MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Determinantal Point Process Priors for Bayesian Variable Selection in Linear Regression

Mutsuki KOJIMA and Fumiyasu KOMAKI

METR 2014–14 June 2014

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

Determinantal Point Process Priors for Bayesian Variable Selection in Linear Regression

Mutsuki KOJIMA* and Fumiyasu KOMAKI*†

*Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo, Tokyo, Japan
†RIKEN Brain Science Institute, Wako-shi, Japan
{mutsuki_kojima,komaki}@mist.i.u-tokyo.ac.jp

June 6th, 2014

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 selected simultaneously because of the repulsion property of discrete DPPs. Three types of DPP priors are proposed. We show the efficiency of the proposed priors through numerical experiments and applications to collinear datasets.

1 Introduction

We consider Bayesian variable selection in linear regression. Suppose we have n observations on a dependent variable y_n ($n\times 1$ matrix) and p predictor variables $X = (x_1, \ldots, x_p)$ ($n \times p$ matrix), for which the normal linear model holds:

$$y_n = X\beta + \varepsilon_n, \tag{1}$$

where $\varepsilon_n \sim \mathcal{N}(0, \sigma^2 I_n)$ ($n \times 1$ matrix) and $\beta = (\beta_1, \ldots, \beta_p)^\top$ ($p \times 1$ matrix). Let $\gamma = (\gamma_1, \ldots, \gamma_p)^\top \in \{0,1\}^p$ be a model parameter, meaning that $\gamma_i = 1$ indicates β_i is nonzero and $\gamma_i = 0$ indicates $\beta_i = 0$. For Bayesian variable selection, we consider 2^p possible submodels of (1). Each submodel is denoted by M_γ. Let X_γ be a $n \times |\gamma|$ design matrix consisting of these columns of X that correspond to the predictors with $\gamma_i = 1$. Here, $|\gamma|$ is the number of

nonzero elements of γ, i.e., $|\gamma| = \sum_{i=1}^{p} \gamma_i$. Under submodel M_γ, y_n follows

$$M_\gamma : \quad y_n = X_\gamma \beta_\gamma + \varepsilon_n, \tag{2}$$

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

The normal linear regression model is simple and useful, but collinearity problem often arises when we apply it to a real data. A serious problem of collinearity is imprecision of the ordinary least squares estimator (OLS). The problem occurs when X is ill-conditioned, i.e., highly correlated predictors exist. Then $(X^\top X)^{-1}$ is numerically unstable and therefore the OLS, $\hat{\beta}_{\mathrm{OLS}} = (X^\top X)^{-1} X^\top y_n$, is not reliable. One way of avoiding the issue is variable selection. However, most existing methods for variable selection do not take into account for the correlations among predictors. In the large p small n setting, in which there are many more predictors than observations, Kwon et al. (2011) exploits the correlations for proposal distribution in a stochastic search method. However, few Bayesian variable selection methods considering the correlations have been proposed. We propose discrete determinantal point processes (DPPs) for the prior distribution on γ, so that submodels including collinear predictors are less likely to be selected. Our Bayesian variable selection method makes use of the correlations among 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, and discrete DPPs 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 requires diversity and quality.

Selected predictors in variable selection should be diverse, which means each pair of selected predictors is nearly uncorrelated. For the selection of diverse predictors, discrete DPPs are efficient priors on γ in Bayesian variable selection. We show that discrete DPPs are useful priors through numerical experiments and applications to real data sets.

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). Next, we propose Bayesian variable selection methods using three types of DPP priors on γ. In Section 4, we show numerical experimental results and we report applications to the two real datasets in Section 5. We conclude the paper in Section 6.

2 Discrete Determinantal Point Processes

First, we give the definition of discrete determinatal point processes (DPPs).

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

Definition 2.1 A random variable $\mathcal{X}$ that takes values in the power set of Λ is called *discrete determinantal point process (DPP) with kernel* L, if

$$P(\mathcal{X} = \gamma) \propto \det(L_\gamma), \tag{3}$$

where $\gamma \in \{0,1\}^p$ and L_γ is the $|\gamma| \times |\gamma|$ matrix whose elements are $(L_{ij})_{i,j \in \gamma}$. We define $\det(L_\emptyset) = 1$.

The normalization constant is given by the next proposition. See Kulesza and Taskar (2012) for the proof.

Proposition 2.2

$$\sum_{\gamma \in \{0,1\}^p} \det(L_\gamma) = \det(L + I_p), \tag{4}$$

where I_p is the $p \times p$ identity matrix, and the sum is taken over all subsets of Λ.

Definition 2.1 and Proposition 2.2 are enough to propose DPP priors. For more detail properties of discrete DPPs, see Kulesza and Taskar (2012). See Hough et al. (2009), for general determinantal point processes.

