Hierarchical Models for Screening of Iron Deficiency Anemia

Technical Report No. 99–14
Department of Information and Computer Science
University of California, Irvine

I. V. Cadez[1], C. E. McLaren[2], P. Smyth[1], and G. J. McLachlan[3]

[1] Department of Information and Computer Science
University of California, Irvine
CA 92697-3425

[2] Division of Epidemiology, Department of Medicine,
University of California, Irvine, CA 92697, U.S.A.

[3] Department of Mathematics,
The University of Queensland, Brisbane, Australia

March 1999

Abstract

We investigate the problem of classifying individuals based on estimated density functions for each individual. Given labelled histograms characterizing red blood cells (RBCs) for different individuals, the learning problem is to build a classifier which can classify new unlabelled histograms into normal and iron deficient classes. Thus, the problem is similar to conventional classification in that there is labelled training data, but different in that the underlying measurements are not feature vectors but histograms or density estimates. We describe a general framework based on probabilistic hierarchical models for modelling such data and illustrate how the model lends itself to classification. We contrast this approach with two other alternatives: (1) directly defining distance between densities using a cross-entropy distance measure, and (2) using parameters of the estimated densities as feature vectors for a standard discriminative classification framework. We evaluate all three methods on a real-world data set consisting of 180 subjects. The hierarchical modeling and density-distance approaches are most accurate, yielding cross-validated error rates in the range of 1 to 2%. We conclude by discussing the relative merits of each approach, including the interpretability of each model from a clinical diagnostic viewpoint.

1 Introduction and Background

Anemia, a reduction in the circulating red cell mass that may diminish the oxygen-carrying capacity of the blood, is one of the most common medical problems. For diagnostic evaluation of anemia and monitoring the response to therapy, blood samples from patients are routinely analyzed to determine the volume of the red blood cells (RBCs) and the amount of hemoglobin, the oxygen-transporting protein of the red cell. In this context it would be highly cost-effective to have the ability to perform automated low-cost accurate diagnostic screening of blood-related disorders using RBC measurements. Many anemia-related diseases manifest themselves via fundamental changes in the univariate volume (V) distribution and the univariate hemoglobin concentration (HC) of RBCs.

Automated techniques have been recently developed which can simultaneously measure both volume and hemoglobin concentration of RBCs from a patient's blood sample. Flow cytometric blood cell counting instruments (Technicon H*1, H*2, H*3; Bayer Corporation, White Plains, NY) make measurements using a laser light scattering system to provide the red cell volume distribution, hemoglobin concentration, and joint red cell volume and hemoglobin concentration distributions. The data we will describe in this paper was generated by such a machine. Typically it takes in about 40,000 blood cells and produces a plot of both the univariate histogram of the V distribution and univariate HC distribution. In addition, it provides a bivariate histogram (on a grid of about 100×100 cells) of the joint V-HC distribution. Figure 1 shows a two-dimensional histogram of the RBC counts in V-HC space for a healthy individual (control) and for an iron deficient patient. Existing diagnostic techniques based on such measurements are largely limited to simple visual examination of such plots and approximate estimates of abnormality based on the skewness or shift of the histogram in various directions in the bivariate space. While this general approach will capture those patients whose distributions are very clearly far removed from the normal pattern, it is relatively insensitive to more subtle changes and is also likely to be relatively insensitive to differential diagnosis among different diseases.

In this paper we will focus on the problem of learning a classification model from the data for the purposes of automated diagnostic screening in a clinical environment. Section 2 will discuss the learning aspects of the problem in general and make connections with relevant prior work. In Section 3 we introduce the notion of a probabilistic hierarchical model for this problem. This is a powerful framework for modeling data sets where there are multiple levels of variation in the data (here we have variation at the individual subject level as well as at the RBC level). We will show that with relatively large amounts of data at the lowest level of the hierarchy (as we have here, with 40,000 RBCs in two-dimensions per subject), that there exists a relatively efficient closed form approximation to the full Bayesian solution. In Section 4 we outline how the Expectation-Maximization (EM) algorithm is used to learn densities from the binned and truncated data. Section 5 discusses two alternative methods for classification: (1) a distance-based approach to classification based on pairwise directed divergence between two densities, and (2) a standard discriminative framework using a classification tree. In Section 6 we compare the empirical classification performance (using cross-validation) of the hierarchical approach with the afore-mentioned alternatives. Section 7 contains a discussion of the merits of the different approaches, including suggestions on how this framework, focusing in particular on the interpretability of the learned models.

