Nonparametric Bayes Classification via Learning of Affine Subspaces

Abhishek Bhattacharya
Indian Statistical Institute

based on the paper *Density Estimation and Classification via Bayesian Nonparametric Learning of Affine Subspaces* jointly with David Dunson & Garritt Page, 2012

January 10, 2013
What are we interested in?

- Build efficient nonparametric Bayes classifiers in presence of many predictors.
What are we interested in?

- Build efficient nonparametric Bayes classifiers in presence of many predictors.
- Different cell probabilities allowed to vary non-parametrically based on a few coordinates expressed as linear combinations of the predictors.
What are we interested in?

- Build efficient nonparametric Bayes classifiers in presence of many predictors.
- Different cell probabilities allowed to vary non-parametrically based on a few coordinates expressed as linear combinations of the predictors.
- Model parameters clearly interpretable and provide insight to which predictors are important in constructing accurate classification boundaries.
What are we interested in?

- Build efficient nonparametric Bayes classifiers in presence of many predictors.
- Different cell probabilities allowed to vary non-parametrically based on a few coordinates expressed as linear combinations of the predictors.
- Model parameters clearly interpretable and provide insight to which predictors are important in constructing accurate classification boundaries.
- Estimated cell probabilities consistent in weak and strong sense.
What are we interested in?

- Build efficient nonparametric Bayes classifiers in presence of many predictors.
- Different cell probabilities allowed to vary non-parametrically based on a few coordinates expressed as linear combinations of the predictors.
- Model parameters clearly interpretable and provide insight to which predictors are important in constructing accurate classification boundaries.
- Estimated cell probabilities consistent in weak and strong sense.
- Data applications support the results.
Let $S$ be an affine subspace of $\mathbb{R}^m$ of dimension $k$ ($k \ll m$).

Let $\theta \in \mathbb{R}^m$ be the projection of the origin in $S$ and $R \in \mathbb{R}^{m \times m}$ the projection matrix of the linear subspace parallel to $S$.

Hence $R = R' = R^2$, rank$(R) = k$, $R\theta = 0$.

Let $R = UU'$, $U \in V_{k,m} = \{U \in \mathbb{R}^{m \times k} : U'U = I_k\}$ - the Steifel manifold.

Any $x \in S$ can be given isometric coordinates $\tilde{x} = U'x \in \mathbb{R}^k$ s.t. $x = U\tilde{x} + \theta$. 

For $x \in \mathbb{R}^m$, its projection $P_S(x) = Rx + \theta$ has coordinates $U'x \in \mathbb{R}^k$.

The residual $R_S(x) = x - P_S(x)$ lies in a linear subspace $S^\perp$ perpendicular to $S$ with projection matrix $I - R = VV'$, $V \in V_{m-k,m}$, $V'U = 0$.

It has coordinates $V'(x - \theta)$ in $\mathbb{R}^{m-k}$.
Joint Density Model

- Let $X$ denote the predictor in $\mathbb{R}^m$ and $Y$ a categorical response taking values in $Y = \{1, \ldots, c\}$.

- Will estimate the conditional class probabilities by modeling the joint of $(X, Y)$ s.t. $Y$ depends on $X$ only through its projection onto $S$.

- $(P_S(X), Y)$ has a nonparametric kernel mixture density in $S \times M_c$ while independently $R_S(X)$ follows a mean zero parametric model on $S^\perp$. 
Say \((U'X, Y) \sim \int \mathbb{R}^k \times S_c N_k(x; \mu, \Sigma_1) M_c(y; \nu) P(d\mu d\nu)\) where

- \(N_k\) denotes the \(k\)-variate Normal kernel,
- \(M_c(y; \nu) = \prod_{l=1}^c \nu_l^{I(y=l)}\) is the multinomial kernel and

\[
S_c = \{ \nu \in [0, 1]^c : \sum_l \nu_l = 1 \}.
\]

Independently \(V'(X - \theta) \sim N_{m-k}(0, \Sigma_2)\).
Then \((X, Y) \sim \int_{\mathbb{R}^k \times S_c} N_m(x; U\mu + \theta, \Sigma)M_c(y; \nu)P(d\mu d\nu)\)
where

\[\Sigma = U\Sigma_1 U' + V\Sigma_2 V'.\]

Wlog can take \(\Sigma_1\) and \(\Sigma_2\) to be diagonal.