Next, we give a brief explanation of the repulsion property of DPPs. The property is a key of our proposal. Let F be a $p \times q$ ($p < q$) matrix, and we denote the rows of F by f_i ($i = 1, 2, \ldots, p$). We assume that $L = FF^\top$. If $\mathcal{X}$ follows the discrete DPP with kernel L, then

$$P(\mathcal{X} = \gamma) \propto (\mathrm{vol}(\{f_i\}_{i \in \gamma}))^2, \tag{5}$$

where $\mathrm{vol}(\{f_i\}_{i \in \gamma})$ means the $|\gamma|$-dimensional volume of the parallelepiped spanned by the rows of $\{f_i\}_{i \in \gamma}$. By considering f_i as the feature vector of item i, $\mathrm{vol}(\{f_i\}_{i \in \gamma})$ is small if there exist similar items in γ. This is because the volume of the parallelepiped is smaller as the angle between two edges decreases. Since the small angle between edges f_i and f_j means that item i is similar to item j, the probabilities that $\mathcal{X}$ includes similar items are small. Thus DPPs are considered to prefer repulsion and diversity.

Finally we give two examples of discrete DPPs. The first example shows the distribution of discrete DPPs with a diagonal matrix kernel corresponds to the Bernoulli distribution. The second example indicates that nonzero off-diagonal elements of the kernel L determine negative correlations between pairs of items.

Table 2.1: The distribution of $\mathcal{X}$ in Example 2.4.

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

Example 2.3 Let $L = (L_{ij})_{i,j=1,\ldots,p}$ be a diagonal matrix whose elements are

$$L_{ij} = \begin{cases} w/(1-w) & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $w \in (0,1)$. Suppose $\mathcal{X}$ follows the discrete DPPs with kernel L, then by proposition 2.2,

$$P(X = \gamma) = \frac{\det(L_\gamma)}{\det(L + I_p)} = w^{|\gamma|}(1-w)^{p-|\gamma|}.$$

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

Example 2.4 Let $\Lambda = \{1,2,3\}$ and

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

Suppose $\mathcal{X}$ follows the discrete DPPs with kernel L. The distribution of $\mathcal{X}$ is as Table 2.1. From Table 2.1, the event that $\mathcal{X}$ equals $\{1,2\}$ or $\{1,2,3\}$ is less likely to occur. This is because the off-diagonal element $L_{12} = 0.9$. In fact,

$$P(\mathcal{X} = \{i,j\}) \propto P(\mathcal{X} = \{i\})P(\mathcal{X} = \{j\}) - \left(\frac{L_{ij}}{\det(L + I_p)}\right)^2. \tag{7}$$

Thus, if off-diagonal elements are nonzero, then the corresponding items are less likely to be included simultaneously.

3 Bayesian Variable Selection Methods

In subsection 3.1, we first review the Bayesian variable selection method proposed by George and Foster (2000). In subsection 3.2, we propose Bayesian variable selection methods using three types of DPP priors on γ.

3.1 Bayesian Variable Selection Method Proposed by George and Foster (2000)

George and Foster (2000) proposed the following Bayesian variable selection.

Zellner's g-prior (Zellner, 1986) is assigned to nonzero regression coefficients β_γ under submodel M_γ:

$$p(\beta_\gamma|g) \sim \mathcal{N}(0, g\sigma^2(X_\gamma^\top X_\gamma)^{-1}), \quad g > 0, \tag{8}$$

where g is the hyperparameter. For the prior distribution on model parameter γ, the Bernoulli distribution

$$p(\gamma|w) = w^{|\gamma|}(1-w)^{p-|\gamma|} \tag{9}$$

with success parameter $w \in (0,1)$ is used. The best model

$$\begin{aligned} \hat{\gamma} &= \underset{\gamma}{\text{argmax}}\ p(\gamma|y_n, g, w) \\ &= \underset{\gamma}{\text{argmax}}\ \exp\left(\frac{g}{2(1+g)}(\text{ss}_\gamma/\sigma^2 - F(g,w)|\gamma|)\right) \\ &= \underset{\gamma}{\text{argmax}}\ (\text{ss}_\gamma/\sigma^2 - F(g,w)|\gamma|), \end{aligned} \tag{10}$$

maximizing the posterior probability, is selected, where

$$\text{ss}_\gamma = y_n^\top X_\gamma X_\gamma^\top y_n, \quad F(g,w) = \frac{1+g}{g}\left(2\log\frac{1-w}{w} + \log(1+g)\right). \tag{11}$$

If we assume σ^2 is known and hyperparameters are appropriately calibrated, the Bayesian variable selection above is identical to 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 is the model maximizing (10)

$$\text{ss}_\gamma/\sigma^2 - 2|\gamma|. \tag{12}$$

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

$$\begin{aligned} (\hat{g}, \hat{w}) &= \underset{g,w}{\text{argmax}}\ p(y_n|g,w) \\ &= \underset{g,w}{\text{argmax}} \sum_{\gamma\in\{0,1\}^p} p(\gamma|w) \int p(y_n|\gamma,\beta_\gamma)p(\beta_\gamma|g)\mathrm{d}\beta_\gamma. \end{aligned} \tag{13}$$

