Nonparametric Bayesian Kernel Models

BY FENG LIANG

Department of Statistics

University of Illinois at Urbana-Champaign, IL 61820, U.S.A.

liangf@uiuc.edu

KAI MAO

Department of Statistical Science

Duke University, Durham NC 27708-0251, U.S.A.

km68@stat.duke.edu

MING LIAO

Marketing Analytics, IMS Health

Philadelphia PA, U.S.A.

liao@stat.duke.edu

SAYAN MUKHERJEE

Department of Statistical Science & Institute for Genome Sciences and Policy

Duke University, Durham NC 27708-0251, U.S.A.

sayan@stat.duke.edu

and

MIKE WEST

Department of Statistical Science

Duke University, Durham NC 27708, U.S.A.

mike@stat.duke.edu

SUMMARY

Kernel models for classification and regression have emerged as widely applied tools in statistics and machine learning. We discuss a Bayesian framework and theory for kernel methods, providing a new rationalisation of kernel regression based on nonparametric Bayesian models. Functional analytic results ensure that such a nonparametric prior specification induces a class of functions that span the reproducing kernel Hilbert space corresponding to the selected kernel. Bayesian analysis of the model allows for direct and formal inference on the uncertain regression or classification functions. Augmenting the model with Bayesian variable selection priors over kernel bandwidth parameters extends the framework to automatically address the key practical questions of kernel feature selection. Novel, customised MCMC methods are detailed and used in example analyses. The practical benefits and modelling flexibility of the Bayesian kernel framework are illustrated in both simulated and real data examples that address prediction and classification inference with high-dimensional data.

Some Key Words: Dirichlet process priors; Kernel parameter estimation; Kernel principal component regression; Reproducing kernel Hilbert space; Semi-supervised learning; Nonparametric Bayesian analysis.

1 INTRODUCTION

Kernel models for regression have a long history in statistics and applied mathematics (Schoenberg, 1942; Parzen, 1963; de Boor and Lynch, 1966; Michelli and Wahba, 1981; Wahba, 1990) and have been used extensively in machine learning for classification and regression problems (Poggio and Girosi, 1990; Vapnik, 1998; Schölkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004). The appeal of these models includes their flexibility and resulting capacity for predictive accuracy if well-calibrated, and simple extension of the underlying ideas to higher-dimensional data analysis. Some widely used statistical models or machine learning algorithms are examples of kernel models, including spline models (Wahba, 1990), regularized logistic regression (O'Sullivan et al., 1986), and support vector machines (SVMs) (Cortes and Vapnik, 1995).

The univariate response regression problem is summarized by the model

$$y = f(x) + \text{error},$$

where y is the measured response, f is an unknown function and $x \in \mathcal{X} \subseteq \mathbb{R}^p$ is the value of the p-dimensional covariate vector corresponding to outcome y. Given data from this model, our objective is to estimate the underlying function f for prediction of future responses. For kernel models the estimate is selected from functions contained in the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ induced by the kernel k. Regularization methods (Tikhonov and Arsenin, 1977) are frequently used to justify the estimate

$$\hat{f} = \arg\min_{f \in \mathcal{H}_k} \; L(f, \text{data}) + \lambda \|f\|^2_{\mathcal{H}_k} \tag{1}$$

where the first term L is a loss function induced from the log-likelihood derived from the assumed form of the error density, the second term is a smoothness penalty on the RKHS norm of the function, and λ is a tuning parameter that balances the trade-off between minimizing the fitting errors and the smoothness. Although the optimization in (1) may be over an infinite dimensional space the optimal solution has the following finite dimensional representation due to the representer theorem (Kimeldorf and Wahba, 1971)

$$\hat{f}(x) = \sum_{i=1}^{n} w_i \, k(x, x_i), \tag{2}$$

where $k(\cdot, \cdot)$ is the kernel function corresponding to the RKHS. This reduces an infinite dimensional optimization problem to one in n variables, which is very attractive for high-dimensional analysis since the optimization is over $n \ll p$ variables and independent of the dimension p.

Access to fully Bayesian formulations of kernel methods would provide a natural framework to further the richness and interpretability of kernel models – a program driving much research in data mining and machine learning. A Bayesian approach would allow for the immediate relaxation of two limitations inherent in classical RKHS models: constraining the smoothness penalty to monotonic functions of the RKHS norm, and requiring additional methods such as bootstrapping or cross-validation to provide confidence intervals and set hyper-parameters. The restriction of the penalty to be monotonic function of the RKHS norm precludes methods based on ℓ_1 penalties such as LASSO (Tibshirani, 1996) since the finite representation of (2) does not hold. Using priors to provide regularization affords greater flexibility.

Bayesian kernel methods have been developed in the context of Gaussian Process (GP) models (Neal, 1997; Bishop and Tipping, 2003; Bishop, 2006; Rasmussen and Williams, 2006) and Bayesian formulations of SVMs have been proposed (Tipping, 2001; Sollich, 2002; Chakraborty et al., 2005). In all these models Bayesian inference is applied directly to the finite representation from equation (2). We propose a more general model that does not start with this finite representation and can result in models with knots at arbitrary points. We develop in detail a particular prior specification under this framework that results in computationally efficient inference that is similar to GP models (Neal, 1997; Chakraborty et al., 2005).

The conceptual novelty and theoretical motivation of this work is to provide priors that do not change with respect to observed covariates and are on the entire RKHS to obtain posterior samples from the RKHS. The practical innovations are efficient procedures to obtain posterior samples from the RKHS. The direct adoption of the finite representation does not provide us with a theoretical framework to satisfy these modelling criteria. For point estimates the direct adoption of equation (2) in a Bayesian analysis is based on the fact that the finite representation is a MAP estimator (Wahba, 1990; Poggio and Girosi, 1990). This argument

does not extend to drawing posterior samples from the RKHS. In addition this justification does not hold for priors that are not a function of the RKHS norm. The Gaussian process approach does not satisfy our modelling criteria either. The duality between RKHS and Gaussian processes suggests placing priors directly on a space of functions by sampling from the paths of the Gaussian process with covariance structure defined by k. The mean of this process is in the RKHS but random functions drawn from the GP are almost surely outside the RKHS induced by k (Kallianpur, 1970; Wahba, 1990). For this reason the GP perspective is natural for point estimates such as the posterior mean but is problematic for posterior samples from the RKHS. There does exist a larger RKHS $\mathcal{H}_R$ induced by a kernel R that contains these functions (Lukić and Beder, 2001; Pillai et al., 2007) and posterior samples from the GP with covariance structure defined by k would be from the RKHS $\mathcal{H}_R$.

We also formulate a procedure for simultaneous dimension reduction in the original input space and in the kernel feature space. Inference of which covariates are most relevant in modelling the response variable for kernel models have been developed in the machine learning literature (Chapelle et al., 2002; Jebara and Jaakkola, 2000; Krishnapuram et al., 2004; Tipping, 2001). Most Bayesian methods for joint inference of variable relevance and the kernel model parameters have been based on variational methods or MAP estimates (Jebara and Jaakkola, 2000; Krishnapuram et al., 2004; Tipping, 2001). We provide an efficient procedure to sample from the posterior distribution of parameters that model the relevance of the covariates. This allows us to obtain estimates of the uncertainty in the relevance of variables.

In summary our approach results in a novel, fully Bayesian framework and theory for kernel regression and classification. Unlike previous approaches we specify priors on the entire RKHS. Our prior specification induces a class of functions that span the RKHS, providing an equivalence between the nonparametric Bayesian model and kernel models used in the penalized loss framework. This implies a Bayesian representer form that results in the finite representation in equation (2) derived from a Bayesian formulation, and that is coherent across samples and sample sizes. This formal model then easily and coherently addresses problems of inference on hyper-parameters, variable selection, and ancillary issues

such as unlabeled data (in semi-supervised learning).

