Determinantal Priors for Variable Selection

A priori basate sul determinante per la scelta delle variabili

Veronika Ročková and Edward I. George

Abstract Determinantal point processes (DPPs) provide a probabilistic formalism for modeling repulsive distributions over subsets. Such priors encourage diversity between selected items through the introduction of a kernel matrix that determines which items are similar and therefore less likely to appear together. We investigate the usefulness of such priors in the context of spike-and-slab variable selection, where penalizing predictor collinearity may reveal more interesting models.

Abstract *I processi di punto basati sul determinante (DPP) rappresentano un formalismo probabilistico per modellare distribuzioni su sottoinsemi di tipo repulsivo. Le distribuzioni a priori basate su tali processi favoriscono la diversit tra gli elementi selezionati attraverso l'introduzione di una matrice nucleo che determina quali elementi sono simili e quindi meno probabili da apparire insieme. Si investiga l'utilit di tali a priori nel contesto della selezione di variabili con ricerca stocastica spike-and-slab dove la penalizzazione della collinearit tra predittori pu rivelare modelli pi interessanti.*

Key words: EMVS, Multicollinearity, Spike-and-Slab

1 Introduction

Suppose observations on y, an $n \times 1$ response vector, and $X = [x_1, \ldots, x_p]$, an $n \times p$ matrix of p potential standardized predictors, are related by the Gaussian linear model

$$f(y \mid \beta, \sigma) = \mathrm{N}_n(X\beta, \sigma^2 \mathrm{I}_n), \tag{1}$$

where $\beta' = (\beta_1, \ldots, \beta_p)$ is a $p \times 1$ vector of unknown regression coefficients and σ is an unknown positive scalar. (We assume throughout that y has been centered at zero to avoid the need for an intercept).

Veronika Ročková and Edward I. George
University of Pennsylvania, Philadelphia (PA), e-mail: vrockova@wharton.upenn.edu

A fundamental Bayesian approach to variable selection for this setup is obtained with a hierarchical "spike-and-slab" Gaussian mixture prior on β. Introducing a latent binary vector $\gamma = (\gamma_1, \ldots, \gamma_p)'$, $\gamma_i \in \{0,1\}$, each component of this mixture prior is defined conditionally on σ and γ by

$$\pi(\beta \mid \sigma, \gamma) = \mathrm{N}_p(0, \sigma^2 D_\gamma), \tag{2}$$

where

$$D_\gamma = \mathrm{diag}\{[(1-\gamma_1)v_0 + \gamma_1 v_1], \ldots, [(1-\gamma_p)v_0 + \gamma_p v_1]\} \tag{3}$$

for $0 \leq v_0 < v_1$, George and McCulloch (1997). Adding a relatively noninfluential prior on σ^2 such as the inverse gamma prior $\pi(\sigma^2) = \mathrm{IG}(\nu/2, \nu\lambda/2)$ with $\nu = \lambda = 1$, the mixture prior is then completed with a prior distribution $\pi(\gamma)$ over the 2^p possible values of γ.

By suitably setting v_0 small and v_1 large in (3), β_i values under $\pi(\beta \mid \sigma, \gamma)$ are more likely to be small when $\gamma_i = 0$ and more likely to be large when $\gamma_i = 1$. Thus variable selection inference can be obtained from the posterior $\pi(\gamma \mid y)$ induced by combining this prior with the data y. For example, one might select those predictors corresponding to the $\gamma_i = 1$ components of the highest posterior probability γ.

The explicit introduction of the intermediate latent vector γ in the spike-and-slab mixture prior allows for the incorporation of available prior information through the prior specification of $\pi(\gamma)$. This can be conveniently done by using hierarchical specifications of the form

$$\pi(\gamma) = \mathsf{E}_{\pi(\theta)} \pi(\gamma \mid \theta) \tag{4}$$

where θ is a (possibly vector) hyperparameter with prior $\pi(\theta)$.

In the absence of structural information about the predictors, i.e., when their inclusion is apriori exchangeable, a useful default choice for $\pi(\gamma \mid \theta)$ is the i.i.d. Bernoulli prior form

$$\pi^B(\gamma \mid \theta) = \theta^{q_\gamma}(1-\theta)^{p-q_\gamma}, \tag{5}$$

where $\theta \in [0,1]$ and $q_\gamma = \sum_i \gamma_i$. Because this $\pi(\gamma \mid \theta)$ is a function only of model size q_γ, any marginal $\pi(\gamma)$ in (4) will be of the form

$$\pi^B(\gamma) = \pi^B_{\pi(\theta)}(q_\gamma)\, \pi^B(\gamma \mid q_\gamma), \quad \pi^B(\gamma \mid q_\gamma) = \binom{p}{q_\gamma}^{-1} \tag{6}$$

where $\pi^B_{\pi(\theta)}(q_\gamma)$ is the prior on model size induced by $\pi(\theta)$, and $\pi^B(\gamma \mid q_\gamma)$ is uniform over models of size q_γ.