Since Zellner's g-prior is the normal distribution, the marginal distribution can is represented in the closed-form:

$$\int p(y_n|\gamma,\beta_\gamma)p(\beta_\gamma|g)\mathrm{d}\beta_\gamma = \frac{(1+g)^{-|\gamma|/2}}{(2\pi)^{n/2}\sigma^n}\exp\left(\frac{g}{1+g}\frac{\mathrm{ss}_\gamma}{2\sigma^2} - \frac{y_n^\top y_n}{2\sigma^2}\right). \quad (14)$$

Therefore the type II likelihood for g and w is

$$p(y_n|g,w) \propto \sum_{\gamma\in\{0,1\}^p} \frac{w^{|\gamma|}(1-w)^{p-|\gamma|}}{\sigma^n(1+g)^{|\gamma|/2}}\exp\left(\frac{g}{1+g}\frac{\mathrm{ss}_\gamma}{2\sigma^2} - \frac{y_n^\top y_n}{2\sigma^2}\right). \quad (15)$$

We refer the method above as EB (empirical Bayes) in the following sections.

3.2 DPP Priors and Proposed Methods

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

$$\tilde{x}_{ij} = \frac{x_{ij}-m_j}{s_j},$$

where

$$m_j \stackrel{\mathrm{def}}{=} \frac{1}{n}\sum_{i=1}^n x_{ij}, \quad s_j \stackrel{\mathrm{def}}{=} \sqrt{\sum_{i=1}^n (x_{ij}-m_j)^2}.$$

We denote the correlation matrix ($p\times p$ matrix) of X by R given by

$$R = \tilde{X}^\top\tilde{X}. \quad (16)$$

The first proposal for prior distribution $p(\gamma|w)$ is

$$p(\gamma|w) \propto \det(wR_\gamma) = w^{|\gamma|}\det(R_\gamma), \quad w>0. \quad (17)$$

By Proposition 2.2,

$$p(\gamma|w) = \frac{\det(wR_\gamma)}{\det(wR+I_p)}. \quad (18)$$

We call this prior *DPP prior* (referred as DPP). This proposal is based on the following consideration. Assume there are three predictors x_1, x_2 and x_3. We also assume that the correlation between x_1 and x_2 is 0.9 and the other pairs are uncorrelated. If $w=1$, then the probabilities of $p(\gamma|w)$ are the same as Table 2.1. From Table 2.1, we see that DPP prior assigns small probabilities to subsets of predictors including collinear predictors (x_1 and x_2). DPPs prefer repulsion and diversity as we see in Section 2. Therefore, when we

put DPP prior on model parameter γ, we can select diverse predictors, which means each pair of selected predictors is nearly uncorrelated. The hyperparameter w controls the expected proportion of nonzero regression coefficients; if $w > 1$ then larger subsets are more preferable, otherwise smaller subsets are more preferable.

From Example 2.3, the Bernoulli distribution is a discrete DPP with a diagonal matrix kernel. Thus DPP prior is a generalization of the Bernoulli distribution that is used for $p(\gamma|w)$ in EB. From this point of view, we propose two types of priors that bridge the Bernoulli distribution and DPP prior:

$$p(\gamma|w,\theta) \propto \det(w(\theta R_\gamma + (1-\theta)I_\gamma)), \quad w > 0, \quad \theta \in [0,1], \tag{19}$$
$$p(\gamma|w,\alpha) \propto \det(w(R^\alpha)_\gamma), \quad w > 0, \quad \alpha \geq 0, \tag{20}$$

where I_γ is the $|\gamma| \times |\gamma|$ identity matrix and R^α is the non-integer powers α of R. We call (19) *linear mixture DPP prior* (referred as LDPP) and (20) *geometric mixture DPP prior* (referred as GDPP). Linear mixture DPP prior corresponds to DPP prior when $\theta = 1$ and corresponds to the Bernoulli distribution when $\theta = 0$. Similarly, geometric mixture DPP prior corresponds to DPP prior when $\alpha = 1$ and corresponds to the Bernoulli distribution when $\alpha = 0$.

Our Bayesian variable selection methods are as follows. We put proposed priors (DPP, LDPP or GDPP) on model parameter γ and g-prior on β_γ. Next, hyperparameters are estimated by maximizing the type II likelihood. Here hyperparameters are g, σ^2 (if not known), w, θ (if using LDPP) and α (if using GDPP). The best model is selected that maximizes the posterior probability $p(\gamma|y_n)$.