In terms of related prior work McCallum et al (1998) and Heskes (1998) used hierarchical models to improve classification and regression performance, but in the more standard manner of using feature vectors rather than densities. Hierarchical models are also widely

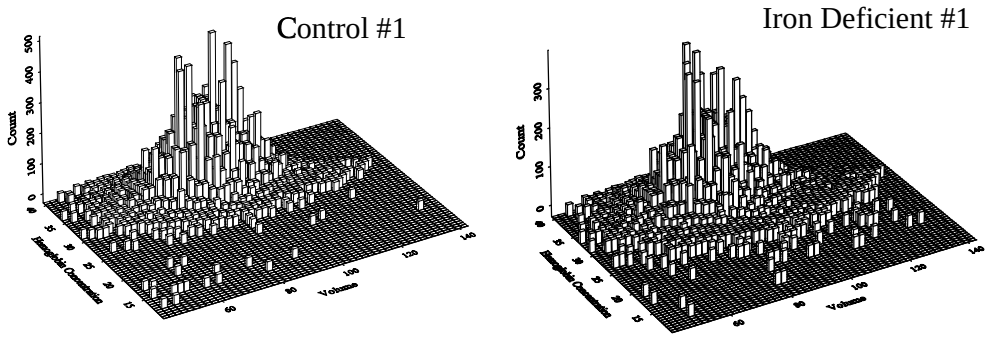

Figure 1: Two-dimensional V-HC histograms of the raw data RBC counts for a Control subject and a subject with Iron Deficient Anemia.

used throughout applied statistics (e.g., see Gelman et al (1995) for a Bayesian perspective). However, we are not aware of any published work in the machine learning, pattern recognition, or statistics literature on using hierarchical models specifically for classifying density functions; in this context, the contribution here is novel. In the context of screening for iron deficiency using RBCs, McLaren and colleagues in a series of papers (McLaren, Brittenham, & Hasselblad, 1986; McLaren et al., 1991; McLaren, 1996) have demonstrated that the distribution of RBC volume can be accurately modeled as a mixture of log-normal density functions. In this paper we extend this work to bivariate V-HC measurements, explicitly model the data-generating process via a hierarchical model, and evaluate the methodology in terms of classification accuracy via cross-validation.

2 A Machine Learning Description of the Problem

Consider a slightly more general description of the problem described above. We have N individuals and each individual belongs to one and only one of K groups, $\{c_1, \ldots, c_K\}$. For each individual i, $1 \leq i \leq N$, in turn we have measurements on a set of n_i "lower-level objects" from that individual (for the RBC problem these lower-level objects are the individual red blood cells). Each of the lower-level objects is itself characterized by a feature-vector (for each patient we have a set of 2-dimensional counts, coming from the the RBCs analyzed by the machine). This type of multi-level structure is not unique to the RBC problem, there are numerous instances of similar scenarios which arise in many practical learning applications, e.g., identifying an individual based on multiple facial images; based on multiple spoken words; based on multiple passages of text written by the individual, and so forth.

This type of data presents two general problems in a learning context:

1. What is an appropriate classification procedure or model for assigning a new individual (described in terms of their RBC measurements) to one of the K classes?, i.e., what

should be the structure of a model which relates the multiple RBC measurements to the class labels?

2. Having defined the general structure, how can we learn such a model from data?

One simple approach is to reduce the multiple measurements for each individual to a standard feature vector of fixed dimensionality. For example, we could take the mean of the V and HC measurements across all 40,000 of an individual's RBCs (or perhaps the variance, or both, or some other summary statistics). However, we don't know in advance whether these features will be discriminative or not and so this approach has somewhat of an ad hoc flavor to it.

A powerful idea in this context is the notion of a *generative probabilistic model* for the data. It is particularly relevant in the context of the RBC problem, where the natural mode of presenting the data to a physician is in the form of a joint distribution. In other words it is the language of joint distributions which forms the existing basis for characterization and description of blood-related diseases given the V and HC measurements.

Let D_i represent the data for the ith individual (i.e., D_i is a set of approximately 40,000 bivariate V-HC measurements made on patient i's blood cells). Thus, we can think of $f(V, HC|D_i, \theta_i)$ as the probability density function for the joint variation of V and HC where θ_i is the set of parameters for the model f. It turns out that, based on prior knowledge of the physical mechanisms of RBC generation and evolution, the functional form of f can be well-determined as being a mixture of log-normal density functions; we will return to this point in Section 4. Thus, assuming the form of f is known, it remains to estimate the parameters θ_i for individual i. We can use standard density estimation techniques (such as EM, e.g., McLachlan and Krishnan, 1997) for this density estimation problem.

Thus, we can reduce the problem of learning from multiple measurements to that of learning from densities, since the density function in principle accurately describes all of the information contained in the original sample D_i. At this point it is not clear yet what we have gained, since it is still not obvious how to model and classify individuals given their density functions, i.e., we are still not in the realm of more familiar classification modeling. In the next section we outline how the hierarchical modeling framework naturally and elegantly will allow us to model both similarities and differences among individuals based on their density functions.

3 Hierarchical Models for Classifying Density Functions

3.1 A Generative Hierarchical Model

Consider the following "generative" model for red blood cell (RBC) production for individuals (a generative model is a model which in a sense can "generate" or "simulate" the observed data: a discriminative model for example is not usually generative since it may only describe the decision boundaries between classes).

1. Choose an individual randomly from the overall population, call this individual i.

2. Assign this individual i to class c_k with probability p_k, $1 \leq k \leq K$ (for the RBC problem we have two classes, $K = 2$, Controls and Iron Deficient).

