IJDAR (2003) 6: 154–166
Digital Object Identifier (DOI) 10.1007/s10032-003-0108-x

International Journal on IJDAR Document Analysis and Recognition

Adaptive classifiers for multisource OCR

Sriharsha Veeramachaneni, George Nagy

Electrical, Computer and Systems Engineering Department, Rensselaer Polytechnic Institute, 110 8th St., Troy, NY 12180 USA

Received: November 14, 2002 / Accepted: March 06, 2003
Published online: September 12, 2003 –

Abstract. When patterns occur in large groups generated by a single source (*style consistent test data*), the statistics of the test data differ from those of the training data, which consist of patterns from all sources. We present a Gaussian model for continuously distributed sources under which we develop adaptive classifiers that specialize in the statistics of style-consistent test data. On NIST handwritten digit data, the adaptive classifiers reduce the error rate by more than 50% operating on one writer (≈ 10 samples/class) at a time.

Keywords: Optical character recognition – Style – Writer consistency – Adaptation – Quadratic discriminant –Interpattern feature dependence – Digit recognition

1 Introduction

Pattern classifiers are generally designed, built, and trained with the tacit assumption that the test patterns encountered in the field are well represented by the training data (i.e., the stochastic process generating the patterns is assumed to be stationary). This assumption is often violated for a variety of reasons. For example, due to the proliferation of fonts and typefaces with the advent of digital font design, previously unseen fonts may be encountered by character recognizers. Also, classifiers are usually designed to recognize patterns from various styles by training them on patterns from a large number of styles. However, in most applications, a given document is rendered in one or at most a few styles. The classification accuracy suffers due to the statistical dissimilarity between the training and test data. It is therefore appealing to consider the possibility of adaptation or modification of the decision regions by learning from the test set.

Correspondence to: G. Nagy (e-mail: nagy@ecse.rpi.edu)

For the purposes of this communication, we assume that the generating process is nonstationary only because each homogeneous test set originates from a single source (whose identity is unknown). Such intrasource homogeneity is called *style consistency*.

When the test set is sufficiently large, general trends in the data can be discovered. This knowledge can then be used by a pattern classifier to improve its decisions on the patterns in the same test set. Although little use of adaptive methods has been reported for OCR, there has been considerable work in the field of communications for operator adaptive Morse code recognition [18], adaptive equalization [13], and, more recently, speech recognition [11]. Gauvain and Lee propose speaker adaptive speech recognition by maximum a posteriori estimation of HMMs [8]. Castelli and Cover explore the relative value of labeled and unlabeled samples for pattern classification from an information theoretic point of view [3].

A heuristic self-corrective character recognition algorithm that adapts to the typeface of the document to increase accuracy is described in [15] and applied to a hundred-font classifier in [1]. Adaptation is performed by retraining the classifier on the test data with the labels previously assigned by the classifier. This process is iterated until the assigned labels are unchanged.

Other approaches to adaptive classification include clustering the test data (identifying groups of similar patterns) and mixture identification [6]. Unsupervised classification algorithms based on mixture identification model the test data as a mixture of patterns from several classes, each distributed according to a parameterized density. The parameters of the mixture distribution are then estimated. The expectation-maximization (EM) algorithm, which is an optimization technique proposed by Dempster et al., is widely used for the maximum-likelihood estimation of mixture distribution parameters [5,17].

After clustering, either linguistic constraints or a labeled training set can be utilized to label the clusters. Linguistic constraints can be exploited by considering the text as a substitution cipher that is decrypted to obtain the recognition result [2,10,16]. Shahshahani and Land-

grebe use coarse information from the training data to identify the clusters [19]. Mathis and Breuel have recently proposed a hierarchical Bayesian approach to adapting to the style of a batch of test data [14]. They assume that the parameters of the test data are drawn according to a known "hyperprior" distribution. The hyperprior they propose is an independent Gaussian distribution of class means. In addition to such style-induced intraclass consistency, our model includes interclass correlations that extend the applicability of adaptive classification even to relatively short fields.