Though we discuss variable selection methods so far, we can estimate the regression coefficients after selecting the best model. The estimator $\hat{\beta}$ is constructed after estimating $\hat{g}$ and selecting the best model $M_{\hat{\gamma}}$:

$$\hat{\beta} = \mathrm{E}[\beta|\hat{\gamma},\hat{g}] = \frac{\hat{g}}{1+\hat{g}}(X_{\hat{\gamma}}^\top X_{\hat{\gamma}})^{-1} X_{\hat{\gamma}}^\top y_n. \tag{21}$$

The representation of the estimator is the same whether the best model $\hat{\gamma}$ is selected by EB or by our methods. This is because the g-prior is put on the regression coefficients in both methods.

4 Numerical Experiments

In this section, we evaluate the risk of estimated regression coefficients as the sample size increases through numerical experiments.

In the numerical experiments, the following settings are considered. First

we sample a 400×6 design matrix X^* whose columns are x_i^* $(i = 1, 2, \ldots, 6)$:

$$\begin{aligned} x_1^*,\ x_2^*,\ x_3^*,\ \varepsilon_4,\ \varepsilon_5,\ \varepsilon_6 &\overset{\text{i.i.d.}}{\sim} \mathcal{N}_{400}(0, I_{400}), \\ x_4^* &= x_1^* + x_2^* + 0.1 \times \varepsilon_4, \\ x_5^* &= x_1^* + x_3^* + 0.1 \times \varepsilon_5, \\ x_6^* &= x_2^* + x_3^* + 0.1 \times \varepsilon_6. \end{aligned}$$

Let X_k^* $(k = 1, 2, \ldots, 20)$ be a $20k \times 6$ submatrix of X^* whose rows correspond to the first $20k$ rows of X^* and $\beta^* = (1, -1, 0, 0, 0, 0)^\top$. For each k, we simulate y_{20k} 10000 times following (1) with $X = X_k^*$, $\beta = \beta^*$ and $\sigma^2 = 0.9^2$. For each y_{20k}, the estimator $\hat{\beta}$ is constructed by each method. A loss of each estimator is averaged over 10000 repetitions.

For the loss function, we employ the maximum loss function $\|\beta^* - \hat{\beta}\|_\infty$ defined by

$$\|\beta^* - \hat{\beta}\|_\infty \overset{\text{def}}{=} \max_i |\beta_i^* - \hat{\beta}_i|, \tag{22}$$

where $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$. A usual loss function for estimators of regression coefficients is quadratic loss (i.e., $\|\beta^* - \hat{\beta}\|_2^2 = \sum_i |\beta_i^* - \hat{\beta}_i|^2$) or predictive loss (i.e., $\|X\beta^* - X\hat{\beta}\|_2^2 = \sum_i |x_i\beta_i^* - x_i\hat{\beta}_i|^2$). However, the maximum loss function is more appropriate than these usual loss functions when influence of collinearity on the estimated regression coefficients is investigated. The reason is described in the last paragraph of this section. Therefore we use the maximum loss function in the numerical experiments.

We denote $\hat{\beta}_{\text{EB}}$, $\hat{\beta}_{\text{DPP}}$, $\hat{\beta}_{\text{LDPP}}$ and $\hat{\beta}_{\text{GDPP}}$ defined by (21) when for $p(\gamma|w)$ using the Bernoulli distribution (EB), DPP prior (DPP), linear mixture DPP prior (LDPP), and geometric mixture DPP prior (GDPP), respectively. We assume that $\sigma^2 = 0.9^2$ is known and estimate other hyperparameters g, w, θ (in LDPP) by maximizing the type II likelihood. For example, when using LDPP, we estimate g, w and θ by maximizing the marginal likelihood

$$\begin{aligned} p(y_n|g, w, \theta) \quad \propto \quad & \sum_{\gamma \in \{0,1\}^p} \left\{ \frac{\det(w(\theta R_\gamma + (1-\theta) I_\gamma))}{(1+g)^{|\gamma|/2} \det(w\theta R + (1 + w - w\theta) I)} \right. \\ & \left. \times \exp\left(\frac{g}{1+g} \frac{\text{ss}_\gamma}{2\sigma^2} \right) \right\}. \end{aligned} \tag{23}$$

For α (in GDPP), we use the parameter $\hat{\alpha}$ that maximizes the type II likelihood $p(y_n|\alpha)$ over $[0, 3]$. Since $X^\top X$ is ill-conditioned in the numerical experiments, we restrict the domain of the optimization.

