Statistica Sinica **26** (2016), 97-117
doi:http://dx.doi.org/10.5705/ss.202014.0161

DETERMINANTAL POINT PROCESS PRIORS FOR BAYESIAN VARIABLE SELECTION IN LINEAR REGRESSION

Mutsuki Kojima[1] and Fumiyasu Komaki[2,3]

[1] *Mitsui Sumitomo Insurance Co., Ltd.*
[2] *The University of Tokyo and* [3] *RIKEN Brain Science Institute*

Abstract: We propose discrete determinantal point processes (DPPs) for priors on the model parameter in Bayesian variable selection. By our variable selection method, collinear predictors are less likely to be simultaneously selected due to the repulsion property of discrete DPPs. Three types of DPP priors are proposed. Our method is an empirical Bayes approach, so hyperparameters are estimated by maximizing the marginal likelihood. We show the efficiency of the proposed priors through numerical experiments and applications to collinear datasets.

Key words and phrases: Collinearity, empirical Bayes, g-prior.

1. Introduction

We consider Bayesian variable selection in linear regression. Suppose we have n observations on a dependent variable $\boldsymbol{y}$ ($n \times 1$ matrix) and p predictor variables $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ ($n \times p$ matrix), for which the normal linear model holds:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ ($n \times 1$ matrix) and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ ($p \times 1$ matrix). Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top \in \{0, 1\}^p$ be a model parameter: $\gamma_i = 1$ indicates β_i is nonzero and $\gamma_i = 0$ indicates $\beta_i = 0$. In Bayesian variable selection, we consider 2^p possible submodels of (1.1). Submodels are denoted by $M_{\boldsymbol{\gamma}}$. Let $\boldsymbol{X}_{\boldsymbol{\gamma}}$ be the $n \times |\boldsymbol{\gamma}|$ design matrix consisting of these columns of $\boldsymbol{X}$ that correspond to the predictors with $\gamma_i = 1$. Here, $|\boldsymbol{\gamma}|$ is the number of nonzero elements of $\boldsymbol{\gamma}$. Under submodel $M_{\boldsymbol{\gamma}}$, $\boldsymbol{y}$ follows

$$M_\gamma : \quad \boldsymbol{y} = \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the $|\boldsymbol{\gamma}|$-dimensional vector of nonzero regression coefficients of $\boldsymbol{\beta}$ with $\gamma_i = 1$. Bayesian variable selection is to identify nonzero components of $\boldsymbol{\beta}$ assigning priors to the parameters. We select the best model that attains the maximum of the posterior probability $p(\boldsymbol{\gamma}|\boldsymbol{y})$.

The normal linear regression model is simple and useful, but the collinearity problem often arises when we apply it to data. Highly correlated predictors can make $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ numerically unstable and the OLS, $\hat{\beta}_{\mathrm{OLS}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$, unreliable. Few Bayesian variable selection methods that consider the correlations have been proposed (Yuan and Lin (2005); Krishna, Bondell and Ghosh (2009)). We propose discrete determinantal point processes (DPPs) for prior distributions on $\boldsymbol{\gamma}$ that discourage the inclusion of groups of collinear predictors.

DPPs have been studied since Macchi (1975) first identified them as a class of point processes. Recently, Borodin and Rains (2005) introduced discrete DPPs that have been applied to machine learning problems by Kulesza and Taskar (2012). Discrete DPPs are elegant probabilistic models of repulsion, and Kulesza and Taskar (2012) considered repulsion as diversity of items. For example, in a document summarization task, modeling the task with discrete DPPs is appropriate because a summary of the document requires diversity.

Selected predictors in variable selection should be diverse in the sense that pairs of selected predictors be nearly uncorrelated. Here, discrete DPPs are efficient priors on $\boldsymbol{\gamma}$ in Bayesian variable selection, and we show that they are useful priors through numerical experiments and applications to datasets.

The remainder of this paper is organized as follows. In Section 2, the definition and examples of discrete DPPs are given. In Section 3, we first review the Bayesian variable selection method proposed by George and Foster (2000), and then propose Bayesian variable selection methods using three types of DPP priors on $\boldsymbol{\gamma}$. In Section 4, we show the results of numerical experiments and we report applications to datasets in Section 5. We conclude the paper in Section 6.

2. Discrete Determinantal Point Processes

Let Λ be $\{1,\dots,p\}$ and let $\boldsymbol{L}$ be a $p\times p$ symmetric positive definite matrix. We identify $\{0,1\}^p$ with the power set of Λ (2^Λ): for $\boldsymbol{\gamma}\in\{0,1\}^p$, $\gamma_i = 1$ indicates $i\in\boldsymbol{\gamma}$ and $\gamma_i = 0$ indicates $i\notin\boldsymbol{\gamma}$.

Definition 1. A random variable $\boldsymbol{\mathcal{X}}$ that takes values in the power set of Λ is called a *discrete determinantal point process (DPP) with kernel* $\boldsymbol{L}$, if $P(\boldsymbol{\mathcal{X}} = \boldsymbol{\gamma}) \propto \det(\boldsymbol{L}_{\boldsymbol{\gamma}})$, where $\boldsymbol{\gamma}\in\{0,1\}^p$ and $\boldsymbol{L}_{\boldsymbol{\gamma}}$ is the $|\boldsymbol{\gamma}|\times|\boldsymbol{\gamma}|$ matrix whose elements are L_{ij} $(i,j\in\{k:\gamma_k = 1\})$. For empty set $\emptyset$, we define $\det(\boldsymbol{L}_\emptyset) = 1$.

The normalization constant is provided by Kulesza and Taskar (2012).

Proposition 1. $\sum_{\boldsymbol{\gamma}\in\{0,1\}^p}\det(\boldsymbol{L}_{\boldsymbol{\gamma}}) = \det(\boldsymbol{L}+\boldsymbol{I}_p)$, where $\boldsymbol{I}_p$ is the $p\times p$ identity matrix, and the sum is taken over all subsets of Λ.

For more detailed properties of discrete DPPs, see Kulesza and Taskar (2012). See Hough et al. (2009) for general determinantal point processes.

Table 1. The distribution of discrete DPPs with kernel $\boldsymbol{L}$ in Example 2.

Subsets	Probabilities	Subsets	Probabilities
$\emptyset$	0.157	$\{1,2\}$	0.030
$\{1\}$	0.157	$\{1,3\}$	0.157
$\{2\}$	0.157	$\{2,3\}$	0.157
$\{3\}$	0.157	$\{1,2,3\}$	0.030

We briefly explain the repulsion property of DPPs, a key of our proposal. Let $\boldsymbol{F}$ be a $p \times q$ ($p < q$) matrix, and denote the rows of $\boldsymbol{F}$ by $\boldsymbol{f}_i$ ($i = 1, 2, \ldots, p$). Assume that $\boldsymbol{L} = \boldsymbol{F}\boldsymbol{F}^\top$. If $\boldsymbol{\mathcal{X}}$ follows discrete DPPs with kernel $\boldsymbol{L}$, then

$$P(\boldsymbol{\mathcal{X}} = \boldsymbol{\gamma}) \propto (\mathrm{vol}(\{\boldsymbol{f}_i\}_{i\in\gamma}))^2, \tag{2.1}$$

where $\mathrm{vol}(\{\boldsymbol{f}_i\}_{i\in\gamma})$ means the $|\boldsymbol{\gamma}|$-dimensional volume of the parallelepiped spanned by the rows of $\{\boldsymbol{f}_i\}_{i\in\gamma}$. When considering $\boldsymbol{f}_i$ as the feature vector of item i, $\mathrm{vol}(\{\boldsymbol{f}_i\}_{i\in\gamma})$ is small if there exist similar items in $\boldsymbol{\gamma}$. Accordingly DPPs favor repulsion, and thus diversity.

The determinant of a symmetric positive definite matrix $\boldsymbol{L}$ is the volume of the parallelepiped spanned by the rows of $\boldsymbol{L}$. From this viewpoint, the off-diagonal elements of $\boldsymbol{L}$ play a key role because they affect the determinant. In fact,

$$P(\boldsymbol{\mathcal{X}} = \{i, j\}) \propto P(\boldsymbol{\mathcal{X}} = \{i\})P(\boldsymbol{\mathcal{X}} = \{j\}) - \left(\frac{L_{ij}}{\det(\boldsymbol{L} + \boldsymbol{I}_p)}\right)^2. \tag{2.2}$$

Therefore, if off-diagonal elements are nonzero, the probabilities of including the corresponding items are small.

We provide two examples of discrete DPPs.

