Determinantal Regularization for Ensemble Variable Selection

Veronika Ročková **Gemma E. Moran** **Edward I. George**
Department of Statistics, University of Pennsylvania, Philadelphia, PA

Abstract

Recent years have seen growing interest in deterministic search approaches to spike-and-slab Bayesian variable selection. Such methods have focused on the goal of finding a global mode to identify a "best model". However, the report of a single model will be a misleading reflection of the model uncertainty inherent in a highly multimodal posterior. Motivated by non-parametric variational Bayes strategies, we move beyond this limitation by proposing an ensemble optimization approach to identify a collection of representative posterior modes. By deploying determinantal penalty functions as diversity regularizers, our approach performs regularization over multiple locations of the posterior. The key driver of these determinantal penalties is a kernel function that induces repulsion in the latent model space domain.

1 Introduction

In the presence of model uncertainty, reporting a set of representative models will be more meaningful than reporting a single "best model". With posterior simulation approaches, such a summary of post-data variable selection uncertainty is conveyed by reporting the most frequent or probable models occurring along an MCMC path. However, the scope of such posterior simulations is ultimately challenged in the domain of high-dimensional data. As a viable alternative for large problems, fast deterministic alternatives (such as EM algorithms) have seen increasing popularity (Ročková and George, 2014; Ormerod et al., 2014). But such approaches have been limited to single mode detection algorithms which, by outputting only a single point estimate, fail to convey underlying model uncertainly. Variational approximations, a further alternative, suffer from local entrapment in multimodal posterior landscapes. Such multimodality is an inevitable manifestation of model uncertainty and should be properly accounted for. Here, we propose a new computational approach that yields an ensemble of representative high-probability models. Motivated by similarities with a non-parametric variational Bayesian approach (Jaakkola and Jordan, 1998; Gershman et al., 2012), our procedure performs a joint regularization over multiple locations of the posterior. This strategy is very different from independently initialized mode hunting algorithms, which do not have the opportunity to mutually interact. Forcing regularization trajectories to repel each other (preventing hill-climbing towards the same modes), our procedure is more effective (a) in finding a diverse/representative set of models and (b) in finding the global mode. The repulsion effect is achieved with new determinantal penalties which track the regions of attraction of individual posterior modes.

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

2 Spike-and-Slab Variable Selection

We consider the classical linear regression model

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is a vector of responses, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ is a matrix of p standardized predictors and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression coefficients. For simplicity of exposition, we assume σ^2 is known and equal to one; the unknown variance case may be easily incorporated as in EMVS (Ročková and George, 2014). We introduce binary latent variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)' \in \{0,1\}^p$, where $\gamma_i = 1$ if $\mathbf{x}_i$ is included in the model.

To avoid the obstinate combinatorial problem of mode finding in the vast model space of 2^p latent states $\boldsymbol{\gamma}$, we will work instead in the continuous space of model parameters $\boldsymbol{\beta}$. To this end, we consider a continuous spike-and-slab mixture of two Gaussian distributions:

$$\pi(\beta_i \mid \gamma_i) = \gamma_i \phi(\beta_i \mid v_1) + (1-\gamma_i)\phi(\beta_i \mid v_0) \tag{2}$$

where $\phi(\beta \mid v)$ is a Gaussian distribution centered at zero with variance v. Here, we have fixed $0 < v_0 < v_1$ so that $\gamma_i = 1$ indicates the β_i which are likely to be large, George and McCulloch (1993); note that one may take v_1 as random as in EMVS. As for $\boldsymbol{\gamma}$, we consider the iid Bernoulli form $\pi(\boldsymbol{\gamma}|\theta) = \theta^{|\gamma|}(1-\theta)^{p-|\gamma|}$ with $\theta \in (0,1)$.

We will infer about the model space posterior $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ indirectly through the posterior $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$. Note that under this spike-and-slab setup, $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$ is a Gaussian mixture indexed by 2^p latent states $\boldsymbol{\gamma}$,

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{Y}) = \sum_{\boldsymbol{\gamma}} \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) \pi(\boldsymbol{\gamma} \mid \boldsymbol{Y}), \tag{3}$$

where $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \boldsymbol{X}'\boldsymbol{Y}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_{\boldsymbol{\gamma}})^{-1}$ and $\boldsymbol{D}_{\boldsymbol{\gamma}} = \text{diag}\{\gamma_i \frac{1}{v_1} + (1-\gamma_i)\frac{1}{v_0}\}_{i=1}^p$. High posterior modes in the parameter domain $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$ are associated with high posterior modes $\boldsymbol{\gamma}$ in the model domain $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$. The transition from the parameter space to the model space can be achieved with the following probabilistic consideration. Having identified a mode $\widehat{\boldsymbol{\mu}}$ of $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$, the most likely model $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p)'$ associated with this mode is

$$\widehat{\gamma}_i = 1 \quad \text{iff} \quad p^\star(\widehat{\mu}_i) \equiv \mathsf{P}(\gamma_i = 1 \mid \widehat{\mu}_i, \theta) > 0.5. \tag{4}$$

As noted in the introduction, deterministic approaches to this problem so far have involved unimodal approximations or single mode hunting algorithms. Such approximations will often be inadequate given the Gaussian mixture form of the posterior. Prior to developing our approach in Section 4, we revisit a multimodal variational approximation to $\pi(\boldsymbol{\beta}|\mathbf{Y})$ in the next section.

