Diversifying Sparsity Using Variational Determinantal Point Processes

N.K. Batmanghelich[1] **G. Quon**[1] **A. Kulesza**[2] **M. Kellis**[1] **P. Golland**[1] **L. Bornn**[3]
[1]Massachusetts Institute of Technology [2]University of Michigan [3]Harvard University

Abstract

We propose a novel diverse feature selection method based on determinantal point processes (DPPs). Our model enables one to flexibly define diversity based on the covariance of features (similar to orthogonal matching pursuit) or alternatively based on side information. We introduce our approach in the context of Bayesian sparse regression, employing a DPP as a variational approximation to the true spike and slab posterior distribution. We subsequently show how this variational DPP approximation generalizes and extends mean-field approximation, and can be learned efficiently by exploiting the fast sampling properties of DPPs. Our motivating application comes from bioinformatics, where we aim to identify a diverse set of genes whose expression profiles predict a tumor type where the diversity is defined with respect to a gene-gene interaction network. We also explore an application in spatial statistics. In both cases, we demonstrate that the proposed method yields significantly more diverse feature sets than classic sparse methods, without compromising accuracy.

1 Introduction

As modern technology enables us to capture increasingly large amounts of data, it is critically important to find efficient ways to create compact, functional, and interpretable representations. Feature selection is a promising approach, since reducing the feature space both improves interpretability and prevents over-fitting; as a result, it has received considerable attention in the literature [e.g., 26, 13]. In this paper, we focus on the problem of *diverse* feature selection, where the notion of diversity can be defined in terms of the features themselves or in terms of available side information.

Diverse feature sets have the potential to be both more compact and easier to interpret, without sacrificing performance. Diversity also plays a more fundamental role in some real-world applications; for example, breast cancer is increasingly recognized to present a highly heterogeneous group of malignancies [8] where subgroups may involve different mechanisms of action. For the common task of identifying gene expression-based biomarkers of different tumor subtypes, maximizing the diversity of selected genes helps identifying these disparate mechanisms of action. Diversity in this case is defined with respect to a separate gene-gene interaction network (see Section 4.1).

Existing techniques for feature selection generally do not explicitly consider feature diversity. From an optimization point of view, feature selection can be viewed as a search over all possible subsets of features to identify the optimal subset according to a pre-specified metric, often balancing model fit and model complexity. To avoid enumerating the entire combinatorial search space, embedded approaches, such as the LASSO [26], relax the problem to a combination of a sparsity term (ℓ_1) and a data fidelity term. However, such methods typically do not encourage diversity explicitly. In fact, LASSO has been shown to be unstable in the face of nearly collinear features [11], with several variants proposed to ameliorate this issue [28].

An alternative approach is to search greedily by successively adding (or removing) the best (or worst) feature [13]. Orthogonal matching pursuit (OMP) proceeds in this way using forward step-wise feature selection, but the selected feature is chosen to be as orthogonal as possible to previously selected features [21]. One can view this orthogonality as a measure to implicitly maximize diversity [6]. In spite of its well-established performance [27], OMP is a procedure that lacks an underlying generative model and, therefore, the flexibility to define diversity other than through

the inner product of features.

In this paper we take a probabilistic view of the problem, assigning a probability measure to feature subsets. We then seek the maximum *a posteriori* (MAP) estimate as the optimal subset. In particular, our probability measure is a variational approximation to the posterior of the spike-and-slab variable selection model. By imposing a particular form on that approximation, we obtain a measure that assigns higher scores to feature subsets that are not only relevant to a regression or classification task but also diverse. The challenge is to find a form that achieves this goal while remaining computationally tractable.

To this end, our variational approximation takes the form of a determinantal point process (DPP). DPPs are appealing in this context since they naturally encourage diversity, defined in terms of a kernel matrix that can be, for example, the feature covariance matrix (discouraging collinearity), or alternatively derived from application-specific notions of similarity [16]. DPPs also offer computationally appealing properties such as efficient sampling and approximate MAP estimation [10, 19]. As a result, not only can we efficiently approximate the optimal feature set, but we can also provide sampling-based credible intervals.

Unlike mean field approximations that fully factorize the posterior distribution, our approximate DPP posterior has a complex dependency structure. This makes fitting the DPP a challenging task. Kulesza *et al.*[16] used an optimization approach to learn conditional DPPs, and Affandi *et al.*[1] proposed to parameterize the kernel of DPPs and learn the parameters using a sampling approach; however, neither approach allows learning a full, unparameterized kernel. In contrast, our algorithm is based on a flexible variational framework proposed by Salimans and Knowles [23] that only requires two basic operations: efficient evaluation of the joint likelihood and efficient sampling from the current estimate of the posterior. Fortunately, for DPPs these operations are efficient. For regression, the marginal likelihood takes a closed form, and for many other models, including classification, it can be approximated efficiently. To the best of our knowledge this work presents the first use of DPPs within a variational framework.

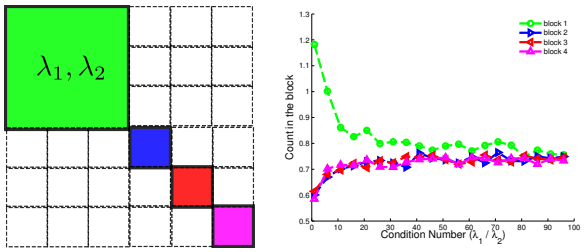
The closest work to ours is by George *et al.*[9]. They suggested variations of so-called *dilution* priors that, instead of assigning prior uniformly across models, assign probability more uniformly across *neighborhoods*, and then distribute the neighborhood probabilities across the models within them. One of the diluting priors proposed in [9] is proportional to the determinant of the features, resembling the form of a DPP. Such a prior does not guarantee diversifying properties on the posteriori selected features. Although possible, instead of prior, we suggest to approximate the posterior with DPP. We investigate this model choice in Section 4.