For comparison, other estimators are also investigated:

$$\hat{\beta}_{\text{RIDGE}} \overset{\text{def}}{=} (X^\top X + \lambda I_p)^{-1} X^\top y_n, \tag{24}$$

$$\hat{\beta}_{\text{OLS}} \overset{\text{def}}{=} (X^\top X)^{-1} X^\top y_n, \tag{25}$$

$$\hat{\beta}_{\text{ORACLE}} \overset{\text{def}}{=} (X_{\gamma^*}^\top X_{\gamma^*})^{-1} X_{\gamma^*}^\top y_n, \tag{26}$$

where $\lambda > 0$ is the hyperparameter in ridge regression (putting $\mathcal{N}(0, \sigma^2\lambda^{-1}I)$ on β) and γ^* is the true subset of nonzero coefficients. We estimate λ by maximizing the type II likelihood

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

Figure 4.1 shows the results of the comparison. From Figure 4.1, DPP, LDPP and GDPP outperform the other estimators. In Section 3, we see 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 predictors X, they perform better than EB. Though RIDGE can reduce the quadratic loss, but its maximum risk is worse than DPP, LDPP and GDPP.

Finally we give the reason why the maximum loss is more appropriate than the usual loss functions. A serious problem of collinearity is the imprecision of OLS. It is well known that the predictive loss is not affected by the collinearity even if collinearity is severe. This is because specific combinations of estimated regression coefficients are well-determined by the ordinary least squares (Belsley et al., 1980). Therefore the 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. However, the quadratic loss may be inappropriate since it summarizes componentwise distances from an estimator to the true parameter. If we want to estimate the all values of regression coefficients precisely, the maximum loss is more appropriate than the quadratic loss since the maximum loss evaluates the furthest distance in all components. For example, assume that three predictors $\{x_i\}_{i=1}^3$ exist and x_3 nearly equals to $x_1 + x_2$. Assume also that the true regression coefficients are $\beta^* = (\beta_1^*, \beta_2^*, \beta_3^*)^\top = (1, -1, 0)^\top$ and two estimators $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ are obtained:

$$\hat{\beta}^{(1)} = (0.5, -1.5, 0.0)^\top, \quad \hat{\beta}^{(2)} = (1.1, -0.9, -0.6)^\top. \tag{27}$$

$\hat{\beta}^{(1)}$ is more favorable than $\hat{\beta}^{(2)}$ since the third component of $\hat{\beta}^{(1)}$ is zero but that of $\hat{\beta}^{(2)}$ is nonzero. However, each quadratic loss is

$$\|\beta^* - \hat{\beta}^{(1)}\|_2^2 = 0.5, \quad \|\beta^* - \hat{\beta}^{(2)}\|_2^2 = 0.38. \tag{28}$$

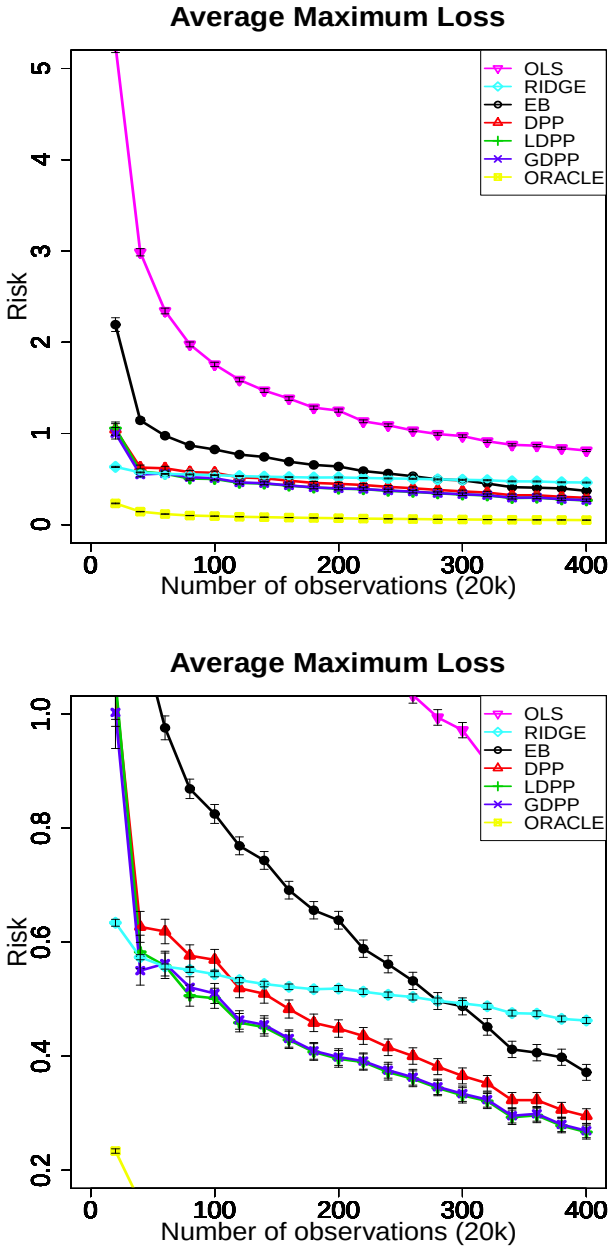

Figure 4.1: Comparison of the maximum risk for each procedure (EB, DPP, LDPP, GDPP, RIDGE, OLS) as the sample size increases: The upper panel shows the result of estimated the maximum risk for each procedure. Each point displays the average maximum risk at $20k$ ($k = 1, 2, \ldots, 20$) observations. Error bars indicate mean $\pm 3\times$standard error. The bottom panel shows an enlargement of the upper panel.

