The University of British Columbia
Department of Statistics
Technical Report #231

Bayesian Variable Selection for Semi-Supervised Learning, with Application to Object Recognition

Paul Gustafson, Natalie Thompson, and Nando de Freitas

July 10, 2007

Department of Statistics
University of British Columbia
333-6356 Agricultural Road
Vancouver, B.C.
V6T 1Z2
CANADA

Bayesian Variable Selection for Semi-Supervised Learning, with Application to Object Recognition

Paul Gustafson
University of British Columbia
gustaf@stat.ubc.ca

Natalie Thompson
University of Toronto

Nando de Freitas
University of British Columbia

July 10, 2007

SUMMARY

The use of Bayesian methods for model-selection and model-averaging has received considerable attention in the literature, particularly in the context of choosing a subset of relevant explanatory variables when modelling the distribution of a response variable given an initial set of explanatory variables. However, similar approaches when modelling the distribution of explanatory variables given a response variable have received less attention. Motivated by an object-recognition problem we consider such techniques. The application, which we describe in detail, involves assigning descriptive labels to parts (segments) of images. The training data have labels at the level of an image, but not at the level of a segment. Thus the data are neither fully labelled nor completely unlabelled, and we describe the learning problem as *semi-supervised*. One emphasis is on a novel and intuitive model whereby the subset of the available explanatory variables regarded as useful in identifying image segments matching label A can differ from the subset used for another label B.

Keywords: Bayesian analysis; mixture models; object recognition; model averaging.

1 Introduction

Consider the problem of classifying units as belonging to one of $K+1$ categories or *concepts*, based on measurements of p variables or *features* on each unit. For a given unit let $X = (X_1, \ldots, X_p)$ be the feature measurements and let $Y \in \{0, 1, \ldots, k\}$ denote the concept to which the unit belongs. The goal is to infer the conditional distribution of $(Y|X)$. If the available training data consist of (Y, X) measurements for n units then we have a *supervised-learning* problem of classification. If the available training data consist of X measurements alone then we have a much harder *unsupervised-learning* or clustering problem, and it may not even be reasonable to take the number of categories K as known. While some of the ideas in this article may be useful in the supervised and unsupervised settings, our main focus is on an application which is probably best described as *semi-supervised*. In this application Y is never observed at the unit level, but there is some direct information about Y in the data.

2 Motivating Application

The motivating application is an object-recognition problem that can be framed as follows. The training data arise from a series of I=68 images. The *normalized cuts* algorithm (Shi and Malik 1997) is used to partition the i-th image into J_i contiguous *segments* or regions that seem homogeneous in terms of properties such as colour and 'texture.' While there are statistical issues surrounding segmentation algorithms, these are not the focus of the present article. For our purposes the segmentation algorithm is viewed as a 'black-box' which defines the units to be classified and the associated unit-level feature measurements. In particular, each segment is a unit, and the features are segment attributes such as colour, texture, position of the segment within the image, and so on. A complete list of the $p = 16$ features available in our dataset appears in Table 1. Each feature is pre-scaled to have mean zero and unit variance across all I images. As examples, the segmentations for three of the images appear in Figure 1. There are

$n = 728$ segments in total across all the images, with an average (SD) of 10.7 (4.3) segments per image, and a range of 5-23 segments per image.

In fact the series of images in our dataset are taken from a larger database in which each image has been annotated with a few *keywords* describing the image content. The annotating words for each image in Figure 1 appear below the image. The vast majority of images have 2-4 keywords, while a handful have either 1 keyword or 5 keywords. There are $K = 17$ distinct keywords across all I images, as listed in Table 2. We hypothesize that every segment in an image belongs to either one of the annotating keywords, or to a 'garbage' concept. We let concept zero be the garbage category, while concepts 1 through K correspond to the K keywords. For the i-th image we let $S_i \subset \{0, \ldots, K\}$ be the union of zero and the indices of the keywords annotating the image. Thus the semi-supervision arises through the partial labelling information: for a segment in the i-th image we observe that $Y \in S_i$.

The introduction of a garbage concept in this application is important for two reasons. First, the annotating words may not comprise an exhaustive description of an image's content. Second, the segmentation procedure is imperfect, and it can produce segments which do not clearly correspond to a well-defined concept in the image.

Computer scientists are interested in such applications because they offer the possibility of automatically annotating images. Once the distribution of labels given features has been estimated from training data, images without annotations could be segmented, and labels could be estimated for these segments using the inferred distribution. Hence the image could be annotated by machine, albeit with some error. Also, in terms of generating training data it is more expedient to ask a human observer to supply some keywords for an image than to ask for keywords on a segment by segment basis. Recent literature on this machine learning problem and variants includes Barnard, Duygulu, de Freitas, Forsyth, Blei, and Jordan (2003), Blei, Jordan, and Ng (2003), Carbonetto, de Freitas, Gustafson, and Thompson (2003), and

Duygulu, Barnard, de Freitas, and Forsyth (2002).

Our present interest in this problem revolves around variable selection. There is no reason to think *a priori* that all p features which are measured will be useful in classifying units, and there is reason to suspect that removing unhelpful features might improve the performance of a classification scheme. There is a large literature on how variable selection schemes can improve regression analysis, but there is relatively little statistical literature on such schemes for classification and clustering scenarios. This may arise because the regression problem is easier, as removing a variable corresponds to simply setting the corresponding regression coefficient to zero. Particularly, there is a substantial literature on the advantages of model selection and model averaging using a Bayesian approach, typically with different possible models corresponding to different subsets of predictor variables being included in a regression model (see Chipman, George and McCulloch 2001 and Clyde and George 2004 for reviews). Also related is the notion of 'automatic relevance determination' (Neal 1998). Liu *et. al.* (2003) pursue Bayesian model selection in an unsupervised clustering context. Indeed, we adopt their approach as a starting point for the present work. Other related recent work in an unsupervised context includes Hoff (2006) and Raftery and Dean (2006) from a Bayesian perspective, and Friedman and Meulman (2004) from a non-Bayesian perspective. We also note that there is more emphasis on variable selection for clustering and classification in the Computer Science literature (for overviews see Guyon and Elisseeff 2003; Guyon, Gunn, Nikravesh, and Zadeh, 2005), though the approaches tend to be quite different than that taken here.

3 The Full Model

Starting with the supposition that all features are relevant for classification, a multivariate normal model would be

$$X|Y \sim N_p(\mu_Y, \Sigma_Y),$$

where (μ_k, Σ_k) for $k = 1, \ldots, K$ are concept-specific means and variances. If $\lambda_k = Pr(Y = k)$ then, without further information on the unit-label Y, inference would be based on

$$Pr(Y = y|X = x) \quad = \quad \frac{\phi_p(x; \mu_y, \Sigma_y)\lambda_y}{\sum_{k=0}^{K} \phi_p(x; \mu_k, \Sigma_k)\lambda_k},$$

where $\phi_p(\cdot; \mu, \Sigma)$ denotes the $N_p(\mu, \Sigma)$ density function.

To adapt this to the structure of the data at hand, let J_i be the number of segments in the i-th image, let X_{ij} be the observed feature vector for the j-th segment in the i-th image, and let Y_{ij} be the corresponding unobserved label. Recall that partial information about the label is available in the form of $Y_{ij} \in S_i$. We also assume that if the k-th word annotates the i-th image, then at least one segment in that image must belong to concept k. We express this as $Y_i \in C_i$, where $Y_i = (Y_{i1}, \ldots, Y_{i,J_i})$, and C_i is the set of all Y_i values such that $\sum_{j=1}^{J_i} I\{Y_{ij} = k\} > 0$ for all $k \in S_i$.