The paper is arranged as follows. Section 2 describes the nonparametric Bayesian approach that allows us to place a coherent prior on the RKHS and recover the parametrisation of the representer theorem as an approximation of the posterior mean. Section 3 provides one approach to complete prior specification over model hyper-parameters and a corresponding MCMC approach to posterior evaluation and inference for both regression and classification settings. Section 4 extends the definition of kernels to allow for variable selection. Examples and discussion are given in Section 5, with summary comments in Section 6.

2 A Class of Non-Parametric Bayesian Kernel Models and a Bayesian Representer Form

The kernel models are based on integral operators placing priors on signed measures rather than directly on the regression function space. We first show why we do not elicit priors directly on the function space.

2.1 Direct Prior Elicitation

Besides Gaussian processes, another natural way to directly elicit priors on a RKHS is based on orthogonal expansions of the RKHS.

Kernel functions $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that are continuous and positive semi-definite on a compact space $\mathcal{X}$ are Mercer kernels for which the RKHS is characterized (Mercer, 1909; König, 1986) as

$$\mathcal{H}_k = \left\{ f \mid f(x) = \sum_{j=1}^{\infty} a_j \phi_j(x) \quad \text{such that } \sum_{j=1}^{\infty} a_j^2/\lambda_j < \infty \right\},$$

where $\{\lambda_j\}$ and $\{\phi_j(x)\}$ are the eigenvalues and eigenfunctions of the integral operator defined by the kernel function

$$\lambda_j \phi_j(x) = \int_{\mathcal{X}} k(x,u)\, \phi_j(u)\, d\mu(u),$$

where μ is a measure. The eigenvalues and RKHS do not depend on the measure (König, 1986) so a prior over the space $\mathcal{A} = \{(a_j)_{j=1}^{\infty} | \sum_{j=1}^{\infty} a_j^2/\lambda_j < \infty\}$ implies a prior on $\mathcal{H}_k$. There are serious computational and conceptual problems

with specifying a prior on the parameter $\mathcal{A}$; it is in general infinite-dimensional, and it is subject to challenging constraints. The crux of the problem is that in this orthonormal expansion model we are working explicitly with eigenfunctions and eigenvalues, and they are inherently challenging to manipulate; many popular kernels do not even lead to eigenfunctions with closed forms, and others are not even computable.

2.2 Priors and Integral Operators

Alternatively, consider the space of functions defined as a convolution of the kernel with a signed (Borel) measure

$$\mathcal{G} = \left\{ f \mid f(x) = \int k(x,u)\ d\gamma(u),\ \ \gamma \in \Gamma \right\}, \tag{3}$$

with $\Gamma(\cdot)$ as a subset of the space of signed Borel measures. Placing a prior on Γ implies a prior on $\mathcal{G}$. The first nonparametric Bayesian kernel developments to exploit this idea were introduced in the unpublished PhD thesis of Liao (2005) using Dirichlet process priors over Γ, and this idea is fully developed here. More recently, it has become clear that this is an example of a more general framework that may utilise any prior over Γ, and equivalences between $\mathcal{G}$ and $\mathcal{H}_k$ exist for appropriate choices of priors on Γ (Pillai et al., 2007) including our Dirichlet process priors.

A variation of the integral operator defined in (3) takes the form

$$f(x) = \int k(x,u)\ d\gamma(u) = \int k(x,u)\, w(u)\, dF(u), \tag{4}$$

where the random signed measure $\gamma(u)$ is decomposed into a probability distribution $F(u)$ and coefficient function $w(u)$; $F(u)$ and $\gamma(u)$ share the same support. In general F denotes the distribution of the location of kernel knots u. Here we set $F = F_X$, the marginal distribution of X. This is a reasonable assumption as long as F_X and γ share the same support. An appealing property of this dependence of f on F_X is that our estimate of $f(x)$ will be locally adaptive in that more knots are allocated in high density regions.

2.3 Dirichlet Process Priors

The Dirichlet process (DP) prior is a natural choice to model uncertainty about the distribution function F. For a specified distribution F_0 having the same support as the uncertain distribution F, and a positive scale parameter α, the notation DP(α, F_0) implies that for any measurable partition of the sample space $(B_1, B_2, \ldots, B_k)$, the random vector $(F(B_1), \ldots, F(B_k))$ follows a Dirichlet distribution with parameter $\alpha(F_0(B_1), \ldots, F_0(B_k))$ (Ferguson, 1973, 1974; Sethuraman, 1994). DP priors are very popular in practical nonparametric Bayesian analysis (West, 1992; West et al., 1994; Escobar and West, 1995; Müller et al., 2004; MacEachern and Müller, 1998) due to modelling flexibility and computational advantages.

A fundamental characteristic of the DP model is that, given a sample $X_n = (x_1, ..., x_n)$ drawn independently from (uncertain) distribution F, the posterior is the DP

$$F \mid X_n \sim \text{DP}(\alpha + n, F_n), \quad F_n = (\alpha F_0 + \sum_{i=1}^{n} \delta_{x_i})/(\alpha + n). \tag{5}$$

Consider, then, such a prior for F in equation (4), and choose some fixed new point x_* to predict the function value $f(x_*)$. Based on the sample of n draws X_n from F we see that

$$\mathbb{E}[f \mid X_n] = a_n \int k(x, u)\, w(u)\, dF_0(u) + n^{-1}(1 - a_n) \sum_{i=1}^{n} w(x_i)\, k(x, x_i) \tag{6}$$

where $a_n = \alpha/(\alpha + n)$. Taking the formal limit of $\alpha \to 0$ to represent a non-informative prior leads to the finite-dimensional *Bayesian Representer* form

$$\hat{f}_n(x) = \sum_{i=1}^{n} w_i\, k(x, x_i), \tag{7}$$

where $w_i = w(x_i)/n$ depends on the "knot" x_i and sample size n. The two finite representations, equations (2) and (7), take the same form although they are derived from two fundamentally different approaches: the solution of a Tikhonov regularization functional versus formal process-prior Bayesian modelling.

A result of this prior specification is that we obtain a representation that is used in both the GP approach to kernel methods as well as the direct adoption of the the

finite representation. However our method is coherent and places a prior on the entire RKHS. Using another process such as a Lévy process or not take the limiting case of a non-informative prior we would obtain knots in the expansion not located at sample points. In addition, the marginalization in (6) ensures that each sample is included in the expansion so the order that observations are obtained does not matter. These choices are fundamentally for computation efficiency.

3 Estimation and Inference

3.1 Likelihood and Prior Specification for Hyper-Parameters

The Bayesian representer form leads to the usual linear regression on the kernel values as covariates with regression parameters w_i. Adding an intercept and a normal error model assumption we have the standard form

$$y_i = w_0 + f(x_i) + \varepsilon_i = w_0 + \sum_{j=1}^{n} w_j \, k(x_i, x_j) + \varepsilon_i, \qquad (i = 1, ..., n), \tag{8}$$

where $\varepsilon_i \sim N(0, \sigma^2)$. In vector form, the model is

$$Y \sim N(w_0 \iota + Kw, \sigma^2 I) \tag{9}$$

where $\iota = (1, \ldots, 1)'$, K is the $n \times n$ design matrix having entries $k(x_i, x_j)$, $Y = (y_1, \ldots, y_n)'$ and the regression parameter vector is $w = (w_1, \ldots, w_n)'$. Since w_0 and w are often treated differently, we orthogonalized the two sets of parameters by centering the kernel matrix. That is, $k(\cdot, \cdot)$ is replaced by a centred kernel $\tilde{k}(\cdot, \cdot)$ with

$$\tilde{k}(x_i, x_j) = k(x_i, x_j) - \bar{k}_{i\cdot} - \bar{k}_{\cdot j} + \bar{k},$$

where $\bar{k} = \sum_{i,j=1}^{n} k(x_i, x_j)/n^2$, $\bar{k}_{i\cdot} = \sum_{j=1}^{n} k(x_i, x_j)/n$ and $\bar{k}_{\cdot j} = \sum_{i=1}^{n} k(x_i, x_j)/n$.