Of particular interest for this formulation has been the beta prior $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$, $a, b > 0$, (5) which yields model size priors of the form

$$\pi^B_{a,b}(q_\gamma) = \frac{\mathrm{Be}(a+q_\gamma, b+p-q_\gamma)}{\mathrm{Be}(a,b)} \binom{p}{q_\gamma} \tag{7}$$

where $\mathrm{Be}(\cdot,\cdot)$ is the beta function. For the choice $a = b = 1$, under which $\theta \sim U(0,1)$, this yields the uniform model size prior

$$\pi^{B}_{1,1}(q_\gamma) \equiv \frac{1}{p+1}. \tag{8}$$

An attractive alternative is to choose a small and b large in order to be more effective for targeting sparse models in high-dimensions. For example, Castillo and van der Vaart (2012) show that the choice $a = 1$ and $b = p$ yields optimal posterior concentration rates in sparse settings.

2 Determinantal Priors for $\pi(\gamma)$

The main thrust of this paper is to propose new model space priors $\pi(\gamma)$ based on the hierarchical representation (4) with the conditional form

$$\pi^{D}(\gamma \,|\, \theta) = \frac{\left|c_\theta X_\gamma{}' X_\gamma\right|}{|c_\theta X'X + I|} \propto |X_\gamma{}' X_\gamma|\; \theta^{q_\gamma}(1-\theta)^{p-q_\gamma} \tag{9}$$

where $c_\theta = \frac{\theta}{1-\theta}$ and X_γ is the $n \times q_\gamma$ matrix of predictors identified by the active elements in γ. The first expression for $\pi^{D}(\gamma \,|\, \theta)$ reveals it to be a special case of a determinantal prior, as discussed below, while the second expression reveals it to be a reweighted version of the Bernoulli prior (5) as in George (2010). Thus, this prior downweights the probability of γ for the predictor collinearity measured by the determinant $|X_\gamma{}' X_\gamma|$, which quantifies the volume of the space spanned by the selected predictors in the γth subset. Intuitively, collinear predictors are less likely to be selected under this prior, due to ill conditioning of the correlation matrix. As will be seen, the use of $\pi^{D}(\gamma \,|\, \theta)$ can provide cleaner posterior inference for variable selection in the presence of multicollinearity, when the correlation between the columns of X makes it difficult to distinguish between predictor effects.

In general, a probability measure $\pi(\gamma)$ on the 2^p subsets of a discrete set $\{1,\ldots,p\}$, indexed by the binary indices γ, is called a *determinantal point process* (DPP) if there exists a positive semidefinite matrix K, such that

$$\pi(\gamma) = \det(K_\gamma), \quad \forall \gamma, \tag{10}$$

where K_γ is the restriction of K to the entries indexed by the active elements in γ. The matrix K is referred to as a marginal kernel as its elements lead to the marginal inclusion probabilities and anti-correlations between the pairs of variables, i.e.

$$P(\gamma_i = 1) = K_{ii}; \qquad \mathsf{P}(\gamma_i = 1, \gamma_j = 1) = K_{ii}K_{jj} - K_{ij}K_{ji}$$

Given any real, symmetric, positive semidefinite $p \times p$ matrix L, a corresponding DPP can be obtained via the L-ensemble construction

$$\pi(\gamma) = \frac{\det(L_\gamma)}{\det(L+\mathrm{I})}, \tag{11}$$

where L_γ is the sub matrix of L given by the active elements in γ and I is an identity matrix. That this is a properly normalized probability distribution follows from

the fact that $\sum_\gamma \det(L_\gamma) = \det(L+\mathrm{I})$. The marginal kernel for the K-ensemble DPP representation (10) corresponding to this L-ensemble representation is obtained by letting $K = (L+\mathrm{I})^{-1}L$. The first expression for $\pi^D(\gamma|\,\theta)$ in (9) can be now seen as a special case of (11) by letting $L = c_\theta X'X$ and $L_\gamma = c_\theta X_\gamma{}'X_\gamma$.