Another peculiarity of our specialization to this problem is that we fix $\lambda = (K+1)^{-1}(1, \ldots, 1)$ rather than treating λ as an unknown parameter vector. This requires some explanation. If we treat λ as unknown then an estimate of λ_k will reflect knowledge about the proportion of *segments* (across all images) which belong to the k-th category. Presumably this will be driven in large part by the proportion of *images* which have the k-th word as an annotation. Inference about Y_{ij} would be based on

$$Pr(Y_{ij} = y|X_{ij} = x) \quad = \quad \frac{\phi_p(x; \mu_y, \Sigma_y)\lambda_y}{\sum_{k \in S_i} \phi_p(x; \mu_k, \Sigma_k)\lambda_k}. \tag{1}$$

Now say that both k_1 and k_2 are in S_i. Then (1) supposes that if the k_1-th word appears in more images than the k_2-th word, then a given segment in the i-th image is *a priori* more likely to belong to concept k_1 than concept k_2. This seems unsatisfactory, since the annotating information is simply that both words are represented somewhere in the image. Moreover, this problem is exacerbated when we consider that 'garbage' (concept zero) is treated as annotating every image. Presumably then a large estimate of λ_0 can ensue, biasing upward the probability

that a given segment belongs to the garbage concept. To avoid these problems we fix $\lambda = (K+1)^{-1}(1,\ldots,1)$, so that a uniform prior over S_i is used when inferring labels for segments in the i-th image via (1).

To complete the model specification we apply standard conjugate prior distributions for (μ_k, Σ_k), namely

$$\begin{aligned} \Sigma_i &\sim IW_a\left(V^{-1}\right), \\ \mu_i|\Sigma_i &\sim N(\mu_0, b^{-1}\Sigma_i), \end{aligned}$$

where $IW_a(V^{-1})$ denotes the inverse Wishart distribution with degrees of freedom a and scale V, leading to a prior mean of $E(\Sigma_i) = (a-p-1)^{-1}V$, provided $a > p+1$. We take $a = p+2$, and $b = 1$, which are common choices when a relatively flat prior is desired (subject to having a finite first moment for Σ). In light of the pre-scaling of the data we take $\mu_0 = (0,\ldots,0)$. Along these same lines initially we tried $V = I_p$, but found that this does not alway provide sufficient 'regularization' if very few units are assigned to the concept in question. Thus we take $V = (p/2)I_p$, so that the identity matrix gets more weight when combined with the sample variance in constructing the posterior distribution for Σ_i.

Computation can proceed according to the joint posterior distribution of the unobserved segment labels Y and the concept-specific means and covariances (μ_k, Σ_k) for $k = 0,\ldots,p$. In particular,

$$\begin{aligned} \pi(y, \mu, \Sigma|\text{data}) \;\propto\; & \prod_{k=0}^{K}\left\{\prod_{\{(i,j):y_{ij}=k\}}\phi_p(x_{ij};\mu_k,\Sigma_k)\right\}\pi(\mu_k|\Sigma_k)\pi(\Sigma_k) \;\times \\ & \prod_{i=1}^{I}\left[I\{y_i \in C_i\}\prod_{j=1}^{J_i} I\{y_{ij} \in S_i\}\right]. \end{aligned} \tag{2}$$

Computational aspects of inference under this model will be dealt with as special cases of the more general models presented forthwith.

One point to emphasize is that without the constraints on y, (2) would correspond to a standard Bayesian mixture model for unsupervised learning. As such, it would be prone to

the well-documented difficulties associated with fitting such models (Celeux, Hurn and Robert 2000; Stephens 2000), most notably *label-switching*. However, the constraints arising from the semi-supervision should largely obviate this problem. In particular, the labels for two concepts would be formally nonidentified (and could be switched without changing the likelihood) only if the two concepts appear as labels for precisely the same images. To elaborate, say that a pair of concepts is *discordant* for an image if exactly one concept from the pair annotates the image. Thus label-switching is formally a problem if a pair of concepts is not discordant for any of the images. Moreover, presumably having more pairs which are discordant for more images will be helpful in terms of MCMC mixing and the avoidance of local maxima.

In the present data, every pair of concepts is discordant for at least two images, with a large majority (100 out of $17 \times 6/2 = 136$) being discordant for at least 10 of the 68 images, and about half the pairs (64 out of 136) being discordant for at least 20 of the images. Also, it should be noted that the few pairs of concepts which are rarely discordant arise from concepts which in fact rarely appear at all. For instance, the three pairs of concepts which are discordant on only two images are (church, horse), (church, snow), and (horse, snow). From Table 2 we see that the constituent concepts are precisely those which appear as annotations in only a single image (and in fact these pairs don't appear simultaneously in any images).

4 Feature Selection

To introduce variable or feature selection, let $M = (M_1, \ldots, M_p)$ where M_j is a binary indicator taking the value 1 to indicate that the j-th feature is relevant for inferring the label, and zero otherwise. Loosely we will also write $M \subset \{1, \ldots, p\}$ to denote the relevant features and M^C to denote the irrelevant features. Further, let $d(M) = \sum_{j=1}^{p} M_j$ be the number of relevant features. The postulated structure is now that the distribution of the relevant features varies across categories, while the conditional distribution of the irrelevant features given the relevant

features does not. That is,

$$\begin{aligned}(X_{M^C}|X_M, Y, M) &\sim N_{p-d(M)}\left\{\tilde{\mu}\left(X_M, M^C, M, \delta, \Omega\right), \tilde{\Sigma}\left(M^C, M, \Omega\right)\right\}, \\ (X_M|Y, M) &\sim N_{d(M)}(\mu_Y, \Sigma_Y), \end{aligned} \tag{3}$$

where $\tilde{\mu}()$ and $\tilde{\Sigma}()$ are the conditional mean and variance arising from the $N_p(\delta, \Omega)$ distribution. That is, if A and B index disjoint subsets of $\{1, \ldots, p\}$, then

$$\begin{aligned}\tilde{\mu}(z, A, B, \delta, \Omega) &= \delta_A - \Omega_{AB}\Omega_{BB}^{-1}(z - \delta_B), \\ \tilde{\Sigma}(A, B, \Omega) &= \Omega_{AA} - \Omega_{AB}\Omega_{BB}^{-1}\Omega_{BA}.\end{aligned}$$

Thus the model structure says that given M and Y, the distribution of irrelevant features given relevant features is the appropriate conditional distribution of the $N_p(\delta, \Omega)$ distribution, regardless of Y.

Note that this model structure is motivated by the observation that the distribution of $(Y|X)$ will not depend on X_{M^C} if and only if the distribution of $(X_{M^C}|X_M, Y)$ does not depend on Y. Thus there is a direct interpretation that X_j does (does not) contribute information about Y if $M_j = 1$ ($M_j = 0$). Note also that X_M and X_{M^C} are *not* assumed to be independent given Y. That is, at the cost of increased computation we allow that two features could be correlated while only one of them is useful in inferring unit labels. This differs from Liu *et. al.* (2003), who do make such an independence assumption. We call the model defined by (3) the *Feature Selection* (FS) model.

Note that now the dimension of (μ_i, Σ_i) depends on M, hence we take prior distributions of the form

$$\pi(\mu, \Sigma, M) = \left\{\prod_{i=1}^{m} \pi(\mu_i|\Sigma_i, M)\pi(\Sigma_i|M)\right\}\pi(M).$$

The normal and inverse-Wishart prior structure from Section 3 is retained, except now the degrees of freedom for the IW prior on each Σ_i is taken to be $d(M) + 2$. A uniform prior is assigned over the 2^p possible values for M.

It should be noted that for several reasons we treat (δ, Ω), the parameters governing the distribution of irrelevant features given relevant features, as fixed and known. Particularly we fix these quantities to be the sample mean and variance of all n feature vectors (so, in light of the pre-scaling of the features, $\delta = 0$ and $\Omega_{jj} = 1$ for $j = 1, \ldots, p$). Thus given M we assume the distribution of $X_{M^C}|X_M$, which is postulated to not vary Y, is identical to the conditional distribution estimated from all the data under a normality assumption. We are reasonably confident in making this assumption, as the uncertainty about this conditional distribution associated with all the data should be small relative to the combined uncertainty about (i) which features are relevant (M), and (ii) the unit labels (Y). Also, inference about (δ, Ω) would be complicated by the trans-dimensional aspect of the structure: given M the data are only informative about the part of (δ, Ω) corresponding to the M^C given M conditional.