This work makes the following contributions. We propose to use a DPP as an approximate posterior for Bayesian feature selection. To fit the variational approximation, we draw a connection between DPPs and the exponential family using the decomposition proposed by Kulesza *et al.*[16]. This connection makes many tools developed for the exponential family available for DPPs, including variational learning methods for distributions that are not fully factorizable. Our proposed method brings a number of advantages, including the ability to: (i) propose multiple sets of relevant and diverse features which can be viewed as alternative feature selection solutions; (ii) characterize feature selection uncertainty through posterior sampling; (iii) flexibly define feature set diversity based on side information, rather than just covariance; and (iv) compute the marginal probability for inclusion of new features conditioned on the presence of an existing feature set, thanks to the computational properties of DPPs.

The remainder of the paper is organized as follows. We first review DPPs in Section 2. The idea of diverse sparsity is illustrated in the context of Bayesian variable selection in Section 3.1. In Section 3.2, we show how to learn the parameters of such models. In the Section 4, we apply the method to identify a diverse set of genes to predict a tumor type while the diversity is defined with respect to a gene-gene interaction network. Finally, in Section 4.2, we explore application to learning an optimal distribution of grid points in a spatial process convolution model.

2 Determinantal Point Process

The determinantal point process (DPP) defines distribution over configurations of points in space. If the space is finite, say $[M] := \{1, \cdots, M\}$; it defines probability mass over all 2^M subsets. Specifically, the probability of a subset $\boldsymbol{\gamma} \in \{0,1\}^M$ is proportional to the determinant of $[\mathbf{L}]_{\boldsymbol{\gamma}}$ where $[\cdot]_{\boldsymbol{\gamma}}$ denotes the submatrix containing the columns and rows i for which $\gamma_i = 1$ and $\mathbf{L} \in \mathbb{R}^{M \times M}$ is a positive semidefinite kernel matrix. Strictly speaking, this representation defines a subclass of DPPs called L-ensembles. If L_{ij} is a measurement of similarity between elements i and j, then the DPP assigns higher probabilities to subsets that are diverse. More precisely, if $L_{ij} = \phi(i)^T \phi(j)$ for a feature function $\phi : [M] \to \mathbb{R}^d$, then the probability of a set $\boldsymbol{\gamma}$ is proportional to the squared volume spanned



(a) (b)

Figure 1: (a) L-ensemble of DPP for a toy problem of six items. The items are dissimilar except the first three items. This leads to a block diagonal kernel $\mathbf{L}$. (b) Empirical average number of elements selected from each block. Although the green block is bigger than the other blocks, as it becomes more collinear ($\frac{\lambda_1}{\lambda_2} \to \infty$), the probability of selecting an item from the first block converges to that of the other blocks.

by $\{\phi(i) \mid \gamma_i = 1\}$. Elements with orthogonal feature functions will tend to span larger volumes, and hence have a higher probability of co-occurrence. For more in-depth review of DPP and its applications in machine learning see [16].

In addition to having computationally appealing properties such as efficient marginalization and sampling, repulsive interactions are also modeled elegantly through a DPP. To illustrate this preference for diversity, suppose we would like to choose a subset from six items. The items are dissimilar except for the first three items. Therefore, their L-ensemble matrix is block diagonal where the first three items form a single group, illustrated as a green block in Figure 1a. We assume that the green block has rank two with eigenvalues of λ_1 and λ_2. As the condition number $\frac{\lambda_1}{\lambda_2}$ increases, the items of the green block become more similar (collinear). We can sample from this DPP and compute the empirical average of samples falling into each block as a function of $\frac{\lambda_1}{\lambda_2}$ (Figure 1b). If there is no interaction, the probabilities are proportional to the sizes of the blocks but as $\frac{\lambda_1}{\lambda_2} \to \infty$, the probability of selecting an item from the first block decreases to that of the rest of the blocks.

3 Methods

3.1 Bayesian Variable Selection

Following standard notation, we consider the regression model $\mathbf{y} = \sum_{m=1}^{M} \mathbf{x}_m \beta_m + \varepsilon$, where the regressors $\{\mathbf{x}_1, \cdots, \mathbf{x}_M\}$ are collected into a design matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_M] \in \mathbb{R}^{N \times M}$ and ε is the residual noise $\varepsilon \sim \mathcal{N}(\cdot; 0, \sigma)$. One can view variable selection as deciding which of the coefficients β_m are nonzero. This is often made explicit in Bayesian variable selection through a latent binary random vector $\boldsymbol{\gamma} \in \{0,1\}^M$ that specifies which predictors are included:

$$\mathbf{y} = \mathbf{X}(\boldsymbol{\gamma} \odot \boldsymbol{\beta}) + \varepsilon, \tag{1}$$

where $\odot$ is the element-wise product. Assuming an exchangeable Bernoulli prior for $\boldsymbol{\gamma}$ and a conjugate prior for $\boldsymbol{\beta}$ with covariance $\sigma^2 \Lambda_0^{-1}$, *i.e.,* $\boldsymbol{\beta} \sim \mathcal{N}(\cdot; 0, \sigma^2 \Lambda_0^{-1})$, random variable $\gamma_m \beta_m$ defines the so-called "spike-and-slab" prior [4, 18]. It is drawn from the spike with probability α and from the slab with probability $1-\alpha$. Assuming an inverse Gamma prior with parameters (a_0, b_0) for variance σ^2, the joint likelihood of the model can be written as follows:

$$p(\mathbf{y}, \pi; \mathbf{X}, \rho) = \tag{2}$$
$$p(\mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma; \mathbf{X}) p(\boldsymbol{\beta}|\sigma; \Lambda_0) p(\sigma; a_0, b_0) p(\boldsymbol{\gamma}; \alpha),$$

where $\pi = \{\boldsymbol{\beta}, \sigma, \boldsymbol{\gamma}\}$ is the set of latent random variables and $\rho = \{\Lambda_0, a_0, b_0, \alpha\}$ is the set of fixed parameters. We set $\Lambda_0 = c\mathbf{I}$.