Example 1. Let $\Lambda = \{1, \ldots, p\}$ and let $\boldsymbol{L}$ be a diagonal matrix with $L_{ii} = w/(1-w)$, $i = 1, \ldots, p$, where $w \in (0, 1)$. Suppose $\boldsymbol{\mathcal{X}}$ follows the discrete DPPs with kernel $\boldsymbol{L}$. By Proposition 1,

$$P(\boldsymbol{\mathcal{X}} = \boldsymbol{\gamma}) = \frac{\det(\boldsymbol{L}_{\boldsymbol{\gamma}})}{\det(\boldsymbol{L} + \boldsymbol{I}_p)} = w^{|\boldsymbol{\gamma}|}(1-w)^{p-|\boldsymbol{\gamma}|}.$$

In this setting, the distribution of $\boldsymbol{\mathcal{X}}$ is the Bernoulli distribution with success probability w.

Example 2. Let $\Lambda = \{1, 2, 3\}$ and let

$$L = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Suppose $\boldsymbol{\mathcal{X}}$ follows the discrete DPPs with kernel $\boldsymbol{L}$. The distribution of $\boldsymbol{\mathcal{X}}$ is given in Table 1. One finds there that $\boldsymbol{\mathcal{X}}$ equals $\{1, 2\}$ or $\{1, 2, 3\}$ is less likely to occur because of the off-diagonal element $L_{12} = 0.9$.

3. Bayesian Variable Selection Methods

3.1. Bayesian variable selection method proposed by George and Foster

George and Foster (2000) proposed the following Bayesian variable selection. Zellner's g-prior (Zellner (1986)) is assigned to nonzero regression coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ under submodel $M_{\boldsymbol{\gamma}}$:

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|g) \sim \mathcal{N}(0, g\sigma^2(\boldsymbol{X}_{\boldsymbol{\gamma}}^\top \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}), \quad g > 0, \tag{3.1}$$

where g is the hyperparameter. For the prior distribution on model parameter $\boldsymbol{\gamma}$, the Bernoulli distribution with success parameter $w \in (0,1)$ is used:

$$p(\boldsymbol{\gamma}|w) = w^{|\boldsymbol{\gamma}|}(1-w)^{p-|\boldsymbol{\gamma}|}.$$

The best model is that which maximizes the posterior probability.

$$\begin{aligned}\hat{\boldsymbol{\gamma}} &= \underset{\boldsymbol{\gamma}}{\operatorname{argmax}}\ p(\boldsymbol{\gamma}|\boldsymbol{y}, g, w)\\ &= \underset{\boldsymbol{\gamma}}{\operatorname{argmax}}\ \exp\left(\frac{g}{2(1+g)}\Big(\frac{\mathrm{ss}_{\boldsymbol{\gamma}}}{\sigma^2} - F(g,w)|\boldsymbol{\gamma}|\Big)\right)\\ &= \underset{\boldsymbol{\gamma}}{\operatorname{argmax}}\ \Big(\frac{\mathrm{ss}_{\boldsymbol{\gamma}}}{\sigma^2} - F(g,w)|\boldsymbol{\gamma}|\Big), \end{aligned} \tag{3.2}$$

where

$$\mathrm{ss}_{\boldsymbol{\gamma}} = \boldsymbol{y}^\top \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{X}_{\boldsymbol{\gamma}}^\top \boldsymbol{y}, \quad F(g,w) = \frac{1+g}{g}\left(2\log\frac{1-w}{w} + \log(1+g)\right).$$

If σ^2 is known and the hyperparameters are appropriately calibrated, this Bayesian variable selection is that of selecting the best model by the typical penalized sum of squares criteria, such as AIC (Akaike (1973)), BIC (Schwarz (1978)), or RIC (Foster and George (1994)). For example, if we set g and w such that $F(g,w) = 2$, then the highest posterior model maximizes (3.2), $\mathrm{ss}_{\boldsymbol{\gamma}}/\sigma^2 - 2|\gamma|$. In this setting, the highest posterior model exactly corresponds to the best model with the lowest AIC.

For hyperparameters g and w, George and Foster (2000) used type-II maximum likelihood estimators $\hat{g}$ and $\hat{w}$, given by

$$\begin{aligned}(\hat{g}, \hat{w}) &= \underset{g,w}{\operatorname{argmax}}\ p(\boldsymbol{y}|g,w)\\ &= \underset{g,w}{\operatorname{argmax}} \sum_{\boldsymbol{\gamma}\in\{0,1\}^p} p(\boldsymbol{\gamma}|w)\int p(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}})p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|g)\mathrm{d}\boldsymbol{\beta}_{\boldsymbol{\gamma}}.\end{aligned}$$

Since the g-prior is normal, the marginal distribution can be calculated in the closed-form:

$$\int p(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}})p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|g)\mathrm{d}\boldsymbol{\beta}_{\boldsymbol{\gamma}} = \frac{(1+g)^{-|\boldsymbol{\gamma}|/2}}{(2\pi)^{n/2}(\sigma^2)^{n/2}}\exp\left(\frac{g}{1+g}\frac{\mathrm{ss}_{\boldsymbol{\gamma}}}{2\sigma^2} - \frac{\boldsymbol{y}^\top\boldsymbol{y}}{2\sigma^2}\right).$$

Therefore, the marginal likelihood for g and w is

$$p(\boldsymbol{y}|g,w) \propto \sum_{\boldsymbol{\gamma}\in\{0,1\}^p} \frac{w^{|\boldsymbol{\gamma}|}(1-w)^{p-|\boldsymbol{\gamma}|}}{(\sigma^2)^{n/2}(1+g)^{|\boldsymbol{\gamma}|/2}} \exp\left(\frac{g}{1+g}\frac{\mathrm{ss}_{\boldsymbol{\gamma}}}{2\sigma^2} - \frac{\boldsymbol{y}^\top\boldsymbol{y}}{2\sigma^2}\right).$$

We henthforth refer to this model as EB (empirical Bayes).

3.2. DPP priors and proposed methods

Let x_{ij} be the (i,j) element of design matrix $\boldsymbol{X}$ and let $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1,\ldots,\tilde{\boldsymbol{x}}_p)$ ($n\times p$ matrix) be the standardized matrix defined by

$$\tilde{x}_{ij} = \frac{x_{ij}-m_j}{s_j}, \quad m_j \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n x_{ij}, \quad s_j \overset{\text{def}}{=} \sqrt{\sum_{i=1}^n (x_{ij}-m_j)^2}.$$

We denote the correlation matrix ($p\times p$ matrix) of $\boldsymbol{X}$ by $\boldsymbol{R} = \tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}$. The first proposal for prior distribution $p(\boldsymbol{\gamma}|w)$ is

$$p(\boldsymbol{\gamma}|w) \propto \det(w\boldsymbol{R}_{\boldsymbol{\gamma}}) = w^{|\boldsymbol{\gamma}|}\det(\boldsymbol{R}_{\boldsymbol{\gamma}}), \quad w>0. \tag{3.3}$$

By Proposition 1,

$$p(\boldsymbol{\gamma}|w) = \frac{\det(w\boldsymbol{R}_{\boldsymbol{\gamma}})}{\det(w\boldsymbol{R}+\boldsymbol{I}_p)}.$$

We call this the *DPP prior*. When we put it on the model parameter $\boldsymbol{\gamma}$, we can select diverse predictors. The hyperparameter w controls the expected proportion of nonzero regression coefficients; if $w>1$ then larger subsets are preferable, but otherwise smaller subsets are preferable.

The DPP prior is a generalization of the Bernoulli distribution, used for $p(\boldsymbol{\gamma}|w)$ in the method proposed by George and Foster (2000). We propose two types of priors to bridge the Bernoulli and DPP priors:

$$p(\boldsymbol{\gamma}|w,\theta) \propto \det(w(\theta\boldsymbol{R}_{\boldsymbol{\gamma}} + (1-\theta)\boldsymbol{I}_{\boldsymbol{\gamma}})), \quad w>0, \quad \theta\in[0,1], \tag{3.4}$$

$$p(\boldsymbol{\gamma}|w,\alpha) \propto \det(w(\boldsymbol{R}^\alpha)_{\boldsymbol{\gamma}}), \quad w>0, \quad \alpha\geq 0, \tag{3.5}$$

where $\boldsymbol{I}_{\boldsymbol{\gamma}}$ is the $|\boldsymbol{\gamma}|\times|\boldsymbol{\gamma}|$ identity matrix and $\boldsymbol{R}^\alpha$ is the α power of $\boldsymbol{R}$. We call (3.4) the *linear mixture DPP prior* (referred to as LDPP) and (3.5) the *geometric mixture DPP prior* (referred to as GDPP). LDPP is the DPP prior when $\theta=1$ and is the Bernoulli when $\theta=0$. Similarly, GDPP is the DPP prior when $\alpha=1$ and the Bernoulli when $\alpha=0$.