3. Choose parameters θ_i for individual i using a prior density function $\pi_k(\theta_i)$ on parameters for class k (thus, each class has a density function π_k describing the variation in *parameters* for that class).

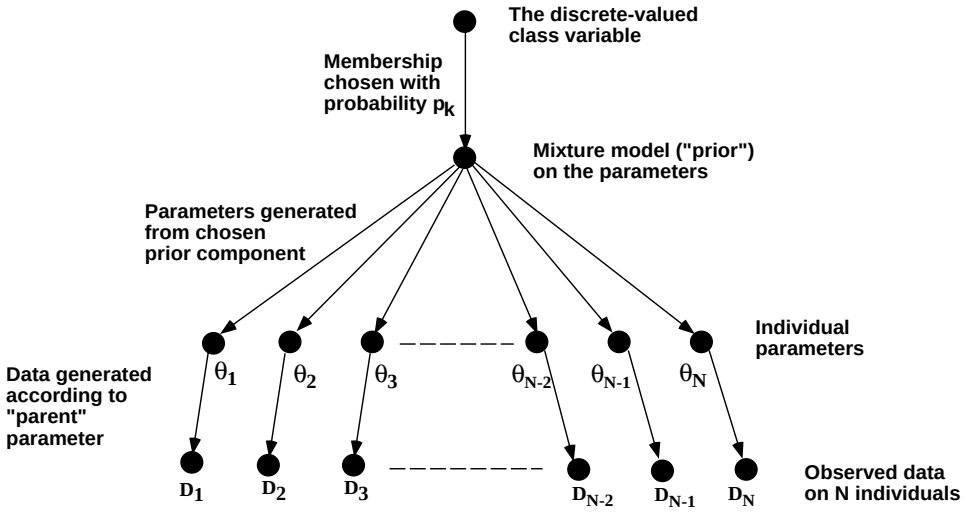

Figure 2: A graphical model (acyclic directed graph) for the hierarchical model for RBC generation.

4. Generate observations (data) D_i for this individual (namely, n_i multivariate measurements, such as V-HC measurements on the individual's red blood cells), conditioned on the parameters θ_i, i.e., we have a data-generating model $p(D_i|\theta_i)$.

Figure 2 summarizes the overall framework graphically. In fact, from a probabilistic viewpoint this is a formal graphical model for the problem and we can read off the relevant conditional independence relations. For example, the data set D_i for any individual i is independent of all other data sets conditioned on the parameters θ_i which generated data D_i. The hierarchical model is quite plausible as a generative mechanism for the RBC data. Homogeneity within a class (Control or Normal) is captured by the assumption that the *parameters* of the density functions for individuals belonging to that class (namely the θ_i's) are themselves parametrized by a common class density function. It is in this parameter space that we will perform classification: if there is relatively little overlap between the different class densities in parameter space, then accurate classification should be possible.

Figure 3 illustrates the general concept of how we model variability in parameter space. Here we have fitted each individual at the RBC level with a two-component mixture density function (full details in Section 4). Figure 3 is a scatter plot of the location of the estimated two-dimensional V-HC bivariate mean for the larger (in probability mass) of the two components in the fitted density function for each individual. The Controls and the Iron Deficient individuals are given different symbols. Furthermore, we have fit a Gaussian model to the parameters of each class (the covariance ellipses of these components are shown); these correspond to the "prior" density functions $\pi_k(\theta)$ on the parameters for class c_k.

3.2 Interpretation of the Model

This type of hierarchical model setup is common in Bayesian statistics. In the Bayesian framework, the densities $\pi_k(\theta)$ on parameters are (naturally) referred to as priors. In fact in a full Bayesian analysis we would put another level of priors in the model ("hyperpriors"), but as we shall see this is not necessary for accurate classification of the RBC data.

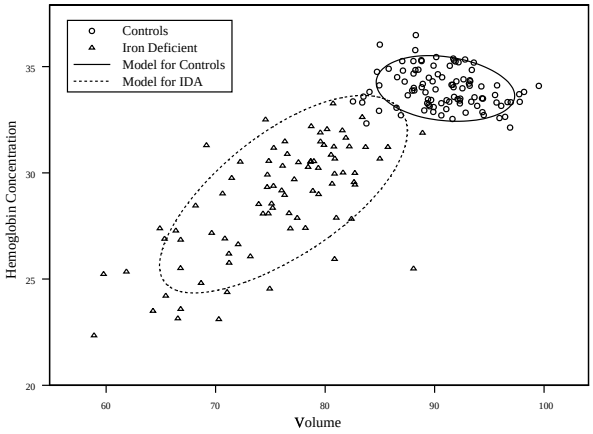

Figure 3: Scatter plot of the estimated mean parameters for individuals from the Iron Deficient and Control groups, with Gaussian "parameter variation models" superposed.