This means $\hat{\beta}^{(2)}$ is more preferable than $\hat{\beta}^{(1)}$. In contrast, each maximum loss is

$$\|\beta^* - \hat{\beta}^{(1)}\|_\infty = 0.5, \quad \|\beta^* - \hat{\beta}^{(2)}\|_\infty = 0.6, \tag{29}$$

which means $\hat{\beta}^{(1)}$ is more preferable. Even though the third component of $\hat{\beta}^{(2)}$ is far from the true parameter β_3^*, the distances of the other components from $\hat{\beta}^{(2)}$ to the true parameter are closer than those of $\hat{\beta}^{(1)}$. Therefore, in total, $\hat{\beta}^{(2)}$ is more preferable than $\hat{\beta}^{(1)}$ with respect to the quadratic loss. This happens because the quadratic loss function summarizes component-wise distances from an estimator to the true parameter. For this reason, the maximum loss is more appropriate than the quadratic loss when we are interested in the impact of collinearity on all components of estimated regression coefficients.

5 Applications to Real Datasets

Let x^k be the k-th row of the design matrix X. In this section, we call x^k the k-th observation. Note that x_i is the i-th predictor.

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

In practice, since the mean of the dependent variable y_n is almost always nonzero, we assume that the constant term $1_n = (1, \ldots, 1)^\top$ ($n \times 1$ matrix) is included. Therefore we consider the following linear regression model:

$$y_n = \mu 1_n + X\beta + \varepsilon_n, \tag{30}$$

where μ is an unknown intercept parameter. In Bayesian variable selection, we consider the following submodels:

$$M_\gamma : y_n = \mu 1_n + X_\gamma \beta_\gamma + \varepsilon_n. \tag{31}$$

We also assume that the columns of X is standardized.

We compare proposed methods and the usual methods (EB, RIDGE and OLS). For Bayesian variable selection (EB, DPP, LDPP and GDPP), hyperparameters μ, σ^2, g, w, θ (if using LDPP) are estimated by maximizing the type II likelihood. For α (in GDPP), we use the parameter $\hat{\alpha}$ that maximizes the type II likelihood $p(y_n|\alpha)$ over $[0, 3]$ since $X^\top X$ is ill-conditioned. The best model is selected that maximizes the posterior probability $p(\gamma|y_n)$. The estimator of regression coefficients $\hat{\beta}$ is constructed following (21). For ridge regression, we put the normal distribution $\mathcal{N}(0, \sigma^2\lambda^{-1}I)$ on the regression coefficients β. We estimate hyperparameters μ, σ^2 and λ by maximizing the marginal liklihood.

We investigate the prediction accuracy of each procedure (EB, DPP, LDPP, GDPP, RIDGE, OLS). In particular, our interest is the robustness of the predictive performance of each method when the value of predictors X in the training dataset and the test dataset are very different. In this setting, collinearity influences on the prediction. In order to investigate the robustness, we divided the observations into two parts according to the values of predictors X. The first part is the candidate for the test dataset, and the second part is for the training dataset. To be precise, we divide the datasets as follows. Let $\tilde{X}$ be the design matrix before standardization and let $\tilde{x}_{ij}$ be the (i,j) element of $\tilde{X}$. We calculate the mean $\tilde{m} = (\tilde{m}_1, \ldots, \tilde{m}_p)^\top$ and the sample covariance matrix $\tilde{\Sigma}$ of $\tilde{X}$:

$$\tilde{m} \stackrel{\text{def}}{=} \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{x}_{i1}, \ldots, \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_{ip} \right)^\top, \quad \tilde{\Sigma} \stackrel{\text{def}}{=} \frac{1}{n-1} \tilde{A}^\top \tilde{A}, \tag{32}$$

where

$$\tilde{A} \stackrel{\text{def}}{=} \tilde{X} - \left(\tilde{m}_1 1_n, \ldots, \tilde{m}_p 1_n \right). \tag{33}$$

Using $\tilde{m}$ and $\tilde{\Sigma}$, we calculate the Mahalanobis distance from $\tilde{m}$ to each observation $\tilde{x}^k$ that is the k-th row of $\tilde{X}$. The furthest 10 observations from $\tilde{m}$ is assigned to the first part. The second part consists of the remaining observations excluding the furthest 20 (for the Air Pollution Data) or 50 (for the Body Fat Data) observations from the $\tilde{m}$. Note that 10 (for the Air Pollution Data) or 40 (for the Body Fat Data) observations are not included in either part. This is because our purpose is to investigate the prediction accuracy when the values of predictors in the training and the test dataset are very different. For $l = 1, 2, \ldots, n$ ($n = 60$ for the Air Pollution Data and $n = 100$ for the Body Fat Data), we randomly sample 1 observation $y_{\text{test}}^{(l)}$ from the fist part and m ($m = 20$ for the Air Pollution Data and $m = 30$ for the Body Fat Data) observations $X_m^{(l)}$ and $y_m^{(l)}$ from the second part. Then the prediction accuracy for each procedure is evaluated by the absolute loss function:

$$|y_{\text{test}}^{(l)} - \hat{y}_l|, \tag{34}$$

where $\hat{y}_l$ is the prediction value for $y_{\text{test}}^{(l)}$ based on $X_m^{(l)}$ and $y_m^{(l)}$.