Properties of DPP priors

We first study the collinearity penalty. Suppose that predictors in S_1 ($|S_1| = q-s$) are mutually uncorrelated and predictors in S_0 ($|S_0|=s$) are correlated with a particular predictor in S_1. Let $\boldsymbol{R}$ be the correlation matrix of $\boldsymbol{X}$. We divide $\boldsymbol{R}_{S_0 \cup S_1}$ as

$$\boldsymbol{R}_{S_0 \cup S_1} = \begin{pmatrix} \boldsymbol{R}_{00} & \boldsymbol{R}_{01} \\ \boldsymbol{R}_{01}^\top & \boldsymbol{R}_{11} \end{pmatrix},$$

where $\boldsymbol{R}_{00}$, $\boldsymbol{R}_{01}$, and $\boldsymbol{R}_{11}$ are the correlation matrices of predictors in S_0 and S_1. Here $\boldsymbol{R}_{11} = \boldsymbol{I}_{q-s}$. Then, for determinants of block matrices, we can write

$$\begin{aligned} \det(\boldsymbol{R}_{S_0 \cup S_1}) &= \det(\boldsymbol{R}_{11}) \det(\boldsymbol{R}_{00} - \boldsymbol{R}_{01}\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{01}^\top) \\ &= \det(\boldsymbol{R}_{S_1}) \det(\boldsymbol{R}_{00} - \boldsymbol{R}_{01}\boldsymbol{R}_{01}^\top). \end{aligned}$$

Thus, the ratio of the prior probability of $S_0 \cup S_1$ over the probability of S_1 is

$$\frac{\det(\boldsymbol{R}_{S_0 \cup S_1})}{\det(\boldsymbol{R}_{S_1})} = \det(\boldsymbol{R}_{00} - \boldsymbol{R}_{01}\boldsymbol{R}_{01}^\top) \leq \det(\boldsymbol{R}_{S_0}).$$

Since predictors in S_0 are correlated with a particular predictor in S_1, the upper bound of the ratio, $\det(\boldsymbol{R}_{S_0})$, is quite small. Therefore, we see that the DPP prior discourages the inclusion of groups of collinear predictors.

We next investigate the induced priors on the model size. By a result of Kulesza and Taskar (2012), we have

$$p(|\boldsymbol{\gamma}| = k) = \frac{e_k(\lambda_1, \ldots, \lambda_p)}{\det(\boldsymbol{L} + \boldsymbol{I}_p)}, \quad k = 0, 1, \ldots, p, \tag{3.6}$$

where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\boldsymbol{L}$ and e_k is the k-th elementary symmetric polynomial:

$$e_k(t_1, \ldots, t_p) \overset{\text{def}}{=} \sum_{\substack{\boldsymbol{\gamma} \subset \{1, \ldots, p\}, \\ |\boldsymbol{\gamma}| = k}} \prod_{l \in \boldsymbol{\gamma}} t_l.$$

Generally speaking, when using DPP priors, large submodels are less likely to be preferable. If the collinearity is severe then, for large k, the prior probability that the model size is k is small because most k-products of the eigenvalues are quite small (see (3.6)). We illustrate the property with a simple example. Take $p = 6$ and

$$\begin{aligned} \boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \boldsymbol{x}_3,\ \boldsymbol{\varepsilon}_4,\ \boldsymbol{\varepsilon}_5,\ \boldsymbol{\varepsilon}_6 &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_{20}), \\ \boldsymbol{x}_4 &= \boldsymbol{x}_1 + \boldsymbol{x}_2 + 0.1 \times \boldsymbol{\varepsilon}_4, \\ \boldsymbol{x}_5 &= \boldsymbol{x}_1 + \boldsymbol{x}_3 + 0.1 \times \boldsymbol{\varepsilon}_5, \\ \boldsymbol{x}_6 &= \boldsymbol{x}_2 + \boldsymbol{x}_3 + 0.1 \times \boldsymbol{\varepsilon}_6. \end{aligned}$$

Here, the correlation matrix, $\boldsymbol{R}$, has three small eigenvalues. Consider the DPP prior with $\boldsymbol{R}$ and $w = 1$. Then the induced prior on the model size is as in Figure

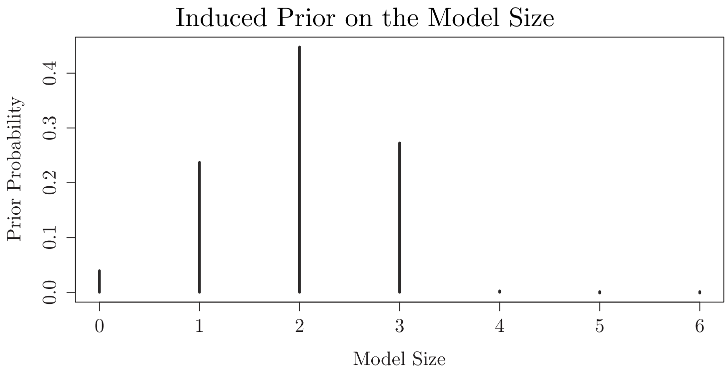

Figure 1. Induced prior on the model size with $w = 1$.

1. Since the correlation matrix has three small eigenvalues, the prior probabilities that the model size exceeds 3 are quite small. Generally if $\boldsymbol{R}$ is almost of rank q, the induced prior assigns small probability to a model whose size is larger than q.

The θ in LDPP controls the collinearity penalty of the DPP prior and therefore the penalty for the model size. This is verified numerically in Figure 2, where $\boldsymbol{R}$ is the same as above and the kernel matrix is $\theta\boldsymbol{R} + (1-\theta)\boldsymbol{I}_p$. From Figure 2, the penalty for the model size decreases as θ gets smaller. In fact, the eigenvalues of $\theta\boldsymbol{R} + (1-\theta)\boldsymbol{I}_p$ are $\theta\lambda_i + (1-\theta)$ and for $\lambda_i < 1$, they increase as θ decreases. Hyperparameter α in GDPP plays a similar role.

Proposed variable selection methods

Our methods proceed as follows. We put the proposed DPP priors (DPP, LDPP, or GDPP) on model parameter $\boldsymbol{\gamma}$ and the g-prior on $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. Hyperparameters are estimated by maximizing the marginal likelihood

$$p(\boldsymbol{y}|\boldsymbol{\xi}) \propto \sum_{\boldsymbol{\gamma}\in\{0,1\}^p} p(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\xi})p(\boldsymbol{\gamma}|\boldsymbol{\xi}),$$

where $\boldsymbol{\xi}$ denotes the hyperparameters to be estimated. Here, they are g, σ^2 (if unknown), w, θ (if using LDPP), and α (if using GDPP). The selected model maximizes the posterior probability $p(\boldsymbol{\gamma}|\boldsymbol{y})$, i.e., the maximum a posteriori (MAP) model.

Although our interest is in variable selection methods, we can estimate the regression coefficients after selecting the best model. The estimator $\hat{\boldsymbol{\beta}}$ is constructed after estimating $\hat{g}$ and selecting the best model $M_{\hat{\boldsymbol{\gamma}}}$:

$$\hat{\boldsymbol{\beta}} = \mathrm{E}[\boldsymbol{\beta}|\hat{\boldsymbol{\gamma}},\hat{g}] = \frac{\hat{g}}{1+\hat{g}}(\boldsymbol{X}_{\hat{\boldsymbol{\gamma}}}^\top\boldsymbol{X}_{\hat{\boldsymbol{\gamma}}})^{-1}\boldsymbol{X}_{\hat{\boldsymbol{\gamma}}}^\top\boldsymbol{y}. \tag{3.7}$$