Conditioned on $\boldsymbol{\gamma}$, the marginal likelihood of the restricted regression can be expressed in closed-form [5]:

$$\log p\left(\mathbf{y}|\boldsymbol{\gamma}; \mathbf{X}, \rho\right) = \tag{3}$$
$$-\frac{N}{2} \log 2\pi + \frac{1}{2} \left(\log\det\left(\Lambda_0\right) - \log\det\left(\Lambda_N\right)\right) +$$
$$\left(a_0 - a_N\right)\left(\log b_0 - \log b_N\right) + \log\Gamma\left(a_N\right) - \log\Gamma\left(a_0\right)$$
$$\Lambda_N = \mathbf{X}^T\mathbf{X} + \Lambda_0, \qquad \boldsymbol{\mu}_N = \Lambda_N^{-1}\mathbf{X}^T\mathbf{y}$$

where $\Gamma(\cdot)$ is the Gamma function, $a_N = a_0 + N/2$, and $b_N = b_0 + \frac{1}{2}\left(\mathbf{y}^T\mathbf{y} - \boldsymbol{\mu}_N^T \Lambda_N\right)$.

Exact inference of the posterior inclusion probability of the regressors, *i.e.,* $p(\boldsymbol{\gamma}|\mathbf{y}; \mathbf{X}, \rho)$, is computationally prohibitive since it entails a sum over all possible subsets of $[M] := \{1, \cdots, M\}$. We therefore resort to an approximation. Variational approaches approximate the form of the posterior; for example, the mean field approach employs a fully factorized function as the approximating distribution [3, 4]. Marginalizing $\boldsymbol{\beta}$ out, the mean field approximates the posterior probability of variable inclusion as

$$q(\boldsymbol{\gamma}; \boldsymbol{\theta}) = \prod_{m=1}^{M} q(\gamma_m; \theta_m) = \prod_{m=1}^{M} \theta_m^{\gamma_m} (1-\theta_m)^{1-\gamma_m} \tag{4}$$

However, this form of posterior does not account for interaction between regressors, diversity being one such form.

To encourage diversity, one needs to model the interactions between features. We propose to use a DPP in an elegant way to define probability mass over all possible subsets of $[M]$. As a naming convention, "`XX-YY`" refers to prior specification `XX` and variational

distribution YY; hence we will refer to the setting as Bernoulli-DPP. It is possible to define the DPP as a prior for $\boldsymbol{\gamma}$ and approximate the posterior with a fully factorized mean field method, *i.e.,* Bernoulli (referred as DPP-Bernoulli). However, such a prior does not guarantee that the posterior exhibits the diversifying property. It is also straightforward to have DPP as prior and posterior (*i.e.,* DPP-DPP) but here we focus on investigating how effective DPP is as prior versus posterior.

Following the formulation of Kulesza *et al.*[16], we propose the following variational posterior distribution:

$$q(\boldsymbol{\gamma};\boldsymbol{\theta}) = \frac{1}{Z_{\boldsymbol{\theta}}} \det\left[\mathbf{L}\right]_{\boldsymbol{\gamma}} = \frac{1}{Z_{\boldsymbol{\theta}}} \det\left[\operatorname{diag}(e^{\frac{\boldsymbol{\theta}}{2}})\boldsymbol{\Phi}\boldsymbol{\Phi}^T\operatorname{diag}(e^{\frac{\boldsymbol{\theta}}{2}})\right]_{\boldsymbol{\gamma}} = \frac{1}{Z_{\boldsymbol{\theta}}} e^{\boldsymbol{\theta}^T\boldsymbol{\gamma}} \det\left[\boldsymbol{\Phi}\boldsymbol{\Phi}^T\right]_{\boldsymbol{\gamma}} \quad (5)$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are parameters and latent random variables respectively, and $Z_{\boldsymbol{\theta}} = \det(\mathbf{I}+\mathbf{L})$ is the normalization constant [16]. $\boldsymbol{\Phi} \in \mathbb{R}^{M\times d}$ is a given matrix of similarity features whose row m, $\phi(m)$, is the similarity feature vector for item m, and L-ensemble matrix is $\mathbf{L} = \operatorname{diag}(e^{\frac{\boldsymbol{\theta}}{2}})\boldsymbol{\Phi}\boldsymbol{\Phi}^T\operatorname{diag}(e^{\frac{\boldsymbol{\theta}}{2}})$. For example if $\boldsymbol{\Phi} = \mathbf{X}$, the DPP discourages collinearity. $\boldsymbol{\Phi}$ can also be defined via side information (see Section 4).

Note that *Eq.*(5) reduces to *Eq.*(4) if the similarity features are indicator vectors, *i.e.,* $\boldsymbol{\Phi}\boldsymbol{\Phi}^T = \mathbf{I}$. In this case, the DPP approximation proposed here reverts to a mean field approximation.

3.2 Learning

We now propose an algorithm to fit the variational approximation for both DPP-Bernoulli and Bernoulli-DPP. Learning with Bernoulli-DPP is more challenging than DPP-Bernoulli. To see why, note that the variational approach minimizes the divergence

$$\mathrm{KL}\left(q_{\boldsymbol{\theta}}|p(\boldsymbol{\gamma},y)\right) = \mathbb{E}_{q_{\boldsymbol{\theta}}}\left[\log q_{\boldsymbol{\theta}}\left(\boldsymbol{\gamma}\right) - \log p\left(\boldsymbol{\gamma},y\right)\right] \quad (6)$$

When $q_{\boldsymbol{\theta}}$ is a DPP, computing the first term, which is the entropy of the DPP entails 2^M summands which, to the best of our knowledge, does not have any closed-form. We focus on approximating Bernoulli-DPP and show how a straightforward modification to the resulting algorithm can effectively approximate the posterior of DPP-Bernoulli.