5.1 Air Pollution Data

We apply our methods to the Air Pollution Data which is a widely known collinear dataset. 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 of the USA, and 15 predictors. The dataset is available from R package SMPracticals (Davison, 2013).

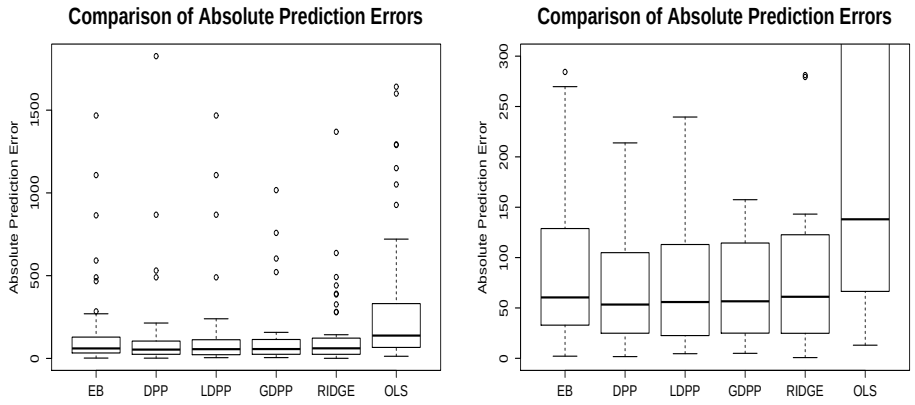

Figure 5.1: Comparison of absolute prediction errors for EB, DPP, LDPP, RIDGE and OLS: The left panel shows the result of prediction by each procedure for the Air Pollution Data when we separate the dataset according to the Mahalanobis distances of X. The right panel shows an enlargement of the left panel.

In order to reduce the computational burden of estimating the hyperparameters in Bayesian variable selection (EB, DPP, LDPP and GDPP), we select 10 important predictors by least angle regression (Efron et al., 2004) beforehand. To be precise, for $l = 1, 2, \ldots, 60$, we select 10 predictors $x_{j_1}, \ldots, x_{j_{10}}$ by least angle regression and hyperparameters are estimated by maximizing the sum

$$\sum_{\gamma_{j_1}=\{0,1\}} \sum_{\gamma_{j_2}=\{0,1\}} \cdots \sum_{\gamma_{j_{10}}=\{0,1\}} p(y_n|\gamma,\xi)p(\gamma|\xi), \tag{35}$$

where ξ denotes all hyperparameters to be estimated. Each evaluation of the type II likelihood $p(y_n|\xi)$ needs to sum of $p(y_n,\gamma|\xi)p(\gamma|\xi)$ 2^p times. Since this computation is a heavy task even if p is moderately large, we approximate the marginal likelihood $p(y_n|\xi)$ by the partial sum (35).

Figure 5.1 shows the result of prediction by each method. DPP, LDPP, GDPP and RIDGE perform better than EB and OLS. For comparison of the predictive performances of EB and DPP, we conduct a Wilcoxon signed rank test for paired samples. The alternative hypothesis is that DPP outperforms EB. As a result, the p-value of the test is 0.058, and we consider DPP outperforms EB when the values of X in the training and the test dataset are very different in the Air Pollution Data.

5.2 Body Fat Data

Next, we report the Body Fat Data application. The dataset consists of estimates of the percentage of body fat determined by underwater weighing and

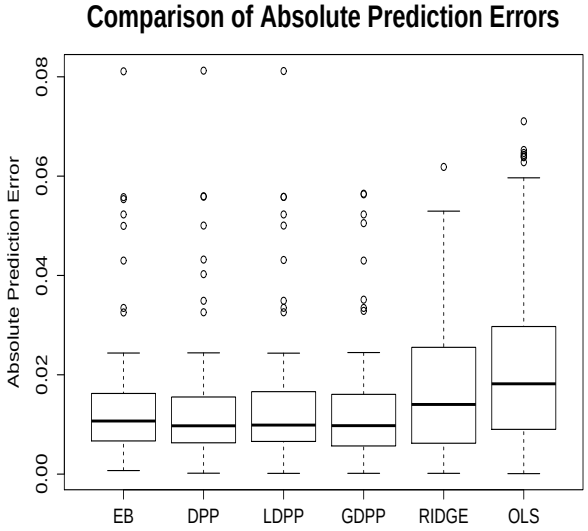

Figure 5.2: Comparison of absolute prediction errors for EB, DPP, LDPP, RIDGE and OLS: The figure shows the result of prediction when we separate the dataset according to the Mahalanobis distances of X.