For sparsity assume \(\Sigma_2 = \sigma_0^2 I_{m-k}\), i.e. the \(X\) residuals are homogeneously distributed.

Let \(\Sigma_1 = \text{diag}(\sigma_1^2, \ldots, \sigma_k^2)\).
Then $\Sigma = U(\Sigma_1 - \sigma_0^2 I_k)U' + \sigma_0^2 I_m$ and the model parameters are

- $k$, $U \in V_{k,m}$, $\theta \in \mathbb{R}^m$ satisfying $U'\theta = 0$, $\underline{\sigma} = (\sigma_0, \sigma_1, \ldots, \sigma_k)$ - a positive vector and $P$ - a probability on $\mathbb{R}^k \times S_c$.

For Bayesian n.p. inference set priors on the parameters s.t. the induced prior on the joint density has full support and the posterior estimate is consistent.
Prior Choice on $\Theta$

- Common prior choice on $\Theta = (k, U, \theta, \sigma, P)$ that preserves conjugacy can be

- a discrete prior on $k$ and given $k$,

- a matrix Bingham-von Mises-Fisher density on $U$ which has the form proportional to $\exp \text{Tr}(UA + UBU'C)$,

- a $m$-variate Normal on $\theta$ restricted to the space of vectors orthogonal to $U$,

- inverse-Gamma priors on the elements of $\sigma$, and,
a Dirichlet process (DP) prior on $P$: $P \sim DP(w_0(P_0 \otimes Q_0))$, where $P_0$ is a $k$-variate Normal and $Q_0$ a Dirichlet distribution on $S_c$.

When $P$ is discrete, say, $P = \sum_{j=1}^{\infty} w_j \delta(\mu_j, \nu_j)$, then

$$P(Y = y|X = x; \Theta) = \sum_{j=1}^{\infty} \tilde{w}_j(U'x)M_c(y; \nu_j)$$

where $\tilde{w}_j(x) = \frac{w_jN_k(x; \mu_j, \Sigma_1)}{\sum_{i=1}^{\infty} w_iN_k(x; \mu_i, \Sigma_1)}$, $x \in \mathbb{R}^k$.

Markov chain Monte Carlo (MCMC) methods can be employed to draw from the posterior.

Choice of o.n. basis leads to rapid convergence and avoids large dimensional matrix inversion.
Consistency of the Conditional Class Probabilities

To show that the conditional density of $Y$ given $X$ under the posterior is consistent.

Assume the following on $f_t$ - the true joint density of $(X,Y)$.

1. $0 < f_t(x, y) < A$ for some constant $A$ for all $(x, y) \in \mathbb{R}^m \times \mathcal{Y}$.
2. $E_t|\log\{f_t(X, Y)\}| < \infty$.
3. For some $\delta > 0$, $E_t \log \frac{f_t(X,Y)}{f_\delta(X,Y)} < \infty$, where $f_\delta(x, y) = \inf_{\tilde{x}: \|\tilde{x} - x\| < \delta} f_t(\tilde{x}, y)$.
4. For some $\alpha > 0$, $E_t \|X\|^{2(1+\alpha)m} < \infty$.

Here $E_t$ denotes expectation under $f_t$. 
Define probability $\tilde{P}_t$ on $\mathbb{R}^m \times S_c$ as

$$\tilde{P}_t(d\mu d\nu) = \sum_{j=1}^{c} f_t(\mu, j)d(\mu)\delta_{e_j}(d\nu)$$

where $e_j$ is the vector with 1 as $j$th coordinate and zeros elsewhere.

Set priors on the parameters such that given $k; (U, \theta), \sigma$ and $P$ are conditionally independent.

Let $(X_n, Y_n) = (X_1, Y_1), \ldots, (X_n, Y_n)$ iid $f_t$. 
Let $Pr(k = m) > 0$ and the conditional priors on $\sigma$ and $P$ given $k = m$ contain $0$ and $\tilde{P}_t$ in their weak supports respectively. Then under assumptions 1-4 on $f_t$, the Kullback-Leibler (KL) condition is satisfied by the induced prior on $f$ at $f_t$.

The proof runs on the same lines of the proof of Theorem 3.1. *Bhattacharya, Page & Dunson 2012*. 
Weak Posterior Consistency (WPC)

Theorem (Weak Posterior Consistency (WPC))

Let $Pr(k = m) > 0$ and the conditional priors on $\sigma$ and $P$ given $k = m$ contain $0$ and $\tilde{P}_t$ in their weak supports respectively.