Traditional priors can be taken for (w_0, σ^2). To minimize the number of hyper-parameters, we use the standard reference prior component $\pi(w_0, \sigma^2) \propto 1/\sigma^2$. Though it is improper, the corresponding posterior is still proper as long as the sample size $n \geq 2$ (Berger et al., 1998; Liang and Barron, 2004).

Specifying priors over the w_i can be done by defining sample size independent priors for values $w(x_i)$ at arbitrary knots. As an alternative, we induce appropriate

sample size dependence and address key questions of inducing regression shrinkage appropriately coupled to the structure of the kernel design space by using ridge regression or g-prior modelling (Zellner, 1986). West (2003) defined and exemplified the use of a flexible and practically very effective class of generalised g-priors that allow for different degrees of shrinkage estimation of regression parameters in different principal component directions on the induced design space for any regression model, and we adopt that strategy here. This is particularly relevant when dealing with many covariates, as it provides an ability to "shrink away" the effects of many irrelevant component dimensions while highlighting those of predictive value. This class of priors explicitly models the distribution $p(w|K)$, so that the sample size dependence is directly induced and the class of priors adapts as the sample size changes.

Specifically, a generalised g-prior is induced by independent normal priors on the regression parameters of the equivalent principal component regression transformation of the model. The kernel matrix K is symmetric and positive semi-definite, so has spectral decomposition $K = F\Delta F'$ where F is the $n \times n$ orthogonal *factor* matrix, and $\Delta = \text{diag}(\lambda_1^2, \ldots, \lambda_n^2)$. In the orthogonal representation the regression maps from Kw to $F\beta$ with $w = F\Delta^{-1}\beta$. Assume conditionally independent normal priors for the elements of β, so that $\beta \sim N(0, T)$ for some $T = \text{diag}(\tau_1, \ldots, \tau_n)$. The induced generalised g-prior for w is then

$$(w \mid K, T) \;\sim\; N\left(0, F\Delta^{-1}T\Delta^{-1}F'\right). \tag{10}$$

Following West (2003), we further specify hyper-priors over the n prior variances τ_i – that play roles as shrinkage parameters – as independent inverse gammas,

$$\tau_i \sim \text{InvGa}(a_\tau/2, b_\tau/2), \qquad (i = 1, \ldots, n),$$

inducing heavier-tailed t-priors on the w_i when we marginalise over the τ_i.

Viewed as hyper-parameters to be estimated, the τ_j's are the prior variances for each factor regression parameter and allow for a varying degree of shrinkage in each of the orthogonal factor dimensions, as discussed. It may be tempting to set hyper-parameter values $a_\tau = b_\tau = 0$ to obtain a non-informative prior on τ_i, namely, $\pi(\tau_i) \propto 1/\tau_i$, which unfortunately correspond to an improper posterior distribution (Hill, 1965). In Section 5, we choose $a_\tau = b_\tau = 2$ which correspond to Cauchy distributions on the β_j, a very natural, highly diffuse though

proper prior specification. In practice, for dimension reduction and also for computational stability, we may choose to truncate the spectral decomposition by rejecting factors with very small eigenvalues λ_i; that is, in such a case we may choose to replace F by its first m columns, Δ by its first m diagonals, and set $T = \text{diag}(\tau_1, \ldots, \tau_m)$, where $m < n$ and the eigenvalues $\lambda_{m+1}, \ldots, \lambda_n$ are less than some pre-specified threshold.

3.2 Model Fitting and Prediction via MCMC

Given the likelihood and the prior distributions a standard Gibbs sampler can be used to simulate the posterior $p(w_0, w, \sigma^2 \mid \text{data})$. After initialization, samples of parameters and hyper-parameters are drawn sequentially from the complete conditional posterior distributions. At each iteration, with all relevant conditioning parameters fixed at their most recent values in the iterates, we update as follows:

1. Update w_0: w_0 is drawn from the normal posterior with mean $n^{-1}\iota'(Y - F\beta)$ and variance σ^2/n.

2. Update w: Simply via β, generate $\beta \sim N(b, V)$ where $V = \text{diag}(V_1, \ldots, V_m)$ with $V_i = \sigma^2\tau_i/(\tau_i + \sigma^2)$, and $b = VF'(Y - w_0)/\sigma^2$; then set $w = F\Delta^{-1}\beta$.

3. Update T: For $j = 1, .., m$, $\tau_j^{-1} \sim \text{Ga}((a_\tau + 1)/2, (b_\tau + \beta_j^2)/2)$.

4. Update σ^2: $\sigma^{-2} \sim \text{Ga}(n/2, s/2)$ with $s = e'e$ where $e = Y - w_0 - F\beta$.

For prediction at a specified new point x_*, any aspect of the predictive distribution for y_* can be included for sampling in the MCMC. Given sampled parameter values at each iterate, we can simply evaluate the mean and variance of the conditional normal distribution $p(y_*|x_*, w_0, w, \sigma^2)$ to generate MCMC samples of posterior predictive quantities of interest. This way we compute MC approximations to predictive means $\mathbb{E}(y^*|x^*, \text{data})$, for example, and can do this across a range of x^* values to map out the predicted non-linear regression function for predictive uses.

3.3 Binary Regression for Classification

The approach developed above for regression models can of course be easily extended to a classification setting using probit regression, or other binary regressions. The standard latent variable imputation extensions of MCMC lead directly to posterior samplers for probit and other binary link functions that are representable as mixtures of normals (Albert and Chib, 1993; Johnson and Albert, 1999). Metropolis-Hastings variants for logistic regression are also trivial modifications. These are practically very relevant extensions for kernel classification problems.

By way of notation and basic structure in the probit model, the responses y_i in the kernel model (9) are now latent and the normal errors are standard, i.e., $\sigma^2 = 1$. We observe binary responses $Z = (z_1, z_2, ..., z_n)'$ generated by $z_i = 1(0)$ if $y_i \geq 0 (< 0)$. The MCMC extensions simply include the latent Y values at each iterate of the simulation. The traditional Gibbs sampler iterates between sampling conditional posteriors for Y given the regression parameters, and vice-versa. Though often effective, this vanilla Gibbs sampler can suffer from very slow mixing due to high correlations between successive draws of latent variables (Liu, 2001; Nobile, 1998); the problem is of course shared by all binary regression models. Proposed solutions in these last two references will not, however, apply in our model (since the kernel matrix, the analog of the design matrix in their case, changes in each iteration of the chain after we introduce a variable selection component in the following section) so we have developed a novel and effective solution. Rather than the Gibbs sampler we use a Metropolis-Hastings method that samples the kernel model parameters jointly with the latent variable Y. This is summarised in the following section, in an extended model that incorporates additional kernel parameters to address variable and feature selection.

4 Variable and Feature Selection

4.1 Kernel Model Extension

Variable selection and feature selection are important problems in high-dimensional regression. The standard formulation of variable selection is to select a relatively

small subset of the p covariates without loss of predictive accuracy. In the problem of feature selection a small subset of combinations of the p covariates are selected. Principle components regression with a few principle components is an example of a feature selection method.