To learn $\boldsymbol{\theta}$ in *Eq.*(5), we borrow the stochastic approximation algorithm from [23], which allows one to approximate any distribution that is given in a closed-form. In *structured* or *fixed-form* variational Bayes [15], the posterior distribution is chosen to be a specific member of an exponential family, namely $q(\boldsymbol{\gamma};\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^T T(\boldsymbol{\gamma}) - U\left(\boldsymbol{\theta}\right)\right)\nu\left(\boldsymbol{\gamma}\right)$ where $T(\boldsymbol{\gamma})$ is the sufficient statistic, $U(\boldsymbol{\theta})$ is the normalizer and $\nu(\boldsymbol{\theta})$ is a base measure. The DPP in its general form is not a member of exponential family but parameterizing DPPs as *Eq.*(5) allows us to employ the framework. We first summarize the algorithm in [23] in our context where $T(\boldsymbol{\gamma}) := \boldsymbol{\gamma}$, $\nu(\boldsymbol{\gamma}) := \det([\boldsymbol{\Phi}\boldsymbol{\Phi}^T]_{\boldsymbol{\gamma}})$, and $U(\boldsymbol{\theta}) := \det(\mathbf{L}+\mathbf{I})$.

For notational convenience, we define $\tilde{q}_{\tilde{\boldsymbol{\theta}}} := \exp(\tilde{\boldsymbol{\gamma}}^T\tilde{\boldsymbol{\theta}})$ where $\tilde{\boldsymbol{\theta}}^T = \left[\boldsymbol{\theta}^T, \theta_0\right]$ and $\tilde{\boldsymbol{\gamma}}^T = \left[\boldsymbol{\gamma}^T, 1\right]$. If $\theta_0 = -U(\boldsymbol{\theta})$, then $\tilde{q}$ is the normalized posterior, otherwise it is an unnormalized version [23]. Taking the gradient of the unnormalized KL-divergence $\mathcal{D}(\tilde{q}_{\tilde{\boldsymbol{\theta}}}|p(\boldsymbol{\gamma},\mathbf{y}))$, with respect to $\tilde{\boldsymbol{\theta}}$ we obtain:

$$\nabla_{\tilde{\boldsymbol{\theta}}}\mathcal{D}\left[q_{\tilde{\boldsymbol{\theta}}}|p\left(\boldsymbol{\gamma},y\right)\right] = \nabla_{\tilde{\boldsymbol{\theta}}}\mathbb{E}_{\tilde{q}}\left[\log\tilde{q}_{\tilde{\boldsymbol{\theta}}}\left(\boldsymbol{\gamma}\right) - \log p\left(\boldsymbol{\gamma},y\right)\right] = \int \tilde{q}_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\gamma})\left[\tilde{\boldsymbol{\gamma}}\tilde{\boldsymbol{\gamma}}^T\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\gamma}}\log p(\boldsymbol{\gamma},y)\right] d\nu(\boldsymbol{\gamma}) \quad (7)$$

By setting *Eq.*(7) to zero, Salimans and Knowles [23] linked linear regression and the variational Bayes method. Namely, at the optimal solution, $\tilde{\boldsymbol{\theta}}$ should satisfy the linear system : $\mathbf{C}\tilde{\boldsymbol{\theta}} = \mathbf{g}$ where $\mathbf{C} := \mathbb{E}_q\left[\tilde{\boldsymbol{\gamma}}\tilde{\boldsymbol{\gamma}}^T\right]$ and $\mathbf{g} := \mathbb{E}_q\left[\tilde{\boldsymbol{\gamma}}\log p\left(\boldsymbol{\gamma},y\right)\right]$. $\mathbf{C}$ and $\mathbf{g}$ are estimated via weighted Monte Carlo sampling by drawing a single sample, $\boldsymbol{\gamma}_t$, from the current posterior approximation $q_{\tilde{\boldsymbol{\theta}}_t}$,

$$\begin{aligned}\mathbf{g}_{t+1} &= (1-w)\,\mathbf{g}_t + w\tilde{\boldsymbol{\gamma}}_t\log p\left(\boldsymbol{\gamma}_t,y\right)\\ \mathbf{C}_{t+1} &= (1-w)\,\mathbf{C}_t + w\tilde{\boldsymbol{\gamma}}_t\tilde{\boldsymbol{\gamma}}_t{}^T\end{aligned} \quad (8)$$

where $w \in [0,1]$ is the step size.

Interestingly, $\mathbb{E}_q\left[\boldsymbol{\gamma}\boldsymbol{\gamma}^T\right]$ is the DPP marginal kernel, $\mathbf{K}$, which has the closed form $\mathbf{K} = \mathbf{L}(\mathbf{L}+\mathbf{I})^{-1}$. Nevertheless, in our experiments, we did not see any clear advantage in substituting the current estimate for $\mathbf{K}$ directly into *Eq.*(8) versus using the empirical estimate.

Pseudo-code for our algorithm is shown in Algorithm 1 in Appendix A. We set $p(\boldsymbol{\gamma}) = \prod_m \alpha^{\gamma_m}(1-\alpha)^{1-\gamma_m}$, and as suggested in [23], $w := \frac{1}{\sqrt{N}}$. We further set the initial L-ensemble to $\mathbf{L} = (e^{\theta_0/2}\boldsymbol{\Phi})(\boldsymbol{\Phi}^T e^{\theta_0/2})$, where θ_0 is adjusted to make sure that the initial samples from the DPP are not empty sets. To do so, we note that the diagonal elements of $\mathbf{K}$ are the marginal probability of inclusion of element i. Therefore the expected cardinality of the subset is $tr(\mathbf{K}) = \sum_i \frac{e^{\theta_0}\lambda_i}{1+e^{\theta_0}\lambda_i}$, where λ_i is the i'th eigenvector of $\boldsymbol{\Phi}\boldsymbol{\Phi}^T$. To set the expected cardinality of subsets to a preset value κ, we solve the equation for θ_0.

