

Wavelet-based Variable Selection for Discriminant Analysis with Markov Random Tree Priors

Francesco Claudio Stingo, Marina Vannucci and Gerard Downey

Abstract Discriminant analysis is an effective tool for classification of experimental units into groups. When the number of variables is much bigger than the number of observations it is necessary to include a variable selection procedure into the inferential process. Typical examples are the classification of subjects according to phenotypes via gene expressions and the classification of different types of food into species via near infrared spectroscopy. We build upon a conjugate normal discriminant model, either linear and quadratic, and include a stochastic search variable selection procedure via an MCMC algorithm. For the spectroscopy example we show how to perform discriminant analysis in the wavelet domain and include a Markov random tree (MRT) prior that maps the connections among wavelets coefficients. We show good performances of the method in simulation studies, gene expression analysis and food authenticity and compare results to several other procedures.

Key words: Classification and pattern recognition, Bayesian variable selection, Markov chain Monte Carlo, Markov random tree prior, Wavelet-based modeling

1 Introduction to Discriminant Analysis

Discriminant analysis, sometimes called supervised pattern recognition, is a statistical technique used to classify observations into groups. For each case in a given training set a $q \times 1$ vector of observations, X_i , and a known assignment to one of G

Francesco Claudio Stingo
Department of Statistics, Rice University, Houston (USA), e-mail: fcs1@rice.edu

Marina Vannucci
Department of Statistics, Rice University, Houston (USA), e-mail: marina@rice.edu

Gerard Downey
TEAGASC, Ashtown Food research Center, Ashtown, Dublin 15 (Ireland), e-mail: Gerard.Downey@teagasc.ie

groups are available. The group indicators are stored in a $n \times 1$ vector $\mathbf{y}$. On the basis of these data, we wish to derive a classification rule that will assign further cases to their correct groups. If the distribution of $\mathbf{X}$ conditional on the group membership is assumed to be a multivariate normal then the appropriate statistical methodology is that of discriminant analysis. When the number of observed variables is considerably large it is necessary to exclude from the model the set of noisy variables that could lead to a high misclassification rate. The method for variable selection we propose can be used both for linear discriminant analysis, where all the groups share the same covariance matrix, and for quadratic discriminant analysis, where different groups are allowed to have different covariance matrices. In this paper we illustrate the model for the quadratic discriminant analysis case.

2 The Model

2.1 Bayesian Discriminant Analysis

We assume that each observation comes from one of the G possible groups with distribution $N(\mu_g, \Sigma_g)$. We represent the data from each group by the $n_g \times q$ matrix $\mathbf{X}_g - \mathbf{1}_{n_g} \mu_g^T \sim \mathcal{N}(I, \Sigma_g)$, with $g = 1, \dots, G$, where the vector μ_g and the matrix Σ_g are the mean and the covariance matrix of the g -th group. With $\mathcal{N}(\cdot, \cdot)$ we indicate a matrix normal variate, see [3]. Taking a conjugate Bayesian approach, we impose a multivariate normal distribution on $\mu_g \sim N(m_g, h_g \Sigma_g)$ and an Inverse-Wishart prior on the covariance matrix $\Sigma_g \sim IW(\delta_g, \Omega_g)$. This parametrization, beside being the standard setting in the Bayesian approach, allows us to create a computationally efficient variable selection algorithm, as shown in Section 3.1. The predictive distribution of a new observation X_j^f ($1 \times q$) is used to classify every new sample into one of the possible G groups. This distribution, see [1] among others, is a multivariate T-student:

$$X^f - \tilde{\mu}_g^T \sim \mathcal{T}(\delta_g^*, a_g, \Omega_g^*), \quad (1)$$

where $\tilde{\mu}_g = \pi_g m_g + (1 - \pi_g) \bar{X}_g$, $\delta_g^* = \delta_g + n_g$, $a_g = 1 + (1/h_g + n_g)^{-1}$ and $\Omega_g^* = \Omega_g + S_g + (h_g + 1/n_g)^{-1} (\bar{X}_g - m_g)^T (\bar{X}_g - m_g)$ with $\pi_g = (1 + h_g n_g)^{-1}$ and $S_g = (\mathbf{X}_g - \mathbf{1}_{n_g} \bar{X}_g)^T (\mathbf{X}_g - \mathbf{1}_{n_g} \bar{X}_g)$. The probability that a future observation, given the observed data, belong to the group g is then given by:

$$\pi_g(y^f | \mathbf{X}) = p(y^f = g | X_g^f, \mathbf{X}) \quad (2)$$

where y^f is the group indicator and $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_G\}$ are the observed data. By estimating the prior probability that one observation comes from group g as $\hat{\pi}_g = n_g/n$, the previous distribution can be written in closed form as:

$$\pi_g(y^f | \mathbf{X}) = \frac{p_g(X^f) \hat{\pi}_g}{\sum_{i=1}^G p_i(X^f) \hat{\pi}_i},$$

where $p_g(X^f)$ indicates the predictive distribution defined in equation (1). A new observations is then assigned to the group with the highest posterior probability.

2.2 Likelihood and prior setting for variable selection

In order to perform variable selection we include a $(q \times 1)$ latent binary vector γ , where the elements equal to 1 indicate the selected variables, *i.e.*, the variables able to classify the units into the correct groups. Extending the approach of [13], to avoid any independence assumptions, we define the following likelihood that allows us to separate the discriminant variables from the noisy ones as:

$$L(X, y; \eta, \Omega, \mu_g, B, \Sigma_g, \gamma) = \prod_{i=1}^n p(X_{i(\gamma)} | X_{i(\gamma)}) \prod_{g=1}^G \prod_{j=1}^{n_g} w_g^{n_g} p_g(X_{j(\gamma)}). \quad (3)$$

The first factor of the likelihood refers to the non important variables, while the second is formed by variables able to classify observations into the correct groups. Under the normality assumption the likelihood becomes:

$$\prod_{i=1}^n N_{|\gamma^c|}(X_{i(\gamma^c)} - BX_{i(\gamma)}; \eta_{(\gamma^c)}, \Sigma_{(\gamma^c)}) \prod_{g=1}^G \prod_{j=1}^{n_g} w_g^{n_g} N_{|\gamma|}(X_{j(\gamma)}; \mu_{g(\gamma)}, \Sigma_{g(\gamma)}). \quad (4)$$

Note that [11] use the same likelihood in a frequentist framework. For the parameters corresponding to the non selected variables we use the following conjugate priors:

$$\begin{aligned} \eta_{(\gamma^c)} | \Sigma_{(\gamma^c)} &\sim N(\mu_0(\gamma^c), h_0 \Sigma_{(\gamma^c)}) \\ B - B_0 | \Sigma_{(\gamma^c)} &\sim \mathcal{N}(H_\gamma, \Sigma_{(\gamma^c)}) \\ \Sigma_{(\gamma^c)} &\sim IW(\delta, \Omega_0(\gamma^c)). \end{aligned} \quad (5)$$

Without loss of generality, at least for the inferential procedure described in Section 3.1, we can assume that the set of non selected variables is formed by only one variable so that the prior parametrization can be simplified using the scalar σ^2 instead of $\Sigma_{(\gamma^c)}$ and the $(p \times 1)$ vector β instead of the $q \times p$ matrix B with $\sigma^2 \sim \text{Inv} - \text{Gamma}(\delta/2, k_0/2)$ and $\beta \sim N(\beta_0, \sigma^2 H_\gamma)$. To obtain this parametrization the commonly used assumption $\Omega_0(\gamma^c) = k_0 I_{|\gamma^c|}$ is needed, see for example [13] for a model based clustering context, and [5] for a graphical model context. We then define an improper non-informative prior on the vector $w = (w - 1, \dots, G)$ using a Dirichlet distribution, $w \sim \text{Dirichlet}(0, \dots, 0)$. Note that, with the inclusion of the variable selection mechanism, the predictive distribution does not change because it depends only on the selected variables.