We present strategies to improve recognition accuracy of a specific class of multisource pattern classifiers (Gaussian quadratic discriminant classifiers) by adapting to the statistics of a style-consistent test set. We will derive a family of adaptive classifiers under the assumption of normally distributed styles and features, explore their characteristics through a detailed study of a simple example, and demonstrate their efficacy on NIST handwritten data.

2 Adaptive quadratic discrimination

We now formulate precisely the problem of classifier adaptation and show that under some reasonable assumptions adaptive methods are effective in exploiting style consistency in pattern fields.

We consider the problem of classifying the patterns in a large test set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L\}$, where each $\boldsymbol{x}_i$ is a d-dimensional feature vector, into one of N classes $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$. The samples in the test set $\mathcal{X}$ are independently drawn according to the probability density $p(\boldsymbol{x}) = \sum_{i=1}^{N} p(c_i)p(\boldsymbol{x}|\, c_i)$, where the class-conditional feature distributions are $p(\boldsymbol{x}|\, c_i) = f_i(\boldsymbol{x}) \sim \mathcal{N}(\mu_i, \Sigma_i)$, $i = 1, \ldots, N$ with a priori class probabilities $p(c_i) = \alpha_i$, $i = 1, \ldots, N$. Thus, $p(\boldsymbol{x})$ can be written as the mixture density

$$p(\boldsymbol{x}; \Theta) = \sum_{i=1}^{N} \alpha_i p(\boldsymbol{x}; \theta_i) \tag{1}$$

where $\theta_i = (\mu_i, \Sigma_i)$ parameterizes the class-conditional density and Θ is the vector of all the parameters (the values α_i and θ_i, $i = 1, \ldots, N$).

We postulate the existence of a training set for estimating the class-conditional feature distributions given by

$$\hat{f}_i^{(0)}(\boldsymbol{x}) \sim \mathcal{N}(\hat{\mu}_i^{(0)}, \hat{\Sigma}_i^{(0)}),\ i = 1, \ldots, N \text{ and}$$

$$\hat{p}^{(0)}(c_i) = \hat{\alpha}_i^{(0)},\ i = 1, \ldots, N$$

If $\mathcal{X}$ is classified using quadratic discriminant functions constructed with the estimated parameters, generally the expected error rate of the classifier on $\mathcal{X}$ will be higher than the lowest achievable error rate because of the discrepancies between the estimated parameters and the true parameters (i.e., the parameters of the test set). Under the assumption that the estimated parameters are "sufficiently" close to the true parameters and that the test set $\mathcal{X}$ is sufficiently large, we wish to adapt the classifier to the true parameters of the test set $\mathcal{X}$.

The general method for adaptive Gaussian quadratic discriminant classification is to estimate the parameter Θ from the test set $\mathcal{X}$ and use the estimate $\hat{\Theta}$ to classify $\mathcal{X}$.

The maximum-likelihood estimator $\hat{\Theta}_{ML}$ for the Θ that parameterizes the mixture density in Eq. 1 is

$$\begin{aligned}\hat{\Theta}_{ML} &= \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{L} \log p(\boldsymbol{x}_i; \Theta) \\ &= \underset{(\alpha_1, \ldots, \alpha_N, \theta_1, \ldots, \theta_N)}{\operatorname{argmax}} \sum_{i=1}^{L} \log \left(\sum_{j=1}^{N} \alpha_j p(\boldsymbol{x}_j; \theta_j) \right)\end{aligned}$$

which presents a maximization problem that is analytically intractable. We utilize the well-known *expectation-maximization* (EM) algorithm, which is an iterative technique to solve this optimization problem. We propose two approaches: (i) initialize the EM algorithm with the parameter values obtained from the training set and (ii) use the parameters of the training set as priors in a MAP formulation of the EM algorithm.