Then under assumptions 1-4 on $f_t$, the Kullback-Leibler (KL) condition is satisfied by the induced prior on $f$ at $f_t$.

The proof runs on the same lines of the proof of Theorem 3.1. 
*Bhattacharya, Page & Dunson 2012.*

This in turn implies a.s. WPC which implies $\forall \epsilon > 0$,

$$\Pi_n \{ |P(Y = y|X \in U; \Theta) - P_t(Y = y|X \in U)| > \epsilon \} \rightarrow 0 \text{ a.s. } P_t$$

where $\Pi_n$ denotes the posterior of $\Theta$ given $(X_n, Y_n)$. 
Strong Posterior Consistency (SPC)

Theorem (Strong Posterior Consistency (SPC))

Assume the conditions for WPC hold. Pick positive constants $a, b, \{\tau_k\}_{k=1}^m$ and $A$ and set the prior s.t. for $k \leq m-1$, $\|\theta\|^a$ follows a Gamma density, $\max(\sigma) \leq A^{1/b}$, and $Pr(\min(\sigma) < n^{-1/b}|k)$ decays exponentially with $n$. This holds for e.g. with $\sigma_j$’s all equal and $\sigma_j^{-b}$ following a Gamma density truncated to $[A^{-1}, \infty)$. For the DP $(w_k(P_k \otimes Q_0))$ prior on $P$, $k \geq 1$, choose $P_k$ to be a Normal density on $\mathbb{R}^k$ with variance $\tau_k^2I_k$. Then a.s. SPC results if the constants satisfy $\tau_k^2 > 4A^2$, $a < 2(1 + \alpha)m$ and $1/a + 1/b < 1/m$. 
Proof follows from the proof of Theorem 3.5. Bhattacharya, Page & Dunson 2012.
Proof follows from the proof of Theorem 3.5. Bhattacharya, Page & Dunson 2012.

SPC implies

\[ \Pi_n \left\{ \int_{\mathbb{R}^m} \left| P(Y = y|X = x; \Theta) - P_t(Y = y|X = x) \right| g_t(x) dx > \epsilon \right\} \rightarrow 0 \text{ a.s. } P_t \forall y \]

with \( g_t \) the density of \( X \) under \( P_t \).
A Inverse Gamma prior on $\sigma$ satisfies the requirements for weak but not strong posterior consistency.
A Inverse Gamma prior on $\sigma$ satisfies the requirements for weak but not strong posterior consistency.

In *Bhattacharya & Dunson 2011*, a gamma prior is proved eligible when $k = m$ as long as the hyperparameters are allowed to depend on sample size $n$ in a suitable way.

However there it is assumed that $f_t$ has a compact support.

The result is expected to hold true in this context too.
The marginal density of $X$ is

$$X \sim g(x; \Theta) = \int_{\mathbb{R}^k} N_m(x; \phi(\mu), \Sigma) P_1(d\mu),$$

$$\phi(\mu) = U\mu + \theta, \quad \Sigma = U\Sigma_1 U' + V\Sigma_2 V',$$

$P_1$ is the $\mu$ marginal of $P$.

The $X$ component on which $Y$ depends is the $k$-principal component of $X$ if the eigenvalues of $\Sigma_1$ are greater than or equal to those of $\Sigma_2$ (and $P$ is non-degenerate).

This holds if $\Sigma = \sigma_0^2 I$. 
In some sense the model can be considered a Bayesian nonparametric extension of the probabilistic PCA of *Tipping & Bishop 1999* and *Nyamundanda et. al. 2010*.

The model could also be thought of as a nonparametric extension of the Bayesian Gaussian process latent variable models of *Titsias & Lawrence 2010* and SVD models of *Hoff 2007*. 
Estimating $S$

- To obtain a Bayes estimate for the subspace $S$, choose an appropriate loss function and minimize the Bayes risk w.r.t. the posterior distribution.

- $S$ is characterized by its projection matrix $R$ and origin $\theta$, i.e. the pair $(R, \theta)$.

- $R \in \mathbb{R}^{m \times m}$, $\theta \in \mathbb{R}^m$ satisfy $R = R' = R^2$ and $R\theta = 0$. We use $S_m$ to denote the space of all such pairs.
One particular loss function on $S_m$ is

$$L((R_1, \theta_1), (R_2, \theta_2)) = \|R_1 - R_2\|^2 + \|\theta_1 - \theta_2\|^2, \ (R_i, \theta_i) \in S_m,$$

where $\|A\|^2 = \sum_{ij} a_{ij}^2 = \text{Tr}(AA')$.