For our problem, since we have K different classes, the "marginal prior" on parameters is a mixture model of the form:

$$\pi(\theta|\phi) = \sum_{k=1}^{K} p(c_k)\pi_k(\theta|\phi_k) \tag{1}$$

where the $p(c_k)$ are the prior probabilities on each class ($\sum_k p(c_k) = 1$), and the $\pi_k(\theta|\phi_k)$ are "component priors" on the parameters θ in a class k and θ here is a parameter *vector*. For example, if we have univariate measurements at the RBC level which we wish to model by a Gaussian density function for each individual, then we would have $\theta = \{\mu, \sigma\}$. π_k would then be a bivariate density function on μ and σ representing the way μ and σ vary within class k. Let ϕ represent the overall set parameters which describe the prior on θ, and let ϕ_k be the parameters for component k of the prior. This appears to be quite a reasonable and parsimonious description of a generative mechanism for the data. For example, in Figure 3, we can see both systematic within- and between-class variability for the means (the μ's).

3.3 Learning the Prior Parameters ϕ from Labelled Data

Given a set of RBC measurements from N individuals, denote the data by $D = \{D_1, \ldots, D_N\}$, where D_i is the set of measurements on blood cells for the ith individual, $1 \leq i \leq N$. Conditioned on the parameters, and given the hierarchical graphical model of Figure 2, the likelihood of the data can be written as:

$$p(D|\theta_1, \ldots, \theta_N, \phi) = \prod_{i=1}^{N} p(D_i|\theta_i). \tag{2}$$

If we don't know the parameters θ_i (as will be the case in practice), we get the *marginal likelihood* as a function of the priors ϕ:

$$L(\phi) = p(D_i|\phi) = \int_{\theta} p(D_i|\theta_i)\pi(\theta_i|\phi)d\theta \tag{3}$$

Given that we have fairly large numbers of blood cells and are fitting a relatively parsimonious bivariate model to their distribution, it seems reasonable to assume that the integral above will be peaked around $\hat{\theta}$ where θ is the maximum a posteriori (MAP) value relative to the prior $p(\theta_i|\phi)$. (If we further assume that this prior is relatively flat in this part of parameter space, then the MAP and maximum likelihood values will be very close). We will simply denote the value around which the integral is peaked as $\hat{\theta}$.

Furthermore, assume that our data are labelled, i.e., we know which of the K classes each individual belongs to. (The generalization to unlabelled, or partially labeled data is straightforward and useful, but is not discussed in this paper). Thus, we can focus on estimating ϕ_k, the parameters describing θ conditioned on membership in class k. For simplicity of notation let us assume (temporarily) that we only have data from class k, and that $D = \{D_1, \ldots, D_N\}$ is the data for the kth class of individuals. Let $\hat{\theta}_i$ be the estimated density parameters for the ith individual in this class, $1 \leq i \leq N$. We have that

$$\begin{aligned}
\arg\max_{\phi} p(\phi_k|D) &= \arg\max_{\phi} p(D|\phi_k)p(\phi_k) \\
&= \arg\max_{\phi} \left(\prod_{i=1}^{N} p(D_i|\phi_k) \right) p(\phi_k) \\
&= \arg\max_{\phi} \left(\prod_{i=1}^{N} \int_{\theta_i} p(D_i|\theta_i)\pi_k(\theta_i|\phi_k)d\theta_i \right) p(\phi_k) \\
&\approx \arg\max_{\phi} \left(\prod_{i=1}^{N} p(D_i|\hat{\theta}_i)\pi_k(\hat{\theta}_i|\phi_k) \right) p(\phi_k) \\
&= \arg\max_{\phi} \left(\prod_{i=1}^{N} \pi_k(\hat{\theta}_i|\phi_k) \right) p(\phi_k). \qquad (4)
\end{aligned}$$

This equation is a standard likelihood (under a conditional independence assumption given the model) times a prior. In other words, to find the MAP value for ϕ, we can just first find the $\hat{\theta}$'s for each individual, and then find the ϕ_k's which maximize the MAP expression above (given a parametric density model for $\pi_k(\theta|\phi_k)$). If the prior $p(\phi_k)$ is relatively flat, it reduces to maximum likelihood estimation of the ϕ_k parameters based on the $\hat{\theta}_i$ "observations," $1 \leq i \leq N$.

For the RBC data, we will be fitting 2-component bivariate log-normal mixtures to the original measurements: this will result in roughly 2 means and 3 covariance parameters per component, for a total of $2(2+3)+1 = 11$ independent parameters in total (i.e., each $\hat{\theta}_i$ will be an 11-dimensional vector). Details on how these parameters are fit (at the individual level) will be provided in Section 4.