2.3 Prior distribution for network information integration

Although the model allows for dependencies among variables, it is not straightforward to specify dependence structures known a priori on the covariance matrix prior. When prior information is available we can easily incorporate it into the prior distribution of γ . In order to define this prior we need to introduce a $q \times q$ matrix R . This matrix describes relationships between variables, with elements $r_{ij} = 1$ if variable i and j have a direct link in the network, and $r_{ij} = 0$ otherwise. As already described by [10] and [2], these relations can be modeled using a Markov random field (MRF), where variables are represented by nodes and relations between them by edges. A MRF is a graphical model in which the distribution of a set of random variables follows Markov properties that can be described by an undirected graph. Relations on the MRF are represented by the following probabilities

$$\pi(\gamma_j | \eta, \gamma_i, i \in N_j) = \frac{\exp(\gamma_j F(\gamma_j))}{1 + \exp(F(\gamma_j))}, \quad (6)$$

where $F(\gamma_j) = \mu + \eta \sum_{i \in N_j} (2\gamma_i - 1)$ and N_j is the set of direct neighbors of variable j in the MRF. The global distribution on the MRF is given by

$$\pi(\gamma | \mu, \eta) \propto \exp(\mu n_1 - \eta n_{01}), \quad (7)$$

where n_1 is the number of selected variables and n_{01} is the number of edges linking nodes with different values of γ_j (*i.e.*, edges linking included and non-included nodes). The parameter μ controls the sparsity of the model, while higher values of η result in neighboring variables taking on the same γ_j value. If a variable does not have any neighbor, its prior distribution reduces to an independent Bernoulli with parameter $p = \exp(\mu) / [1 + \exp(\mu)]$, which is a logistic transformation of μ . In Section(4.2), following [12], we use a specific network structure, called Markov Random Tree (MRT), where the connections among wavelet coefficients depend on their position along the space dimension, *i.e.* every coefficient at a given scale can be related to a set of coefficients at the next finer scale corresponding to the same location. When network information is not available q independent Bernoulli priors can be specified for the elements of γ :

$$\pi(\gamma) = \prod_{j=1}^p p^{\gamma_j} (1-p)^{1-\gamma_j}$$

where p represents the expected number of important variables a priori.

2.4 Discriminant analysis in wavelet domain for functional data

As noted by [7], an alternative approach to variable selection is to apply principal component analysis to $\mathbf{X}$ and then perform discriminant analysis on the first k principal components. This dimension reduction approach is motivated by the fact that, with functional data, working with transformed variables, such as principal components or, in our case, wavelet coefficients, can improve the classification performance. Note that in our framework we incorporate both selection and synthesis of the data by transforming the data into wavelet coefficients and then selecting the important ones. We transform the data to the wavelet domain by $\mathbf{Z}_g = \mathbf{X}_g W'$, where W is an orthogonal matrix ($W'W = I$) representing the discrete wavelet transform (DWT). We rewrite the model as:

$$\mathbf{Z}_g - \hat{\mu}'_g \sim \mathcal{N}(I, W \Sigma_g W'),$$

$$\tilde{\Sigma}_g \sim IW(\delta, \tilde{\Omega}_g),$$

with $\tilde{\Sigma}_g = W \Sigma_g W'$ and $\tilde{\Omega}_g = W \Omega_g W'$. Once the wavelet transform is applied we can use the model presented in the previous sections taking into account the different meaning of the new hyperparameters.

3 Posterior Inference

3.1 MCMC for variable selection

Variable selection is performed via the posterior distribution of the latent indicator vector γ . This distribution cannot be obtained in closed form and a MCMC is needed. However the inferential process can be simplified integrating over β , σ^2 , w_g , μ_g and Σ_g and obtaining the following marginal likelihood:

$$\begin{aligned} p(X|y, \gamma) &\propto (k_0 + (\mathbf{X}_{(\gamma^c)} - 1_{p_\gamma} \mu_{0(\gamma^c)} - \mathbf{X}_{(\gamma)} \beta)^T (I_n + h_0 \mathbf{1}_n \mathbf{1}_n^T + \mathbf{X}_{(\gamma)} H_\gamma \mathbf{X}_{(\gamma)}^T)^{-1} \\ &\quad (\mathbf{X}_{(\gamma^c)} - 1_{p_\gamma} \mu_{0(\gamma^c)} - \mathbf{X}_{(\gamma)} \beta))^{-\frac{n+\delta}{2}} \\ &\quad \prod_{g=1}^G K_g(\gamma) |Q_{1(\gamma)}|^{(\delta+p_\gamma-1)/2} |Q_{1(\gamma)} + S_g(\gamma)|^{-(n_g+\delta+p_\gamma-1)/2} \end{aligned} \quad (8)$$