Then a point estimate for $(R, \theta)$ is the $(R_1, \theta_1)$ minimizing the posterior expectation of loss $L$ over $(R_2, \theta_2)$, provided there is a unique minimizer.
Theorem (Subspace Estimator)

Let \( f(R, \theta) = \int_{(R_2, \theta_2)} L((R, \theta), (R_2, \theta_2)) dP_n(R_2, \theta_2), (R, \theta) \in S_m. \)

This function is minimized by \( R = \sum_{j=1}^{k} U_j U_j' \) and \( \theta = (I - R)\tilde{\theta} \)

where \( \bar{R} \) and \( \bar{\theta} \) are the posterior means of \( R_2 \) and \( \theta_2 \) respectively,

\[
2\bar{R} - \bar{\theta}\tilde{\theta}' = \sum_{j=1}^{m} \lambda_j U_j U_j', \quad \lambda_1 \geq \ldots \geq \lambda_m
\]

is a s.v.d. of \( 2\bar{R} - \bar{\theta}\tilde{\theta}' \), and \( k \) minimizes \( k - \sum_{j=1}^{k} \lambda_j \). The minimizer is unique iff there is a unique minimimizer \( k \) and \( \lambda_k > \lambda_{k+1} \) for that \( k \).
Proof follows from *Bhattacharya et. al. 2012* and *Bhattacharya, A. & Bhattacharya, R. 2012*. 
Proof follows from *Bhattacharya et. al. 2012* and *Bhattacharya, A. & Bhattacharya, R. 2012*.

The relative importance of different features \( \{X_1, \ldots, X_m\} \) in explaining \( Y \) can then be judged by the magnitude of the corresponding diagonal entry of \( R \).

The magnitudes can also be used to group the features according to their relative importance.
Identifiability of $S$

- $X \sim N_m(0, \Sigma) \ast (P_1 \circ \phi^{-1})$, with “$\ast$” denoting convolution.

- The characteristic function of $X$ is
  \[
  \Phi_X(t) = \exp(-1/2t'\Sigma t)\Phi_{P_1 \circ \phi^{-1}}(t), \quad t \in \mathbb{R}^m.
  \]

- If a discrete $P$ is employed, then $\Sigma$ and $P_1 \circ \phi^{-1}$ can be uniquely determined from the marginal of $X$.

- $P_1 \circ \phi^{-1}$ is a distribution on $\mathbb{R}^m$ supported on $S = \phi(\mathbb{R}^k)$. 
Define the *affine support* of a probability $Q$, $\text{asupp}(Q)$ as the intersection of all affine subspaces having prob. 1. It contains the support $\text{supp}(Q)$ (but may be larger).

To identify $S$ and $k$ we assume that $\text{asupp}(P_1)$ is $\mathbb{R}^k$.

Then $\text{asupp}(P_1 \circ \phi^{-1})$ is an affine subspace of $\mathbb{R}^m$ of dimension equal to that of $\text{asupp}(P_1) = k$. 
Since $\text{asupp}(P_1 \circ \phi^{-1})$ is identifiable, this implies that $k$ is also identifiable as its dimension.

Since $S$ contains $\text{asupp}(P \circ \phi^{-1})$ and has dimension equal to that of $\text{asupp}(P \circ \phi^{-1})$, $S = \text{asupp}(P \circ \phi^{-1})$.

Then $R = UU'$ and $\theta$ are identifiable as the projection matrix and origin of $S$. 
The classifier built (PSC) is used in real data examples and its performance compared with other well known classification methods.
The classifier built (PSC) is used in real data examples and its performance compared with other well known classification methods.

Three such competitors considered are $k$ nearest neighbor (KNN), mixture discriminant analysis (MDA), and support vector machine (SVM).
KNN is algorithmic based and classifies well in a variety of settings. A range of neighborhood sizes are considered with the one producing the best out of sample prediction ultimately used.
KNN is algorithmic based and classifies well in a variety of settings. A range of neighborhood sizes are considered with the one producing the best out of sample prediction ultimately used.

MDA is a flexible model based Gaussian mixture classifier (see Hastie & Tibshirani 1996). The number of components in the Gaussian mixture chosen to produce the best out of sample prediction.
KNN is algorithmic based and classifies well in a variety of settings. A range of neighborhood sizes are considered with the one producing the best out of sample prediction ultimately used.