Standard practise in kernel regression allows each coordinate of x to be scaled (Jebara and Jaakkola, 2000; Chapelle et al., 2002; Krishnapuram et al., 2004)

$$k_\nu(x,u) = k(\sqrt{\nu} \otimes x, \sqrt{\nu} \otimes u)$$

where $a \otimes b$ is the element-wise product of two vectors and $\nu = (\nu_1, ..., \nu_p)$ is a p-dimensional vector with $\nu_k \in [0, \infty]$ as an individual scale parameter for the k-th dimension. This approach can be applied to most kernels and for the linear, polynomial, and Gaussian kernels the resulting adaptive kernels are

$$\begin{aligned} k_\nu(x,u) &= \sum_{k=1}^{p} \nu_k \, x_k \, u_k, \\ k_\nu(x,u) &= \left(1 + \sum_{k=1}^{p} \nu_k \, x_k \, u_k\right)^d, \\ k_\nu(x,u) &= \exp\left(-\sum_{k=1}^{p} \nu_k (x_k - u_k)^2\right). \end{aligned}$$

We will focus on the Gaussian kernel for which ν_k can be regarded as the reciprocal of the bandwidth for the k-th variable which determines the neighbourhood size for that dimension. When $\nu_k = 0$, the neighbourhood size is infinity and the corresponding variable is irrelevant in predicting the response variable. Variable/bandwidth selection is then a problem of estimation or selection of the parameter ν, now explicitly in the context of allowing for zero values. This invites analysis using standard Bayesian variable selection/model uncertainty strategies based on "point mass, mixture prior" over these parameters.

For each ν_k independently, we adopt the prior

$$\begin{aligned} \nu_k &\sim (1-\gamma)\delta_0 + \gamma \, \text{Ga}(a_\nu, a_\nu s), \quad (k = 1, \ldots, p), \\ s &\sim \text{Exp}(a_s), \quad \gamma \sim \text{Be}(a_\gamma, b_\gamma), \end{aligned}$$

where $(a_\nu, a_s, a_\gamma, b_\gamma)$ are specified hyperparameters, $\text{Be}(\cdot,\cdot)$ represents the beta distribution and $\text{Exp}(\cdot)$ the exponential.

4.2 Overall MCMC

The MCMC analysis can now be extended to include the ν parameters. These parameters are treated with appropriate Metropolis-Hastings steps since their complete conditionals are not of standard forms. Our overall MCMC sampler uses a Metropolis-Hastings step to jointly sample the kernel bandwidth and regression parameters; in the case of binary outcomes when the y_i are latent, this sampling step is extended to jointly sample these parameters and Y together. As mentioned in the previous section, this novel MCMC – that has been tested successfully in a number of examples – is designed to mix faster than the traditional Gibbs sampler in binary models, and now also provides an overall approach for the kernel variable selection extension in both linear and binary outcomes cases.

The full hybrid sampler for the posterior of $(w_0, \nu, \beta, Y, s, \gamma, T)$ in the case of binary probit regression is detailed here. The changes to this to generate the corresponding MCMC for the linear model simply adds in the sampling of the residual variance σ^2 at each step and removes the imputation of the y_i that are, in the linear case, known; these details are left to the reader.

The sequence of steps per iteration in the full binary kernel model with feature selection are as follows:

1. Update w_0 as in section 3.2.

2. Update (ν, β, Y) jointly, in the following two steps.

   2.1. Update (ν, β):

   2.1.1. Propose ν^*: Let p_g, p_l, p_u denote the probabilities for a *global move*, *local move*, or *update move* respectively.

   - For the *global move*, draw ν^* from the prior.
   - For the *local move*, set $\nu* = \nu$ then randomly pick a dimension k. If $\nu_k \neq 0$, set $\nu_k^* = 0$; otherwise draw $\nu_k^* \sim \text{Ga}(a_\nu, a_\nu s)$, the continuous part of the prior.
   - For the *update move*, set $\nu* = \nu$ and then, for all dimensions k where $\nu_k \neq 0$, draw $\nu_k^* \sim \text{Ga}(a_\nu, a_\nu s)$.

   Our proposals use $p_g = .25, p_l = .5, p_u = .25$.

2.1.2. Propose β^*: Compute the proposed kernel matrix K^* with entries $\tilde{k}_{\nu^*}(x_i, x_j)$ and its spectral factors F^* and Δ^*. Set $\hat{Y} = w_0 + F^*\beta$ and simulate Y^* via, for each $i = 1, \ldots, n$,

$$y_i^* \sim \begin{cases} N(\hat{y}_i, 1)^+, & \text{if } z_i = 1, \\ N(\hat{y}_i, 1)^-, & \text{if } z_i = 0. \end{cases}$$

Then, propose $\beta^* \sim N(b^*, V)$ where $V = \text{diag}(V_1, \ldots, V_m)$ with $V_i = \tau_i/(1 + \tau_i)$, and $b^* = VF^{*\prime}(Y^* - w_0)$.

2.1.2. Acceptance ratio to compare and test (ν^*, β^*) against the current values (ν, β): The Metropolis-Hastings acceptance ratio is

$$r = \frac{p(Z \mid \nu^*, \beta^*, w_0)\, \pi(\nu^*, \beta^* \mid s, \gamma) q(\nu, \beta \mid T, w_0, s, \gamma)}{p(Z \mid \nu, \beta, w_0)\, \pi(\nu, \beta \mid s, \gamma) q(\nu^*, \beta^* \mid T, w_0, s, \gamma)}$$

where the terms $p(Z|\cdots)$ are likelihood evaluations from the binary regression and $\pi(\cdot), q(\cdot)$ denote the prior distribution function and proposal distribution function, respectively. With probability $\min\{r, 1\}$ accept the proposed values and hence set $\nu = \nu^*$ and $\beta = \beta^*$; otherwise, retain the current values.

Denote the accepted or retained values by $\{\nu, \beta, K, F, \Delta\}$, and set $w = F\Delta^{-1}\beta$.

2.2. Update Y: $\hat{Y} = w_0 + F_\nu \beta$ and resample Y via, for each $i = 1, \ldots, n$,

$$y_i \sim \begin{cases} N^+(\hat{y}_i, 1), & \text{if } z_i = 1, \\ N^-(\hat{y}_i, 1), & \text{if } z_i = 0. \end{cases}$$

where N^+ and N^- denote the positive and negative parts of a truncated normal.

3. Update hyper-parameters: (s, γ, T)

3.1. Update s: $s \sim \text{Ga}(a_\nu + 1, a_s + a_\nu \sum \nu_k)$.

3.2 Update γ: $\gamma \sim \text{Be}(a_\gamma + p_1, b_\gamma + p - p_1)$ where p_1 is the number of nonzero elements in ν.

3.3 Update T as in section 3.2.

This MCMC combines variable and feature selection. It is a variable selection method since only those variables with nonzero ν_k will be selected. It is also a feature selection method since we are weighting each variable that is selected by ν_k. The parameter $m \leq n$ was introduced in Section 3.1 to allow for the opportunity to truncate the expansion of the kernel matrix for numerical stability. Reducing m as dimension reduction is often criticised since principal components of the kernel matrix that dominate variation in the kernel design space may not necessarily correspond to the factors most relevant in prediction of the response variable. In the current setting that now includes variable and feature selection over elements of ν, this problem is obviated: the weight for each dimension is adjusted in MCMC steps such that the top m kernel principle components are indeed the ones most relevant to the response variable. Nevertheless, with larger samples it is still generally desirable to consider restricting to $m < n$ for numerical and computational reasons.

5 EXAMPLES

5.1 Synthetic Data Sets

A simulated example considers binary classification with variable selection, using two synthetic data sets to illustrate different aspects of the model. For the MCMC in this subsection, we used 5000 iterations including an initial 2500 iterations for burn-in.