Applying $\pi(\gamma) = \mathsf{E}_{\pi(\theta)}\pi(\gamma|\,\theta)$ to $\pi^D(\gamma|\,\theta)$ with the beta prior $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$, we obtain

$$\pi^D(\gamma) = h_{a,b}(q_\gamma)|X_\gamma{}'X_\gamma|, \tag{12}$$

where

$$h_{a,b}(q_\gamma) = \frac{1}{\mathrm{Be}(a,b)}\int_0^\infty |cX'X+I|^{-1}\frac{c^{q_\gamma+a-1}}{(1+c)^{a+b}}\,dc. \tag{13}$$

Although not in closed form, $h_{a,b}(q_\gamma)$ is an easily computable one dimensional integral.

For comparison with the exchangeable beta-binomial priors $\pi^B(\gamma)$, it is useful to reexpress (12) as

$$\pi^D(\gamma) = \pi^D_{\pi(\theta)}(q_\gamma)\,\pi^D(\gamma|\,q_\gamma), \tag{14}$$

where

$$\pi^D_{\pi(\theta)}(q_\gamma) = W(q_\gamma)\,h_{a,b}(q_\gamma), \quad \pi^D(\gamma|\,q_\gamma) = \frac{|X_\gamma{}'X_\gamma|}{W(q_\gamma)}, \quad W(q) = \sum_{q_\gamma=q}|X_\gamma{}'X_\gamma|. \tag{15}$$

Thus, to generate γ from $\pi^D(\gamma)$ one can proceed by first generating the model size $q_\gamma \in \{0,\ldots,p\}$from $\pi^D_{\pi(\theta)}(q_\gamma)$, and then generating γ conditionally from $\pi^D(\gamma|\,q_\gamma)$. Note that the model size prior $\pi^D_{\pi(\theta)}(q_\gamma)$ may be very different from the beta-binomial prior $\pi^B_{\pi(\theta)}(q_\gamma)$. For example, it is not uniform when $a=b=1$. Therefore, one might instead prefer, as is done in Section 4 below, to consider the alternative obtained by substituting a prior such as $\pi^B_{\pi(\theta)}(q_\gamma)$ for the first stage draw of q_γ, but still use $\pi^D(\gamma|\,q_\gamma)$ for the second stage draw of γ to penalize collinearity.

Lastly, note that the computation of the normalizing constant $W(q)$ can be obtained as a solution to Newton's recursive identities for elementary symmetric polynomials (Kulesza and Taskar 2013). This is better seen from the relation

$$\sum_{q_\gamma=q}|X_\gamma{}'X_\gamma| = e_k(\lambda) := \sum_{q_\gamma=q}\prod_{i=1}^p \gamma_i\lambda_i, \tag{16}$$

where $e_q(\lambda)$ is the qth elementary symmetric polynomial evaluated at $\lambda = \{\lambda_1,\ldots,\lambda_p\}$, the spectrum of $X'X$. Defining $p_q(\lambda) = \sum_{i=1}^p \lambda_i^q$, the qth power sum of the spectrum, we can obtain normalizing constants $e_1(\lambda),\ldots,e_p(\lambda)$ as solutions to the recursive system of equations

$$q\,e_q(\lambda) = p_q(\lambda) + \sum_{j=1}^{q-1}(-1)^{j-1}e_{q-j}(\lambda)p_j(\lambda). \tag{17}$$

3 Implementing Determinantal Priors with EMVS

EMVS (Rockova and George 2014) is a fast deterministic approach to identifying sparse high posterior models for Bayesian variable selection under spike-and-slab priors. In large high-dimensional problems where exact full posterior inference must be sacrificed for computational feasibility, deployments of EMVS can be used to find subsets of the highest posterior modes. We here describe a variant of the EMVS procedure which incorporates the determinantal prior $\pi^D(\gamma \mid \theta)$ in (9) to penalize predictor collinearity in variable selection.

At the heart of the EMVS procedure is a fast closed form EM algorithm, which iteratively updates the conditional expectations $\mathsf{E}[\gamma_i \mid \psi^{(k)}]$, where here $\psi^{(k)} = (\beta^{(k)}, \sigma^{(k)}, \theta^{(k)})$ denotes the set of parameter updates at the k^{th} iteration. The determinantal prior induces dependence between inclusion probabilities so that conditional expectations cannot be obtained by trivially thresholding univariate directions