3.4 Classification given the Fitted Model

Assume that we have estimated all the parameters of the hierarchical model as described above. Now consider a new individual for whom we have RBC measurements, say D_{N+1}. We would like to know the class probabilities for this individual:

$$\begin{aligned}
p(c_k|D_{N+1}) &\propto p(D_{N+1}|c_k)p(c_k) \\
&= \left(\int p(D_{N+1}|\theta_{N+1})\pi_k(\theta_{N+1}|c_k)d\theta_{N+1} \right) p(c_k) \\
&\propto p(D_{N+1}|\hat{\theta}_{N+1})\pi_k(\hat{\theta}_{N+1}|c_k)p(c_k) \\
&\propto \pi_k(\hat{\theta}_{N+1}|c_k)p(c_k). \qquad (5)
\end{aligned}$$

Thus, based again on the MAP approximation, the posterior class probabilities have a very simple intuitive form: find the parameters $\hat{\theta}_{N+1}$ to fit the RBC data for the unclassified individual, and use Bayes' rule to classify these parameters relative to the prior $\pi_k(\theta|\phi_k)$ (in the equation above, $\pi_k(\theta|c_k)$ is the same as $\pi_k(\theta|\phi_k)$).

Thus, we have arrived at a method for performing probabilistic classification given the hierarchical model, as well as determining how that model can be learned from labelled data. To summarize, the procedure is as follows:

Training: For each class k, $1 \leq k \leq K$:

1. Estimate density function parameters $\hat{\theta}_i$ for each individual i labelled as class k,
2. Estimate the parameters ϕ_k of the"prior component" $\pi_k(\theta|\phi_k)$ using the estimated θ_i's,
3. Estimate the mixture weights in the prior, $p(c_k)$ (e.g., simply as the proportion of individuals belonging to class k).

Classification: For a new unlabelled individual, with observed data D_{N+1}:

1. Estimate the density parameters $\hat{\theta}_{N+1}$ for this individual
2. Find the maximum (over k) of $p(c_k|D_{N+1})$ as described in Equation 5.

4 Density Modeling at the RBC Level

4.1 The Functional Form of the RBC Density Model

The first step in modeling the data is to characterize the two-dimensional V-HC distribution. It can be shown that the marginal volume distribution of a single population of RBC is theoretically lognormal. The lognormality comes from the biological mechanism governing the manner by which cells are produced (McLaren, Brittenham, and Hasselblad; 1986). At each "production step" cells divide and have normal variations in their respective volumes. Since the process is repetitive and the effect is multiplicative (i.e. cells divide), the resulting distribution is lognormal. For iron-deficient subjects, the argument follows that the RBC density can be well-approximated as a two-component log-normal mixture (McLaren, et al. , 1991; McLaren, 1996). Specifically, there are two biological processes that are constantly occurring in a body: 1) red blood cells are produced in the bone marrow; 2) these cells are extruded into the bloodstream and die after about 120 days. For a healthy individual a single population of red blood cells is produced with a mean cell volume and mean hemoglobin concentration within the normal range. In iron deficiency anemia the red blood cells that are produced have decreased volume and hemoglobin, below normal for that of a healthy individual. Thus with development of the disease, gradually, a second subpopulation of red blood cells begins to emerge and over a period of time, the relative ratio of subpopulations of red cells changes.

4.2 Implementation of Mixture Modelling of RBC Count Data

The actual values of volumes and hemoglobin concentrations are not observed for each blood cell (they are quantized). In addition, the counts outside the measurable range are also missing. Model fitting under these conditions (binned and truncated data) has been studied in McLachlan and Jones (1988) and McLachlan and Jones (1990) in connection with

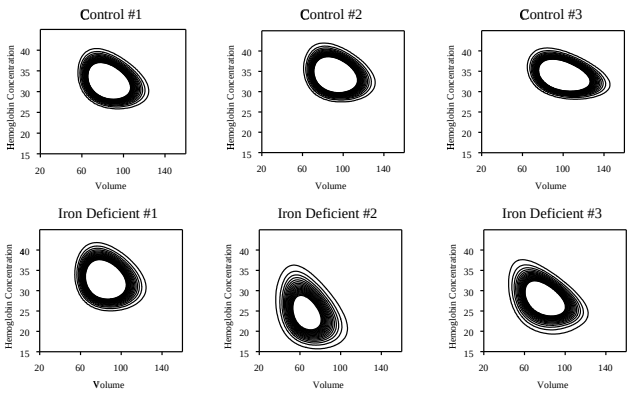

Figure 4: Contour plots from estimated density estimates for three typical Control patients and three typical Iron Deficient Anemia patients. The lowest 10% of the probability contours are plotted to emphasize the systematic difference between the two groups.

fitting univariate volume distributions of RBCs. The theoretical results in the cited papers only handle the one-dimensional case; the extensions to multiple dimensions are relatively straightforward but are not described in any detail here due to space considerations. The workhorse of the fitting procedure is the EM algorithm (McLachlan and Krishnan, 1997). For this problem the M-step does not have a closed-form solution as is the case when fitting mixture models to non-binned data. The main quantities that need to be evaluated at each EM step are now integrals over the bins (McLachlan and Jones, 1988).