Algorithm 1 only requires sampling from a DPP with parameters $\boldsymbol{\theta}_t$ and computing $p(y,\boldsymbol{\gamma}_t) = p(y|\boldsymbol{\gamma}_t)p(\boldsymbol{\gamma}_t)$. For example, in the regression problem *Eq.*(3), $p(y|\boldsymbol{\gamma}_t)$

has the closed-form solution in *Eq.*(4). In a linear logistic regression case (classification), $p(y|\boldsymbol{\gamma}_t)$ does not have a closed-form but conditioned on $\boldsymbol{\gamma}_t$, its computation is the equivalent of computing the marginal likelihood for linear kernel Gaussian Process model, which can be approximated using expectation propagation (EP).Joint distribution $p(y, \boldsymbol{\gamma}_t)$ can encode more involved models as long as $p(y|\boldsymbol{\gamma}_t)$ can be approximated efficiently (see the supplementary material for an example). One side benefit of the algorithm is the automated selection of the number of features included in the model. If fixed model size is desired, the algorithm can be easily extended to employ $k-$DPPs, where the cardinality of the subset is fixed, by replacing the sampling part of the algorithm.

After learning the DPP, we compute the MAP estimate using [10, 19] to find the most relevant and diverse set. Other than MAP, we can easily compute a credible interval of our approximation of y by drawing samples from the approximate posterior, predicting y for each draw, and computing the variance of the prediction.

In Algorithm 1, we focused on having Bernoulli prior and DPP posterior, *i.e.,* `Bernoulli-DPP`. To adapt the algorithm for `DPP-Bernoulli`, we modify $p(\mathbf{y}, \boldsymbol{\gamma})$ by changing the prior to $p(\boldsymbol{\gamma}) = \frac{\det([\mathbf{L}]_{\boldsymbol{\gamma}})}{\det(\mathbf{L}+\mathbf{I})}$ and setting the $\boldsymbol{\Phi}$ for the posterior to the identity matrix which results in *Eq.*(4). The rest of the algorithm stays intact.

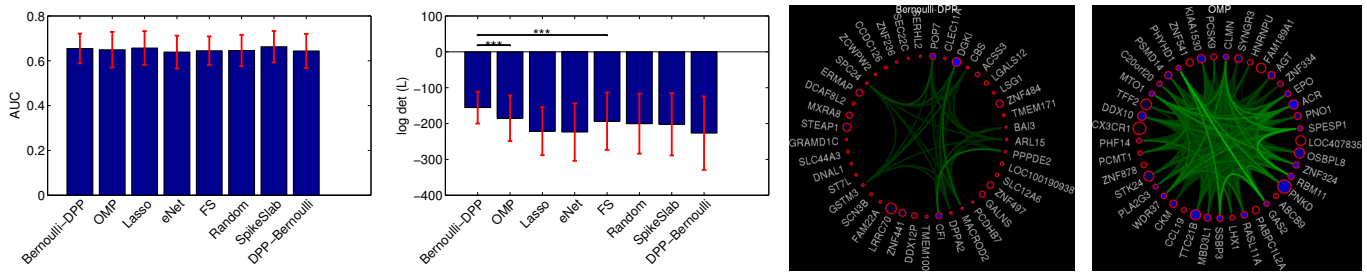
Computational Complexity: To perform the inversion in line 10 of Algorithm 1, we use conjugate gradient(CG) which has the complexity of $O(m\sqrt{k})$, where k is the condition number and m is the number of nonzero entries. We initialize the solver with warm initialization $\boldsymbol{\theta}^{t-1}$ which helps greatly (in our experience, CG converges very quickly). We currently rely on a MATLAB implementation to prove the concept; a low-rank approximation of $\mathbf{C}$ (similar to LBFGS method) should alleviate the memory complexity. If $\boldsymbol{\Phi}\boldsymbol{\Phi}'$ is low-rank (which is the case in this paper) the complexity of sampling from the DPP is reduced to $O(d^2M)$ per iteration where d is the rank of the similarity matrix and M is the number of elements. Computing the marginal likelihood for regression has a closed form that involves inversion of a matrix ($O(J^3)$ where J is the number of selected elements in each iteration). For classification, we use expectation propagation to approximate the marginal likelihood and the complexity of that is defined by J (number of selected elements in each iteration). With smart initialization from the previous iteration, EP converges very quickly. In addition, smart bookkeeping from previous iterations can reduce the number of marginal likelihood computations.

4 Experiments

We show the results for two experiments covering both classification (Section 4.1) and regression (Section 4.2); more experiments are provided in the supplementary material. While in Section 4.2, the features themselves are used to define diversity ($\boldsymbol{\Phi} = \mathbf{X}$) to penalize collinearity, in Section 4.1 we define the diversity through side information ($\boldsymbol{\Phi} \neq \mathbf{X}$). In all of our experiments, we fit the parameters of the posterior DPP and compute the MAP subset, $\mathcal{S}^*$, using [19]. Since $\mathbf{L}$ is low rank and due to the numerical scale of the optimal quality score, $\exp(\boldsymbol{\theta})$, the local optimization strategy of [10] failed, hence we use the greedy approach proposed in [19] to approximate the MAP. We compared our models to six other baseline methods: orthogonal matching pursuit (`OMP`), generalized linear model (GLM) Lasso (`Lasso`), GLM elastic net (`eNet`), forward selection (`FS`), spike-and-slab (`SpikeSlab`) [4], and using DPP as prior with a fully factorized mean field (`DPP-Bernoulli`). `Lasso` and `FS` are standard approaches for feature selection using convex and greedy optimization. Elastic net was chosen since the extra ℓ_2 norm better copes with collinearity better. `OMP` was chosen since the orthogonality procedure induces some notion of diversity for $\boldsymbol{\Phi} = \mathbf{X}$. `SpikeSlab` and `DPP-Bernoulli` assume Bernoulli and DPP priors for the inclusion of the features respectively but both deploy mean field to approximate the posterior. Parameters of the methods are adjusted to match the number of selected features with the cardinality of $\mathcal{S}^*$.