3 Non-parametric Variational Bayes

Variational methods seek to approximate the posterior $\pi(\boldsymbol{\beta}|\mathbf{Y})$ with a variational density $q(\boldsymbol{\beta})$. The approach stems from the following lower-bound argument:

$$\log \pi(\mathbf{Y}) \geq \int_{\boldsymbol{\beta}} q(\boldsymbol{\beta}) \log \frac{\pi(\mathbf{Y}, \boldsymbol{\beta})}{q(\boldsymbol{\beta})} d\boldsymbol{\beta} \equiv \mathcal{F}[q(\cdot)] \tag{5}$$

where $\mathcal{F}[q(\cdot)]$ is known as the evidence lower bound. The form of $q(\boldsymbol{\beta})$ is typically constrained to a family that ensures (5) is tractable. The parameters of $q(\boldsymbol{\beta})$ are then found by maximizing (5), which is equivalent to minimizing the Kullback-Leibler distance between $\pi(\boldsymbol{\beta}|\mathbf{Y})$ and $q(\boldsymbol{\beta})$. For our variable selection problem, $q(\cdot)$ is often augmented to include also the latent indicators $\boldsymbol{\gamma}$ and constrained to be of the product form $q(\boldsymbol{\beta}, \boldsymbol{\gamma}) = q(\boldsymbol{\beta}) \prod_{j=1}^p q_j(\gamma_j)$, also known as a mean field approximation. Such strategy is very similar to the EMVS algorithm. However, this variational approximation hinges on the often inadequate assumptions that the density $q(\boldsymbol{\beta})$ is unimodal and that the $\gamma_j's$ are independent.

Jaakkola and Jordan (1998) considered a more flexible non-parametric variational approach using mixtures as approximating distributions. This approach was further developed by Gershman et al. (2012) who used Gaussian mixtures, noting that any density can be approximated arbitrarily closely with a sufficient number of Gaussian densities.

Here, we implement the non-parametric variational Bayes approach for spike-and-slab variable selection. Following Gershman et al. (2012), we seek to approximate $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$ by the variational density:

$$q_M(\boldsymbol{\beta}) = \frac{1}{K} \sum_{k=1}^K \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_k, \sigma_k^2 \mathbf{I}_p). \tag{6}$$

Note that here we have a whole matrix of K unknown location vectors, $\mathbf{M} = [\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K] = (\mu_{jk})_{j,k=1}^{p,K}$. Additionally, we have margined out the K binary indicator vectors $\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_K$; these indicators will be reintroduced later to facilitate optimization of the objective.

With (6), the evidence lower bound (5) becomes

$$\begin{aligned} \mathcal{F}[q_M(\cdot)] &= \int_{\boldsymbol{\beta}} q_M(\boldsymbol{\beta}) \log \left[\frac{\pi(\boldsymbol{Y}, \boldsymbol{\beta})}{q_M(\boldsymbol{\beta})} \right] \mathrm{d}\boldsymbol{\beta} \\ &= E[f(\cdot)] + \mathcal{H}[q_M], \end{aligned} \tag{7}$$

where $Ef(\cdot) = \int_{\boldsymbol{\beta}} \log \pi(\mathbf{Y}, \boldsymbol{\beta}) q_M(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta}$ and $\mathcal{H}[q_M] = -\int_{\boldsymbol{\beta}} q_M(\boldsymbol{\beta}) \log q_M(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta}$ is the entropy of the mixture distribution (6). The goal is then to maximize the lower bound $\mathcal{F}[q_M(\cdot)]$ with respect to the unknown modes $\mathbf{M}$ and variances $\sigma_1^2, \ldots, \sigma_K^2$.

The first summand in (7) represents the contribution of the posterior, where trajectories $\boldsymbol{\mu}_j$ should gravitate towards regions with high posterior mass. The entropy $\mathcal{H}[q_M]$ serves as a diversifying penalty, moving the trajectories far away from each other. However, there is no closed form expression for either of these terms; the next sections focus on discussing suitable approximations.

3.1 The Posterior Contribution

We now take a closer look at the first term of the evidence lower bound (7). We note that

$$E[f(\cdot)] = \frac{1}{K} \sum_{k=1}^K \int_{\boldsymbol{\beta}} \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_k, \sigma_k^2 \mathbf{I}_p) \log \pi(\mathbf{Y}, \boldsymbol{\beta}) d\boldsymbol{\beta}.$$

Following Gershman et al. (2012), we approximate each integral with a second order Taylor expansion

around $\boldsymbol{\mu}_k$ as follows:

$$\log \pi(\mathbf{Y}, \boldsymbol{\beta}) \approx \log \pi(\mathbf{Y}, \boldsymbol{\mu}_k) + \nabla \log \pi(\mathbf{Y}, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\mu}_k}(\boldsymbol{\beta} - \boldsymbol{\mu}_k) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_k)'\mathbf{H}_k(\boldsymbol{\beta} - \boldsymbol{\mu}_k). \quad (8)$$

where $\mathbf{H}_k = \nabla^2 \log \pi(\mathbf{Y}, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\mu}_k}$ is the Hessian matrix of second derivatives. The approximate expectation is

$$E[f(\cdot)] \approx \frac{1}{K}\sum_{k=1}^{K}\left[\log \pi(\boldsymbol{Y}, \boldsymbol{\mu}_k) + \frac{\sigma_k^2}{2}\mathrm{tr}(\mathbf{H}_k)\right] \quad (9)$$

The calculation of the Hessian can be found in the supplementary material.

The first summand in the approximation (9) is a separable function of unknown location vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$. Finding $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ which maximize this term is equivalent to finding K highest local modes of the posterior $\log \pi(\boldsymbol{\beta}|\boldsymbol{Y})$. The second term in (9) penalizes high posterior peaks with low volume (Gershman et al., 2012).

3.2 Entropy as a Diversifying Penalty

The entropy term $\mathcal{H}[q_M]$ in the lower bound (7) imposes extra regularization on $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, forcing the newfound modes to be far apart. This repulsion avoids unnecessary overlap between the approximating mixture components.