3 Model for continuous styles

We formulate the problem of classifying *fields* of *style-consistent* patterns in a way that leads naturally to adaptive strategies. We classify a field $\boldsymbol{y} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_L^T)^T$ (each $\boldsymbol{x}_i \in \mathbb{R}^d$ represents one of the L patterns in the field) into one of the field classes in $\mathcal{C}^L$, where $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$. We assume the existence of a "hidden" Nd-dimensional random vector $\mathbf{s} = (\mathbf{m}_1^T, \ldots, \mathbf{m}_N^T)^T$, where each $\mathbf{m}_i$ is a d-dimensional random vector, $\forall i = 1, \ldots, N$. Let the random vector $\mathbf{s}$ represent the style whose identity is entirely determined by its class-conditional means. We will assume that the a priori class probabilities are known and independent of $\mathbf{s}$. We make the following assumptions on the feature distributions:
1. $\mathbf{s} \sim \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$, where

$$\boldsymbol{\mu}_s = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_N \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_s = \begin{pmatrix} \mathrm{C}_{11} & \mathrm{C}_{12} & \ldots & \mathrm{C}_{1N} \\ \mathrm{C}_{21} & \mathrm{C}_{22} & \ldots & \mathrm{C}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{C}_{N1} & \mathrm{C}_{N2} & \ldots & \mathrm{C}_{NN} \end{pmatrix}$$

where $\mathrm{C}_{ij} = E\{(\mathbf{m}_i - \mu_i)(\mathbf{m}_j - \mu_j)^T\}$
2. $(\mathbf{y}|\, \mathsf{c}_1 = c_i, \ldots, \mathsf{c}_L = c_k, \mathbf{s} = (\boldsymbol{m}_1^T, \ldots, \boldsymbol{m}_N^T)^T)$ $\sim \mathcal{N}((\boldsymbol{m}_i^T, \ldots, \boldsymbol{m}_k^T)^T, \boldsymbol{\Sigma}_{i,\ldots,k})$, where

$$\boldsymbol{\Sigma}_{i,\ldots,k} = \begin{pmatrix} \Sigma_i & \ldots & 0_{d\times d} \\ \vdots & \ddots & \vdots \\ 0_{d\times d} & \ldots & \Sigma_k \end{pmatrix}$$

The covariance matrix $\boldsymbol{\Sigma}_s$ specifies the amount of style variability, while the matrices $\{\Sigma_1, \ldots, \Sigma_N\}$ specify the amount of intrastyle variance in the patterns.

Under this model, a style is generated by selecting the class means. The mean for a particular class is a random translation from the overall mean (μ_i) of that class (grand mean over all styles). The translation vector for each class is Gaussian distributed with zero mean. Furthermore, the translations for different classes are correlated (given by $\boldsymbol{\Sigma}_s$). After the style is chosen, the features are generated for each class independently, according to the style-specific means and a covariance matrix (Σ_i) that depends on the class but not on the style.

In such a scenario it can be shown that the field-class-conditional feature distributions $(\mathbf{y}|\, c_i, \ldots, c_k)$ are Gaussian [20]. Therefore, a style-conscious quadratic discriminant function (SQDF) field classifier can be constructed in the field-feature space to yield the minimum field misclassification rate.

The field-class conditional means and covariance matrices can be shown to be

$$\mu_{i,j,\ldots,k} = E\{\mathbf{y}|\, c_i, c_j, \ldots, c_k\} = \begin{pmatrix} \mu_i \\ \mu_j \\ \vdots \\ \mu k \end{pmatrix} \tag{2}$$

$$\begin{aligned} \mathsf{K}_{i,j,\ldots,k} &= E\{(\mathbf{y} - \mu_{i,j,\ldots,k})(\mathbf{y} - \mu_{i,j,\ldots,k})^T |\, c_i, c_j, \ldots, c_k\} \\ &= \begin{pmatrix} \Sigma_i + \mathrm{C}_{ii} & \mathrm{C}_{ij} & \ldots & \mathrm{C}_{ik} \\ \mathrm{C}_{ji} & \Sigma_j + \mathrm{C}_{jj} & \ldots & \mathrm{C}_{jk} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{C}_{ki} & \mathrm{C}_{kj} & \ldots & \Sigma_k + \mathrm{C}_{kk} \end{pmatrix} \end{aligned} \tag{3}$$

Example 1.