4.1 Breast cancer prognosis prediction

In this section, we turn to the motivating application of our method – finding a diverse set of genes (features) that distinguish stage I and II breast tumors. Constructing accurate classifiers will help identify important biomarkers of breast cancer progression, and furthermore, increasing the functional diversity of selected genes will more likely identify a comprehensive set of cancer-related pathways. The main idea is as follows. Gene expression profiles are the most readily available data for predicting breast cancer prognosis. However, correlation of gene expression is a relatively poor predictor of correlation in gene function [17], and so $\mathbf{X}$ is a poor feature to define functional similarity of genes. Distance between gene pairs in a protein-protein interaction (PPI) network is predictive of gene function [17]: PPI networks tend to form communities, and genes belonging to the same community perform similar functions. However, since community detection is very challenging [7], using network distance to define similarity for DPP avoids a community detection step.



(a) (b) (c) (d)

Figure 2: (a) Area under curve (AUC) performance averaged over 100 train/test repeats of classification of tumor stages for different methods. (b) Diversity of selected features, quantified as the determinant of $\mathbf{L}_{\mathcal{S}}$ where $\mathcal{S}$ is the subset of genes selected by each method (higher values indicate more diversity). DPP yields more diverse subset without compromising accuracy. The asterisks in (b) indicate statistical significance (based on p-value) using a Wilcox rank sum test. (c), (d) Networks for top 40 genes for `Bernoulli-DPP` and `OMP` respectively. The genes are sorted according to the number of times they present in the optimal set in 100 repeats. The radius is proportional to number of times the gene is selected while the color indicates the sum of L_{ij} in that gene.

We first collected 668 subjects from The Cancer Genome Atlas [20] with stage I and II breast cancer. We computed normalized expression levels for 13,876 genes for which at least one physical protein-protein interaction was found in the BioGRID database [24], then focused on the top 2,000 genes with smallest p-value (according to a likelihood ratio test for a univariate logistic regression model) with respect to the tumor stages. We then used the BioGRID gene interaction network to compute pairwise similarities between genes (features) as follows: Given the scale-free nature of the network, we first identified hubs of the network as those nodes with total degree higher than 100. For each gene i, we then defined its network profile $\mathbf{f}_i$ as a 300-dimensional vector, where each component specifies the shortest path from that gene to a hub. Our feature similarity matrix, $\mathbf{L} = \mathbf{\Phi}\mathbf{\Phi}^T$, measures similarity between genes i and j as $L_{ij} = \exp\left(-||\mathbf{f}_i - \mathbf{f}_j||^2/\sigma^2\right)$ where σ is set to 3, approximately the average pairwise distance between genes.

Figure 2a and Figure 2b show that `Bernoulli-DPP` identifies gene sets significantly more diverse than all other methods, without compromising prediction accuracy. We note that imposing DPP as an approximate posterior leads to more diverse set than having DPP as prior in `DPP-Bernoulli`. We also randomly select genes with low p-value to see if it leads to diverse set (*i.e.,* `Random` in Figure2b). Although AUC of Random is in the same range as the other method, the diversity is below `Bernoulli-DPP` and `FS`. `SpikeSlab` produces good accuracy but the selected genes are basically top genes according to p-value (not shown in the figure) and the gene set is not diverse.

We next assessed whether the diverse set of genes identified by `Bernoulli-DPP` pinpoints pathways involved in breast cancer. We first divided the 2,000 genes into 410 communities [2] using the BioGRID network. Based on 20 different cross-validation runs using all five methods, we identified 18 communities within the network that were selected more often (in at least 20% of the runs) by the `Bernoulli-DPP` than any other method. We found these DPP-preferred communities were enriched in genes related to cell cycle checkpoints, metabolism, DNA repair, predicted breast cancer gene modules, and interactors of several known cancer genes such as BRCA2, AATF, ANP32B, HDAC1, and PRKDC, among others [25]. We also observed enrichment in genes down-regulated in activated T-cells relative to naive T-cells and other immune cell types. Although the role of T-cells in tumor immunity is not fully understood, recent work has implicated immune cell activity (and T-cell infiltration in general) with breast cancer survival [12, 22] and our results both support these findings and potentially widen the set of genes that may need to be investigated further for anti-tumor properties.

4.2 Optimal Gridding of Spatial Process Convolution Models

We now demonstrate the method applied to a problem in spatial statistics, namely constructing a non-stationary Gaussian process (GP) model in a computationally efficient way. One way to construct a GP is to convolve white noise $x(s)$ by a continuous function: $z(s) = k(s) * x(s)$ $(s \in \Omega \subset \mathbb{R}^d)$. The resulting GP has covariance $\int_\Omega k(u-d)k(u)du$. Higdon *et al.*[14] suggested to define $z(s)$ to be zero mean GP and instead of defining the covariance directly, determine it implicitly through a latent white noise process $x(s)$ and smoothing kernel $k(s)$. $x(s)$ are restricted