Following Liu *et. al.* (2003) we can analytically integrate out (μ, Σ) from the joint posterior density $\pi(Y, M, \mu, \Sigma|\text{data})$, leaving an expression for $\pi(Y, M|\text{data})$. This is particularly nice, as it takes us from a posterior distribution of varying dimension to one of fixed dimension. Some details on this, and on MCMC updates for Y and M, are given in the Appendix.

5 Concept-Varying Feature Selection

A further refinement of the FS model is motivated by several observations. First, the FS model space is arguably rather coarse, and this does tend to be manifested empirically in the sense that small changes in M (i.e., flipping one component from zero to one or vice-versa) given Y can cause very large changes in posterior density, and consequently poor MCMC mixing. Indeed, in the unsupervised context Liu *et. al.* (2003) note that the addition of feature selection worsens MCMC performance, and they consider tempering methods (Geyer 1991) in an effort to alleviate this problem.

Moreover, in the present application and others it is easy to imagine that the features which

are useful for identifying one concept could differ from the features which are useful for identifying another concept. For instance, it is easy to imagine that the subset of features which are helpful in identifying 'sky' might be quite different than the subset helpful for 'lion.' In part for this reason, Kueck, Carbonetto and de Freitas (2004) tackle the object recognition problem in what computer scientists would describe as a *discriminative* rather than *generative* manner. Specifically, they define $\tilde{Y}^{(k)} = I\{Y = k\}$ as a binary indicator taking the value 1 to indicate that the segment belong to the k-th concept, and zero otherwise. Then they model $(\tilde{Y}^{(k)}|X)$ using a binary regression model with variable selection from a particular set of basis functions $b_1(X), \ldots, b_r(X)$. This is repeated for each k, yielding the advantage of selecting a potentially different set of basis functions when trying to identify each different concept. A disadvantage, however, is that in fitting the $K+1$ models separately one obtains incompatible classification probabilities. That is

$$\begin{aligned} \sum_{k=0}^{K} \hat{Pr}(Y = k|X) &= \sum_{k=0}^{K} \hat{Pr}(\tilde{Y}^{(k)} = 1|X) \\ &\neq 1. \end{aligned} \tag{4}$$

While an obvious classification scheme still exists, i.e., $\hat{Y} = \text{argmax}_k Pr(\tilde{Y}^{(k)} = 1|X)$, at best (4) is unsettling, and at worst it speaks to a loss of information.

In light of this we consider an extension of the FS model which has a finer model space, and which aims for the advantage but not the disadvantage of the discriminative approach. Essentially we let each concept have its own choice of relevant features. To do so, let M be a $(K+1) \times p$ matrix of binary indicators, with M_{ki} taking the value 1 if the i-th feature is relevant for the k-th concept, and 0 otherwise. Also, let $M_k = (M_{k1}, \ldots, M_{kp})$ be the row of M corresponding to the k-th concept, and let $d_k(M) = \sum_{i=1}^{p} M_{ki}$ be the number of features which are relevant for the k-th concept.

Armed with this notation we generalize (3) to

$$(X_{M_Y^C}|X_{M_Y}, Y, M) \sim N_{p-d_Y(M)}\left\{\tilde{\mu}\left(X_{M_Y}, M_Y^C, M_Y, \delta, \Omega\right), \tilde{\Sigma}\left(M_Y^C, M_Y, \Omega\right)\right\},$$

$$(X_{M_Y}|Y, M) \quad \sim \quad N_{d_Y(M)}(\mu_Y, \Sigma_Y). \tag{5}$$

Thus we still assume that the $N(\delta, \Omega)$ distribution describes the parts of the distribution of X which are not relevant for classifying units, i.e., the distribution of $(X_{M_Y^C}|X_{M_Y}, Y)$ for each Y.

In the present application the 'garbage' concept is thought of as all segments not belonging to any keyword. Thus it does not seem sensible to think of particular features as being useful for detecting 'garbage'. Consequently we add the constraint that no features are relevant for the 'garbage' concept, i.e., $M_{0i} = 0$ for $i = 1, \ldots, p$.

We can still apply the conjugate prior structure for $\pi(\mu_k|\Sigma_k, M)$ and $\pi(\Sigma_k|M)$, as before. We simply adapt the degrees of freedom in the inverse Wishart prior for Σ_k to be $d_k(M) + 2$. It is less clear however, whether a uniform prior distribution for M is still appropriate. Particularly we are concerned that unfettered flexibility for M may promote overfitting, so we adopt a prior for M which gives some favouritism to 'simpler' values of M which are closer to the earlier FS model. In particular we take a hierarchically structured prior of the form

$$\pi(M|\alpha) \quad = \quad \prod_{j=1}^{p} \alpha_j^{M_{\cdot j}} (1 - \alpha_j)^{K+1-M_{\cdot j}},$$

where $M_{\cdot j} = \sum_{k=0}^{K} M_{kj}$. The hierarchy is completed by specifying $\pi(\alpha)$ according to

$$\alpha_1, \ldots, \alpha_p \quad \overset{iid}{\sim} \quad \omega Beta(1, \gamma) + (1 - \omega) Beta(\gamma, 1).$$

Thus the elements of the j-th column of M are conditionally independent given α_j, with the prior distribution on α_j thereby inducing positive dependence for the column elements. This reflects some tendency for a feature to be generally useful or generally not useful, without the rigidity of making the column elements identical, i.e., reverting back to the FS model. The U-shaped nature of the prior on α_j encourages the dependence by encouraging either a low or high proportion of concepts for which the j-th feature is relevant. We take $\omega = 0.5$ and $\gamma = 10$ in the analysis of the next section.

In Section 4 we argued that treating (δ, Ω) as known and equal to an overall sample mean and variance is reasonable in the context of the FS model. The same approach seems more dubious in the present context, since (δ, Ω) no longer encode a conditional distribution of irrelevant features given relevant ones across all concepts. Thus we now treat these parameters as unknown. As formulated there is not an obvious MCMC updating scheme for (δ, Ω), since observations which are currently assigned to one concept contribute information about a different part of (δ, Ω) than do observations which are currently assigned to another concept. We circumvent this problem with a data-augmentation scheme described in the Appendix. Under this scheme, the augmented data are such that every observation contributes information about all of (δ, Ω), so that a standard Gibbs sampling update can be implemented.

6 Synthetic Data Results

Before presenting results for the image data, we consider a much simplified synthetic data problem as 'proof of concept' for the feature selection techniques. Data are generated on $p = 5$ features for $K = 3$ concepts. The k-th concept is taken as characterized by elevated values of the k-th feature, while the fourth and fifth features are noise features which are correlated with the other features. Particularly, equi-correlated data are generated, i.e., $Cor(X_i, X_j|Y = y) = \rho$ for all y, with mean vectors $E(X|Y = 1) = \Delta(1, -.5, -.5, 0, 0)'$, $E(X|Y = 2) = \Delta(-.5, 1, -.5, 0, 0)'$, and $E(X|Y = 3) = \Delta(-.5, -.5, 1, 0, 0)'$. The within-concept variances are set as $Var(X_i|Y) = 1 - \Delta^2/2$ for $i = 1, 2, 3$, and $Var(X_i|Y) = 1$ for $i = 4, 5$. It is easy to check that with a uniform distribution on $Y \in \{1, 2, 3\}$ this leads to standardized data, i.e., unconditionally $E(X_i) = 0$ and $Var(X_i) = 1$ for $i = 1, \ldots, 5$. A synthetic dataset is generated with 100 observations for each concept, using $\rho = 0.5$ and $\Delta = 0.47$. We treat the labels as fully known (i.e., we are doing supervised learning), and investigate the FS and CVFS schemes for feature selection.

A MCMC run of length 2500 for the FS model gives $Pr(M_i = 1|\text{data}) = (1, 1, 0.23, 0, 0)$.