MDA is a flexible model based Gaussian mixture classifier (see Hastie & Tibshirani 1996). The number of components in the Gaussian mixture chosen to produce the best out of sample prediction.

SVM is a very accurate classifier and is therefore included.
- KNN is algorithmic based and classifies well in a variety of settings. A range of neighborhood sizes are considered with the one producing the best out of sample prediction ultimately used.

- MDA is a flexible model based Gaussian mixture classifier (see Hastie & Tibshirani 1996). The number of components in the Gaussian mixture chosen to produce the best out of sample prediction.

- SVM is a very accurate classifier and is therefore included.

- Out of sample prediction error rates used to compare PSC to the 3 competitors.
Brain Computer Interface (BCI) Data

- The BCI dataset consists of a single person performing 400 trials in each of which he imagined movements with either the left hand or the right hand.

- For each trial, EEG recorded from 39 electrodes.

- An autoregressive model of order 3 was fit to each of the resulting 39 time series.
The trial is then represented by the total of $117 = 39 \times 3$ dimensional feature space.

Goal is to classify each trial as left or right hand movements using the 117 features.

200 observations randomly selected to serve as testing data.

Posterior combinations done with dimension $k$ fixed.
To select a $k$ the out of sample prediction error rates and area under the receiver operating characteristic (ROC) curve are employed.

Since low out of sample prediction error rates and large areas under the curve are desirable, a $k$-value at-most 25 that maximized the difference between them is selected.

Following this criteria, $k = 3$ chosen.

PSC produces an out of sample prediction error rate of 0.205 compared to 0.51 for KNN, 0.25 for MDA and 0.23 for SVM.
Wisconsin Breast Cancer (WBC) data set

- In this data set the response is breast cancer diagnosis while the covariates consists of 9 nominal variables describing some type of breast tissue cell characteristic.

- Although this data set is not high dimensional, it provides a nice illustration of the type of information the PSC can provide regarding associations between covariates and response.

- Similar to what was done with the BCI data set $k = 3$ is selected.

- This results in an out of sample prediction error rate of 0.017 which is smaller than the error rate for KNN (0.035), MDA (0.028) and SVM (0.028).
Even though the PSC classifies more accurately than the other methods, what is of particular interest is how each of the 9 tumor attributes influence classification.

The 9 attributes (clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitosis) are all related to a lump being benign or not.

From the theorem on subspace estimation the estimated principal directions are found in the Table below.
Theorem (Subspace Estimator)

Let $f(R, \theta) = \int_{(R_2, \theta_2)} L((R, \theta), (R_2, \theta_2)) dP_n(R_2, \theta_2)$, $(R, \theta) \in S_m$. This function is minimized by $R = \sum_{j=1}^k U_j U_j'$ and $\theta = (I - R)\bar{\theta}$ where $R$ and $\bar{\theta}$ are the posterior means of $R_2$ and $\theta_2$ respectively,

$$2\bar{R} - \bar{\theta}\bar{\theta}' = \sum_{j=1}^m \lambda_j U_j U_j', \quad \lambda_1 \geq \ldots \geq \lambda_m$$

is a s.v.d. of $2\bar{R} - \bar{\theta}\bar{\theta}'$, and $k$ minimizes $k - \sum_{j=1}^k \lambda_j$. The minimizer is unique iff there is a unique minimimizer $k$ and $\lambda_k > \lambda_{k+1}$ for that $k$. 
Table: The $k = 3$ principal directions of the Breast Cancer data set along with the row norms

| Variable                        | $U_{[1]}$         | $U_{[2]}$         | $U_{[3]}$         | norm   |
|---------------------------------|-------------------|-------------------|-------------------|--------|
| clump thickness                 | -0.294            | 0.233             | 0.453             | 0.588  |
| uniformity of cell size         | -0.399            | -0.132            | -0.189            | 0.460  |
| uniformity of cell shape        | -0.395            | -0.102            | 0.0172            | 0.408  |
| marginal adhesion               | -0.314            | -0.007            | -0.477            | 0.571  |
| single epithelial cell size     | -0.231            | -0.181            | -0.307            | 0.424  |
| bare nuclei                     | -0.450            | 0.713             | 0.101             | 0.849  |
| bland chromatin                 | -0.295            | -0.032            | -0.194            | 0.354  |
| normal nucleoli                 | -0.376            | -0.587            | 0.543             | 0.883  |
| mitosis                         | -0.121            | -0.173            | -0.305            | 0.371  |
A way to assess the relative importance of each variable and also provide a means of grouping the variables is to calculate the norm associated with each row of $U$ (i.e. the norm of the corresponding diagonal entry of $R = UU^T$).