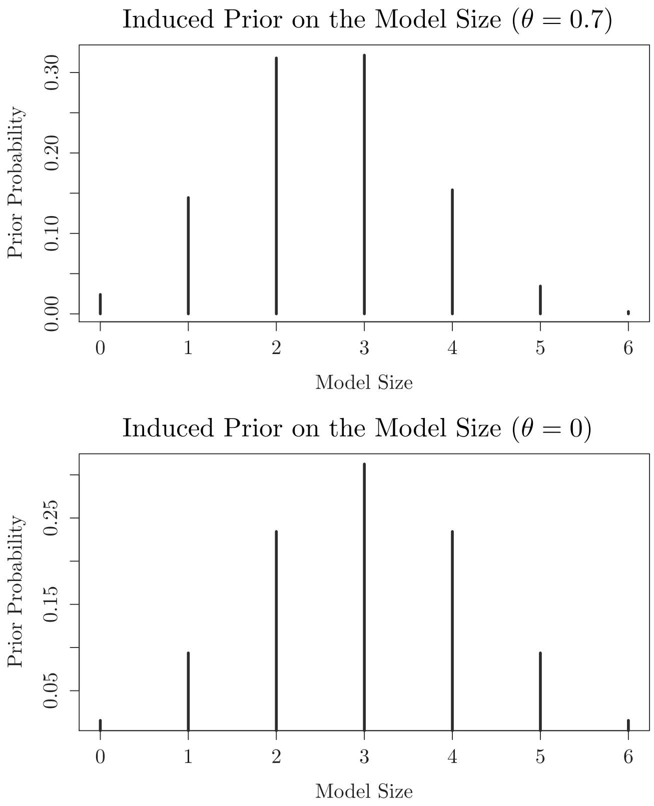

Figure 2. Induced prior on the model size with $\theta = 0.7$ (the upper panel) and $\theta = 0$ (the lower panel).

The representation of the estimator is the same whether the best model $\hat{\boldsymbol{\gamma}}$ is selected by EB or by our methods.

Remark 1. Although we select the MAP model, one could choose the median probability model or apply ad-hoc hard thresholding methods. When dealing with collinear data, the correlations among the predictors should be taken into consideration and it is preferable to utilize the joint, not marginal, posterior distribution of $\boldsymbol{\beta}$. Recently, Hahn and Carvalho (2015) provided posterior summary selection methods that explicitly account for the collinearity. In addition, for high-dimensional regression, Bondell and Reich (2012) proposed a variable selection method, using posterior credible regions of $\boldsymbol{\beta}$, that does not need MCMC iterations.

4. Numerical Experiments

In this section, we evaluate the risk of estimated regression coefficients as the sample size increases through simulations. The following settings were considered. First we sampled a 400×6 design matrix $\boldsymbol{X}^*$ with columns $\boldsymbol{x}_i^*$ ($i = 1, 2, \dots, 6$):

$$\begin{aligned}
&\boldsymbol{x}_1^*,\ \boldsymbol{x}_2^*,\ \boldsymbol{x}_3^*,\ \boldsymbol{\varepsilon}_4^*,\ \boldsymbol{\varepsilon}_5^*,\ \boldsymbol{\varepsilon}_6^* \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_{400}),\\
&\boldsymbol{x}_4^* = \boldsymbol{x}_1^* + \boldsymbol{x}_2^* + 0.1 \times \boldsymbol{\varepsilon}_4^*,\\
&\boldsymbol{x}_5^* = \boldsymbol{x}_1^* + \boldsymbol{x}_3^* + 0.1 \times \boldsymbol{\varepsilon}_5^*,\\
&\boldsymbol{x}_6^* = \boldsymbol{x}_2^* + \boldsymbol{x}_3^* + 0.1 \times \boldsymbol{\varepsilon}_6^*.
\end{aligned}$$

Let $\boldsymbol{X}_k^*$ ($k = 1, 2, \dots, 20$) be the $(20k) \times 6$ submatrix of $\boldsymbol{X}^*$ whose rows correspond to the first $20k$ rows of $\boldsymbol{X}^*$, and let $\boldsymbol{\beta}^* = (1, -1, 0, 0, 0, 0)^\top$. For each k, we simulated $\boldsymbol{y}_k^{(l)}$ ($(20k) \times 1$ matrix), $l = 1, \dots,$ 10,000, following (1.1) with $\boldsymbol{X} = \boldsymbol{X}_k^*$, $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, and $\sigma^2 = 0.9^2$. From each $\boldsymbol{y}_k^{(l)}$ and $\boldsymbol{X}_k^*$, the estimator $\hat{\boldsymbol{\beta}}^{(l)}$ was constructed by each procedure. The loss of each estimator was averaged over 10,000 repetitions for each k.

For the loss function, we employed the maximum loss function

$$\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_\infty \overset{\text{def}}{=} \max_i |\beta_i^* - \hat{\beta}_i|,$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^\top$. The usual loss function for estimators of regression coefficients is either the quadratic loss $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 = \sum_i |\beta_i^* - \hat{\beta}_i|^2$ or the predictive loss $\|\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2 = \sqrt{\sum_i \|\boldsymbol{x}_i\beta_i^* - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_i\|_2^2}$. But, the maximum loss function seems more appropriate than these usual loss functions when the influence of collinearity on the estimated regression coefficients is investigated.

We took $\hat{\boldsymbol{\beta}}_{\text{EB}}$, $\hat{\boldsymbol{\beta}}_{\text{DPP}}$, $\hat{\boldsymbol{\beta}}_{\text{LDPP}}$, and $\hat{\boldsymbol{\beta}}_{\text{GDPP}}$, defined by (3.7), for $p(\boldsymbol{\gamma}|w)$ using the Bernoulli distribution (EB), DPP prior (DPP), linear mixture DPP prior (LDPP), and geometric mixture DPP prior (GDPP), respectively. We took $\sigma^2 = 0.9^2$ and estimated hyperparameters g, w, and θ (LDPP) by maximizing the marginal likelihood. Thus, for example, we estimated g, w, and θ, when using LDPP, by maximizing

$$p(\boldsymbol{y}|g, w, \theta) \propto \sum_{\boldsymbol{\gamma} \in \{0,1\}^p} \frac{\det(w(\theta \boldsymbol{R}_{\boldsymbol{\gamma}} + (1-\theta)\boldsymbol{I}_{\boldsymbol{\gamma}})) \exp\left([g/(1+g)][\mathrm{ss}_{\boldsymbol{\gamma}}/2\sigma^2]\right)}{(1+g)^{|\boldsymbol{\gamma}|/2} \det(w\theta\boldsymbol{R} + (1 + w - w\theta)\boldsymbol{I}_p)}.$$

For α, we used the parameter $\hat{\alpha}$ that maximizes the marginal likelihood $p(\boldsymbol{y}|\alpha)$ over $[0, 3]$. Since $\boldsymbol{X}^\top\boldsymbol{X}$ is ill-conditioned here, we restricted the domain of the optimization.

For comparison, other estimators were also investigated:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{RIDGE}} &\overset{\text{def}}{=} (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{y},\\
\hat{\boldsymbol{\beta}}_{\text{OLS}} &\overset{\text{def}}{=} (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y},\\
\hat{\boldsymbol{\beta}}_{\text{ORACLE}} &\overset{\text{def}}{=} (\boldsymbol{X}_{\boldsymbol{\gamma}^*}^\top\boldsymbol{X}_{\boldsymbol{\gamma}^*})^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}^*}^\top\boldsymbol{y},
\end{aligned}$$

where $\lambda > 0$ is the hyperparameter in ridge regression (putting $\mathcal{N}(0, \sigma^2\lambda^{-1}\boldsymbol{I}_p)$ on $\boldsymbol{\beta}$) and $\boldsymbol{\gamma}^*$ is the true subset of nonzero coefficients, $\boldsymbol{\gamma}^* = \{1, 2\}$. We estimated λ by maximizing the marginal likelihood

$$\begin{aligned} p(\boldsymbol{y}|\lambda) &= \int p(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\lambda)\mathrm{d}\boldsymbol{\beta} \\ &\propto \lambda^{p/2} \int \exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{\lambda}{2\sigma^2}(\boldsymbol{\beta}^\top\boldsymbol{\beta})\Big)\mathrm{d}\boldsymbol{\beta}. \end{aligned}$$

In addition, the Bayesian elastic net (BEN) was studied. The model is

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n), \\ \boldsymbol{\beta} \mid \sigma^2 &\sim \exp\Big\{ -\frac{1}{2\sigma^2}(\lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2)\Big\}, \\ \sigma^2 &\sim \frac{1}{\sigma^2}, \end{aligned}$$

where $\|\cdot\|_p$ is the L^p norm. Hyperparameters λ_1 and λ_2 were estimated by the EM algorithm. See Li and Lin (2010) for detailed MCMC sampling schemes.