That is, the MCMC sampler stays at $M_1 = M_2 = 1$ and $M_4 = M_5 = 0$, and only mixes on M_3. In the present context this seems good, given that the non-mixing components of M are fixed at correct values, i.e., features 1 and 2 are indeed relevant while features 4 and 5 are not. It seems unlikely, however, that the posterior distribution is really putting all its mass on these points, given the modest sample size and the correlated features. A more plausible conclusion is that the MCMC scheme has difficulty moving 'through' the posterior distribution over M.

A MCMC run of the same length for the CVFS model gives the values of $Pr(M_{ki} = 1|\text{data})$ appearing in Table 3. The 'finer' structure in the real data-generating mechanism is indeed reflected in these inferences, i.e., for $k = 1, 2, 3$, $Pr(M_{ki} = 1|\text{data})$ is much higher for $i = k$ than for $i \neq k$. Also, in contrast to the FS model, there is some MCMC mixing for all components of M. For instance, this results in low (but non-zero) posterior probabilities of relevance for the indicators associated with the fourth and fifth features. The extent to which mixing over M in the 'finer' CVFS model space improves upon mixing in the 'coarser' FS model space is examined in more detail for our main example in the next section.

For these synthetic data we also computed predictive label distributions for a validation data set, using the the FULL, FS, and CVFS models. In fact these predictive distributions are very similar under the three models, and we do not see improved predictive performance as sometimes arises from model selection or model averaging. For these data at least, the feature selection schemes do provide qualitative insight into the $X-Y$ relationship, but they do not offer improved prediction. Again we consider this issue in more detail for the image data analysis.

7 Image Data Results

We fit the three models (FULL, FS, CVFS) to the image data in the following manner. The I=68 images are randomly split into $I_T = 50$ images for training data and $I_V = 18$ images for validation data, subject to the constraint that each of the $K = 17$ keywords must annotate at

least one of the training data images. The particular split obtained yields $n_T = 529$ segments in the training data and $n_V = 199$ segments in the validation data. Each model is fit using 2200 MCMC iterations, with the first 200 of these used as burn-in. As discussed in Section 8, computational concerns do necessitate quite short MCMC runs. However, as discussed in the Appendix, the use of a Rao-Blackwellized estimator for $(Y|X)$ makes it feasible to use short runs.

For the purposes of performance assessment, all the segments in the present data have in fact been labelled by eye. We emphasize that these 'true' segment labels are not used in any of the model fitting procedures, as the goal is to assess how well we can learn the $X - Y$ relationship at the segment level from data comprised of image labels rather than segment labels. We also note that for most (79%) of the segments the human raters assigned a single true label from amongst the keywords annotating the encompassing image. But in some instances (11%) the raters assigned two or more true labels to a segment, since the segment cuts across two or more concepts in the image. And in some instances (10%) the raters assign no labels to a segment, since the segment simply does not match with any of the annotating keywords. Thus we carry out performance assessment as follows. For each segment in the training data we take the posterior mode of Y_{ij} as the estimated label (which will necessarily belong to S_i). The validation images are treated as unannotated (i.e., we pretend we have no image annotations), and Y_{ij} is estimated as the concept with the largest predictive probability (which will not necessarily belong to S_i). In either case, the estimate is judged to be correct if either (i) the estimate matches the segment's single true label, (ii) the estimate matches one of the segment's multiple true labels, or (iii) the estimate is zero (the garbage concept) and the segment has no true labels.

This performance assessment yields both a correct classification rate on the training data, which reflects how well we can learn segment labels given the corresponding image labels, and a correct classification rate on the validation data, which reflects how well we can learn segment

labels in unannotated images given a series of annotated images. Of course the latter is a much harder learning task than the former. Random guessing in the former case would yield say a 25% correct classification rate in images with $|S_i| = 4$ (three annotating words plus 'garbage'), but a $(K+1)^{-1} \approx 6\%$ correct classification rate in the latter situation. In interpreting results we emphasize that since the training data involve image labels rather than segment labels, it is not the case that we can make the training data classification rate arbitrarily good by using more complex models.

We obtain correct classification rates for the training segments of 45% under the FULL model, 43% under the FS model, and 40% under the CVFS model. The corresponding rates for the validation segments are 30% (FULL), 25% (FS), and 29% (CVFS). Thus neither feature selection scheme appears to offer a predictive benefit, though the relatively small sample sizes n_T and n_V indicate that the variation in classification rates ought not to be over-interpreted. We return to the stability of our results across different training/validation data splits and different MCMC runs presently.

To glean some sense of the classification results, Figure 2 plots the number of segments (training and validation combined) with true concept j against the proportion of these segments that are correctly classified using the CVFS model. Focussing on concepts with more than a handful of segments, some of the worst and best cases are marked on the plot.

Upon examining the MCMC output for the FS model we see that in fact there is no mixing of M beyond the burn-in period. That is, the sampler is 'stuck' at a single value of M which corresponds to features (1,4,6,12,13) being relevant. (More precisely, 15 of the 16 feature indicators are completely frozen, the indicator M_{13} makes a single switch from zero to one during the run). An obvious concern is that this may be a local maxima of the posterior density from which the sampler cannot escape, rather than a situation where the posterior distribution really puts all its mass on a single value of M. Conversely, the components of M under the CVFS model

do seem to mix tolerably well. For each of the $K \times p$ elements of M (feature indicators) in the CVFS model, a *thinned switch rate* (TSR) is plotted in the upper-left panel of Figure 3. The TSR is based on the value of the indicator M_{ki} at every 4-th MCMC iteration (i.e., the output is thinned to conserve storage space); particularly, TSR is taken to be the proportion of times that the indicator changes from zero to one or vice-versa in the thinned output. We compare this to the estimated *independence switch rate* (ISR) which would arise under *iid* sampling from the posterior distribution, i.e., $ISR = 2q(1-q)$, where $q = Pr(M_{ki} = 1|\text{data})$. Thus we plot TSR against ISR to glean a sense of how well the MCMC scheme mixes for M relative to *iid* sampling (which corresponds to the identity line indicated on the plots). Clearly the sampling is much less efficient than *iid* sampling, but most of the indicators are mixing to some extent, in contrast to the FS model.

To assess whether the FS and CVFS models are in rough agreement about which features are important, let $r_i = E\left(K^{-1}\sum_{k=1}^{K} I\{M_{ki} = 1\}|\text{data}\right)$, be the estimated proportion of concepts for which the i-th feature is relevant under the CVFS model. The upper-right panel of Figure 2 plots $Pr(M_i = 1|\text{data})$ under the FS model against r_i for the CVFS model. Above and beyond the discreteness that arise because all but one of the estimated posterior probabilities under the FS model is zero or one, there is apparently no agreement between the two models concerning which features are important.

The posterior distribution on M under the CVFS model is displayed via a greyscale plot in the bottom-left panel of Figure 3. As favoured by the prior distribution on M, there is some tendency for features to be generally useful or not useful across concepts.

Figure 4 examines the MCMC mixing of the segment labels Y_{ij} for the training data, again using plots of thinned switching rates against estimated switching rates under *iid* sampling. Note that now the ISR is given as $\sum_{k=0}^{K} q_k(1-q_k)$, where $q_k = Pr(Y_{ij} = k|\text{data})$. Not surprisingly the mixing for some labels is substantially worse than under *iid* sampling, though there is not an

indication of pathological mixing problems. The label-mixing performance seems comparable under the three models, with a suggestion of slightly worse mixing under the CVFS model.

7.1 Stability Study

To assess the extent to which the results in the previous section are typical, we re-do the analysis for three different random training/validation divisions of the data. For each data split we use three MCMC runs of 1000 iterations after 200 burn-in iterations for each model. In particular, each of the three runs uses a different random initialization of the segment labels Y_{ij}.