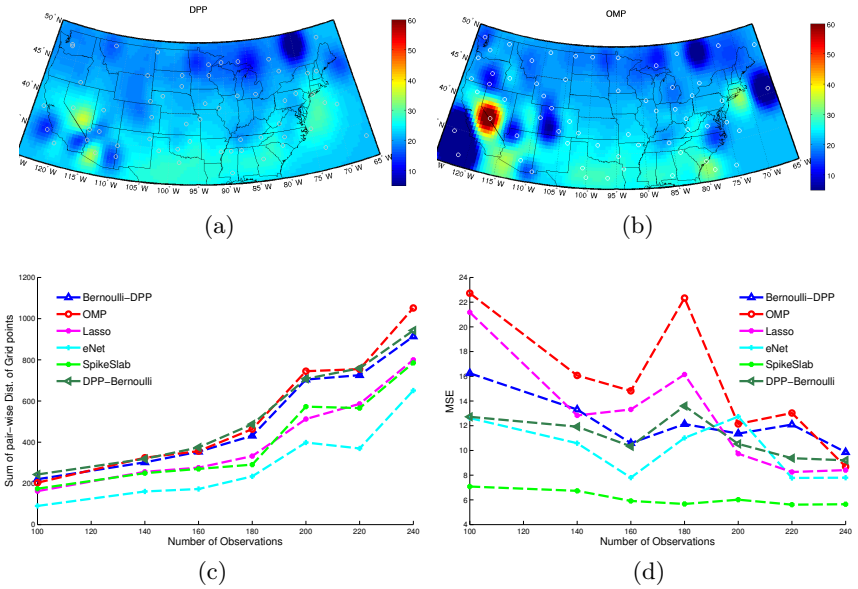


Figure 3: (a) and (b) show examples of gridding (white circles) and the prediction for `Bernoulli-DPP` and `OMP`. The grid points are spread out in both methods but `OMP` suffers from overshooting (or undershooting) prediction (*e.g.,* over California in (b)). (c) and (d) show the average of pairwise distance between selected grid points and MSE respectively for different number of measurements. Diversity promoting methods (*i.e.,* `Bernoulli-DPP`, `DPP-Bernoulli`, `OMP`) performs similarly in term of diversity while DPP-related method are slightly better in term of MSE. `SpikeSlab` and `eNet` outperform other methods in term of MSE but as shown in (c) the selected grid points are much closer to each other. `OMP` and DPP-related methods seem to have a better balance between MSE and having distributed grids particularly for a large number of measurements.

to be nonzero at the spatial sites $\omega_1, \ldots, \omega_M \in \Omega$ and each is drawn from $\mathcal{N}(\cdot; 0, \sigma_x^2)$. The resulting GP is $z(s) = \sum_{j=1}^{M} x_j k(s - \omega_m)$. One can view ω_m as (irregularly spaced) grid points. Assuming z is observed with some noise, the problem is equivalent to regression, where feature selection is equivalent to finding the optimal locations for the spatial bases.

Our objective is to find the optimal gridding of spatial domain for prediction, while also ensuring a broad spread across the spatial domain. Each grid point covers an areas but here in a 2-D domain. The grid points are the centers of the basis vectors which are isotropic Gaussian bumps on three different scales. Notice that having spatially spread out basis functions boils down to having basis functions with little overlap. Hence diversity may simply be computed as the inner product between basis vectors, i.e. $\mathbf{\Phi} = \mathbf{X}$. For this experiment, the temperature is measured for the month of July at 476 sensors located across the United States. We randomly choose varying number of sensors as a training set and evaluate the performance on the left out sensors for all methods. This procedure was repeated 20 times and we report the average performance (MSE) in Figure 3d. Figure 3c reports the average pairwise distance between the selected points. In term of MSE, both `DPP-Bernoulli` and `Bernoulli-DPP` perform better than `OMP` and slightly better than `Lasso` but `SpikeSlab` outperforms the other methods. However, as is evident in Figure (c), `SpikeSlab` does not produce spread out grid points which was the main objective. In contrast, `DPP-Bernoulli`, `Bernoulli-DPP` and `OMP` strike a good balance between prediction and diversity. Examples of the reconstructions are shown in Figure 3a for `Bernoulli-DPP` and Figure3b for `OMP`. `OMP` tends to overshoot in areas with few measurements – a trait also observed in the simulation (see the supplementary material). It is also interesting to see that when $\mathbf{\Phi} = \mathbf{X}$ having DPP as prior or approximate posterior perform similarly.

5 Conclusions

In this paper, we have proposed a probabilistic method for diverse feature selection. We proposed to approximate the posterior distribution with DPPs as a computationally elegant way to encourage diversity. Our approach selects the most representative items in communities of relevant items. Similarity between items can be encoded through the inner product between features to discourage collinearity (similar to OMP) or

may be defined based on side information (*e.g.,* Section 4.1). Our model therefore allows features and similarities to be different ($\mathbf{\Phi} \neq \mathbf{X}$). When $\mathbf{\Phi} = \mathbf{X}$, in the experiments in Section 4.2, using DPP as an approximate posterior performs similarly to using DPP as prior with mean-field approximation.

While learning the parameters of DPPs is an active research area, we have shown a computationally efficient strategy for learning the parameters in our variational approach. As far as we know, our method is the first variational method used to learn the parameters of a DPP distribution.

Our algorithm relies on sampling from the DPP, which involves a singular value decomposition (SVD) in each iteration. SVD is not very stable for matrices with very large condition numbers, hence it would be interesting to explore other parametrization of DPPs, such as those in [1]. An alternative parametrization can hopefully improve the condition number of the optimal kernel matrix $\mathbf{L}$ and improve the performance of the MAP approximation [10].

In conclusion, imposing DPP as an approximate posterior selects more diverse features without compromising the accuracy; further, it allows for sampling-based quantification of uncertainty. If the posterior distribution is multi-modal, sampling from the model can provide an alternative solution - something not possible with OMP. In fact, the simulated examples demonstrated that DPP is more robust than OMP (see supplement).

A Algorithm for Posterior Approximation

The Algorithm 1 can be used to approximate the posterior for both `DPP-Bernoulli` and `Bernoulli-DPP`. For `Bernoulli-DPP`, we set $p(\boldsymbol{\gamma}) = \prod_{m=1}^{M} \alpha^{\gamma_m}(1-\alpha)^{1-\gamma_m}$.

In case of `DPP-Bernoulli`, DPP and Bernoulli (*i.e.,* fully factorized posterior) are deployed as the prior and the approximate posterior respectively. We just need to modify $p(\boldsymbol{\gamma}) = \frac{\det([\mathbf{L}]_{\boldsymbol{\gamma}})}{\det(\mathbf{L}+\mathbf{I})}$ and $\mathbf{\Phi} = \mathbf{I}$.