To simplify manipulations with $\mathcal{H}[q_M]$, Gershman et al. (2012) derived the lower-bound

$$\mathcal{H}[q_M] \geq -\frac{1}{K}\sum_{k=1}^{K}\log h_k, \quad (10)$$

where $h_k = \frac{1}{K}\sum_{l=1}^{K}\mathcal{N}(\boldsymbol{\mu}_k; \boldsymbol{\mu}_l, (\sigma_k^2 + \sigma_l^2)\mathbf{I})$. We use this lower bound in our implementation of the Non-Parametric Variational Bayes Variable Selection (NPVS) strategy introduced below.

We first illuminate the repulsive penalization effect of the entropy with a one-dimensional example where $p = 1$ and $K = 2$ (Figure 1(a)). For simplicity, we assume an isotropic mixture with $\sigma_1 = \sigma_2 = 1$. The entropy $\mathcal{H}[q_M]$ is depicted as a function of the first location μ_1 while keeping $\mu_2 = 2$. Expectedly, the entropy has a unique minimum at the point $\mu_1 = \mu_2$, encouraging μ_1 to move away from μ_2 by penalizing the basin centered around μ_2. More generally, with K mixture components, the entropy as a function of the first argument has $K-1$ global minima located at $\mu_2, \ldots, \mu_K$. Figure 1(a) also displays the lower bound (10) and a first order approximation, both of which have a similar monotonicity pattern.

Interestingly, a similar repulsive effect is achieved within the probabilistic framework of determinantal point processes, which we will discuss in Section 4.

3.3 Optimization Strategy

With the proposed approximations to the two terms in (7), we are ready to outline the NPVS strategy. We now seek to maximize the approximate evidence lower bound

$$\mathcal{L}[q_M(\cdot)] = \frac{1}{K}\sum_{k=1}^{K}\left[\log \pi(\mathbf{Y}, \boldsymbol{\mu}_k) + \frac{\sigma_k^2}{2}\mathrm{tr}(\mathbf{H}_k) - \log h_k\right].$$

Note that both h_k and $\mathrm{tr}(\mathbf{H}_k)$ depend on $\boldsymbol{\mu}_k$. Maximizing $\mathcal{L}[q_M(\cdot)]$ with respect to $\boldsymbol{M}$ is hampered by the unavailability of $\pi(\mathbf{Y}, \boldsymbol{\mu}_k)$. The enumeration of this term is infeasible due to the summation over all the 2^p hidden states of $\boldsymbol{\gamma}$. As in Ročková and George (2014), we proceed indirectly by augmenting $\mathcal{L}[q_M(\cdot)]$ with latent indicators $\boldsymbol{\gamma}$. But unlike Ročková and George (2014), here we have a matrix of hidden indicators $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K]$, one for each trajectory $\boldsymbol{\mu}_k$. Maximizing $\mathcal{L}[q_M(\cdot)]$ proceeds iteratively, updating $\boldsymbol{\mu}_k$ while keeping all the other vectors $\boldsymbol{\mu}_j, j \neq k$ fixed. Using an EM step within the variational approximation, we proceed to maximize a lower bound

$$\mathcal{Q}_{NPV}[q_M] \equiv \frac{1}{K}\sum_{k=1}^{K}\Big\{E_{\boldsymbol{\gamma}_k}[\log \pi(\mathbf{Y}, \boldsymbol{\mu}_k, \boldsymbol{\gamma}_k)] + \frac{\sigma_k^2}{2}\mathrm{tr}(\mathbf{H}_k) - \log h_k\Big\}, \quad (11)$$

where $E_{\boldsymbol{\gamma}_k}$ is the conditional expectation w.r.t $\boldsymbol{\gamma}_k$ given $\boldsymbol{\mu}_k$. This expectation greatly simplifies by noting

$$E_{\boldsymbol{\gamma}_k}[\log \pi(\mathbf{Y}, \boldsymbol{\mu}_k, \boldsymbol{\gamma}_k)] = -\frac{1}{2}\left[\|\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_k\|^2 + \boldsymbol{\mu}_k'\mathbf{D}_k^\star\boldsymbol{\mu}_k\right]$$

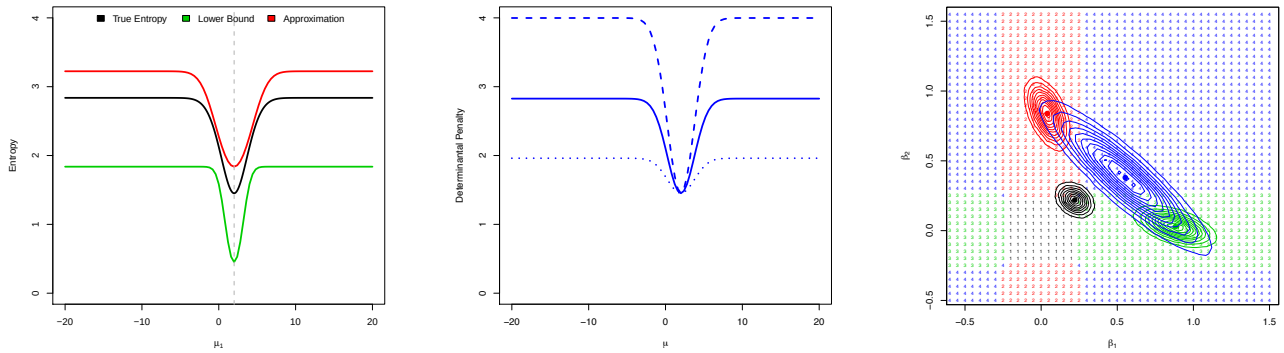
where $\mathbf{D}_k^\star = \mathrm{diag}\{d_{jk}^\star\}_{j=1}^p$ with

$$d_{jk}^\star = \frac{1}{v_1}p^\star(\mu_{jk}) + \frac{1}{v_0}[1 - p^\star(\mu_{jk})]$$

and where $p^\star(\mu)$ was defined in (4). The surrogate objective function (11) now simplifies to

$$\mathcal{Q}_{NPV}[q_M] = \frac{1}{K}\sum_{k=1}^{K}\Big\{-\frac{1}{2}\left[\|\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_k\|^2 + \boldsymbol{\mu}_k'\mathbf{D}_k^\star\boldsymbol{\mu}_k\right] + \frac{\sigma_k^2}{2}\mathrm{tr}(\mathbf{H}_k) - \log h_k\Big\}. \quad (12)$$