The following example illustrates our model of continuous styles for classifying test fields of length L (each field is generated by a single source). The possible singlet-class labels are $\mathcal{C} = \{A, B\}$. For simplicity, we assume that the classes are equally likely (i.e., $p(A) = p(B) = 1/2$) and no linguistic dependence is present (i.e., $p(AA) = p(A)p(A)$ etc.).

The class-and-source conditional singlet-feature distributions are

$$(\mathsf{x}|\, A, \mathsf{s} = s) \sim N(s, \sigma^2), \;\; (\mathsf{x}|\, B, \mathsf{s} = s) \sim N(d_c + s, \sigma^2)$$

and the sources are distributed according to

$$\mathsf{s} \sim N(0, d_s^2/4) \tag{4}$$

The distribution of sources and the feature distributions are shown in Fig. 1. Here, the source is identified by only one number (instead of the two class-conditional means) because, given the mean of A (denoted m_A) of a particular source, we can obtain the source-specific mean of B (denoted m_B) by $m_B = m_A + d_c$. That is, the mean of A and the mean of B are maximally correlated (in a multiwriter handwriting recognition problem the correlation, although not maximal, arises because of intrawriter consistency). Here the single-style variable just specifies the "shift" of the mean of A from the origin.

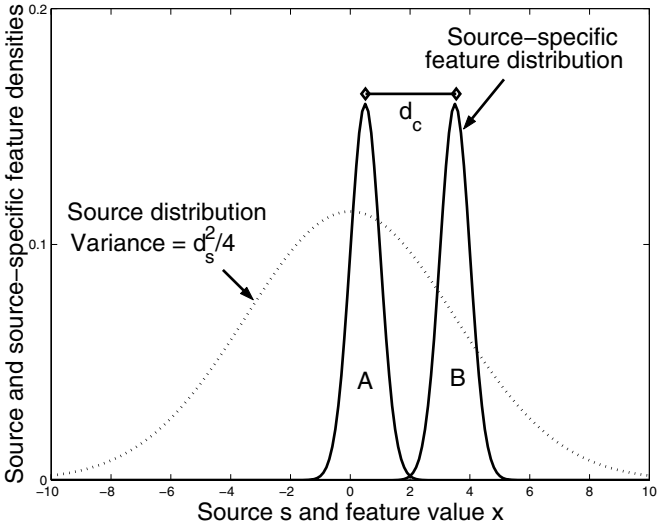


Fig. 1. The source-and-class-conditional feature distributions – normally distributed sources

3.1 EM algorithm

We now briefly summarize the EM algorithm. Suppose that the observations (the so-called *incomplete* data) $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L\}$ are drawn independently according to the probability law $p(\boldsymbol{x}; \Theta)$. We posit the existence of the *complete* data $\mathcal{Z} = (\mathcal{X}, \boldsymbol{C})$, distributed with the probability law

$$p(\mathbf{z}; \Theta) = p(\mathbf{x}, \mathbf{c}; \Theta) = p(\mathbf{c}|\, \mathbf{x}, \Theta) p(\mathbf{x}; \Theta) \tag{5}$$

Therefore,

$$\log p(\mathcal{X}; \Theta) = \log p(\mathcal{Z}; \Theta) - \log p(\boldsymbol{C}|\, \mathcal{X}, \Theta)$$

Since $E\{\log p(\mathcal{X}; \Theta)|\, \mathcal{X}, \Theta^k\} = \log p(\mathcal{X}; \Theta)$, for any fixed Θ^k, we have

$$\begin{aligned} \log p(\mathcal{X}; \Theta) = & \underbrace{E\{\log p(\mathcal{Z}; \Theta)|\, \mathcal{X}, \Theta^k\}}_{Q(\Theta|\, \Theta^k)} \\ & - \underbrace{E\{\log p(\boldsymbol{C}|\, \mathcal{X}, \Theta)|\, \mathcal{X}, \Theta^k\}}_{H(\Theta|\, \Theta^k)} \end{aligned} \tag{6}$$

The above expectations are over the unobserved variables. It can be shown using Jensen's inequality that $H(\Theta|\, \Theta^k) \leq H(\Theta^k|\, \Theta^k)$. Thus, in order to maximize the log-likelihood function we need to maximize only $Q(\Theta|\, \Theta^k)$. The EM algorithm has the following two iterated steps and is guaranteed to converge to a local optimum.