Figure 3 shows the results of the comparison. From Figure 3, DPP, LDPP, and GDPP outperform the other estimators. In Section 3, we observed that the best model selected by EB corresponds to the best model selected by the typical penalized sum of squares criteria, such as AIC, BIC or RIC. Therefore EB is considered to evaluate complexity of a submodel by its dimension. Since DPP, LDPP, and GDPP penalize a submodel not only by its dimension but also by the correlations among included predictors, they perform better than EB. Although the ridge estimator reduces the quadratic loss, its maximum risk is worse than DPP, LDPP, and GDPP. The risk of BEN is similar to that of RIDGE. Since the elastic net (Zou and Hastie (2005)) has the grouping effect, BEN is considered not to reduce the maximum risk.

Table 2 shows the medians of estimated hyperparameters $\hat{\theta}$ and $\hat{\alpha}$. From Table 2, estimators $\hat{\theta}$ are nearly 0 and $\hat{\alpha}$ decreases as the sample size increases. In the simulations, since likelihood $p(\boldsymbol{y}|\boldsymbol{\gamma})$ of submodels $\boldsymbol{\gamma}$ that include collinear predictors are small, $\hat{\theta}$ is nearly 0. From Figure 4, since the likelihood of $\boldsymbol{\gamma} = \{1, 2\}$, $\{1, 4\}$, $\{2, 4\}$ and submodels with three predictors are large when the sample size is 20, estimators $\hat{\alpha}$ are large. When sample size is 400, $\hat{\alpha}$ is nearly 1 since the likelihood of $\boldsymbol{\gamma} = \{1, 2\}$ is particularly large and the prior probability of $\boldsymbol{\gamma} = \{1, 2\}$ attains its maximum around $\alpha = 1$.

We provide the reason here why the maximum loss is more appropriate than the usual loss functions. It is well known that the predictive loss is not affected by collinearity even if it is severe, since specific combinations of estimated regression coefficients are well determined by ordinary least squares

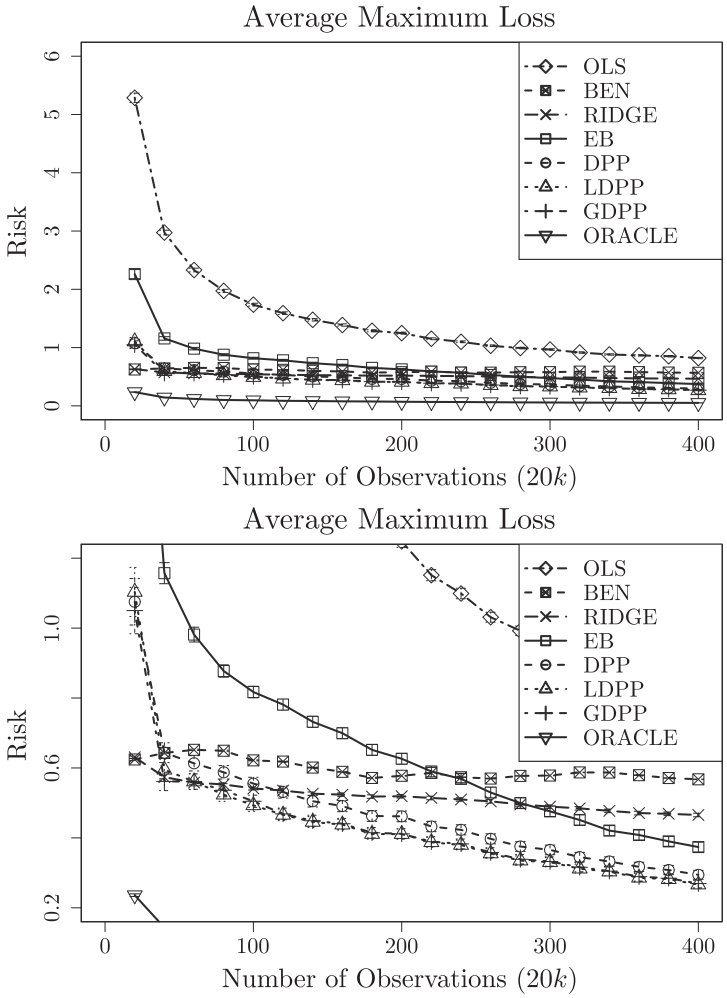

Figure 3. Comparison of the maximum risk for each procedure (EB, DPP, LDPP, GDPP, RIDGE, BEN, and OLS) as the sample size increases. The upper panel shows the result of the estimated maximum risk for each procedure. Each point displays the average maximum risk at $20k$ ($k = 1, 2, \ldots, 20$) observations. Error bars indicate mean ±3×(standard error). The bottom panel shows an enlargement of the upper panel.

(Belsley, Kuh, and Welsch (1980)). Thus predictive loss is not appropriate for investigation of the influence of collinearity on estimated regression coefficients. Since the quadratic loss is mathematically tractable, it has been used when dealing with correlated predictors. Ridge regression (Hoerl and Kennard (1970)) is the method of constructing estimators with the quadratic penalty for estimated coefficients. Quadratic loss summarizes componentwise distances from an estimator to the true parameter, while maximum loss evaluates the furthest distance

Table 2. Median of estimators $\hat{\theta}$ and $\hat{\alpha}$ in the numerical experiments.

Sample size (20k)	θ	α	Sample size (20k)	θ	α
20	9.3e-17	3.00	220	6.2e-15	0.94
40	5.4e-16	3.00	240	7.3e-15	0.90
60	1.1e-15	3.00	260	8.4e-15	0.90
80	1.7e-15	3.00	280	9.3e-15	0.93
100	2.3e-15	1.47	300	1.0e-14	0.94
120	2.7e-15	0.96	320	1.1e-14	0.96
140	3.4e-15	0.94	340	1.1e-14	0.94
160	3.9e-15	0.93	360	1.1e-14	0.93
180	4.2e-15	0.90	380	1.1e-14	0.92
200	4.9e-15	0.90	400	1.1e-14	0.92

in all components. Thus, for example, assume three predictors $\{\boldsymbol{x}_i\}_{i=1}^3$ exist and $\boldsymbol{x}_3$ is nearly equal to $\boldsymbol{x}_1 + \boldsymbol{x}_2$. Suppose that the true regression coefficients are $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, \beta_3^*)^\top = (1, -1, 0)^\top$ and estimators $\hat{\boldsymbol{\beta}}^{(1)} = (0.5, -1.5, 0.0)^\top$, $\hat{\boldsymbol{\beta}}^{(2)} = (1.1, -0.9, -0.6)^\top$ are obtained. $\hat{\boldsymbol{\beta}}^{(1)}$ is prefered to $\hat{\boldsymbol{\beta}}^{(2)}$ since the furthest distance from $\hat{\boldsymbol{\beta}}^{(1)}$ to $\boldsymbol{\beta}^*$ is $|\hat{\beta}_1^{(1)} - \beta_1^*| = 0.5$ but that from $\hat{\boldsymbol{\beta}}^{(2)}$ is $|\hat{\beta}_3^{(2)} - \beta_3^*| = 0.6$. The quadratic losses are

$$\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{(1)}\|_2^2 = 0.5, \quad \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{(2)}\|_2^2 = 0.38. \tag{4.1}$$

Hence, in total, $\hat{\boldsymbol{\beta}}^{(2)}$ is prefered to $\hat{\boldsymbol{\beta}}^{(1)}$ with respect to the quadratic loss. Since one of the serious problems of collinearity is the imprecision of OLS, it is important to investigate the estimation accuracy of every component.

5. Applications to Datasets

Let $\boldsymbol{x}^k$ be the k-th row of the design matrix $\boldsymbol{X}$. In this section, we call $\boldsymbol{x}^k$ the k-th observation, with $\boldsymbol{x}_i$ as the i-th predictor.

Before we report the results of applications to the Air Pollution Data and the Body Fat Data, we summarize assumptions and analysis methods for the datasets.

In practice, since the mean of dependent variable $\boldsymbol{y}$ is almost always nonzero, we assume that the constant term $\boldsymbol{1}_n = (1, \ldots, 1)^\top$ ($n \times 1$ matrix) is included and consider Bayesian variable selection in $M_{\boldsymbol{\gamma}} : \boldsymbol{y} = \mu \boldsymbol{1}_n + \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}$, where μ is an unknown intercept parameter. We take the columns of $\boldsymbol{X}$ as standardized.