The first data set is in $\mathbb{R}^{30}$ but only the first two dimensions influence the classification. The x data for two classes are sampled from Gaussian mixture models with

$$(x|y=0) \sim 0.5N(\mu_{01}, \Sigma) + 0.5N(\mu_{02}, \Sigma)$$
$$(x|y=1) \sim 0.5N(\mu_{11}, \Sigma) + 0.5N(\mu_{12}, \Sigma)$$

where $\Sigma = \text{diag}(.38, .38, 1, ..., 1)$ and $\mu_{01} = (1, 1, 0, ..., 0)$, $\mu_{02} = (-1, -1, 0, ..., 0)$, $\mu_{11} = (-1, 1, 0, ..., 0)$ and $\mu_{12} = (1, -1, 0, ..., 0)$. We drew 30 samples from each class as training data, and a further 100 samples from each class as test data. The data on the first two x dimensions are plotted in Figure 1.

To provide an initial, baseline comparison, we fitted a binary model, with a Gaussian kernel on the first two dimensions and $\nu = (1.5, 1.5)$, to the test data alone; no feature selection was used here. The value of ν was chosen to be the one that produces the smallest test error. In terms of posterior means of the resulting classification probabilities, the resulting test error was $.5\%$ (only one sample being misclassified). The predictive probability of $(y_* = 1 \mid x_*)$ with respect to the first two dimensions of x_* is displayed in Figure **??**(a).

We compared the kernel model analysis with and without variable selection to this baseline kernel model. For the kernel model without variable selection, we set the bandwidth parameter to be constant in all dimensions, $\nu = (\nu, ..., \nu)$. For a variety of choices of ν the test error never fell below 33.5% and the training error was 0. This poor performance is illustrated in the prediction plot in Figure **??**(b), where we plot $(y_* = 1 \mid x_*)$ again with respect to the first two dimensions of x_*. We then applied the kernel model with variable selection to this data with hyper-parameters

$$a_\tau = b_\tau = 2,\ a_\gamma = b\gamma = 5,\ a_\rho = 1, a_s = 1, m = 5. \tag{11}$$

The test error of $.5\%$ was comparable to the "optimal" model results as in Figure **??**(a); the prediction plot in Figure **??**(c) shows the efficacy of the variable selection component of the analysis in honing in on the truly predictive variables and adapting the non-linear predictive model appropriately.

The second synthetic data set is analysed to explore variable selection further as well as to provides a sense of scale for each of the x variables. The data set is in $\mathbb{R}^{20}$ but only the fist two dimensions are relevant. The two classes are sampled from Gaussian mixture models with

$$(x|y = 0) \sim \frac{1}{3}N(\mu_{01}, \Sigma) + \frac{1}{3}N(\mu_{02}, \Sigma) + \frac{1}{3}N(\mu_{03}, \Sigma)$$
$$(x|y = 1) \sim \frac{1}{3}N(\mu_{11}, \Sigma) + \frac{1}{3}N(\mu_{12}, \Sigma) + \frac{1}{3}N(\mu_{13}, \Sigma)$$

where $\Sigma = \text{diag}(.38, .38, 1, ..., 1)$ and $\mu_{01} = (-1, -1, 0, ..., 0)$, $\mu_{02} = (0, 1, 0, ..., 0)$, $\mu_{03} = (1, -1, 0, ..., 0)$, $\mu_{11} = (-1, 1, 0, ..., 0)$ and $\mu_{12} = (0, -1, 0, ..., 0)$, $\mu_{13} = (1, 1, 0, ..., 0)$. The training data consist of 45 points from each class. The first two dimensions are plotted in Figure **??**(a). This plot as well as the generative distributions suggest that the first two dimensions should scale differently and

this should be reflected in posterior draws of the corresponding bandwidth parameters ν_1, ν_2. Specifically, we should expect $\nu_1/\nu_2 \approx 2/3$. We applied the kernel model with variable selection to this data with the same hyper-parameter values. Figure **??**(b) displays the predictive probability as a function of the first two relevant variables. Figure **??**(c) displays the 90% credible interval for ν_k for $k = 1 : 20$. Examining the posterior distribution of the elements of ν we found that $P(\nu_1 \neq 0 \mid \text{data}) = P(\nu_2 \neq 0 \mid \text{data}) = 1$ and $P(\nu_k = 0 \mid \text{data}) \geq 86\%$ for the irrelevant dimensions $k = 3, .., 20$ where P stands for the empirical posterior probability estimated from the MCMC outputs. Meanwhile, the posterior mean and median are $3.01, 2.84$ for ν_1 and $1.78, 1.42$ for ν_2. This illustrates how the analysis is capable of inferring appropriate scales of variables in addition to their relative inclusion probabilities.

5.2 Real Data: The MNIST Data Set

A standard data set used in the machine learning community is the MNIST data set[1]. This data set contains $60,000$ images of handwritten digits $\{0, 1, 2, ..., 9\}$, where each image consists of $p = 28 \times 28 = 784$ gray-scale pixel intensities.

We considered all pairwise comparisons among the 10 different digits resulting in 45 binary classification problems. For each classification problem we randomly selected 50 training samples from each class as training data and 50 samples from each class as test data. This was repeated 5 times and the average test error was computed.

Since the 784 pixels in the image are strongly correlated we pre-processed the data by projecting the training and test data onto the first 50 principle components computed on the training data. We then applied the kernel model analysis twice – with and without variable selection. We used a linear kernel and the same hyper-parameter values as above with the exception that we restricted to m kernel principal components and the reported analysis summaries are based on choosing m to optimise 5-fold cross-validation classification errors within the training data set in each analysis. Note that for linear kernel model without variable selection, using $\nu = (\nu, \nu, \ldots, \nu)$ is equivalent to setting $\nu = 1$. We ran the MCMC for

[1] Available at *http://yann.lecun.com/exdb/mnist/*

5000 iterations after an initial 5000 iterations for burn-in for each experiment. The results for the 45 comparisons are reported in Figure **??**. The performance of the kernel model with variable selection is substantially superior to that without selection for all 45 classification problems.

We further explored variable selection by focusing on the task of classifying "3" vs "5", one of the most challenging comparisons. We ordered the variables by their approximate posterior model inclusion probabilities averaged over the 5 repeat experiments. Due to the image processing underlying the raw data, each variable is not precisely a single pixel from the original image; rather, it is a locally-weighted linear combination of all 784 pixels. We visualize each variable by plotting the corresponding 784 weights on the 28×28 grid. In Figure **??** we plot a few apparently relevant variables corresponding to ν_k with high posterior probabilities (upper panel), together with a few apparently irrelevant variables corresponding to ν_k with low posterior probabilities (lower panel). Visually, it is clear that the relevant variables capture geometric differences between "3" and "5", while the irrelevant variables do not. Since the training and test data sets vary in the 5 experiments, we then randomly select a new data set with 100 samples from each class. Projections of this new data set onto sets of two relevant variables are displayed in Figure **??**; similar projections onto sets of two irrelevant variables are in Figure **??**. It is clear that the two classes show some separation in the relevant variables but not in the irrelevant variables.

6 Summary Comments

With the growth of interest in statistical classification and prediction methods in the machine learning communities, and an escalation of interest in applications among practitioners, there is a consequent need for refined theoretical understanding of the underlying statistical models as well as improved methodology and algorithms. We address each of these issues here. The theoretical foundation of our Bayesian kernel models is based on the equivalence between a class of functions induced by a nonparametric prior specification and a reproducing kernel Hilbert space. This Bayesian framework of the model allows for coherent inference, assessment of uncertainty, and access to the posterior distributions via Markov chain Monte Carlo sampling. Practical issues such as choice of hyper-parameters and

variable selection are automatically incorporated into the Bayesian modelling and inference.