The details of the optimization strategy are given in Algorithm 1. The optimization is done using standard L-BFGS. It is interesting to note that with $K = 1$, the procedure resembles the EMVS procedure of Ročková



(a) Entropy **(b)** Determinantal penalty **(c)** Regions of attraction

Figure 1: A plot of the entropy $\mathcal{H}(q_M)$ as a function of μ_1 together with the first order approximation and lower bound (left). A plot of a determinantal penalty (middle). Convergence regions for EMVS.

and George (2014). We use Algorithm 1 for comparisons with our new approach developed in the next section.

The NPVS strategy boils down to performing an EM-like computation simultaneously for K trajectories, forcing them to be far apart with the entropy function. Although intuitively appealing, additional computation of the Hessian and approximations to the entropy are required. Moreover, the isotropic approximating mixture is at odds with the known topography of the posterior. From (3) we know that the posterior modes are ellipsoids with a covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\gamma}}$. An illustration of the posterior is in Figure 1(c), where we have $p = 2$ highly collinear predictors and 4 true posterior modes. NPVS will be prone to wasting many small variance mixture components approximating a single ellipsoidal peak in the posterior. Ideally, we would like to spend only one mixture component for each peak/model. We now proceed to develop an alternative regularization approach for multimodal posterior exploration which makes use of the geometry of the posterior.

Algorithm 1: NPVS

Input: Data: $\boldsymbol{Y}, \mathbf{X}$; Tuning parameters: K, θ
Initialize: $[\boldsymbol{\mu}_1^{(0)}; \cdots \boldsymbol{\mu}_K^{(0)}] \in \mathbb{R}^{p\times K}, \sigma_1^{(0)}, \ldots, \sigma_K^{(0)}$.
Repeat
  for $k = 1$ **to** K,
    E-step: for each $i = 1$ to p, calculate $d^\star(\mu_{jk})$.
    M-step: calculate
    $\boldsymbol{\mu}_k = \arg\max_{\boldsymbol{\mu}_k \in \mathbb{R}^p} \mathcal{Q}_{NPV}[q_M]$[1]
    $\sigma_k^2 = \arg\max_{\sigma_k^2 \in \mathbb{R}^p} \mathcal{Q}_{NPV}[q_M]$
Until change in $\mathbf{M}$ is less than $\varepsilon = 10^{-5}$.

[1]Following Gershman et al. (2012), the term $\frac{\sigma_k^2}{2}\text{tr}(\mathbf{H}_k)$ is omitted in this step.

4 Determinantal Point Processes as Repulsive Priors

Zou and Adams (2012) presented strategies for diversified latent variable modelling using determinantal process priors as regularizers in MAP estimation. Here, we pursue this strategy further in the context of Bayesian variable selection. Robert and Mengersen (2003) implemented an MCMC variant of our approach with different repulsive priors.

As before, we want to estimate a set of location parameters $\mathbf{M} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K]$ and we want them to be far apart in some suitable sense. Having a $K \times K$ positive definite kernel matrix $K(\boldsymbol{M})$ which quantifies pairwise distances between $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$, a useful aggregate measure of diversity encoded in $K(\boldsymbol{M})$ is the determinant $|K(\boldsymbol{M})|$. The determinant quantifies the volume of a parallelepiped delineated by $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, where large volumes are associated with more diverse sets of vectors. The kernel $K(\cdot)$ can be chosen based on various context-specific considerations, encouraging repulsion in different metric spaces.

4.1 Repulsion in the Parameter Space

Regarding $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ as location parameters of a Gaussian mixture, a suitable kernel function is the probability product Gaussian kernel (Affandi et al., 2014). This kernel was used for Gaussian mixture modeling by Zou and Adams (2012). In the case of an isotropic covariance $\mathbf{\Sigma}_k = \mathrm{I}_p$, this kernel simplifies to

$$k(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) \propto \exp\left(-||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||^2/4\right) \qquad (13)$$

The corresponding determinant of the kernel can be rescaled to become a prior distribution over $\boldsymbol{M} =$

$[\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K] \in \mathbb{R}^{p \times K}$

$$\pi_{DET}(\boldsymbol{M}) \propto |K(\boldsymbol{M})|. \tag{14}$$

It is worthwhile to note that the prior (14) is closed under conditioning. To see this, we first define

$$K([\mathbf{x}, \boldsymbol{\mu}]) = \begin{bmatrix} K_{\mathbf{xx}} & K_{\mathbf{x}\boldsymbol{\mu}}^T \\ K_{\mathbf{x}\boldsymbol{\mu}} & K(\boldsymbol{\mu}) \end{bmatrix} \tag{15}$$

for $\mathbf{x} \in \mathbb{R}^p$ and $\boldsymbol{\mu} = [\boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K] \in \mathbb{R}^{p \times (K-1)}$, where $K_{\mathbf{xx}}$ is a scalar, $K_{\mathbf{x}\boldsymbol{\mu}}^T = [k(\mathbf{x}, \boldsymbol{\mu}_2), \cdots, k(\mathbf{x}, \boldsymbol{\mu}_K)]$ is a $1 \times (K-1)$ matrix. Using the Schur complement identity for determinants, we immediately obtain

$$\pi_{DET}([\mathbf{x}, \boldsymbol{\mu}]) \propto |K(\boldsymbol{\mu})|(K_{\mathbf{xx}} - K_{\mathbf{x}\boldsymbol{\mu}}^T K(\boldsymbol{\mu})^{-1} K_{\mathbf{x}\boldsymbol{\mu}}). \tag{16}$$

Therefore, $\pi_{DET}(\mathbf{x} \,|\, \boldsymbol{\mu})$ is proportional to the determinant of the Schur complement.