Algorithm 1: Variational Learning for Diverse Variable Selection

Input: Similarity features $\mathbf{\Phi}$, a function to compute/approximate the restricted marginal likelihood $p(y|\boldsymbol{\gamma})$, $p(\boldsymbol{\gamma})$, initial cardinality of DPP κ , number of iterations: N.

Output: Parameters of the posterior (q): $\boldsymbol{\theta}$

Adjust expected cardinality of the DPP by solving for θ_0 in $\sum_i \frac{e^{\theta_0}\lambda_i}{1+\lambda_i e^{\theta_0}} = \kappa$;

Initialize $\boldsymbol{\theta} = \theta_0 \mathbb{1}$, $\mathbf{L} = e^{\theta_0/2}\mathbf{\Phi}\mathbf{\Phi}^T e^{\theta_0/2}$, and $\mathbf{K} = \mathbf{L}(\mathbf{L}+\mathbf{I})^{-1}$;

Set $(\mathbf{C}_1)_{ii} = K_{ii}$, $\mathbf{g}_1 = \mathbf{C}_1\tilde{\boldsymbol{\theta}}$, and $\bar{\mathbf{C}}, \bar{\mathbf{g}} = 0$;

for $t \leftarrow 1$ ***to*** N **do**

  Draw a set from current posterior approximation DPP: $\boldsymbol{\gamma}_t^* \sim q_{\tilde{\boldsymbol{\theta}}_t}$;

  Set $\hat{\mathbf{g}}_t = \tilde{\boldsymbol{\gamma}_t^*} \log p(y|\boldsymbol{\gamma}_t^*) p(\boldsymbol{\gamma}_t^*)$;

  Set $\hat{\mathbf{C}}_t = \tilde{\boldsymbol{\gamma}_t^*}\tilde{\boldsymbol{\gamma}_t^*}^T$ or current estimate of $\mathbf{K}_{\theta_t}$;

  Set $\mathbf{g}_{t+1} = (1-w)\mathbf{g}_t + w\hat{\mathbf{g}}_t$;

  Set $\mathbf{C}_{t+1} = (1-w)\mathbf{C}_t + w\hat{\mathbf{C}}_t$;

  Solve $\tilde{\boldsymbol{\theta}}_{t+1} = \mathbf{C}_{t+1}^{-1}\mathbf{g}_{t+1}$;

  if $t > N/2$ **then**

    Set $\bar{\mathbf{g}} = \bar{\mathbf{g}} + \hat{\mathbf{g}}_t$;

    Set $\bar{\mathbf{C}} = \bar{\mathbf{C}} + \hat{\mathbf{C}}_t$;

return $\boldsymbol{\theta} = \bar{\mathbf{C}}^{-1}\bar{\mathbf{g}}$;

*Bibliography

[1] R. H. Affandi, E. B. Fox, R. P. Adams, and B. Taskar. Learning the Parameters of Determinantal Point Process Kernels, February 2014.

[2] YY Ahn, JP Bagrow, and S Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2006.

[3] M. Beal and Z. Ghahramani. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, 7:1–10, 2003.

[4] P. Carbonetto, M. Stephens, et al. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108, 2012.

[5] B. P Carlin and T. A Louis. *Bayes and empirical Bayes methods for data analysis*. CRC Press, 2008. ISBN ISBN 1-58488-697-8.

[6] A. Das and D. Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

[7] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[8] M. Garcia-Closas, P. Hall, H. Nevanlinna, K. Pooley, J. Morrison, D. A Richesson, S. E Bojesen, B. G Nordestgaard, C. K Axelsson, J. I Arias, et al. Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS genetics*, 4(4):e1000054, 2008.

[9] Edward I George et al. Dilution priors: Compensating for model space redundancy. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pages 158–165. Institute of Mathematical Statistics, 2010.

[10] J. Gillenwater, A. Kulesza, and B. Taskar. Near-Optimal MAP Inference for Determinantal Point Processes. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *NIPS*, pages 2744–2752, 2012.

[11] E. Grave, G. Obozinski, F. Bach, et al. Trace Lasso: a trace norm regularization for correlated designs. In *NIPS*, volume 2, page 5, 2011.

[12] C et al. Gu-Trantien. CD4+ follicular helper T cell infiltration predicts breast cancer survival. *J. Clin. Invest.*, 123(7):2873–2892, 2013.

[13] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[14] D. Higdon. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, 5(2):173–190, 1998.

[15] A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen. Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *The Journal of Machine Learning Research*, 9999:3235–3268, 2010.

[16] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

[17] I. Lee, U M. Blom, P. I Wang, J. Eun Shim, and E. M Marcotte. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research*, 21(7):1109–1121, 2011.

[18] T. J Mitchell and J. J Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

[19] G. L Nemhauser, L. A Wolsey, and M. L Fisher. An analysis of approximations for maximizing submodular set functions - I. *Mathematical Programming*, 14 (1):265–294, 1978.

[20] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[21] Y. Chandra Pati, R. Rezaiifar, and PS Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44. IEEE, 1993.

[22] I Peguillet, M Milder, et al. High numbers of differentiated effector CD4 T cells are found in patients with cancer and correlate with clinical response after neoadjuvant therapy of breast cancer. *Cancer Res.*, 74(8):2204–2216, 2014.

[23] T. Salimans and D. A. Knowles. Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression. *CoRR*, abs/1206.6679, 2012.

[24] C Stark, BJ Breitkreutz, T Reguly, L Boucher, A Breitkreutz, and A Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34(D): 535–539, 2006.

[25] A Subramanian, P Tamayo, VK Mootha, S Mukherjee, BL Ebert, MA Gillette, A Paulovich, SL Pomeroy, TR Golub, ES Lander, and JP Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.

[26] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[27] J. A Tropp and A. C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53 (12):4655–4666, 2007.

[28] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2):301–320, 2005.