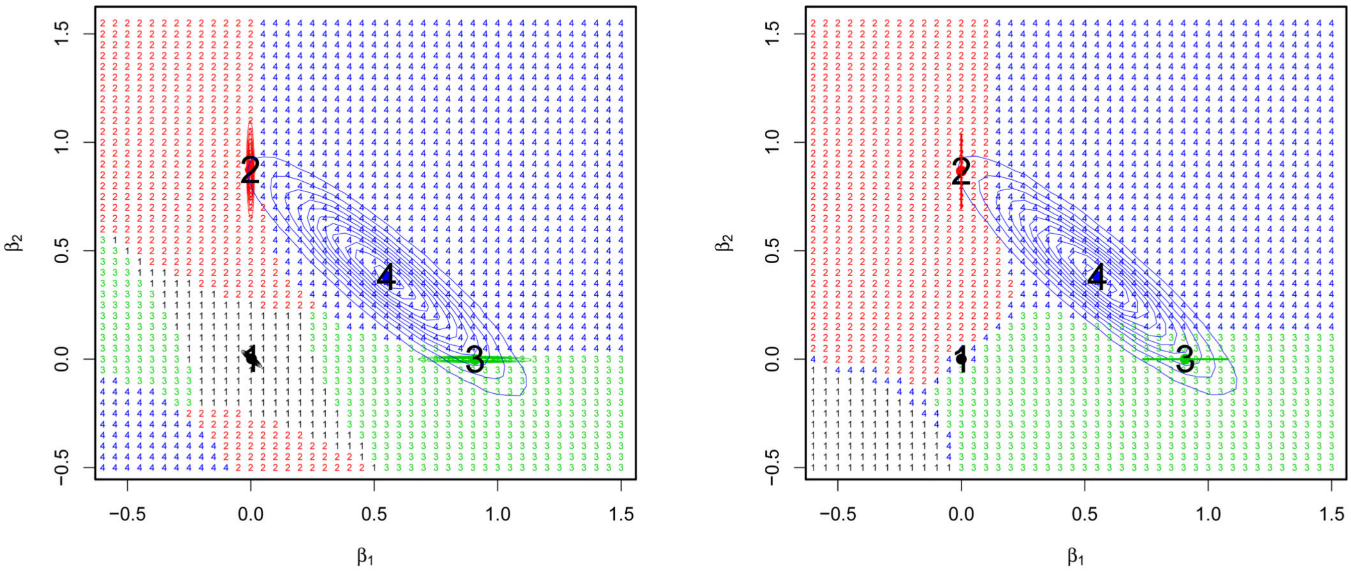


(a) $v_0 = 0.005,\ v_1 = 1000,\ \boldsymbol{R} = \mathrm{I}_2,\ t = 0.2$

(b) $v_0 = 0.005,\ v_1 = 1000,\ \boldsymbol{R} = \mathrm{I}_2,\ t = 0.1$

Figure 4.

Geometry of convergence regions using deterministic annealing.



(a) $v_0 = 0.005$, $v_1 = 1000$, $\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, $t = 0.00001$

(b) $v_0 = 0$, $v_1 = 1000$, $\boldsymbol{R} = \mathrm{I}_2$, $t = 0.1$

Figure 5.

Geometry of convergence regions using deterministic annealing.