Here, we will use the determinantal point process prior (14) to define a penalty for repulsive regularization. For $\boldsymbol{M} \in \mathbb{R}^{p \times K}$, we have

$$Pen_K(\boldsymbol{M}) = \lambda \log |K(\boldsymbol{M})|, \tag{17}$$

where $K(\boldsymbol{M})$ is the $K \times K$ positive definite kernel matrix. To illuminate the role of the determinantal penalty, it is useful to regard $Pen_K([\mathbf{x}, \boldsymbol{\mu}])$ as a function of the first vector for a fixed set $\boldsymbol{\mu}$, using again the Schur formula (16).

Figure 1(b) depicts the conditional determinantal distribution (16) for the isotropic kernel (13) with $K = 2$. The penalty is plotted as a function of the first argument μ_1, keeping $\mu_2 = 2$. For comparison with Figure 1(a), we rescaled the determinantal penalty so that it has the same minimal value as the entropy $\mathcal{H}(q_M)$. Figure 1(b) depicts three penalties obtained with different scaling factors λ. This parameter determines the amount of penalization for values far away from the minimal point $\mu_1 = \mu_2$ and can be matched so that the level equals the asymptote of the entropy term (Figure 1(a)). The solid curve in the Figure 1(b) shows one particular value λ, which yields a very close approximation to the entropy term (black curve in Figure 1(a)). Similar close approximations can be obtained also in higher dimensions.

It is known that the entropy of a *single* Gaussian distribution with a covariance matrix $\boldsymbol{\Sigma}$ is a linear function of $\log |\boldsymbol{\Sigma}|$. However, the interesting similarity between the log determinant of the Gaussian kernel (13) and the entropy of the Gaussian mixture is new to us, as we could not find any such connection in the literature.

The repulsive penalty $\lambda \log |K(\boldsymbol{M})|$ will ultimately be deployed to locally distort the objective function $\log \pi(\boldsymbol{\beta}|\boldsymbol{Y})$, erasing its modes from the posterior landscape. The calculation will again proceed iteratively, updating each location $\boldsymbol{\mu}_j$ conditionally on the location of $\boldsymbol{\mu}_k, k \neq j$. To further illustrate the penalty effect, assume $p = 2$. Conditionally on the location $\boldsymbol{\mu}_1 = (0, 2)'$, the determinantal penalty $\log \pi_{DET}(\mathbf{x} \mid \boldsymbol{\mu}_1)$ (Figure 2(a)) creates a circular hole around $\boldsymbol{\mu}_1$. With $K = 2$ and $\boldsymbol{\mu}_1 = (2, 0)', \boldsymbol{\mu}_2 = (2, 0)'$, the penalty now contains two holes and so on.

4.2 Repulsion in the Model Space

The "Swiss-cheese" Gaussian-kernel determinantal penalty in Figure 2(a) will erase mass in circular areas around the posterior modes. Similarly as with the entropy of an isotropic Gaussian mixture, this will not be ideal with ellipsoidal modes such as in Figure 1(c). Since we are ultimately interested in models $\boldsymbol{\gamma}_k$ that underpin each mode $\boldsymbol{\mu}_k$, it may be more appropriate to use a kernel $K(\boldsymbol{M})$ that reflects the distance between models as opposed to modes.

From the geometry of the spike-and-slab posterior (Figure 1(c)), we know that there is only one mode for each model. Each model is associated with a box-shaped region in $\mathbb{R}^p$, a domain of attraction of EMVS from where all initializations land in the same mode. Due to the binary nature of the indicators $\boldsymbol{\gamma}$, the parameter domain $\mathbb{R}^p$ can be regarded as a tessellation of boxes. Instead of considering elliptical or circular penalties, we use box-shaped penalties, reflecting the geometry of the convergence regions.

Our determinantal penalty in the probability domain is again (17) with a new kernel defined as

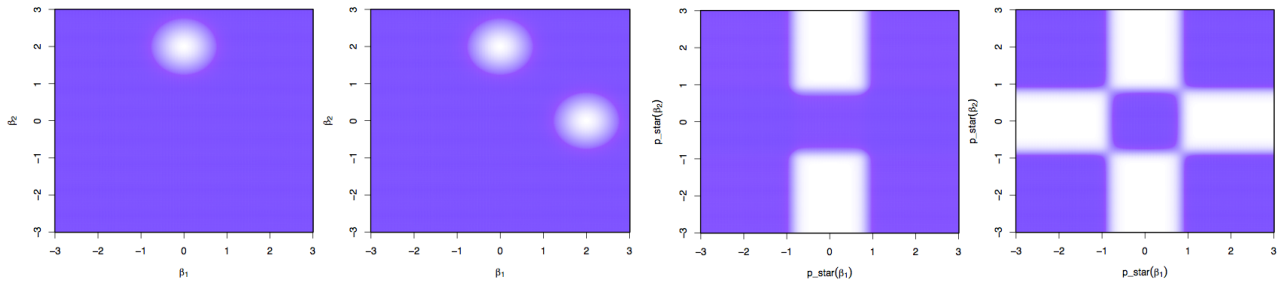
$$\widetilde{k}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) \propto \exp\left(-||\boldsymbol{p}^*(\boldsymbol{\mu}_i) - \boldsymbol{p}^*(\boldsymbol{\mu}_j)||^2/4\right), \tag{18}$$

where $\boldsymbol{p}^*(\boldsymbol{\mu}_i) = [p^*(\mu_{i1}), \ldots, p^*(\mu_{i1})]'$ and $p^*(\cdot)$ is defined in (4). The vectors $\boldsymbol{p}^*(\boldsymbol{\mu}_i)$ are a continuous version of the underlying model $\boldsymbol{\gamma}_i$. Thus (18) reflects the distance between models rather than the distance between parameters. Figure 2(b) exemplifies this effect.