Figure 5 compares the feature relevance under the FS model to that under the CVFS model, for each data split and each MCMC run separately, in the same format used in Figure 3. Again there is virtually no mixing of M under the FS model, so that $Pr(M_i = 1|\text{data})$ is estimated to be zero or one in most instances. Moreover, these estimates are not stable across different MCMC runs (with the same data split), indicating that the FS model is indeed prone to a problem of local maxima. Conversely, the estimated feature relevances under the CVFS model are rather stable across MCMC runs and even somewhat stable across data splits. This stability certainly appears to be a strength of the CVFS model over the FS model. We also consider combining the three MCMC runs for each data split, i.e., forming a larger posterior sample over (Y, M) by concatenating the three individual samples. The last column of Figure 5 compares feature relevance under the two models using the combined MCMC runs. Here we do start to see some agreement between the two models about which features are most helpful. Of course under the FS model we are presumably averaging three local maxima for M, so we cannot regard this as posterior sampling in a meaningful sense.

Figure 6 compares correct classification rates across for the three data splits and three MCMC runs. The dashed lines give rates for training and validation segments as computed under the three MCMC runs separately. The solid lines give rates based on combining the three runs.

Combining the runs seems most beneficial for the classification rates under the FS model. It is not surprising that averaging results based on three local maxima might be particularly helpful, given the extensive literature on improvements to classification schemes by combining different classifiers (see, for example, Hastie, Tibshirani, and Friedman 2001, Ch. 10). The overall impression from Figure 6 is that the results are very mixed. There is not clear evidence on the question of whether feature selection is helpful in the present application, nor is there clear evidence on whether the FS or CVFS model tends to produce better classification rates. The small validation sample size coupled with the presumably small impact (either positive or negative) of feature selection does not yield clear conclusions about this impact.

8 Discussion

Certainly there are limitations associated with the methods we have described. First, we have not been able to demonstrate a clear benefit (nor a clear detriment) to Bayesian feature selection in a semi-supervised context for the particular data at hand. Second, the approaches described are rather computationally intensive. Consequently we have limits on the length of MCMC runs that can be employed, and particularly we do not have an attractive scaling of the computational requirements with the number of available features. The latter is obviously a limitation when considering the classification and clustering efforts are focussed on bioinformatics where enormous numbers of features may be available.

More positively, the CVFS model seems to be quite promising in several regards. First, by providing a much 'finer' collection of models whose elements are closer together than in the FS model we make MCMC sampling from the posterior distribution over the model space feasible. Second, the CVFS model has the conceptual appeal of being consistent with human classification processes whereby the features of an object that lead to its identification as X may be quite different from the features that reveal an object to be Y. In light of this it seems worthwhile to

consider variants of the CVFS model in a range of classification and clustering contexts.

References

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., and Jordan, M.I. (2003). Matching words and pictures. *Journal of Machine Learning Research* **3**, 1107-1135.

Blei, D.M., Jordan, M.I., and Ng, A.Y. Hierarchical Bayesian models for applications in information retrieval. In *Bayesian Statistics 7*, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West (Eds.) Oxford University Press.

Carbonetto, P., de Freitas, N., Gustafson, P., and Thompson, N. (2003) Bayesian feature weighting for unsupervised learning, with application to object recognition. *Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics*, C. M. Bishop and B. J. Frey (Eds.) Society for Artificial Intelligence and Statistics.

Celeux, G., Hurn, M., and Robert, C.P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* **95**, 957-970.

Chipman, H., George, E.I., and McCulloch, R.E. (2001). The practical implementation of Bayesian model selection (with discussion). In *Model Selection*, P. Lahiri (Ed.), Institute of Mathematical Statistics Lecture Notes-Monograph Series, Vol. 38, 65-134.

Clyde, M. and George, E.I. (2004). Model uncertainty. *Statistical Science* **19**, 81-94.

Duygulu, P., Barnard, K., de Freitas, N., and Forsyth, D. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision, Part 4* A. Heyden, G. Sparr, M. Nielsen, P. Johansen (Eds.) Springer, 97-112.

Friedman, J.H. and Meulman, J.J. (2004). Clustering objects on subsets of attributes (with

discussion). *Journal of the Royal Statistical Society, Series B* **66**, 815-840.

Geyer, C.J. (1991). Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics: The 23rd symposium on the interface* E. Keramigas (Ed.) Fairfax: Interface Foundation, 156-163.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**, 1157-1182.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. (2005). *Feature Extraction, Foundations and Applications* (including the Proceedings of the NIPS 2003 Workshop on Feature Extraction). New York: Springer.

Hastie, T., Tibshirani, T., and Friedman, J. (2001). *The Elements of Statistical Learning.* New York: Springer.

Hoff, P.D. (2006) Model-based subspace clustering. *Bayesian Analysis* **1**, 321-344.

Kueck, H., Carbonetto, P., and de Freitas, N. (2004). A constrained semi-supervised learning approach to data association. In *Proceedings of the 8th European Conference on Computer Vision, Part 3* T. Pajdla, J. Matas (Eds.), Springer, 1-12.

Liu, J.S., Zhang, J.L., Palumbo, M.J., and Lawrence, C.E. (2003). Bayesian clustering with variable and transformation selection. In *Bayesian Statistics 7*, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West (Eds.) Oxford University Press, pp. 249-275.

Neal, R.M. (1998). Assessing relevance determination methods using DELVE. In *Neural Networks and Machine Learning*, C.M. Bishop (Ed.) Springer Verlag, pp. 97-129.

Raftery, A.E. and Dean N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association* **101**, 168-178.

Shi, J., and Malik, J. (1997). Normalized cuts and image segmentation. In *IEEE Conference*

on Computer Vision and Pattern Recognition, 731-737.

Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society B*, **62**, 795–809.

Appendix

Following Liu *et. al.* (2003), the analytic integration of means and variances can be expressed as follows. If

$$W_1, \ldots, W_n | \mu, \Sigma \overset{iid}{\sim} N_p(\mu, \Sigma)$$

while $(\mu|\Sigma) \sim N(\mu_0, \rho_0^{-1}\Sigma)$ and $\Sigma \sim IW_{\nu_0}(S_0^{-1})$, then the marginal density of W is given as $f(w) = g(p, \nu_0, \rho_0, S_0, n, SS)$, where

$$g_p(\nu_0, \rho_0, \mu_0, S_0, n, SS) = \frac{Z(\nu_0, S_0, p)}{Z(\nu_0 + n, S_0 + SS, p)} (2\pi)^{-np/2} \left(\frac{n + \rho_0}{\rho_0} \right)^{-p/2}, \tag{6}$$

with

$$SS = \sum_{i=1}^{n} (w_i - \bar{w})'(w_i - \bar{w}) + \frac{n\rho_0}{n + \rho_0} (\bar{w} - \mu_0)'(\bar{w} - \mu_0), \tag{7}$$

and

$$Z(\nu, S, p) = |S|^{\nu/2} \left\{ 2^{\nu p/2} \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left(\frac{\nu + 1 - i}{2} \right) \right\}^{-1}.$$

Also, it follow that

$$\Sigma | W \sim IW_{\nu_0 + n}\{(S_0 + SS)^{-1}\}, \tag{8}$$

and

$$\mu | \Sigma, W \sim N\left(\frac{\rho_0}{n + \rho_0} \mu_0 + \frac{n}{n + \rho_0} \bar{w}, (n + \rho_0)^{-1} \Sigma \right). \tag{9}$$

Based on (6), the posterior distribution for the CVFS model with the concept-specific means and variances integrated out can be written as

$$\begin{aligned}\pi(Y, M, \delta, \Omega, \alpha | X, S) \quad \propto \quad & \prod_{k=0}^{K} g_{d(M_k)}(d(M_k)+2, \{d(M_k)/2\} I_{d(M_k)}, 1, n_k^*(Y), SS_{Y=k,M_k}) \ \times \\ & \prod_{k=0}^{K} \phi(X_{ij,M_k^C} | X_{ij,M_k}; \delta, \Omega) \times \\ & \prod_{j=1}^{p} \alpha_j^{M_{\cdot j}} (1-\alpha_j)^{K+1-M_{\cdot j}} \pi(\alpha_j) \times \\ & \pi(\delta|\Omega)\pi(\Omega) \times \\ & \prod_{i=1}^{I} \left[I\{y_i \in C_i\} \prod_{j=1}^{n_i} I\{y_{ij} \in S_i\} \right]. \end{aligned} \tag{10}$$

Notationally here $SS_{Y=k,M}$ denotes the sum-of-squares term (7) for the $n_k^* = \sum_{i=1}^{n} I\{Y_i = k\}$ units in the k-th group and the M_k subset of the feature vectors. Also $\phi(x_A|x_B; \mu, \Sigma)$ is taken to be normal density of X_A given X_B arising from $X \sim N(\mu, \Sigma)$.