where

$$\begin{aligned} K_g(\gamma) &= (h_1 n_g + 1)^{-p_\gamma/2} \prod_{j=1}^{p_\gamma} \Gamma(1/2(n_g + \delta + p_\gamma - j)) / \Gamma(0.5(\delta + p_\gamma - j)) \\ S_g(\gamma) &= \sum_{i| \gamma_i=1}^{n_g} (X_{i(\gamma)} - \bar{X}_{g(\gamma)})(X_{i(\gamma)} - \bar{X}_{g(\gamma)})^T + \\ &\quad \frac{n_g}{h_0 n_g + 1} (\mu_{0(\gamma)} - \bar{X}_{g(\gamma)})(\mu_{0(\gamma)} - \bar{X}_{g(\gamma)})^T. \end{aligned} \quad (9)$$

Following [13] and [9] we develop a Stochastic Search Variable Selection (SSVS) algorithm to perform variable selection. This is Metropolis type of algorithm that uses two different moves, either the addition/deletion of one selected variable or the swapping of one selected variable with a non selected one.

The proposed γ^{new} is then accepted with a probability that is the ratio of the relative posterior probabilities of the new versus the current model:

$$\min \left[\frac{p(X|y, \gamma^{new})\pi(\gamma^{new})}{p(X|y, \gamma^{old})\pi(\gamma^{old})}, 1 \right]. \quad (10)$$

Because these moves are symmetric, the proposal distribution does not appear in the previous ratio. As noted by [11] in a different context, the calculation of the acceptance probability in the Metropolis step can be simplified using the following factorization of the marginal likelihood:

$$p(\mathbf{X}_{(\gamma)}|y, \mathbf{X}_{(\gamma)}, \mathbf{X}_{(prop)})p(\mathbf{X}_{(\gamma)}, \mathbf{X}_{(prop)}|y)$$

where $\mathbf{X}_{(prop)}$ represents the variable, for an add/delete move, or the two variables, for a swap move, whose indicator element has been proposed to change.

4 Data Analysis

We show how the model can correctly identify significant variables in a small n large p framework, by analyzing both gene expression data and near infrared spectroscopy data.

4.1 Application to Microarray Data

We illustrate how our model can be applied to DNA microarray data, where thousands of gene expression levels are observed on a few samples. We analyze the widely used leukemia data of [8] using 38 patients as the training set and 34 patients as validation set. In this analysis gene expressions are supposed to discriminate acute lymphoblastic leukaemia (ALL) patients from acute myeloid leukaemia (AML) patients. Following [6], we truncate expression measures beyond the threshold of reliable detection at 100 and 16,000, and remove probe sets with intensities such that $max/min \leq 5$ and $max - min \leq 500$. This leave us with 3,571 genes for the analysis. The expression readings are log-transformed and each variable is rescaled by its range.

This data has been already analyzed by [9] using a similar model for cluster analysis. As done by [9], we assume that the non significant variables are marginally independent of the significant ones. The hyperparameters are taken to $\delta = 3$, $h_1 = 10$, $h_0 = 100$, $\Omega_1 = 0.6^{-1} \cdot I_{|\gamma|}$ and $k_0 = 10^{-1}$; we set the expected value of the Bernoulli

prior on γ equal to $p = 0.03 (\simeq 100/3,571)$ and assume unequal covariance across the groups. The sampler was started with a random selected model with 10 significant variables and 150,000 iterations were run with the first 50,000 used as burn-in. With a threshold of 0.5 we select 69 significant variables that are able to correctly classify 30 of the 34 patients of the validation set.