These values can be found under the header “norm” in the Table.

It appears that a bare nuclei and normal nucleoli form a group.

Another is formed by clump thickness and marginal adhesion.

Finally it appears that uniformity of cell size, uniformity of cell shape and single epithelial cell size form a group.
Summary

- A flexible nonparametric model proposed for classification via feature space dimension reduction.
Summary

- A flexible nonparametric model proposed for classification via feature space dimension reduction.
- The model satisfies large support & consistency conditions.
Summary

- A flexible nonparametric model proposed for classification via feature space dimension reduction.

- The model satisfies large support & consistency conditions.

- A simple Gibbs sampler can be implemented with conjugate sampling steps for posterior sampling.
Summary

- A flexible nonparametric model proposed for classification via feature space dimension reduction.

- The model satisfies large support & consistency conditions.

- A simple Gibbs sampler can be implemented with conjugate sampling steps for posterior sampling.

- Better performance than commonly used machine learning, computer science and parametric statistical methods.
These methods are algorithmic or highly parameterized black boxes and apart from classification, provide no further information specific to the problem being studied.
These methods are algorithmic or highly parameterized black boxes and apart from classification, provide no further information specific to the problem being studied.

In addition to building efficient classifiers, the proposed methodology provides insight regarding predictors that are influential in explaining the response - an information applied scientists often highly value.
These methods are algorithmic or highly parameterized black boxes and apart from classification, provide no further information specific to the problem being studied.

In addition to building efficient classifiers, the proposed methodology provides insight regarding predictors that are influential in explaining the response - an information applied scientists often highly value.

Can easily be extended to other regression setup.
Further Work possible

- Change the joint kernel choice to build better classifier.
Further Work possible

- Change the joint kernel choice to build better classifier.
- Change the notion of inner product to use non-linear predictor transformations to explain the response.
Further Work possible

- Change the joint kernel choice to build better classifier.
- Change the notion of inner product to use non-linear predictor transformations to explain the response.
- A nonparametric model may be fit on the non-signal predictors as well.
Further Work possible

- Change the joint kernel choice to build better classifier.
- Change the notion of inner product to use non-linear predictor transformations to explain the response.
- A nonparametric model may be fit on the non-signal predictors as well.
- Use other priors besides Dirichlet Process.
Further Work possible

- Change the joint kernel choice to build better classifier.

- Change the notion of inner product to use non-linear predictor transformations to explain the response.

- A nonparametric model may be fit on the non-signal predictors as well.

- Use other priors besides Dirichlet Process.

- Extend to nonparametric hypothesis testing on the lines of Bhattacharya & Dunson 2012.
References

Bhattacharya, A. & Bhattacharya, R. (2012). Nonparametric Statistics on Manifolds with Applications to Shape Spaces, IMS Monograph 2, Cambridge University Press.

Bhattacharya, A. & Dunson, D. (2011). Strong consistency of nonparametric Bayes density estimation on compact metric spaces with applications to specific manifolds. Ann Inst Stat Math 64, 687-714.

Bhattacharya, A. & Dunson, D. (2012). Nonparametric Bayes classification and hypothesis testing on manifolds. Jour. Multiv. Analysis 111, 1-19.
References

Bhattacharya, A., Page, G., Dunson, D.B. (2012). Density estimation and classification via Bayesian nonparametric learning of affine subspaces. *JASA*, revision submitted.

Hastie, T. & Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures. *JRSSB* 58, 155-176.

Hoff, P.D. (2007). Model Averaging and Dimension Selection for the Singular Value Decomposition. *JASA* 102: 674-685.

Nyamundanda, G., Brenna, L. & Gormley, I.C. (2010). Probabilistic Principal Component Analysis. *BMC Bioinformatics* 11: 571.

Tipping, M.E. & Bishop, C.M. (1999). Probabilistic Principal Component Analysis. *JRSSB* 61, 611-622.
TITSIAS, M.K. & LAWRENCE, N.D. (2010). Bayesian Gaussian Process Latent Variable Model. *Proc. 13th Int. Workshop on Art. Intelligence & Stat.* 9, 25-32.