The Bayesian kernel model suggests several interesting future directions as well as open problems. The computational challenges of searching high-dimensional parameter space is of utmost importance and for variable selection increasing the efficiency of the MCMC to be able to handle thousands of variables is an open problem of great practical importance. The nonparametric Bayesian kernel model we developed is an example of a more general framework described in Pillai et al. (2007). Further exploration of other process priors from a theoretical, computational, and applied data analysis perspective is of interest.

A striking example of the flexibility and coherence of the Bayesian kernel model is its application to what is referred to as the semi-supervised problem in the machine learning literature, the incorporation of unlabelled data – an example of ancillary design data – in classification and regression problems (Joachims, 1999; Blum and Mitchell, 1998; Szummer and Jaakkola, 2001; Zhu et al., 2003; Belkin et al., 2006). Our Bayesian kernel model incorporates the unlabeled data in a natural way without having to introduce additional penalties to the loss function as is the case for regularization approaches, which is discussed in detail in Liang et al. (2007).

ACKNOWLEDGMENTS

We acknowledge support of the National Science Foundation (DMS-0342172, DMS-0732276 and DMS-0732260) and the National Institutes of Health (NCI U54-CA-112952-01). Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NSF or NIH.

Appendix B: Software

Matlab software implementing the MCMC analysis for binary and linear kernel regression models, and with a range of specified kernel functions, is available to interested readers at the web site *http://www.stat.duke.edu/∼km68/bakerintro.htm*. The examples reported in the current paper are available with the code.

REFERENCES

Albert, J. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc. 88*, 669–679.

Belkin, M., P. Niyogi, and V. Sindhwani (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res. 7*, 2399–2434.

Berger, J., L. Pericchi, and J. Varshavsky (1998). Bayes factors and marginal distributions in invariant situations. *Sankhya, Ser. A 60*, 307–321.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Bishop, C. and M. Tipping (2003). Bayesian regression and classification. In J. Suykens (Ed.), *Advances in Leanring Theory: Methods, Models, and Applications*, pp. 267–288. IOS Press.

Blum, A. and T. Mitchell (1998). Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*, pp. 92–100. Morgan Kaufmann Publishers, San Francisco.

Chakraborty, S., M. Ghosh, and B. Mallick (2005). Bayesian non-linear regression for large p small n problems. *http://www.stat.ufl.edu/ schakrab/svmregression.pdf*.

Chapelle, O., V. Vapnik, O. Bousquet, and S. Mukherjee (2002). Choosing multiple parameters for support vector machines. *Machine Learning 46*(1-3), 131–159.

Cortes, C. and V. N. Vapnik (1995). Support-vector networks. *Machine Learning 20*(3), 273–297.

de Boor, C. and R. E. Lynch (1966). On splines and their minimum properties. *J. Math. Mech. 15*, 953–969.

Escobar, M. and M. West (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc. 90*, 577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist. 1*, 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist. 2*, 615–629.

Hill, B. M. (1965). Inference about variance components in the one-way model. *J. Amer. Statist. Assoc. 60*(311), 806–825.

Jebara, T. and T. Jaakkola (2000). Feature selection and dualities in maximum entropy discrimination. In C. Boutilier and M. Goldszmidt (Eds.), *Proceedings of Uncertainity In Artificial Intellegence 2000*, Stanford, CA, pp. 291–300. Morgan Kaufmann Publishers.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In I. Bratko and S. Dzeroski (Eds.), *Proceedings of ICML-99, 16th International Conference on Machine Learning*, Bled, SL, pp. 200–209. Morgan Kaufmann Publishers, San Francisco.

Johnson, V. and J. Albert (1999). *Ordinal Data Modeling*. Springer-Verlag.

Kallianpur, G. (1970). The role of reproducing kernel Hilbert spaces in the study of Gaussian processes. *Advances in Probability and Related Topics 2*, 49–83.

Kimeldorf, G. and G. Wahba (1971). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist. 41*(2), 495–502.

König, H. (1986). *Eigenvalue distribution of compact operators*, Volume 16 of *Operator Theory: Advances and Applications*. Basel: Birkhäuser Verlag.

Krishnapuram, B., A. J. Hartemink, L. Carin, and M. A. Figueiredo (2004). A bayesian approach to joint feature selection and classifier design. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*(9), 1105–1111.

Liang, F. and A. R. Barron (2004). Exact minimax strategies for predictive density estimation, data compression and model selection. *IEEE Trans. Inform. Theory 50*, 2708–2726.

Liang, F., S. Mukherjee, and M. West (2007). Understanding the use of unlabelled data in predictive modelling. *Statistical Science 22*(2), 198–205.

Liao, M. (2005). *Bayesian models and machine learning with gene expression analysis applications*. PhD dissertation, Duke University, Institute of Statistics and Decision Sciences.

Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Lukić, M. N. and J. H. Beder (2001). Stochasitic processes with sample paths in reproducing kernel Hilbert spaces. *Trans. Amer. Math. Soc. 353*(10), 3945–3969.

MacEachern, S. and P. Müller (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist. 7*, 223–238.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. R. Soc. Lond. Ser. A 209*, 415–446.

Michelli, C. A. and G. Wahba (1981). Design problems for optimal surface interpolation. In Z. Ziegler (Ed.), *Approximation Theory and Applications*, pp. 329–348. New York: Academic Press.

Müller, P., F. Quintana, and G. Rosner (2004). A method for combining inference across related nonparametric Bayesian models. *J. Amer. Statist. Assoc. 66*, 735–749.

Neal, R. (1997). Monte carlo implementation of gaussian process models for bayesian regression and classification. Technical Report 9702, University of Toronoto, Department of Statistics. *http://arxiv.org/abs/physics/9701026*.

Nobile, A. (1998). A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statist. Comp. 8*, 229–242.

O'Sullivan, F., B. Yandell, and W. Raynor (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc. 81*, 96–103.

Parzen, E. (1963). Probability density functionals and reproducing kernel Hilbert spaces. In M. Rosenblatt (Ed.), *Proceedings of the Symposium on Time Series Analysis*, pp. 155–169. New York: Wiley.

Pillai, N., Q. Wu, F. Liang, S. Mukherjee, and R. Wolpert (2007). Characterizing the function space for Bayesian kernel models. *J. Mach. Learn. Res. 8*, 1769–1797.

Poggio, T. and F. Girosi (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science 247*, 978–982.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Schoenberg, I. J. (1942). Positive definite functions on spheres. *Duke Math. J. 9*, 96–108.

Schölkopf, B. and A. J. Smola (2001). *Learning with Kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge: The MIT Press.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica 4*, 639–650.

Shawe-Taylor, J. and N. Cristianini (2004). *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge Univ. Press.

Sollich, P. (2002). Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine Learning 46*, 21–52.

Szummer, M. and T. Jaakkola (2001). Partially labeled classification with markov random walks. In *Advances in Neural Information Processing Systems (NIPS)*, Volume 14, pp. 945–952. MIT Press.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B. 58*(1), 267–288.

Tikhonov, A. N. and V. Y. Arsenin (1977). *Solutions of Ill-posed Problems*. Washington: Winston & Sons.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res. 1*, 211–244.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia: SIAM.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. ISDS Discussion Paper Series 1992-03, Duke University, Institute of Statistics and Decision Sciences. *http://ftp.stat.duke.edu/WorkingPapers/92-A03.ps*.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (Eds.), *Bayesian Statistics 7*, pp. 723–732. Oxford: Oxford University Press.