Extending the McLachlan and Jones (1990) procedure to the bivariate case directly leads to a rather slow EM algorithm (the execution time for a single EM step is rather large). In the results presented here we used a heuristic initialization method which resulted in reducing the computation time for EM by about a factor of 100. A sub-sample of data points (3000 in the results here) are drawn from the histogram counts assuming a uniform distribution in each bin (this uniform assumption is clearly incorrect since the true densities will be smoothly varying in practice, however, the assumption is fine as a rough guess for initialization). Standard EM is then run to convergence on the simulated sample. The resulting density is then used as a starting point for the binned/truncated version of EM which uses all of the data and performs numerical integration over each bin during each EM step. This reduces the number of iterations of the full EM algorithm drastically while not affecting the accuracy of the solution as the convergence criterion is defined by the full (but slow) algorithm. It is guaranteed to return a local maximum of the likelihood function for binned and truncated data. The process is repeated 10 times with different randomly chosen subsamples to avoid poor local maxima.

Figure 4 shows the densities which were fit to the histogram data by applying this EM mixture modeling procedure to 3 individuals from each class, in terms of the lower 10 percent probability contours of the estimated mixture density One can see that there are differences between the two classes, as well as specific within-class variation.

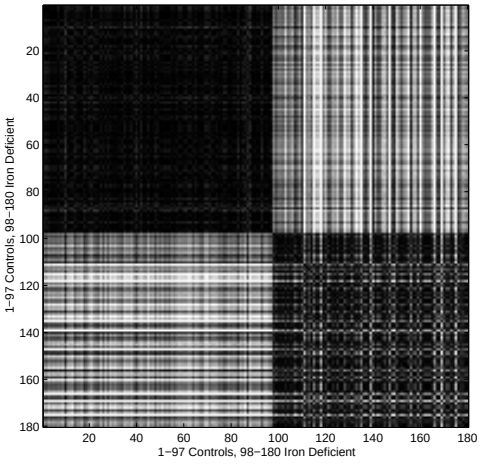

Figure 5: Matrix of Kullback-Leibler divergence scores for the fitted RBC density models.

5 Alternative Classification Techniques

We compared three qualitatively different approaches to classifying RBC data:

- The generative hierarchical modeling approach (described earlier),
- A non-parametric approach based on density distance and multidimensional scaling (MDS), described in Section 5.1, and,
- Discriminative classification using the C5.0 and CART classification tree algorithms (Quinlan, 1997; Breiman et al (1984)), described in Section 5.2.

5.1 Density Distance using K-L Divergence and Multidimensional Scaling (MDS)

The idea here is to define a distance measure directly between the *densities* themselves, and to completely avoid modelling parameters in parameter-space. A well-known distance measure between two densities is the Kullback-Leibler (K-L) distance, defined as the cross-entropy $\int p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta$. This is not a metric, but is strictly positive unless $p = q$ everywhere. It can be made symmetric by defining the *K-L divergence* as

$$KL(p, q) = \frac{1}{2} \left(\int p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta \right) \tag{6}$$

The K-L divergence provides a well-defined notion of density distance. For a subset of the Control and Iron Deficient individuals for whom we have RBC data, we fitted densities to each as described in Section 4 and calculated the pairwise K-L divergence between each pair of individuals. The resulting distance matrix is shown in Figure 5. The separation of the two groups is very clear from the block-diagonal structure of the matrix (darker pixels in the image correspond to smaller K-L values, i.e., the densities are more similar). Furthermore, there appears to be a systematic difference in the two groups in that the Controls (the upper left block) is much more homogeneous (in terms of density distances) than the Iron Deficient group (lower right block) which has greater variability in terms of

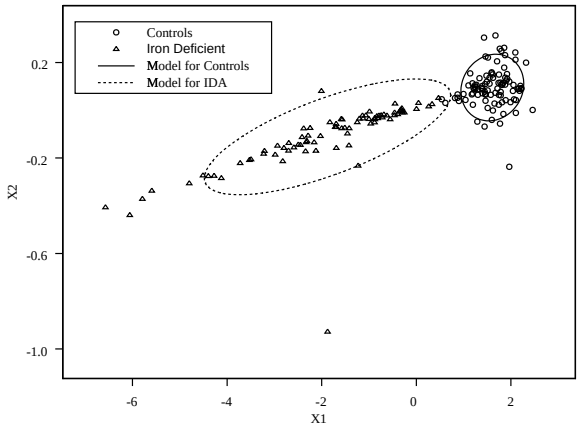

Figure 6: Kullback-Leibler matrix of pairwise density distance scores embedded in a two-dimensional vector space using the MDS technique with fitted Gaussian covariance ellipses for each class.

pairwise distances (i.e., there is greater spread among the densities). This is consistent with Figure 3 where the spread in *component means* is lower for Controls than for Iron Deficient subjects.

The K-L matrix is useful for exploratory analysis but does not easily lend itself to classification in the standard manner. To define a classifier using this data we use multidimensional scaling (MDS) to approximate the distance matrix in a vector space (Borg and Groenen, 1997). MDS places each individual (a row or column from the matrix) in a d-dimensional space such that the pairwise (matrix) distances between points are preserved as closely as possible in the Euclidean vector space. Based on analysis of the fit versus dimensionality, we determined that the best number of dimensions for MDS for this data was 2. The resulting MDS-derived vector space, showing the individuals positioned to reflect as closely as possible their KL divergences, is shown in Figure 6.