4.2 Application to Spectral Data

Discriminant analysis has been frequently used to classify units into groups by observing near infrared spectra, see [7] for a Bayesian approach with variable selection. Recently [11] have shown how variable selection within a discriminant analysis model for classification of different types of foods outperforms other techniques such as Transductive Support Vector Machine (SVM), Random Forest, AdaBoost, Bayesian Multinomial Regression, Factorial Discriminant Analysis (FDA), k-nearest neighbors, discriminant partial least squares (PLS) regression and soft independent modeling of class analogy (SIMCA). Correct identification of food via analysis of near infrared spectroscopy data is an important tool to avoid potential fraud like supplying of cheaper inauthentic food instead of the more expensive original products. In this Section we analyze spectra coming from five different species of meat, Beef, Chicken, Lamb, Pork and Turkey. These data have been already analyzed using a variable selection/discriminant analysis procedure by [11] and using a wavelets-based discriminant analysis by [4], first thresholding the wavelet coefficients to reduce the dimensionality of the data on the wavelets domain and then performing a discriminant analysis only on the selected subset of wavelets coefficients. After removing the first 27 variables, we transform the remaining 1,024 wavelengths into the wavelet domain using Daubechies wavelets with 3 vanishing moments, obtaining 8 scaling coefficients and 1,016 wavelets coefficients. Because the scaling coefficients carry only general features of the signal, that are usually not able to discriminate units, we decided to perform our selection only on the standardized wavelets coefficients. Then we set $\delta = 3$, $h_1 = 100$, $h_0 = 1000$, $\Omega_1 = 3^{-1} \cdot I_{|\gamma|}$, $H_\gamma = 100 \cdot I_{|\gamma|}$ and $k_0 = 10^{-1}$; the hyperparameters of the MRT were set $\mu = -2.5$ and $\eta = 0.3$. We assumed unequal covariance across the groups. The sampler was started first with a random selected model with 2 significant variables and then with 10 significant variables; the two 200,000 iterations chains were then pooled together; the first 1,000 iterations were used as burn-in, because the algorithm seems to quickly select significant variables. With a threshold of 0.5 we select 7 significant variables that are able to correctly classify 102 of the 114 food samples of the validation set while, with a threshold of 0.5 and 14 selected variables, the model is able to classify 110 of the 114 units. With a missclassification rate of 3.4% our method is a better classifier than the model proposed by [11], and, specifically, is able to better separate Turkey and Chicken samples.

5 Conclusion

We have illustrated how to perform variable selection in discriminant analysis following the Bayesian paradigm. We have described how information on the network structure of the variables can be easily incorporated into the prior distribution and how posterior inference can be performed, either in the original data domain or in the wavelet domain, via a fast MCMC. We have shown that our model performs well in two practical important applications.

References

1. Brown, P.J.: Measurement, Regression, and Calibration. Oxford University Press, Oxford (1993)
2. Chen Y.A., Stingo F.C., Tadesse M.G., Vannucci M.: Incorporating Biological Information in Bayesian Models for the Selection of Pathways and Genes. (submitted) (2009)
3. Dawid A.P.: Some Matrix-Variate Distribution Theory: Notational Considerations and a Bayesian Application. *Biometrika*. **68**, 265-274 (1981)
4. Dean N., Murphy T.B., Downey G.: Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society, Series C: Applied Statistics*. **5**, 1-14 (2006)
5. Dobra A., Jones B., Hans C., Nevins J.R., West M.: Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*. **90**, 196-212 (2004)
6. Dudot S., Fridlyand J., Speed T.P.: Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data. *Journal of the American Statistical Association*. **97**, 77-87 (2002)
7. Fearn T., Brown P.J., Besbeas P.: A Bayesian decision theory approach to variable selection for discrimination. *Statistics and Computing*. **12(3)**, 253-260 (2002)
8. Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*. **286**, 531-37 (1999)
9. Kim S., Tadesse M.G., Vannucci M.: Variable selection in clustering via Dirichlet process mixture models. *Biometrika*. **93(4)**, 877-93 (2006)
10. Li F., Zhang N.R.: Bayesian Variable Selection in Structured High-Dimensional Covariate Space with Application in Genomics. *Journal of American Statistical Association* (to appear). , - (2010)
11. Murphy T.B., Dean N., Raftery A.E.: Variable Selection and Updating In Model-Based Discriminant Analysis for High Dimensional Data with Food Authenticity Applications. *Annals of Applied Statistics* (accepted). , - (2010)
12. Shapiro J.M.: Embedded Image Coding Using Zeotrees of Wavelet Coefficients. *IEEE Transactions on Signal Processing*. **41(12)**, 3445-62 (1993)
13. Tadesse M.G., Sha N., Vannucci M.: Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*. **100**, 602-17 (2005)