$$\text{E step: Compute } Q(\Theta|\, \Theta^k) = E\{\log p(\mathcal{Z}; \Theta)|\, \mathcal{X}, \Theta^k\} \tag{7}$$

$$\text{M step: Update } \Theta^{k+1} = \underset{\Theta}{\operatorname{argmax}}\, Q(\Theta|\, \Theta^k) \tag{8}$$

We are interested in Gaussian mixtures. That is, for the test set $\mathcal{X}$ the likelihood function is

$$p(\mathcal{X};\Theta) = \prod_{i=1}^{L}\sum_{j=1}^{N}\alpha_j p(\boldsymbol{x}_i|\, c_j;\theta_j)$$
$$= \prod_{i=1}^{L}\sum_{j=1}^{N}\frac{\alpha_j}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}$$
$$\times exp(-\frac{1}{2}(\boldsymbol{x}_i-\mu_j)^T\Sigma_j^{-1}(\boldsymbol{x}_i-\mu_j))$$

It has been shown for Gaussian mixtures that

$$Q(\Theta|\,\Theta^k) = \sum_{i=1}^{L}\sum_{j=1}^{N}\log(\alpha_j)p(c_j|\,\boldsymbol{x}_i;\Theta^k) + \sum_{i=1}^{L}\sum_{j=1}^{N}\log p(\boldsymbol{x}_i|\,c_j;\theta_j))p(c_j|\,\boldsymbol{x}_i;\Theta^k) \quad (9)$$

To obtain the EM update equation for Θ, $Q(\Theta|\,\Theta^k)$ must be maximized with respect to Θ, subject to the constraints $\alpha_i \geq 0\ \forall i = 1,\ldots,N$ and $\sum_{j=1}^{N}\alpha_i = 1$. Below we present the general update formulae for EM algorithm for Gaussian mixture identification.

$$\alpha_j^{k+1} = \frac{1}{L}\sum_{i=1}^{L}p(c_j|\,\boldsymbol{x}_i;\Theta^k)$$

$$\mu_j^{k+1} = \frac{\sum_{i=1}^{L}\boldsymbol{x}_i p(c_j|\,\boldsymbol{x}_i;\Theta^k)}{\sum_{i=1}^{L}p(c_j|\,\boldsymbol{x}_i;\Theta^k)}$$

$$\Sigma_j^{k+1} = \frac{\sum_{i=1}^{L}p(c_j|\,\boldsymbol{x}_i;\Theta^k)(\boldsymbol{x}_i-\mu_j^{k+1})(\boldsymbol{x}_i-\mu_j^{k+1})^T}{\sum_{i=1}^{L}p(c_j|\,\boldsymbol{x}_i;\Theta^k)}$$

3.2 Maximum a posteriori (MAP) parameter estimation of component means of Gaussian mixtures

Suppose that we are given L observations $\mathcal{X} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_L\}$ drawn independently according to the probability law, $p(\boldsymbol{x}|\,\theta)$, where θ is the realization of the random variable $\boldsymbol{\theta}$ that we wish to estimate. The *maximum a posteriori estimator* of θ is given by

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\ \{\log p(\theta) + \log p(\mathcal{X}|\,\theta)\} \quad (10)$$

The log-likelihood function for MAP estimation is identical to that for ML estimation, save the extra bias term $\log p(\theta)$. As suggested in [5], we note the mathematical equivalence between the optimization problem for MAP estimation and ML estimation to justify the EM algorithm for the solution. This variant of the EM algorithm has been dubbed the *Bayesian EM algorithm* [4].

From Eq. 6 we have

$$\log p(\mathcal{X},\Theta) = Q(\Theta|\,\Theta^k) + \underbrace{\log p(\Theta)}_{B(\Theta)} - H(\Theta|\,\Theta^k)$$

Therefore, at every iteration of the EM algorithm, we improve the estimate of Θ by choosing the value that maximizes the quantity

$$Q'(\Theta|\,\Theta^k) = Q(\Theta|\,\Theta^k) + B(\Theta) \quad (11)$$

$B(\Theta)$ is a bias term that factors in our prior knowledge about the distribution of $\boldsymbol{\Theta}$.