We now describe how MCMC updating with (10) as the target distribution is obtained. Updating schemes for the simpler FS and FULL models are special cases of what follows.

To update the vector M_k of feature indicators for the k-th concept we simply use a Metropolis-Hastings update with a proposal obtained by flipping a randomly selected element of the current value.

To update the unit labels Y_{ij} we generate a 'temporary' set of means and variances (μ_k, Σ_k) for $k \in S_i$ using (8) and (9), and based on the current value of M. These are deterministically extended to dimension p based on the current values of (δ, Ω). Then Y_{ij} can be sampled from (1). After all elements of Y have been updated, the temporary means and variances are discarded. That is, they are not part of the state-space for MCMC simulation of (10). It is straightforward to verify that this update does leave (10) as its stationary distribution. Note that this approach differs from that used by Liu *et. al.* (2003). One advantage of the present approach is that the distribution (1), which is calculated at every iteration, can be averaged across iterations as a Rao-Blackwellized estimate of the posterior distribution of Y_{ij}. That is, we estimate $Pr(Y_{ij} = y|X)$

as a Monte Carlo average of $Pr(Y_{ij} = y|\mu, \Sigma, X)$ rather than as an average of $I\{Y_{ij} = y\}$. This makes it possible to utilize much shorter MCMC runs than might otherwise be contemplated. This approach also improves the computation of predictive distributions for the labels validation segments.

The MCMC updating of α_j involves straightforward Gibbs sampling, as its conditional distribution given all other quantities is a mixture of two beta distributions. Updating of (δ, Ω) is less obvious, since X_{ij} belonging to different concepts (according to the current Y) contribute information about different parts (according to the current M) of (δ, Ω). However, a data-augmentation trick makes the updating feasible. The trick is most easily explained in a 'stripped-down' setting. The contribution of a single feature vector to the posterior density is

$$\phi(x_{M^c}|x_M; \delta, \Omega)\phi(x_M; \mu_k, \Sigma_k).$$

With the addition of x^*_M this can be augmented to

$$\phi(x_{M^c}|x_M; \delta, \Omega)\phi(x^*_M; \delta, \Omega)\phi(x_M; \mu_k, \Sigma_k).$$

Now consider reparameterizing from $(x_M, x^*_M, x_{M^C}, \delta, \Omega)$ to $(x_M, x^*_M, x^*_{M^C}, \delta, \Omega)$, where

$$x^*_{M^C} \;=\; x_{M^C} + \Omega_{M^C M}\Omega^{-1}_{MM}(x^*_M - x_M). \tag{11}$$

The Jacobian of the transformation is one, and the contribution to the posterior density under the new parameterization is simply

$$\phi(x^*_{M^c}|x^*_M; \delta, \Omega)\phi(x^*_M; \delta, \Omega)\phi(x_M; \mu_k, \Sigma_k),$$

such that $(x^*_M, x^*_{M^c})$ given (δ, Ω) follow the $N_p(\delta, \Omega)$ distribution. Thus we have the following steps for our data-augmentation scheme. First, sample the augmenting vector x^*_M for each unit, from the appropriate marginal of the $N(\delta, \Omega)$ distribution. Second, deterministically compute $x^*_{M^C}$ for each unit according to (11), to yield complete x^* vectors of length p for each unit. Third, appealing to the alternate parameterization, carry out standard conjugate updating of (δ, Ω) given the values of x^* for each unit.

X_1	area
$X_2 - X_3$	mean horizontal and vertical position in image
$X_4 - X_5$	standard deviation of horizontal and vertical position in image
X_6	perimeter divided by area
X_7	convexity
$X_8 - X_{10}$	average (lab) color
$X_{11} - X_{13}$	standard deviation of (lab) color
$X_{14} - X_{16}$	skewness of (lab) color

Table 1: The $p = 16$ features in the dataset

airplane	(8)	*bird*	(5)	*church*	(1)	*elephant*	(15)
grass	(34)	*ground*	(7)	*horse*	(1)	*house*	(2)
lion	(12)	*mountains*	(8)	*road*	(2)	*rock*	(11)
sand	(2)	*sky*	(44)	*snow*	(1)	*trees*	(38)
water	(19)						

Table 2: The $K = 17$ annotating keywords. The number of images in which the word appears is given in parentheses.

		feature				
		1	2	3	4	5
concept	1	0.90	0.44	0.17	0.15	0.08
	2	0.28	0.78	0.25	0.11	0.12
	3	0.41	0.46	0.98	0.08	0.07

Table 3: Posterior probabilities that the i-th feature is relevant for the k-th concept, for the synthetic data of Section 6. The $k-th$ row and i-th column give $Pr(M_{ki} = 1|\text{data})$, as computed from 2500 MCMC iterations.

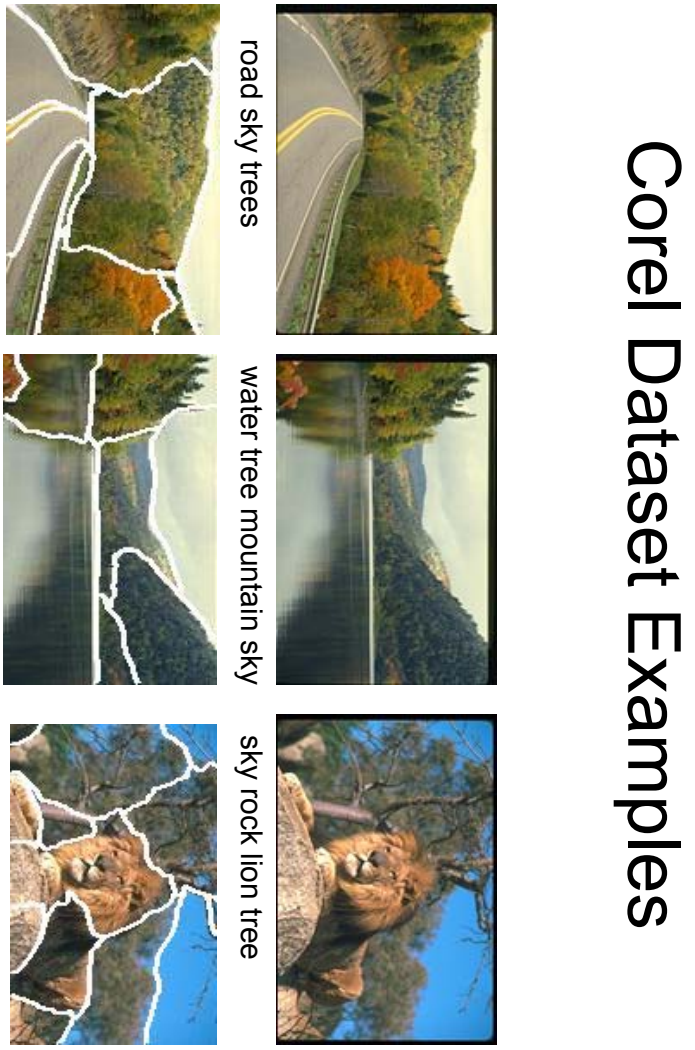

Figure 1: Sample images, before and after segmentation.