In Bayesian variable selection (EB, DPP, LDPP, and GDPP), hyperparameters μ, σ^2, g, w, and θ (LDPP is being used) are estimated by maximizing the marginal likelihood. For α, the parameter $\hat{\alpha}$ is used that maximizes the marginal likelihood $p(\boldsymbol{y}|\alpha)$ over $[0, 3]$. The best model that maximizes the posterior probability $p(\boldsymbol{\gamma}|\boldsymbol{y})$ is selected. The estimator of regression coefficients $\hat{\boldsymbol{\beta}}$ are

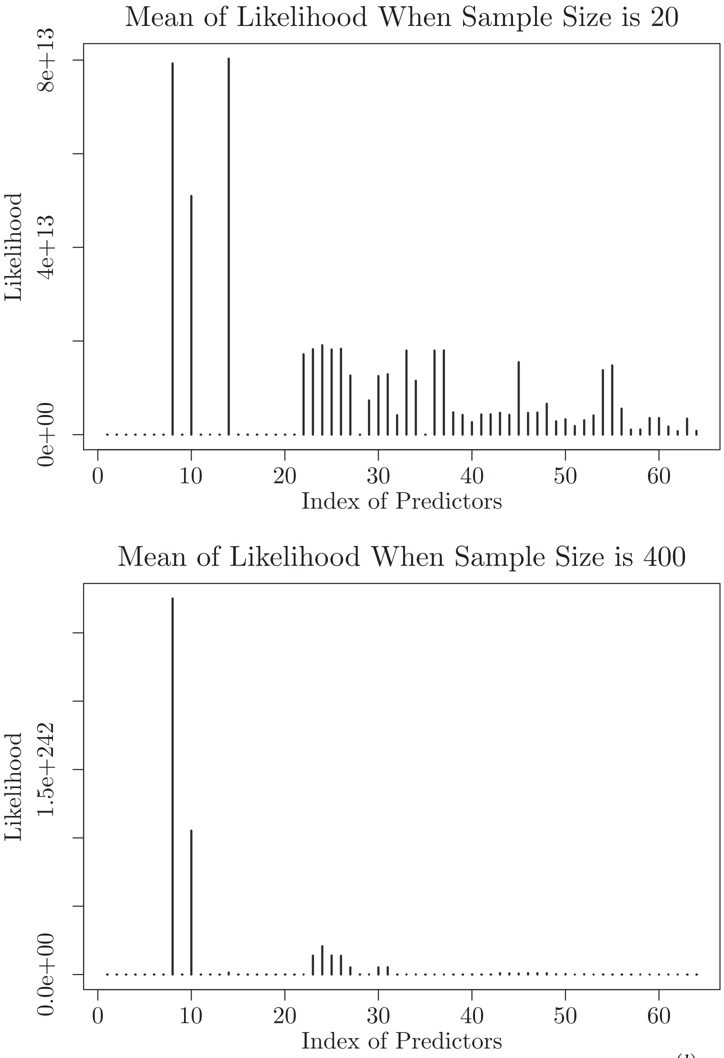

Figure 4. The two panels show that the mean of likelihood $p(\boldsymbol{y}_k^{(l)}|\boldsymbol{\gamma})$ ($l = 1, \ldots,$ 10,000) when sample size is 20 (the upper panel) or 400 (the lower panel). Index of predictors has submodels sorted as $\emptyset < \{1\} < \{2\} < \cdots < \{6\} < \{1,2\} < \{1,3\} < \cdots < \{5,6\} < \{1,2,3\} < \{1,2,4\} < \cdots < \{1,\ldots,6\}$. The likelihood of numbers 8 ($\boldsymbol{\gamma} = \{1,2\}$), 10 ($\boldsymbol{\gamma} = \{1,4\}$), and 14 ($\boldsymbol{\gamma} = \{2,4\}$) are large when sample size is 20. In addition, the likelihood function of submodels with three predictors (numbers: 23-42) are relatively large. Only number 8 ($\boldsymbol{\gamma} = \{1,2\}$) is significantly large when sample size is 400.

constructed following (3.7). For ridge regression, we put the normal distribution $\mathcal{N}(0, \sigma^2\lambda^{-1}\boldsymbol{I})$ on the regression coefficients $\boldsymbol{\beta}$. We estimate hyperparameters μ,

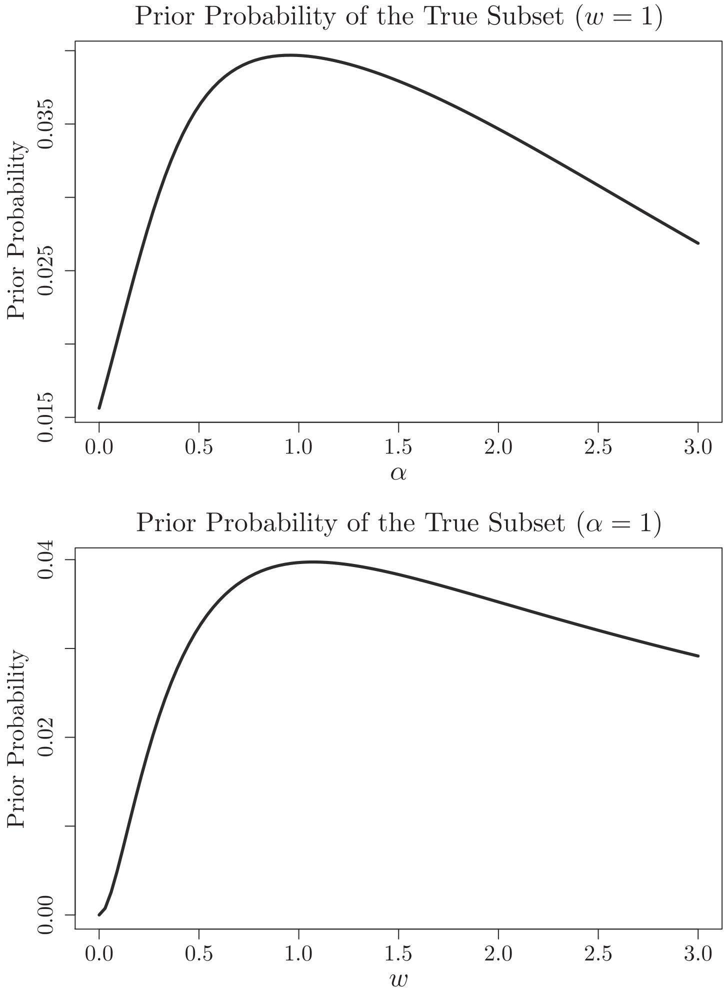

Figure 5. The upper panel shows that prior probability $p(\boldsymbol{\gamma} = \{1, 2\}|w = 1, \alpha)$ when sample size is 400. Similarly, the lower panel shows that prior probability $p(\boldsymbol{\gamma} = \{1, 2\}|w, \alpha = 1)$ when sample size is 400.

σ^2 and λ by maximizing the marginal likelihood.

We investigate prediction accuracy of each procedure (EB, DPP, LDPP, GDPP, RIDGE, BEN, and OLS). In particular, our interest is robustness of the predictive performance of each method when the values of predictors $\boldsymbol{X}$ in the training and test datasets are quite different. In this setting, collinearity influences prediction. To investigate robustness, we divided the observations into two parts according to the values of predictors $\boldsymbol{X}$. The first part is the test dataset and the second part is the candidate for the training datasets. Let $\tilde{\boldsymbol{X}}$ be the design matrix before standardization and let $\tilde{x}_{ij}$ be (i, j) element of $\tilde{\boldsymbol{X}}$. We calculate

$$\tilde{\boldsymbol{m}} \overset{\text{def}}{=} \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_{i1},\ldots,\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_{ip}\right)^{\top}, \quad \tilde{\boldsymbol{\Sigma}} \overset{\text{def}}{=} \frac{1}{n-1}\tilde{\boldsymbol{A}}^{\top}\tilde{\boldsymbol{A}},$$

where $\tilde{\boldsymbol{A}} \overset{\text{def}}{=} \tilde{\boldsymbol{X}} - \big(\tilde{m}_1\mathbf{1}_n,\ldots,\tilde{m}_p\mathbf{1}_n\big)$. Using $\tilde{\boldsymbol{m}}$ and $\tilde{\boldsymbol{\Sigma}}$, we calculate the Mahalanobis distance from $\tilde{\boldsymbol{m}}$ to each observation $\tilde{\boldsymbol{x}}^k$. The furthest 10 observations from $\tilde{\boldsymbol{m}}$ are assigned to the first part (the test dataset). The second part consists of the remaining observations excluding the furthest 20 (in the Air Pollution Data) or 50 (in the Body Fat Data) observations from $\tilde{\boldsymbol{m}}$. Note that 10 (in the Air Pollution Data) or 40 (in the Body Fat Data) observations are not included in either part as our aim is to investigate prediction accuracy when the values of predictors in the training and test datasets are quite different. For $k = 1,\ldots,100$, we randomly sample 20 observations ($\boldsymbol{X}^{(k)}$ ($20\times p$ matrix) and $\boldsymbol{y}^{(k)}$ (20×1 matrix)) from the second part. Then, the prediction accuracy of each procedure is evaluated by the prediction mean squared error (PMSE):

$$\text{PMSE}^{(k)} \overset{\text{def}}{=} \sqrt{\frac{1}{10}\sum_{i=1}^{10}(y_i^{\text{test}} - \hat{y}_i^{(k)})^2},$$

where y_i^{test} ($i = 1,\ldots,10$) is the value of the dependent variable in the test dataset and $\hat{y}_i^{(k)}$ is prediction value of y_i^{test} based on the training dataset ($\boldsymbol{X}^{(k)}$ and $\boldsymbol{y}^{(k)}$), and the values of predictors in the test dataset.