With MDS as a tool to embed RBC densities into a vector space (with a well defined notion of "distance"), we then use a Gaussian mixture-model classifier in this vector space to classify the "points" (representing individuals) in this space.

5.2 Discriminative Classification

We also consider a standard discriminative classifier to classify the fixed-length parameter vectors obtained from the density modeling of Section 4. We chose classification trees as a typical method for discriminative classification using the C5.0 and CART algorithms. The decision tree approach is the most direct approach of the three methods for classification and has the advantage of being relatively interpretable from a clinical viewpoint.

We also use the discriminative approach on "ad hoc" features estimated directly from the histogram, Thus, no attempt is made to model the density of the RBCs but instead an individual's histogram is summarized by a vector of 4 simple features, namely, the univariate mean and standard deviation for each of V and HC, as estimated from the histogram. The tree algorithm is then used to learn a classification model in this 4-dimensional space. This

approach has the advantage of being considerably simpler than any of the afore-mentioned classification methods and we will refer to it as the Baseline method in the rest of the paper.

6 Experimental Methodology and Results

6.1 Experimental Methods for Comparing Classifiers

Data collection for this study was completed during 1995. A reference sample group of healthy individuals was recruited from staff physicians and hospital employees at the Western Infirmary, Glasgow, Scotland. Patients included in the study were seen on the wards and in the outpatient clinics and referred to the hospital laboratories for a complete blood count. The results described here are based on the 97 Control and 83 Iron Deficient subjects in the study.

Each of the three methods begins with the EM estimates of the parameters $\hat{\theta}_i$ for each individual i, $1 \leq i \leq N$. These estimates are obtained by fitting a two-component log-normal mixture density using EM as described in Section 4. Each set of parameters $\hat{\theta}_i$ has an associated class label (Control or Iron Deficient) for individual i.

For each of the three methods we performed 100 cross-validation runs, where in each run the data were divided into a randomly chosen training set of 80% of the data and a test set consisting of the remaining 20%. Overall performance for each method is reported as the mean and standard deviation of classification accuracy on the test sets over the 100 runs. Note that the parameter estimation (the running of EM to determine the $\hat{\theta}_i$'s) is completely independent of any other data or class labels; it is a purely unsupervised procedure performed on each individuals RBC measurements. Thus, the parameters $\hat{\theta}_i$ are estimated once and for all, before any cross-validation takes place.

The hierarchical model takes the 11-dimensional parameters and constructs two density models for the parameters, one for Controls and one for Iron Deficient. We experimented with two different density models for the parameters:

1. **"11-Parameters":** here we modeled all of the parameters with a Gaussian mixture model, and used block covariance matrices which allowed covariance between all 4 means and between the 3 independent covariance parameters for each RBC component.

2. **"9-Parameters":** here we re-parametrized some of the parameters to reflect a more natural scale for modeling. We modeled both the log-odds of the component weights and the log of the eigenvalues of the covariance matrices as Gaussian. We again used a two-component Gaussian mixture model with block diagonal covariance matrix allowing covariance between the 4 means, covariance between the 2 log-eigenvalues for each RBC component, and allowing the log-odds of the weight to be independent of the other parameters.

On the training data, for each cross-validation run, we performed a further "internal" cross-validation run to automatically determine whether a single Gaussian or a 2-component Gaussian mixture provides the best fit (using 20 runs of cross-validated likelihood), i.e., each $\pi_k(\theta)$ is itself modeled as either a 1 or 2-component Gaussian mixture. The motivation for this last internal cross-validation is to allow the model to automatically choose a reasonable structure for the "prior" in parameter space.

Table 1: Means and standard deviations of the cross-validated classification error for each of the different classification models across 100 runs.

Method	Features	Mean Error Rate (%)	Standard Deviation
C5.0	Baseline	3.99	Not available
	9-Parameters	3.49	Not available
	11-Parameters	3.36	Not available
CART	Baseline	4.72	3.05
	9-Parameters	4.11	3.48
	11-Parameters	3.53	3.05
KL-MDS		1.56	2.06
Hierarchical	9-Parameters	2.08	4.98
	11-Parameters	1.42	2.57

For the KL-divergence/MDS method, we project the training data as before using the KL/MDS technique described earlier and then use an internal cross-validation strategy (similar to that described above) to build a Gaussian mixture model for each class in the MDS space, using the training data. Classification is then performed on the projected test data points using this Gaussian mixture model.

6.2 Experimental Results

Table 1 summarizes the mean cross-validated error rates and standard errors across the different methods. The discriminative algorithms (C5.0 and CART) were run on the 11 original parameters, the 9 rescaled parameters, and the 4 "Baseline" features. For both trees, the lowest error rate was obtained with the 11 density parameters, followed by the 9 rescaled parameters, followed by histogram-based Baseline features. This suggests that the density modeling is worthwhile in that it leads to lower error rates than simple histogram-based features.