With the determinantal prior $\pi^D(\gamma \mid \theta)$, the joint conditional posterior distribution is

$$\pi(\gamma \mid \psi) \propto \exp\left(-\frac{\beta D_\gamma \beta}{2\sigma^2}\right) |D_\gamma|^{1/2} |c_\theta X_\gamma' X_\gamma|, \tag{18}$$

where $D_\gamma = \text{diag}\{\gamma_i/v_1 + (1-\gamma_i)/v_0\}_{i=1}^p$. We can then write

$$\pi(\gamma \mid \psi) \propto \exp\left[-\frac{1}{2\sigma^2}\left(\frac{1}{v_1} - \frac{1}{v_0}\right)(\beta \circ \beta)'\gamma\right] |D_\gamma|^{1/2} c_\theta^{q_\gamma} |X_\gamma' X_\gamma|, \tag{19}$$

where $\circ$ denotes the Hadamard product. The determinant $|D_\gamma|$ can be written as

$$|D_\gamma| = \exp\left\{\left[\log\left(\frac{1}{v_1}\right) - \log\left(\frac{1}{v_0}\right)\right]\gamma' 1 + p\log\left(\frac{1}{v_0}\right)\right\},$$

so that the joint distribution in (19) can be expressed as

$$\pi(\gamma \mid \psi) \propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\left(\frac{1}{v_1} - \frac{1}{v_0}\right)(\beta \circ \beta) - \log\left(\frac{v_0}{v_1}\right) 1 - 2\log(c_\theta) 1\right]' \gamma\right\} |X_\gamma' X_\gamma|.$$

Defining the $p \times p$ diagonal matrix

$$A_\psi = \text{diag}\left\{\exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\left(\frac{1}{v_1} - \frac{1}{v_0}\right)\beta_i^2 - \log\left(\frac{v_0}{v_1}\right)\right] - 2\log(c_\theta)\right\}\right\}_{i=1}^p, \tag{20}$$

the exponential term above can be regarded as the determinant of $A_{\gamma,\psi}$, the $q_\gamma \times q_\gamma$ diagonal submatrix of A_ψ whose diagonal elements are correspond to the nonzero elements of γ.

It now follows that the determinantal prior is conjugate in the sense of yielding the updated determinantal form

$$\pi(\gamma \mid \psi) \propto |A_{\gamma,\psi} X_\gamma' X_\gamma|. \tag{21}$$

The marginal quantities from this distribution can be obtained by taking the diagonal of a matrix $K_\theta = (A_\psi X'X + \mathrm{I}_p)^{-1} A_\psi X'X$, namely

$$\mathsf{P}(\gamma_i = 1 \mid \psi) = [K_\psi]_{ii}. \tag{22}$$

4 Mitigating Multicolinearity with Determinantal Priors

In order to demonstrate the redundancy correction of the determinantal model prior we revisit the collinear example of George and McCulloch (1997) with $p = 15$ predictors. The collinearity induces severe posterior multimodality, as displayed in the plot of 32768 posterior model probabilities in Figure 1. Models whose design matrix is "ill-conditioned", i.e. with smallest eigenvalue $\lambda_{min}(\gamma)$ of the gram matrix L_γ below 0.1, are designated in red. The deteminantal prior penalizes such models and puts more posterior weight on diverse covariate combinations, effectively reducing both posterior multi-modality and entropy.

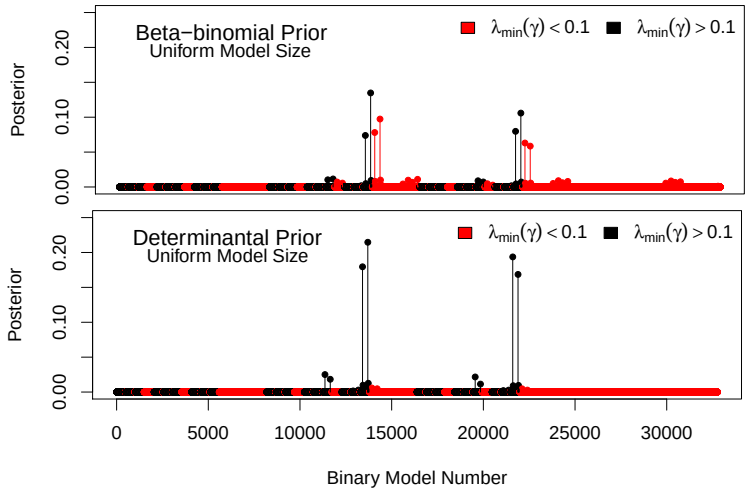

Fig. 1 Posteriors arising from beta-binomial and determinantal priors (uniform on the model size).

References

1. Castillo, I., van der Vaart, A.: Needles and Straw in a Haystack: Posterior Concentration for Possibly Sparse Sequences, Annals of Statistics, 40, 2069-2101. (2012)
2. George, E.I.: Dilution priors: Compensating for model space redundancy. In Borrowing Strength, IMS Collections, Vol. 6, 158-165. (2010)
3. Kulesza, A., Taskar, B.: Determinantal point processes for machine learning. ArXiv: 1207.6083 (2013)
4. Rockova, V., George, E.I.: EMVS: The EM Approach to Bayesian Variable Selection. Journal of the American Statistical Association (to appear) (2014)