Similarly as in Figure 2(a), suppose that we condition on a location $\boldsymbol{\mu}_1 = (0, 2)'$. The underlying model is clearly $\boldsymbol{\gamma}_1 = (0, 1)'$. The determinantal penalty now erases the entire box of associated with $\boldsymbol{\gamma}_1$ (Figure 2(b)): the first coordinate is small and the second coordinate is large in absolute value. Adding the mode $\boldsymbol{\mu}_2 = (2, 0)'$, the penalty deletes another box and so on.

5 Determinantal Regularization for Ensemble Variable Selection

In the previous section, we illustrated how a determinantal penalty in the model space is ideal for diversification of posterior modes for the problem of variable



(a) Gaussian kernel **(b)** Probability kernel

Figure 2: A plot of the conditional determinantal penalty $\log \pi_{DET}(\mathbf{x} \mid \boldsymbol{\mu})$ for different kernels. The darker the area the smaller the penalty. Left graphs assume $\boldsymbol{\mu} = \boldsymbol{\mu}_1 = (0, 2)'$, right graphs assume $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2]$ where $\boldsymbol{\mu}_1 = (0, 2)'$, $\boldsymbol{\mu}_2 = (2, 0)'$.

selection. In the development of the NPVS algorithm, we discussed how the objective function comprises of a posterior contribution and an entropy term encouraging diversity among the modes. In light of this, we now develop a new algorithm whereby we replace the Hessian and entropy terms in (12) with a determinantal penalty to obtain a new objective function:

$$\mathcal{Q}_{DR}(\mathbf{M}) = \frac{1}{K}\sum_{k=1}^{K} -\frac{1}{2}\left[\|\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_k\|^2 + \boldsymbol{\mu}_k' \mathbf{D}_k^{\star} \boldsymbol{\mu}_k\right] + \lambda \log|\widetilde{K}(\mathbf{M})|, \tag{19}$$

where $\lambda|\widetilde{K}(\mathbf{M})|$ is the diversity penalty with the kernel (18). In effect, we search for an estimate of a collection of diverse posterior modes $\widehat{\mathbf{M}}$, which jointly maximize the weighted sum of the posterior peaks, that is $\widehat{\mathbf{M}} = \arg\max_{\mathbf{M}\in\mathbb{R}^{p\times K}} Q_{DR}(\mathbf{M})$. We proceed conditionally, updating one mode at a time, conditioning on the location of the remaining particles. This Determinantal Regularization Ensemble Variable Selection (DREVS) algorithm is outlined in Algorithm 2. Note that with $\lambda = 0$, this strategy corresponds to running EMVS independently from starting locations $\mathbf{M}^{(0)}$.

Algorithm 2: DREVS

Input: Data: $\boldsymbol{Y}, \mathbf{X}$; Tuning parameters: K, θ, λ
Initialize $\mathbf{M}^{(0)} = [\boldsymbol{\mu}_1^{(0)}; \cdots \boldsymbol{\mu}_K^{(0)}] \in \mathbb{R}^{p\times K}$.
Repeat
  for $k = 1$ **to** K,
    E-step: for each $j = 1$ to p, calculate $d^{\star}(\mu_{jk})$.
    M-step: calculate
    $\boldsymbol{\mu}_k = \arg\max_{\boldsymbol{\mu}_k\in\mathbb{R}^p} \mathcal{Q}_{DR}(\mathbf{M})$
Until change in $\mathbf{M}$ is less than $\varepsilon = 10^{-5}$.

The DREVS strategy is illustrated in Figure 3. There we compare three independent EMVS initializations (right) which gravitate towards the same mode, and the ensemble implementation (left) which spreads out towards three different modes.

6 Numerical examples

In this section, we apply both NPVS and DREVS to both simulated and real data.

6.1 Highly correlated predictors

We first examine the performance on a simulated dataset where the predictors are highly correlated in blocks, resulting in a highly multimodal posterior. The dataset consists of $n = 50$ observations with $p = 16$ predictors, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{16}$ and response Y. The predictors are generated as $\mathbf{X} \sim N_p(\mathbf{0}, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \Lambda_{4\times4} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Lambda_{4\times4} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Lambda_{4\times4} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Lambda_{4\times4} \end{pmatrix}$$

with $\Lambda_{ij} = 0.99^{|i-j|}, \quad 1 \le i, j \le 4$. The response is generated as

$$\boldsymbol{Y} = \mathbf{X}_1 + \mathbf{X}_3 + \mathbf{X}_{14} + \mathbf{X}_{16} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$. With $p = 16$, all the posterior model probabilities can be computed for validations of our methodology.

Implementation details

We recall that v_0 specifies the "spikiness" of the spike-and-slab prior. For a larger value of v_0, the posterior landscape is smoother, resulting in easier mode detection. This motivates the following strategy: we implement the methods for a large value of v_0 (generally around 0.75) and then decrease the value of v_0 incrementally to $v_0 = 0.1$, using the results from the previous run as "warm starts". In practice, we have found that a v_0 path of four steps is generally sufficient. The benefit of having a small but non zero value for v_0 is that negligible but non zero estimates of $\boldsymbol{\beta}$ are not included in the model. We take the slab variance to be $v_1 = 100$. We set $\theta = 0.5$.