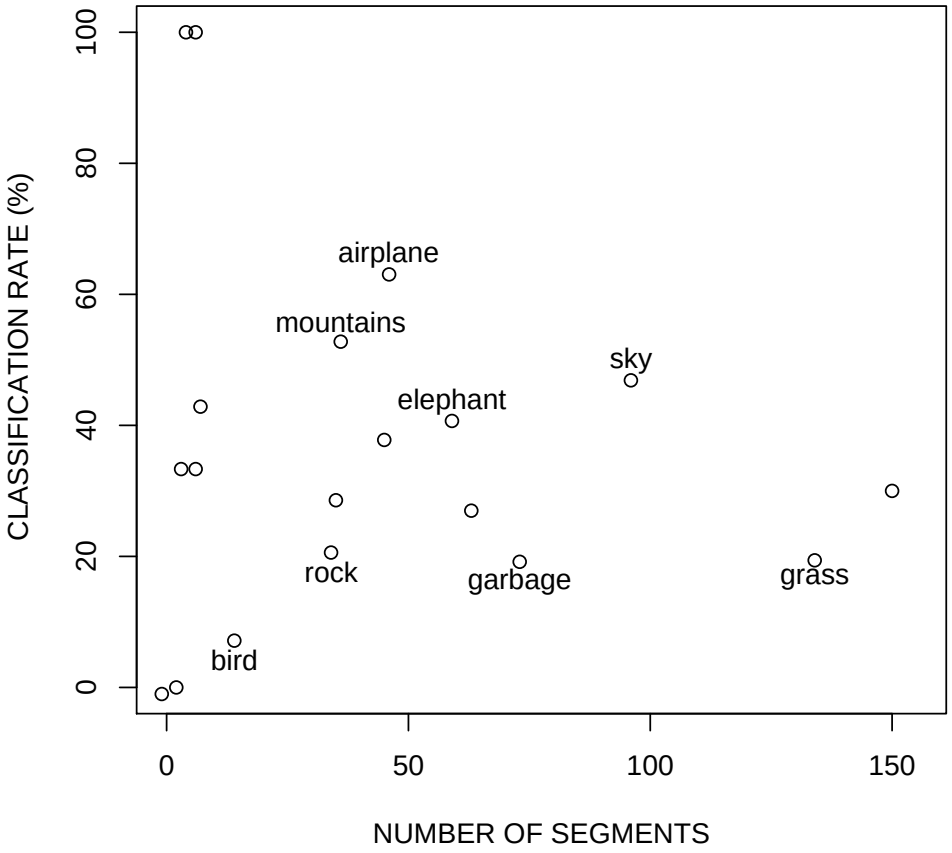

Figure 2: Classification rates by concept, on training and validation data combined. For each concept the number of segments truly labelled by this concept is plotted against the proportion of these segments that are classified correctly. Selected concepts are identified by name.

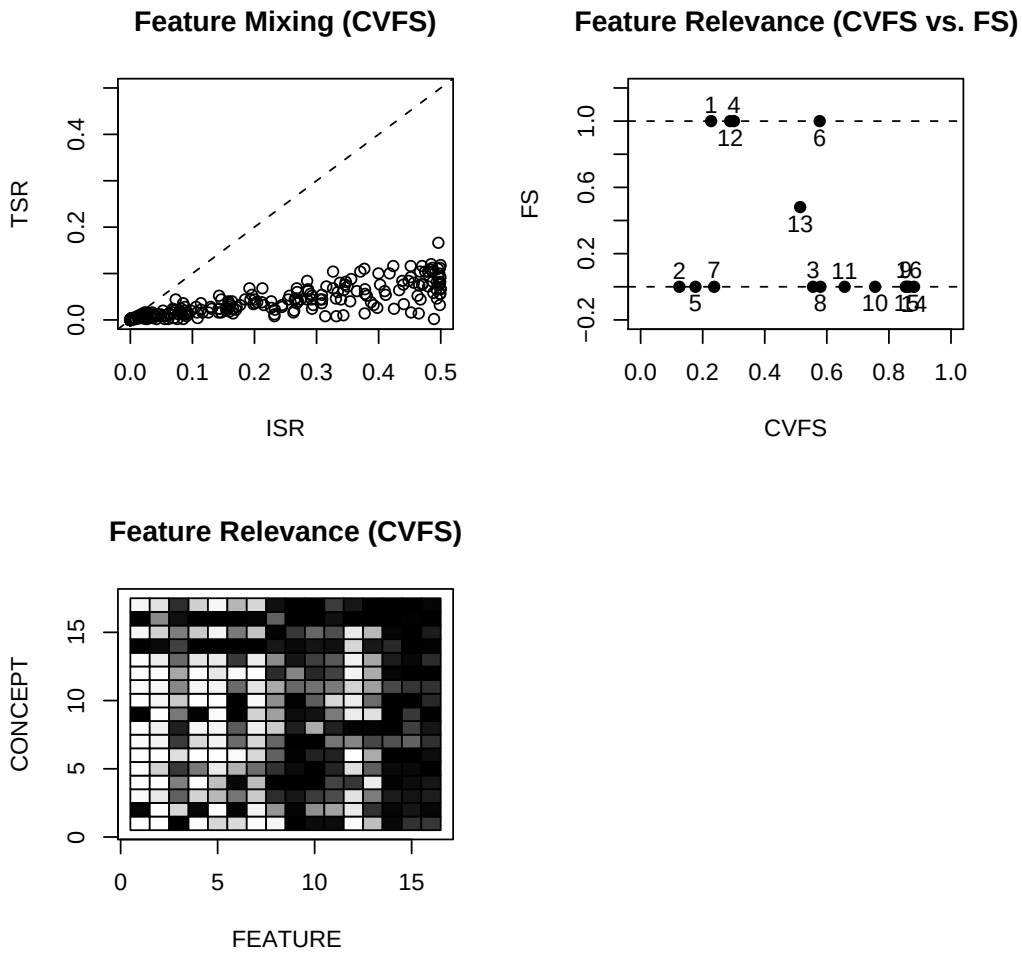

Figure 3: Inferences about features. The top left panel plots thinned switching rate (TSR) against the estimated rate under independent sampling (ISR), as a summary of the MCMC mixing for the feature indicators. The top right panel plots feature relevance under the FS model (posterior probability of feature inclusion) against feature relevance under the CVFS model (posterior expectation of the number of concepts for which the feature is included). The bottom-left panel plots $Pr(M_{ki} = 1|\text{data})$ on a greyscale from zero (white) to one (black).