To reduce the computational burden of estimating hyperparameters, we select 10 important predictors by least angle regression (Efron et al. (2004)) beforehand. Thus for $k = 1, 2,\ldots,100$, we select 10 predictors $\boldsymbol{x}_{k_1},\ldots,$ $\boldsymbol{x}_{k_{10}}$ by least angle regression and estimate hyperparameters by maximizing

$$\sum_{\gamma_{k_1}=\{0,1\}}\sum_{\gamma_{k_2}=\{0,1\}}\cdots\sum_{\gamma_{k_{10}}=\{0,1\}} p(\boldsymbol{y}|\tilde{\gamma},\boldsymbol{\xi})p(\tilde{\gamma}|\boldsymbol{\xi}), \tag{5.1}$$

where $\tilde{\boldsymbol{\gamma}} = (\gamma_{k_1},\ldots,\gamma_{k_{10}})^{\top}$ (10×1 matrix) and $\boldsymbol{\xi}$ denotes the hyperparameter to be estimated. Evaluation of the marginal likelihood needs to sum $p(\boldsymbol{y},\boldsymbol{\gamma}|\boldsymbol{\xi})p(\boldsymbol{\gamma}|\boldsymbol{\xi})$ 2^p times. Since this computation is a heavy task even when p is moderately large, we approximate the marginal likelihood by partial sum (5.1).

5.1. Air pollution data

We applied our methods to the Air Pollution Data. The Air Pollution Data was originally analyzed by McDonald and Schwing (1973). The data consists of daily mortality rates in 60 Standard Metropolitan Statistical Areas in the US, along with 15 predictors. The dataset is available from the R package SMPracticals (Davison (2013)).

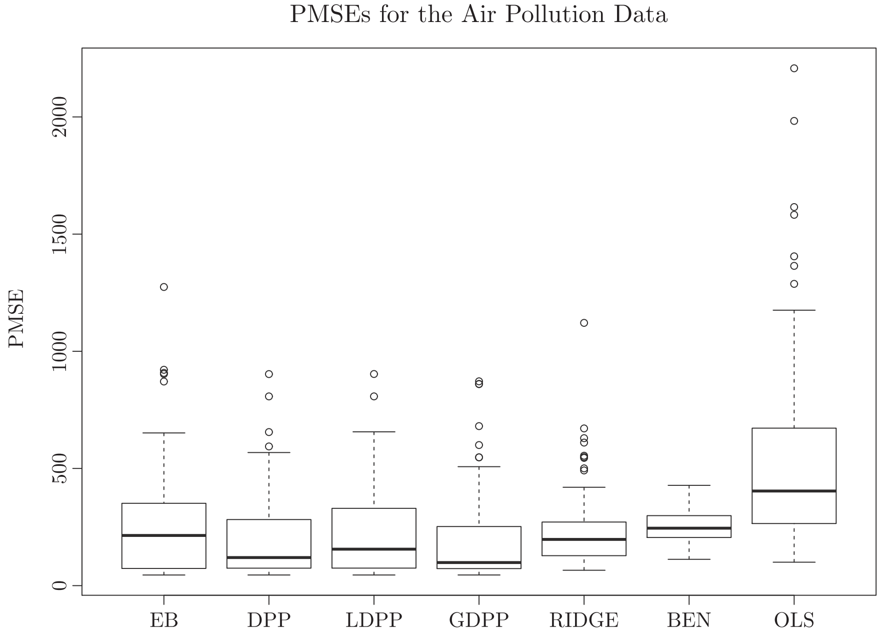

Figure 6. Box plots of prediction mean squared errors (PMSEs) of EB, DPP, LDPP, RIDGE, BEN, and OLS: The panel shows the PMSE of each procedure for the Air Pollution Data when we separate the dataset according to the Mahalanobis distances of $\boldsymbol{X}$.

Table 3. Comparison of median prediction mean squared errors (PMSE) of EB, DPP, LDPP, RIDGE, BEN, and OLS for the Air Pollution Data. Standard errors are given in parentheses.

Method	EB	DPP	LDPP	GDPP	RIDGE	BEN	OLS
Median PMSE	214(23)	119(17)	155(18)	98(18)	197(16)	245(5.8)	403(40)

Table 3 shows the median of PMSE of each method and Figure 6 shows the result of prediction by each method. DPP, LDPP, and GDPP perform better than EB, RIDGE, BEN, and OLS. To compare the predictive performances between EB and DPP, we conducted a paired Wilcoxon signed rank test. The alternative hypothesis is that DPP outperforms EB. Here, the p-value of the test is 3.5×10^{-3}, and we consider DPP to outperform EB.

5.2. Body fat data

The dataset consists of estimates of the percentage of body fat determined by underwater weighing, and 13 body circumference measurements for 252 men. To assess health, it is important to estimate the percentage of body fat. Since accurate evaluation of body fat percentage is inconvenient and expensive, we estimate the percentage from body circumference measurements such as neck

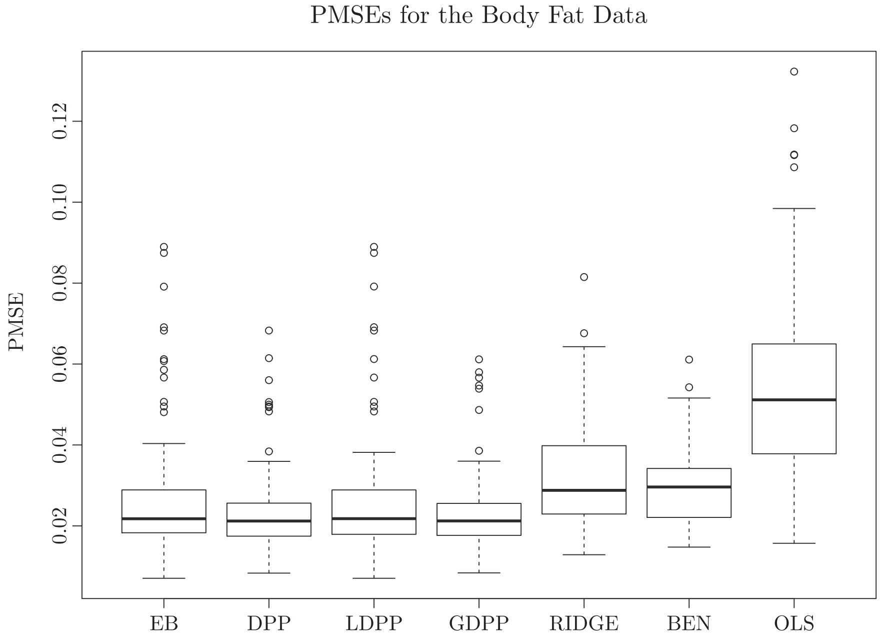

Figure 7. Box plots of prediction mean squared errors (PMSEs) of EB, DPP, LDPP, RIDGE, BEN, and OLS. The panel shows the PMSE of each procedure for the Body Fat Data when we separate the dataset according to the Mahalanobis distances of $\boldsymbol{X}$.

Table 4. Comparison of median of prediction mean squared errors (PMSEs) of EB, DPP, LDPP, RIDGE, BEN, and OLS for the Body Fat Data. Standard errors are given in parentheses.

Method	EB	DPP	LDPP	GDPP
Median PMSE ($\times 10^2$)	2.17(0.16)	2.12(0.10)	2.17(0.15)	2.12(0.10)

Method	RIDGE	BEN	OLS
Median PMSE ($\times 10^2$)	2.87(0.12)	2.96(0.09)	5.11(0.24)

circumference and ankle circumference. We can compute body fat percentages from Siri's equation

$$\text{body fat} = \frac{495}{(\text{body density})} - 450.$$

The dataset is available from Statlib (`http://lib.stat.cmu.edu/datasets/bodyfat`).