West, M., P. Müller, and M. Escobar (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In A. Smith and P. Freeman (Eds.), *Aspects of Uncertainty: A tribute to D. V. Lindley*, pp. 63–386. Wiley.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *J. Amer. Statist. Assoc. 81*, 446–451.

Zhu, X., Z. Ghahramani, and J. D. Lafferty (2003). Semi-supervised learning using gaussian fields and harmonic functions. In T. Fawcett and N. Mishra (Eds.), *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), August 21-24, 2003, Washington, DC, USA*, pp. 912–919. AAAI Press.

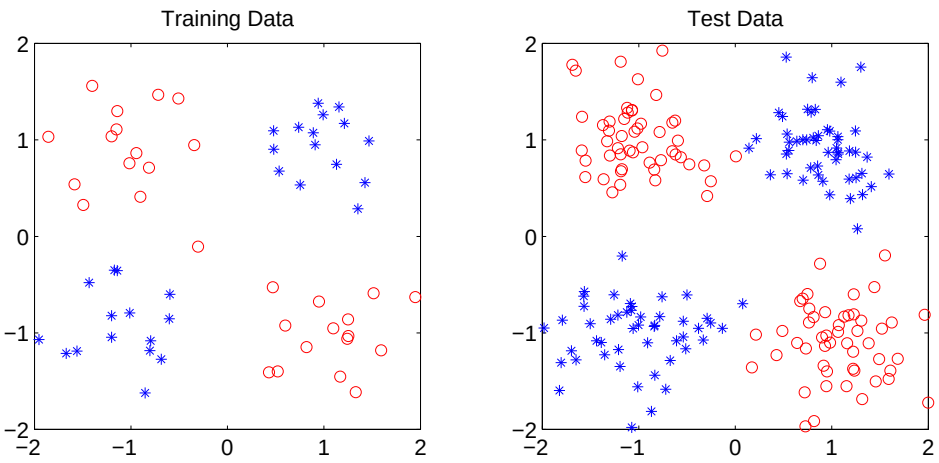
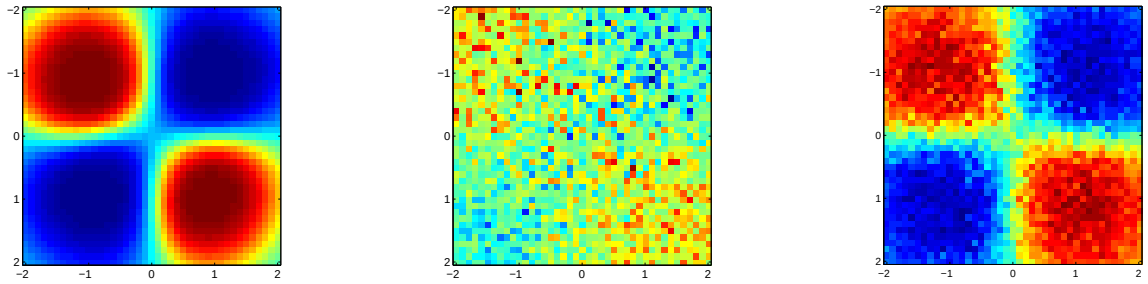
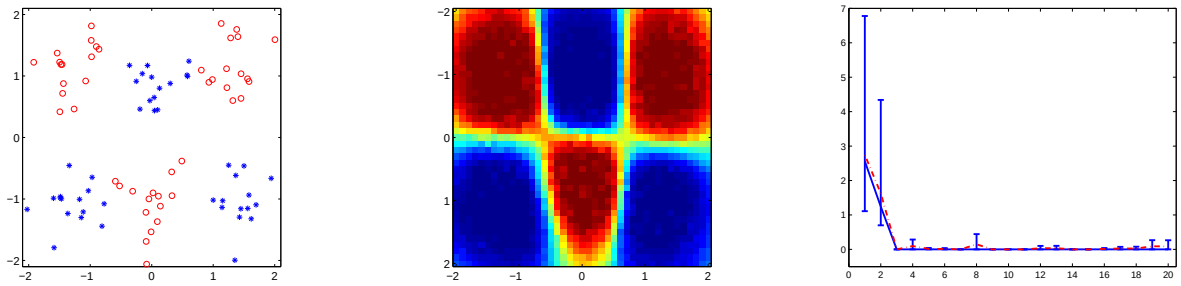


Figure 1: Synthetic data set 1. Scatter plot of the training data (60 observations) and the test data (200 observations) on the first two dimensions, with cases $y_i = 0$ in blue stars and $y_i = 1$ in red circles.



(a) (b) (c)

Figure 2: Synthetic data set 1. The color images represent the posterior predictive probability $\Pr(y_* = 1|x_*, \text{data})$ when the first two dimensions of x_* varies, coded such that the predictive probability of $y_* = 1$ increases from near 0 (blue) to near 1 (red). (a) Only the first two dimensions of the data are used in the classification model and the hyper-parameters are optimized with respect to the test error. (b) All 30 dimensions are used in a kernel classification model without variable selection. (c) All 30 dimensions are used in a kernel classification model with variable selection.



(a) (b) (c)

Figure 3: Synthetic data set 2. (a) Scatter plot of test data on the first two dimensions with cases $y_i = 0$ in blue stars and $y_i = 1$ in red circles. (b) Plot of the posterior predictives $\Pr(y_* = 1|x_*, \text{data})$ as the first two dimensions of x_* varies. (c) Credible interval plot for ν_k for $k = 1, \ldots, 20$. The blue solid line indicates the posterior median and the red dashed line indicates the posterior mean for each dimension.

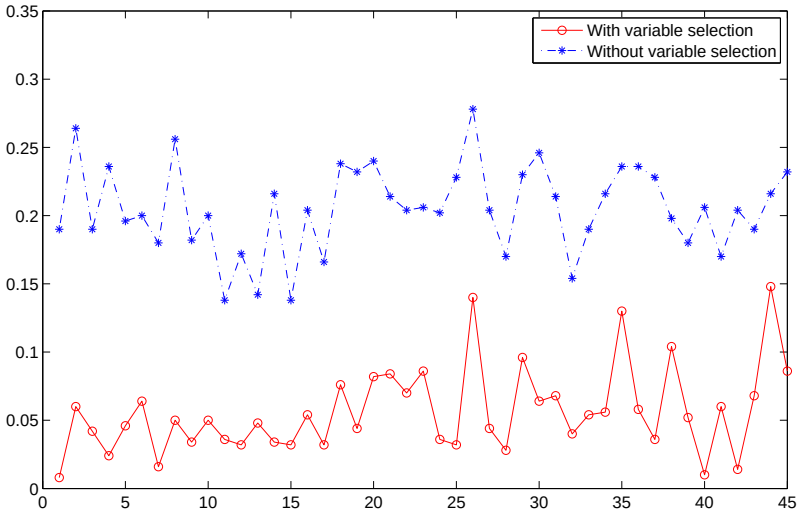


Figure 4: The MNIST data. Plot of the 45 classification errors for the kernel model with variable selection (solid line with circles) and without variable selection (dashed line with stars).