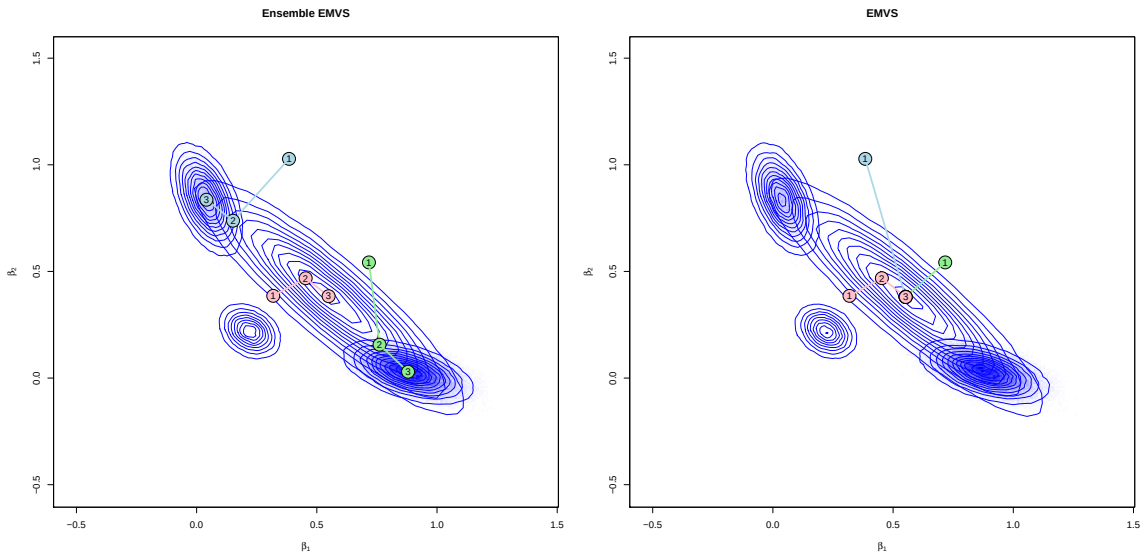


Figure 3: Illustration of DREVS ($\lambda > 0$ on the left) and EMVS ($\lambda = 0$ on the right) on a simulated example with $p = 2$ and two highly collinear predictors. The true model is $\boldsymbol{\gamma} = (1, 0)'$.

We additionally examine the effect of the tuning parameter λ for $\lambda = 0, 1, 10, 50$.

For this example, we take $K = 10$. The choice of K may be adjusted based on knowledge of how correlated the predictors are (and hence how multimodal the posterior will be).

Another important consideration is the choice of starting values. Based on deterministic annealing arguments, Ročková and George (2014) recommend a promising initialization for the EMVS algorithm to be at the ridge regression solution with equal penalties $\frac{v_0+v_1}{2v_0v_1}$:

$$\widehat{\boldsymbol{\beta}}_{RG} = \left[\mathbf{X}'\mathbf{X} + \frac{v_0 + v_1}{2v_0v_1}\mathbf{I}\right]^{-1}\mathbf{X}'\mathbf{Y}.$$

Additionally, if it is known that the true coefficient vector is sparse, a good initialization may be at the null model $\boldsymbol{\beta}_{NULL} = \mathbf{0}$. For this dataset, we examine three different sets of starting values:

(a) $\boldsymbol{\mu}_k^{(0)} \sim N_p(\mathbf{0}, \mathbf{I})$;

(b) $\boldsymbol{\mu}_k^{(0)} \sim N_p(\mathbf{0}, \sigma^2\mathbf{I})$, $\sigma^2 = 0.01$; and

(c) $\boldsymbol{\mu}_k^{(0)} \sim N_p(\widehat{\boldsymbol{\beta}}_{RG}, \sigma^2\mathbf{I})$, $\sigma^2 = 0.01$

for $k = 1, \ldots, K$. We note that (a) is a very naive choice and has been included to test the sensitivity of the methods to the starting vectors. For (b) and (c), we perturb the starting points $\mathbf{0}$ and $\widehat{\boldsymbol{\beta}}_{RG}$ slightly to examine how well NPVS and DREVS diversify the modes.

Results

We implemented NPVS and DREVS for 20 initializations of each of the three types of starting values described above. The total posterior probability found for each run is displayed by the box plots in Figure 4.

We see that in Figure 4a (corresponding to starting values (a)), there is high variance in the performance of both methods, demonstrating their sensitivity to different starting values. For a low dimensional setting such as this, it is feasible for the independent EMVS algorithm ($\lambda = 0$) to find a large number of high posterior probability models, provided the initializations are sufficiently diverse.

For starting vectors (b), we see a dramatic difference in the performance for the different values of λ. As expected, $\lambda = 0$ always finds the same model, whereas when $\lambda > 0$, the determinantal penalty drives trajectories apart to find a diverse set of modes. The clear winner is DREVS with $\lambda = 10$, which finds over 50% of the total posterior probability in all of the 20 runs. As expected, NPVS here performs better than $\lambda = 0$ in finding higher posterior probability modes; however, we see a clear benefit of inducing repulsion in the model space domain as opposed to the parameter domain.