13 body circumference measurements for 252 men. To assess one's health, it is important to estimate the percentage of body fat. However, since accurate evaluation of body fat percentage is costly, we estimate the percentage from body circumference measurements such as neck circumference, ankle circumference and so on. Since we can estimate body fat percentages from body density by Siri's equation

$$\text{body fat} = 495/(\text{body density}) - 450, \tag{36}$$

estimating body density is enough to investigate body fat. Therefore we consider body density as the dependent variable. The dataset is available from Statlib (`http://lib.stat.cmu.edu/datasets/bodyfat`).

Figure 5.2 shows the result of prediction by each procedure. Bayesian variable selection procedures (EB, DPP, LDPP, GDPP) yielded more accurate prediction than RIDGE and OLS. In order to compare the prediction accuracy of RIDGE and DPP, we conduct a Wilcoxon signed rank test for paired samples. The alternative hypothesis is that DPP outperforms RIDGE. As a result, the p-value of the test is 9.9×10^{-4}, and we consider DPP outperforms RIDGE when the values of X in the training and test dataset are very different in the Body Fat Data.

From Figure 5.1, our methods (DPP, LDPP and GDPP) and RIDGE outperform EB and OLS for the Air Pollution Data. On the other hand,

from Figure 5.2, our methods and EB outperform RIDGE and OLS. Therefore we conclude that the predictive performances of our methods are better than the other methods (EB, RIDGE and OLS) in the sense that the prediction by our methods are more accurate and robust. We consider the robustness comes from the repulsion property of DPPs. If the values of predictors X in the training and the test dataset are very different, collinearity influences on prediction. In this setting, usual methods (EB, RIDGE and OLS) are inappropriate because such methods do not take into account for the correlations among predictors X. However, since DPP priors assign small prior probabilities to submodels including collinear predictors, prediction 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 γ. Since the proposed prior (DPP prior) assigns small probabilities to submodels including collinear predictors, collinear predictors are less likely to be selected simultaneously. In Section 2, we see that DPP prior is a generalization of the Bernoulli distribution that is used for $p(\gamma)$ in the method proposed by George and Foster (2000) (EB). Therefore our method is a generalization of EB. From this point of view, we also proposed linear mixture DPP prior (LDPP) and geometric mixture DPP prior (GDPP) that bridge the Bernoulli distribution and DPP prior.

In the numerical experiment, the estimators of regression coefficients constructed by our methods reduce the maximum risk more than the other estimators (EB, the ridge estimator (RIDGE) and the ordinary least squares estimator (OLS)) when collinearity is severe. We also apply our methods to the Air Pollution Data and the Body Fat Data. Our interest is the robustness of the predictive performance of each method when the value of predictors X in the training dataset and the test dataset are very different. For the Air Pollution Data, proposed methods and RIDGE yielded more accurate prediction than EB and OLS. In addition, for the Body Fat Data, proposed methods and EB yielded more accurate prediction than RIDGE and OLS. From these results of the applications, we conclude that prediction of our methods are more accurate and robust comparing to the other methods (EB, RIDGE and OLS). We consider the robustness comes from the repulsion property of DPPs.

Finally, we give a future plan of this work. In the large p small n setting, we intend to use the proposed DPP priors, combining 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.

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281, 1973.

D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley, 1980.

A. Borodin and E. Rains. Eynard–Mehta theorem, Schur process, and their pfaffian analogs. *Journal of Statistical Physics*, 121:291–317, 2005.

A. Davison. *SMPracticals: Practicals for use with Davison (2003) Statistical Models*, 2013. URL `http://CRAN.R-project.org/package=SMPracticals`. R package version 1.4-2.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32:407–499, 2004.

D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22:1947–1975, 1994.

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

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

J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. *Zeros of Gaussian analytic functions and determinantal point processes*. American Mathematical Society, 2009.

A. Kulesza and B. Taskar. *Determinantal point processes for machine learning*. Now Publishers, 2012.

D. Kwon, M. T. Landi, M. Vannucci, H. J. Issaq, D. Prieto, and R. M. Pfeiffer. An efficient stochastic search for Bayesian variable selection with high-dimensional correlated predictors. *Computational Statistics and Data Analysis*, 55:2807–2818, 2011.

O. Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7:83–122, 1975.

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

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–465, 1978.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel and A. Zellner, editors, *Bayesian inference and decision techniques: essays in honor of Bruno de Finetti*, pages 233–243. Elsevier, 1986.