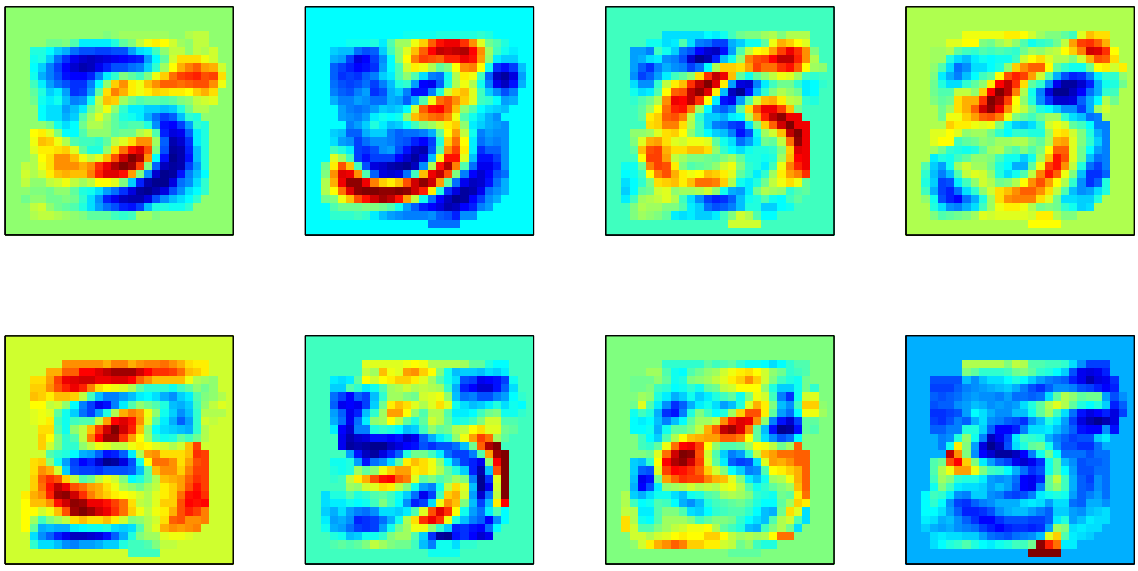


Figure 5: The MNIST data. Upper panel: plot of relevant variables (the 1st, 2nd, 4th and 5th variables). Lower panel: plot of irrelevant variables (the 3rd, 6th, 10th and 11th variables).

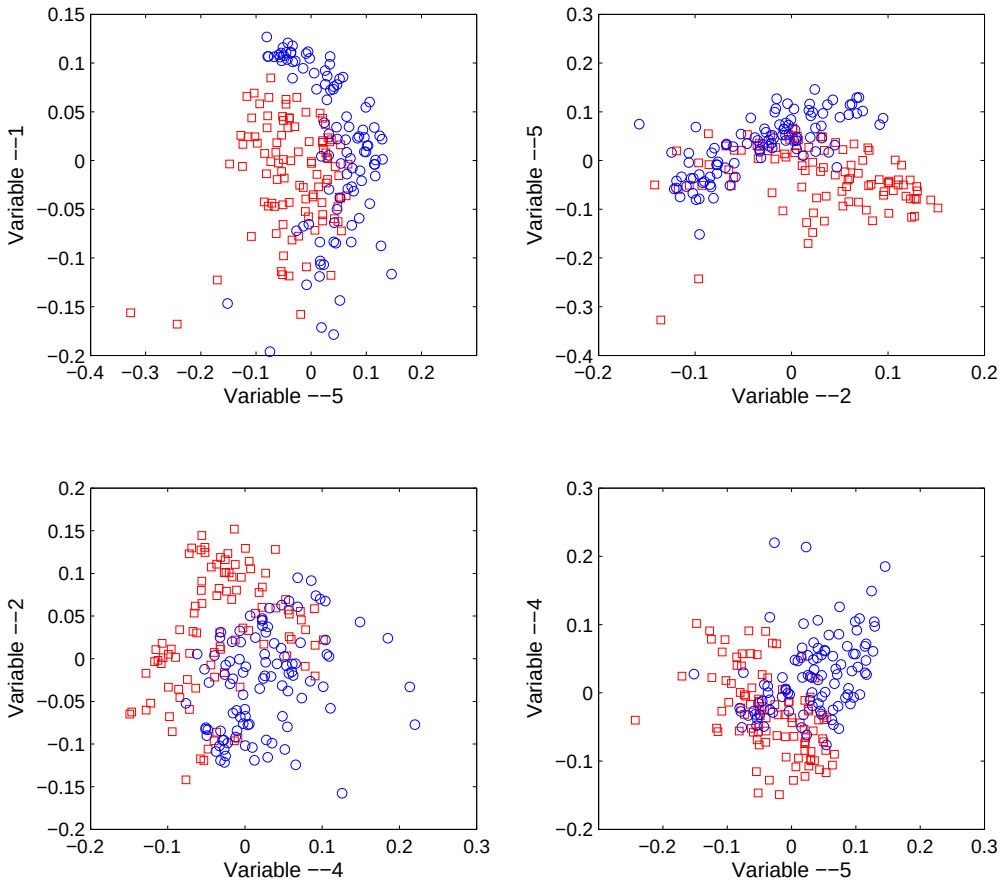


Figure 6: The MNIST data. Plot of projections onto sets of two relevant variables, where circle represents “3” and square represents “5”. The two classes show some separation in the relevant variables.

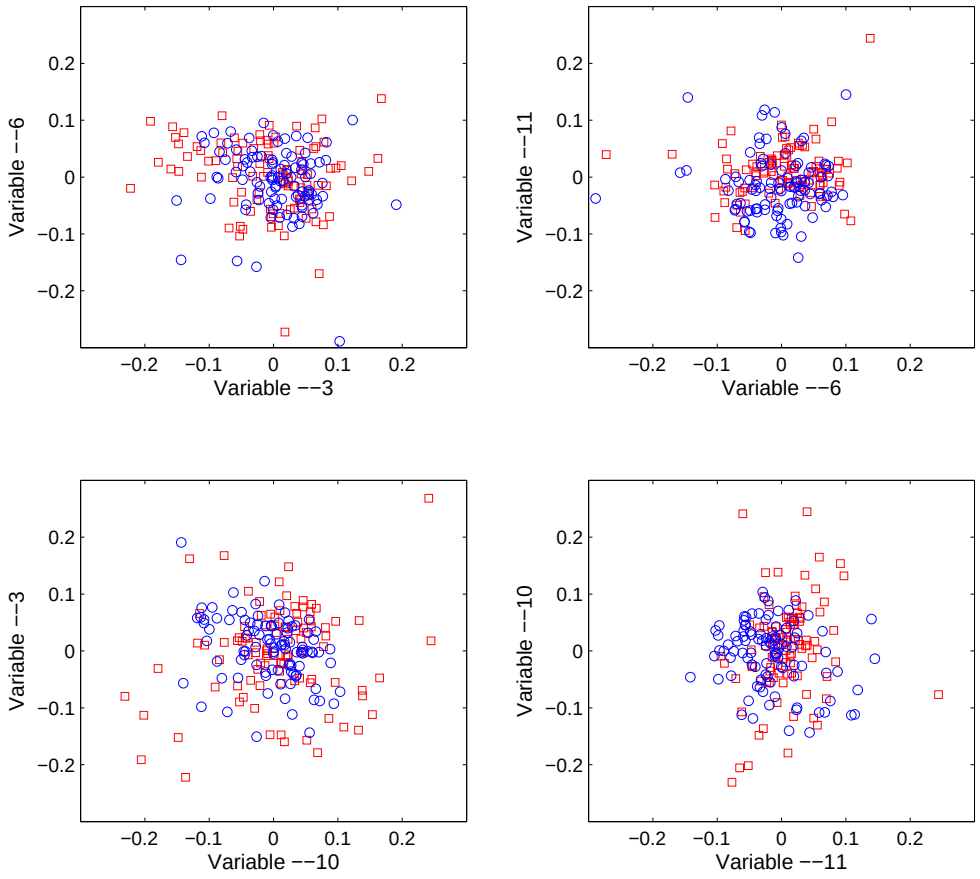


Figure 7: The MNIST data. Plot of projections onto sets of two irrelevant variables, where circle represents “3” and square represents “5”. The two classes are mixed in the irrelevant variables.