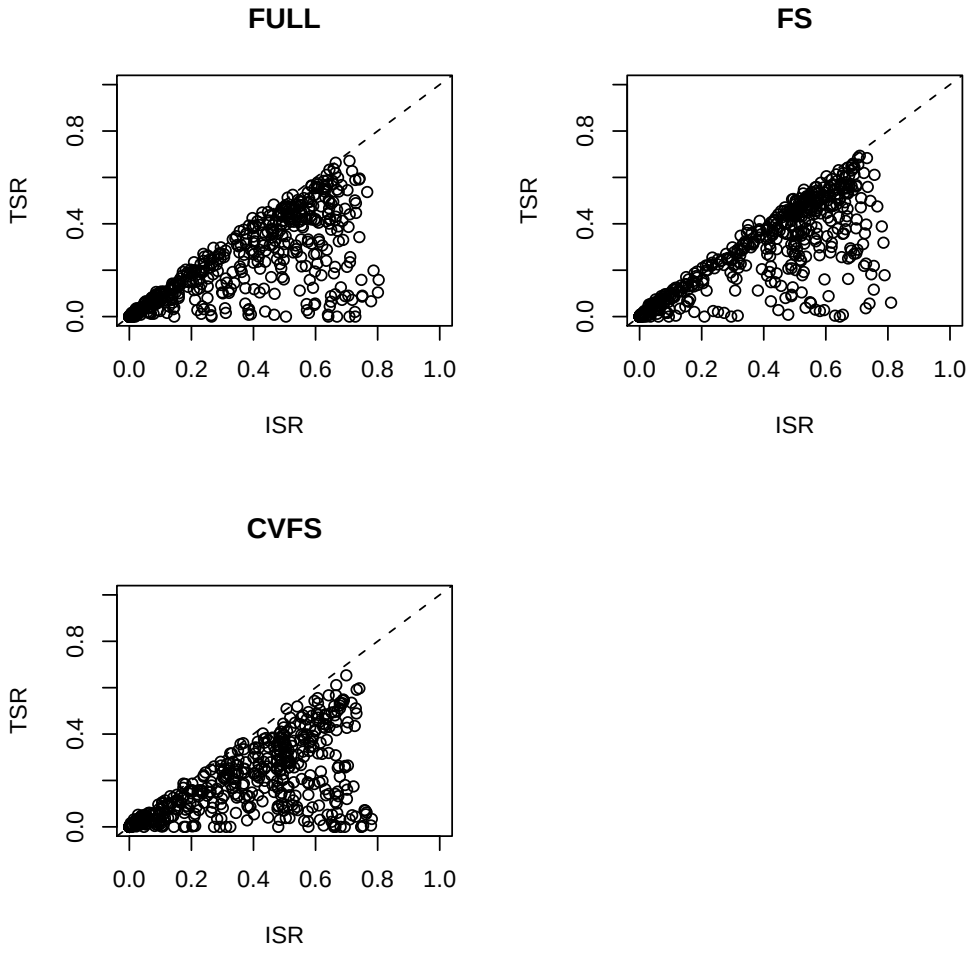

Figure 4: MCMC mixing for the labels Y. The thinned switch rate (TSR) is plotted against the estimated independence switch rate (ISR) for all 728 labels. The three panels correspond to the FULL, FS, and CVFS models.

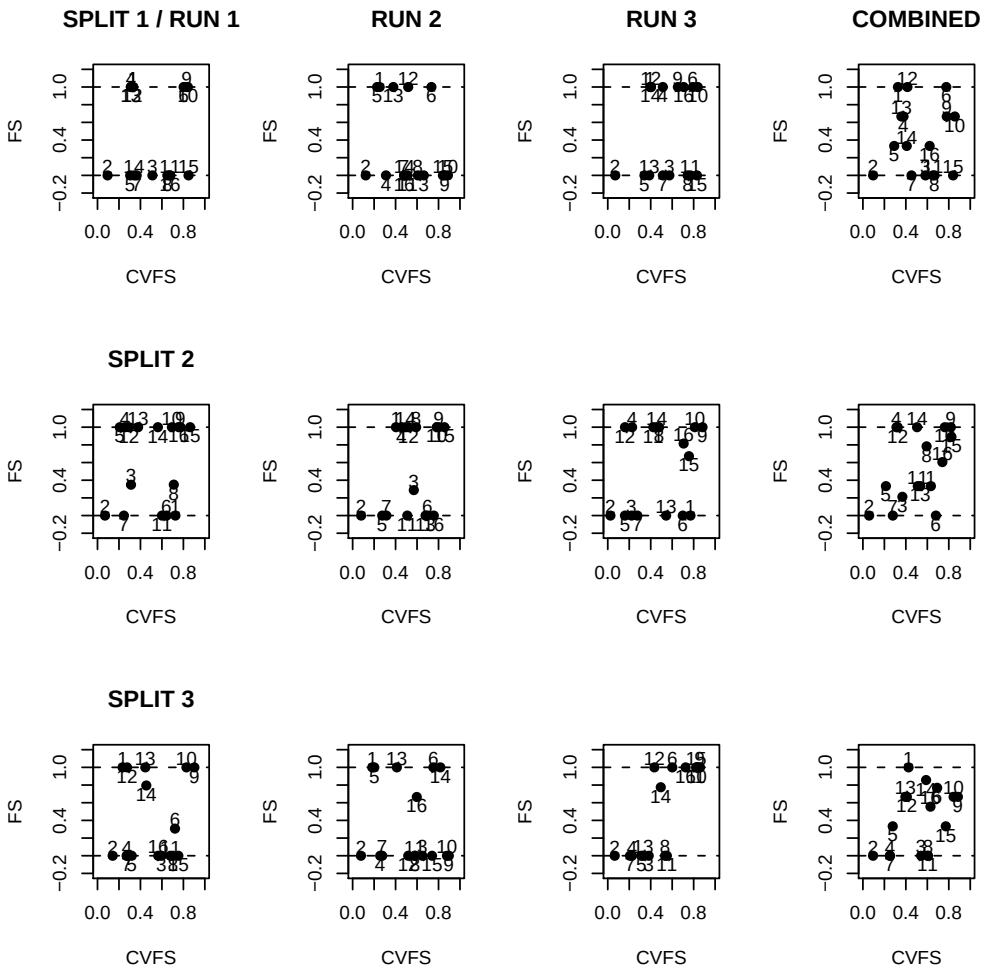

Figure 5: Feature relevance in the stability study. Each panel plots feature relevance for the FS and CVFS models as in Figure 3. The three rows correspond to the three data splits. The first three columns correspond to the three MCMC runs, while the fourth column corresponds to pooling the three runs.

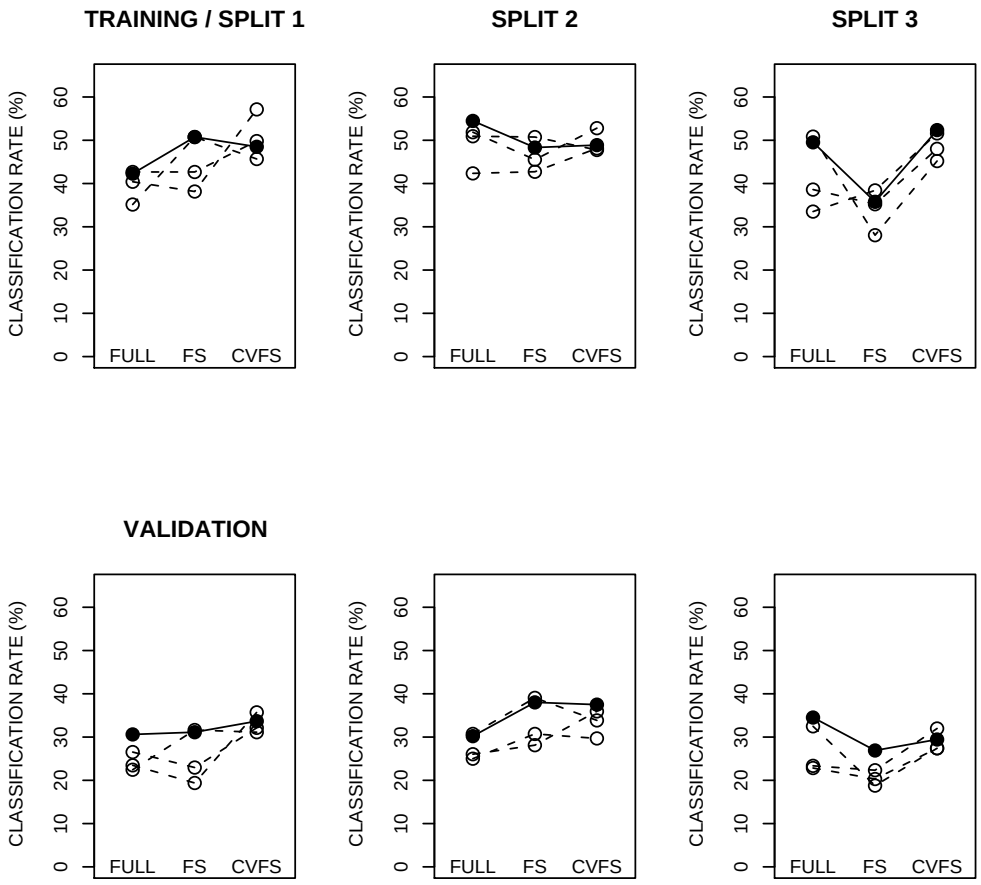

Figure 6: Classification rates in the stability study. Each panel plots the percentage of segments classified correctly under the three models (FULL, FS, CVFS) and under the three MCMC runs. The top (bottom) row plots correspond to classification of training (validation) segments. The three columns of plots correspond to the three data splits. The open circles give rates based on the three MCMC runs separately, while the closed circles correspond to pooling these runs.