Figure 7 shows the result of prediction by each method and Table 4 shows the median of PMSE of each method. EB, DPP, LDPP, and GDPP perform better than RIDGE , BEN, and OLS. As in Section 4, the performance of BEN

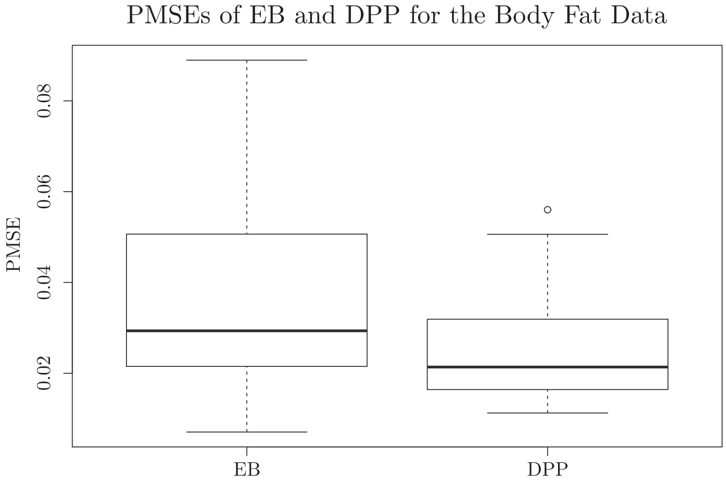

Figure 8. Comparison of prediction mean squared errors (PMSEs) of EB and DPP: The panel shows the PMSEs of EB and DPP for the Body Fat Data when the best model selected by EB differs from that by DPP.

is similar to that of RIDGE under collinearity.

The regression coefficients estimated by DPP are similar to those by EB when the methods select the same best model. For the Body Fat Data, EB and DPP select the same model 71 of 100 times. We consider EB to perform as well as DPP. We compared EB to DPP when they selected a different best model. Figure 8 shows the result. We conclude that DPP performs better than EB when they select different predictors.

From Figure 6, DPP, LDPP, and GDPP outperform EB, RIDGE, BEN, and OLS for the Air Pollution Data. Moreover, from Figure 7, our methods and EB outperform RIDGE, BEN, and OLS. We conclude that the predictive performances of our methods are better than those of EB, RIDGE, BEN, and OLS as prediction by our methods is more accurate and robust. We consider the robustness to arise from the repulsion property of DPPs. If the values of the predictors $\boldsymbol{X}$ in the training and test datasets are quite different, collinearity influences prediction. In this setting, EB, RIDGE, BEN, and OLS are inappropriate because they do not consider correlations among predictors. Since DPP priors assign small prior probabilities to submodels including collinear predictors, predictions by our methods are robust and accurate.

6. Conclusion

We considered Bayesian variable selection in linear regression, and proposed discrete determinantal point processes (DPPs) for prior distributions on model parameter $\boldsymbol{\gamma}$. Since the proposed prior (DPP prior) assigns small probabilities to submodels including collinear predictors, the best model is less likely to include collinear predictors. We observed that the DPP prior is a generalization of the Bernoulli distribution, which is used for $p(\boldsymbol{\gamma})$ in the method proposed by George and Foster (2000) (EB). Therefore, our method is a generalization of EB. We also proposed the linear mixture DPP prior (LDPP) and the geometric mixture DPP prior (GDPP) that bridge the Bernoulli distribution and the DPP prior.

Ročkóva and George (2014a) propose using DPPs for priors on model parameter $\boldsymbol{\gamma}$ in Bayesian variable selection, independent of our study. We propose estimating hyperparameters in the DPP priors by maximizing the marginal likelihood and selecting the best model that maximizes the posterior probability. Ročkóva and George (2014a) propose marginalizing hyperparameters with respect to a hyper prior and combining EMVS (Ročkóva and George (2014b)). Since the EM algorithm is an iterative method, when p (the number of predictors) is not large, our methods are faster than that of Ročkóva and George. However, when p is large, EMVS can find the best model faster than our methods (and other MCMC methods). For approximating the marginal likelihood, our methods would be as fast as EMVS even when p is not small.

In the simulations, the estimators of regression coefficients constructed by our methods reduce the maximum risk more than do EB, the ridge estimator (RIDGE), the Bayesian elastic net (BEN), and the ordinary least squares estimator (OLS) when collinearity is severe. In the experiments, since the correlation matrix of predictors is ill-conditioned, we restricted the domain of optimization when maximizing the marginal likelihood using GDPP. Relaxation of the restriction is a future task.

We also applied our methods to air pollution data and body fat data. Our interest was in the robustness of the predictive performance when the value of predictors $\boldsymbol{X}$ in the training dataset and the test dataset are quite different. For air pollution data, our proposed methods yielded more accurate prediction than did the others. For body fat data, our proposed methods and EB yielded more accurate prediction than did RIDGE, BEN, and OLS. From these results of the applications, we find that predictions by our methods are more accurate and robust compared to others. We consider the robustness to arise from the repulsion property of DPPs.

In the large p small n setting, wherein there exist many more predictors than observations, we intend to combine the proposed DPP priors with the stochastic search method proposed by Kwon et al. (2011). In this setting, since too many predictors exist and collinearity is severe, our methods will be efficient.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 26280005.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the* 2*nd International Symposium on Information Theory* (Edited by B. N. Petrov, and F. Csaki), 267-281. Akademiai Kiado, Budapest.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley, New York.

Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *J. Amer. Statist. Assoc.* **107**, 1610-1624.

Borodin, A. and Rains, E. (2005). Eynard-Mehta theorem, Schur process, and their Pfaffian analogs. *J. Statist. Phys.* **121**, 291-317.

Davison, A. (2013). *SMPracticals: Practicals for use with Davison* (2003) *Statistical Models.* R package version 1.4-2.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-499.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947-1975.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731-747.

Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *J. Amer. Statist. Assoc.* **110**, 435-448.

Hoerl, E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

Hough, B. J., Krishnapur, M., Peres, Y. and Virág, B. (2009). *Zeros of Gaussian Analytic Functions and Determinantal Point Processes.* American Mathematical Society, Providence.

Krishna, A., Bondell, H. D. and Ghosh, S. K. (2009). Bayesian variable selection using an adaptive powered correlation prior. *J. Statist. Plann. Inference* **139**, 2665-2674.

Kulesza, A. and Taskar, B. (2012). Determinantal point processes for machine learning. *Foundations and Trends in Machine Learn.* **5**, 123-286.

Kwon, D., Landi, M. T., Vannucci, M., Issaq, H. J., Prieto, D. and Pfeiffer, R. M. (2011). An efficient stochastic search for Bayesian variable selection with high-dimensional correlated predictors. *Comput. Statist. Data Anal.* **55**, 2807-2818.

Li, Q. and Lin, N. (2010). The Bayesian elastic net. *Bayesian Anal.* **5**, 151-170.

Macchi, O. (1975). The coincidence approach to stochastic point processes. *Adv. Appl. Probab.* **7**, 83-122.

McDonald, G. C. and Schwing, R. C. (1973). Instabilities of regression estimates relating air pollution to mortality. *Technometircs* **15**, 463-481.

Ročková, V. and George, E. I. (2014a). Determinantal priors for variable selection. In *Proceedings of the* 47*th Scientific Meeting of the Italian Statistical Society* (Edited by S. Cabras, T. D. Battista and W. Racugno). Cooperativa Universitaria Editrice Cagliaritana, Cagliari.

Ročková, V. and George, E. I. (2014b). EMVS: The EM approach to Bayesian variable selection. *J. Amer. Statist. Assoc.* **109**, 828-846.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-465.

Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *J. Amer. Statist. Assoc.* **100**, 1215-1224.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distribution. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (Edited by P. K. Goel and A. Zellner), 233-243. North-Holland, Amsterdam.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

Mitsui Sumitomo Insurance Co.,Ltd., Tokyo Sumitomo Twin Building (West Tower) 27-2, Shinkawa 2-Chome, Chuo-ku, Tokyo 104-0033, Japan.

E-mail: m.kojima9067@gmail.com

Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.

RIKEN Brain Science Institute, 2-1 Hirosawa, Wako City, Saitama 351-0198, Japan.

E-mail: komaki@mist.i.u-tokyo.ac.jp

(Received December 2014; accepted May 2015)