The KL-MDS method and the Hierarchical method (with either 9 or 11 features) had lower error rates than any of the discriminative tree methods (usually on the order of a factor of 2 lower). The 11-parameter hierarchical error rate of 1.42% corresponds to about a 70% decrease in error rate from the error rate of 4.72% of CART on the baseline features and about a 64% decrease in error rate from the 3.99% error rate of C5.0 on the baseline features. Thus, for this particular problem, the hierarchical model and the KL-MDS method appear superior in terms of classification accuracy. Note at the time of writing of this paper, paired experiments have not yet been performed to allow formal statistical hypothesis tests.

For routine clinical classification of RBC data, the decision tree approaches are attractive since the operation of the classifier can be explained in simple "rule-like" terms to a clinician. However, as well as being not quite as accurate in our experiments, tree classifiers do not produce particularly accurate posterior class probabilities and do not fully characterize the within and between-class variability. Thus, if clinicians wished to rank subjects based on likelihood of iron deficiency (for example), the hierarchical model approach may be the most useful since it produces much more plausible posterior probabilities than can be produced from the tree model. In addition, the hierarchical model is intrinsically interesting from a medical research viewpoint since it provides a basis for a complete characterization of blood

disorders in H-VC space both in terms of typicality and variability of individuals within each group, as well as full characterization of group differences.

The KL-MDS approach is also an interesting alternative method. It is competitive with the hierarchical method in terms of accuracy but it suffers from a lack of interpretability (the MDS dimensions are not necessarily interpretable). However, it clearly has a strong visual appeal and can be quite useful for exploratory data analysis in problems of this nature.

7 Conclusions

We investigated the problem of automated screening of blood samples of individuals for the purpose of detecting iron deficiency anemia. The problem is more complex than the typical machine learning classification problem since each individual must be classified based on a bivariate histogram (rather than a feature vector). Three different techniques for classifying the individual densities were proposed: (1) a probabilistic hierarchical model, (2) embedding pairwise KL distances between densities in a vector space, and (3) direct discrimination of density or histogram parameters using a classification tree. All methods were accurate in the 96 to 99% range in the cross-validation experiments, with density-based methods outperforming simpler discriminative methods. We conclude that accurate low-cost automated screening of subjects for iron deficiency, using V-HC count data, appears quite feasible.

Acknowledgments

The contribution of CMcL to this paper has been supported in part by grants from the National Institutes of Health (R43-HL46037 and R15-HL48349) and a Wellcome Research Travel Grant awarded by the Burroughs Wellcome Fund. The contributions of IC and PS have been supported in part by the National Science Foundation under Grant IRI-9703120. We thank Thomas H. Cavanagh for providing laboratory facilities. We are grateful to Dr. Albert Greenbaum for technical assistance.

References

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984) *Classification and Regression Trees*, Belmont, CA: Wadsworth Press.

Borg, I. and Groenen, P. (1997) *Modern Multidimensional Scaling: Theory and Application*, New York: Springer.

Gelman, A., Carlin, J., Stern., H., and Rubin, D.B. (1995) *Bayesian Data Analysis*, New York: Chapman and Hall.

Heskes, T. (1998), 'Solving a huge number of similar tasks: a combination of multi-task learning and hierarchical Bayesian modeling,' in *Machine Learning, Proceedings of the Fifteenth International Conference (ICML'98)*, J. Shavlik (ed.), San Francisco: Morgan Kaufmann, 233-241.

Jones, P. N., McLachlan, G. J. (1990) 'Maximum Likelihood Estimation from Grouped and Truncated Data with Finite Normal Mixture Models,' *Applied Statistics-Journal of the Royal Statistical Society Series C*, 39(N2):273-282.

McCallum, A., Rosenfeld, R., Mitchell, T., Ng A.Y. (1998) 'Improving text classification by shrinkage in a hierarchy of classes,' in *Machine Learning, Proceedings of the Fifteenth International Conference (ICML'98)*, J. Shavlik (ed.), San Francisco: Morgan Kaufmann.

McLachlan, G. J. and Jones, P. N. (1988) 'Fitting mixture models to grouped and truncated data via the EM algorithm,' *Biometrics*, 44(2):571-8.

McLachlan, G. J., and Krishnan, T. (1997) *The EM Algorithm and Extensions*, New York: John Wiley and Sons.

McLaren, C. E., Brittenham, G. M., Hasselblad, V. (1986) 'Analysis of the volume of red blood cells: application of the expectation-maximization algorithm to grouped data from the doubly-truncated lognormal distribution,' *Biometrics*, 42(1):143-58.

McLaren, C. E. (1996) 'Mixture models in haematology: a series of case studies,' *Statistical Methods in Medical Research*, 5(2):129-53.

McLaren, C. E., Wagstaff, M., Brittenham, G. M., Jacobs, A. (1991) 'Detection of Two-Component Mixtures of Lognormal Distributions in Grouped, Doubly Truncated Data: Analysis of Red Blood Cell Volume Distributions,' *Biometrics*, 47(2):607-22.

Quinlan, J. R. (1997) 'Using C5.0: An Informal Tutorial,' Sydney, Australia: Rulequest Research.