Figure 5 is a plot of the posterior probabilities for each of the 2^p models, showing the models found for $\lambda = 0$ and $\lambda = 10$ with starting values generated as in (c). Here, we see that having been initialized close to $\widehat{\boldsymbol{\beta}}_{RG}$, $\lambda = 0$ finds the global mode but is unable to capture the remaining posterior probability. Contrast-

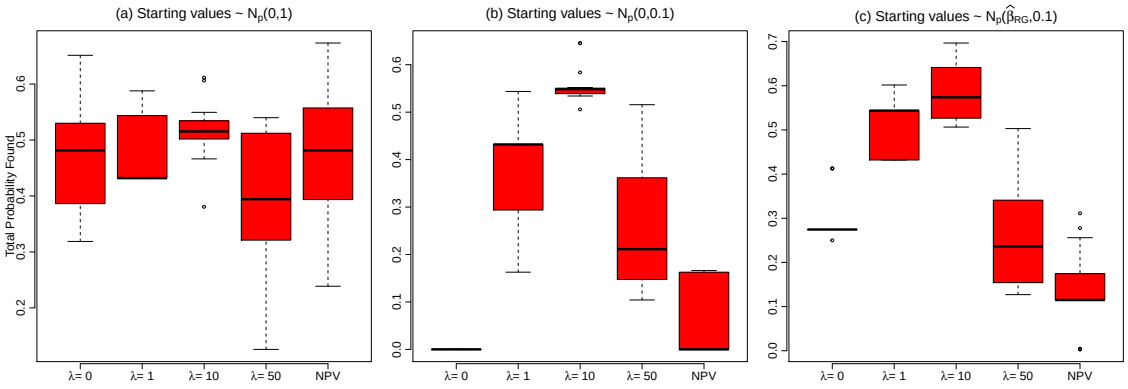


Figure 4: The box plots display the total probability explored by DREVS ($\lambda = 0, 1, 10, 50$) and NPVS over 20 initializations for each of the three types of starting values. The higher the total posterior probability found, the better.

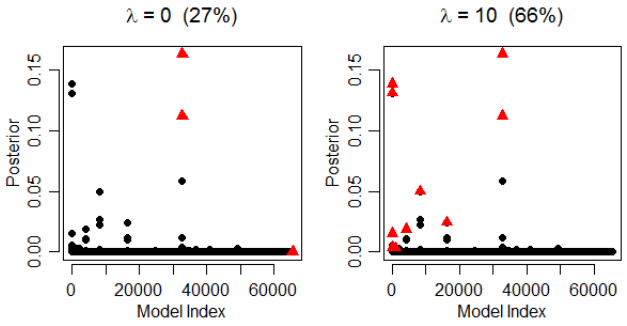


Figure 5: Plots show posterior probability of each of the 2^p possible models. The models found are shown in red. The total amount of posterior probability is displayed in parentheses. The methods found 3, 10 models for $\lambda = 0, 10$ respectively. Starting vectors generated as in (c).

ingly, $\lambda > 0$ finds the global mode and then quickly diversifies to find a set of high posterior modes.

6.2 Protein activity data

In this section, we apply DREVS to the protein activity dataset from Clyde and Parmigiani (1998). Following these authors, we code the categorial variables by indicators, consider main effects and first order interactions and add second order terms for the numerical variables. This results in $p = 88$ predictors. The sample size is $n = 96$. We also take $v_0 = 0.1$ and $v_1 = 1000$. As the model posterior may no longer be completely enumerated for a dataset of this size, we approximate it with a Metropolis-Hastings (MH) sampler. Specifically, we run the MH algorithm with a one-step random scan proposal to simulate from the marginal posterior on $\boldsymbol{\gamma}$ for 10^6 iterations.

We implemented DREVS for a range of different λ values and found again that $\lambda = 10$ performed the best. Using this setting, as well as $K = 50$ and starting at $\widehat{\boldsymbol{\beta}}_{RG}$ for each mode, DREVS found the top three posterior probability models, which corresponded to 16% of the total probability mass. This performance is remarkable given the vastness of the model space (2^{88} models). We also note that the median probability model, computed from DREVS and MH were the same. In this dataset, the median probability model coincided with the highest posterior probability model. Contrastingly, EMVS always found the null model, as $\widehat{\boldsymbol{\beta}}_{RG}$ lies in its domain of attraction.

7 Discussion

In this paper, we applied the nonparametric variational Bayes framework to the problem of Bayesian variable selection. This strategy motivated the development of a new determinantal regularization variant (DREVS), which takes advantage of the geometry of the spike-and-slab posterior. We showed the efficacy of DREVS in finding a set of high posterior modes when the design matrix has blocks of highly correlated predictors. We applied DREVS to a protein activity dataset and showed that it successfully found the top three modes.

DREVS is a prototype deterministic ensemble approach to multimodal posterior discovery. Going forward, Ročková (2016) proposed a fast new Particle EM ensemble approach, a more computationally efficient alternative that operates directly in the model space domain.

Acknowledgements

This research was supported by the NSF Grant DMS-1406563.

References

Affandi, R. H., Fox, E., Adams, R., and Taskar, B. (2014), "Learning the Parameters of Determinantal Point Process Kernels," in *Proceedings of The 31st*

International Conference on Machine Learning, pp. 1224–1232.

Clyde, M. A. and Parmigiani, G. (1998), "Protein construct storage: Bayesian variable selection and prediction with mixtures," *Journal of Biopharmaceutical Statistics*, 8, 431–443.

George, E. I. and McCulloch, R. E. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

Gershman, S., Hoffman, M., and Blei, D. M. (2012), "Nonparametric variational inference," in *Proceedings of the 29th International Conference on Machine Learning*, pp. 663–670.

Jaakkola, T. S. and Jordan, M. I. (1998), "Improving the mean field approximation via the use of mixture distributions," in *Learning Graphical Models*, Springer, pp. 163–173.

Ormerod, J. T., You, C., and Müller, S. (2014), "A variational Bayes approach to variable selection," *Manuscript*.

Robert, C. and Mengersen, K. (2003), "IID sampling using self-avoiding population Monte Carlo: the pinball sampler," *Bayesian Statistics*, 7, 277–292.

Ročková, V. and George, E. I. (2014), "EMVS: The EM approach to Bayesian variable selection," *Journal of the American Statistical Association*, 109, 828–846.

Ročková, V. (2016), "Particle EM for Variable Selection," *Manuscript*, 1–29.

Zou, J. Y. and Adams, R. P. (2012), "Priors for diversity in generative latent variable models," in *Advances in Neural Information Processing Systems*, pp. 2996–3004.