We will use the above model for style-consistent test fields as a prior for the MAP parameter estimation. We model styles by their class means, distributed as described in Sect. 3. We often encounter classification problems where the training data contain patterns drawn from several sources but the test patterns are generated by only one of these sources. It is therefore the objective of the MAP adaptive scheme to adapt or *specialize* to the said source. The samples in the test set $\mathcal{X} = \{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_L\}$ are independently drawn according to the probability law $p(\boldsymbol{x}) = \sum_{i=1}^{N}p(c_i)p(\boldsymbol{x}|\,c_i)$, where the class-conditional feature distributions are $p(\boldsymbol{x}|\,c_i) = f_i(\boldsymbol{x}) \sim \mathcal{N}(\mathbf{m}_i,\Sigma_i),\ i = 1,\ldots,N$ with a priori class probabilities $p(c_i) = \alpha_i,\ i = 1,\ldots,N$. We assume that the means $\{\mathbf{m}_i\}_{i=1}^{N}$ are distributed according to
1. $\mathbf{s} = (\mathbf{m}_1^T,\ldots,\mathbf{m}_N^T)^T \sim \mathcal{N}(\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s)$, where

$$\boldsymbol{\mu}_s = \begin{pmatrix}\mu_1\\ \vdots\\ \mu_N\end{pmatrix} \text{ and } \boldsymbol{\Sigma}_s = \begin{pmatrix}\mathrm{C}_{11} & \mathrm{C}_{12} & \ldots & \mathrm{C}_{1N}\\ \mathrm{C}_{21} & \mathrm{C}_{22} & \ldots & \mathrm{C}_{2N}\\ \vdots & \vdots & \ddots & \vdots\\ \mathrm{C}_{N1} & \mathrm{C}_{N2} & \ldots & \mathrm{C}_{NN}\end{pmatrix}$$

Note that this distribution is identical to the one describing the distribution of styles in Sect. 3. Since we have prior knowledge only about the distribution of the means, we assume for now that the mixture component weights $\{\alpha_i\}_{i=1}^{N}$ and the style-conditional singlet-class covariance matrices $\{\Sigma_i\}_{i=1}^{N}$ are known. We wish to compute only the MAP estimate of $\boldsymbol{s}$, the vector of class means, using the EM algorithm.

From Eqs. 9 and 11 we have

$$Q'(\Theta|\,\Theta^k) = \sum_{i=1}^{L}\sum_{j=1}^{N}\log(\alpha_j)p(c_j|\,\boldsymbol{x}_i,\Theta^k) + \sum_{i=1}^{L}\sum_{j=1}^{N}\log p(\boldsymbol{x}_i|\,c_j,\theta_j)p(c_j|\,\boldsymbol{x}_i,\Theta^k) + \log\underbrace{p(\boldsymbol{m}_1,\ldots,\boldsymbol{m}_N)}_{p(\boldsymbol{s})}$$

Ignoring all the terms that are functionally independent of the class means, we have for our problem

$$
\begin{aligned}
&Q'(\Theta|\,\Theta^k)\\
&=\sum_{i=1}^{L}\sum_{j=1}^{N}-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{m}_j)^T\Sigma_j^{-1}(\boldsymbol{x}_i-\boldsymbol{m}_j)p(c_j|\,\boldsymbol{x}_i,\boldsymbol{s}^k)\\
&\quad-\frac{1}{2}(\boldsymbol{s}-\boldsymbol{\mu}_s)^T\boldsymbol{\Sigma}_s^{-1}(\boldsymbol{s}-\boldsymbol{\mu}_s)\\
&=\sum_{i=1}^{L}-\frac{1}{2}(\boldsymbol{v}_i-\boldsymbol{s})^T\mathsf{D}_i^{-1}(\boldsymbol{v}_i-\boldsymbol{s})\\
&\quad-\frac{1}{2}(\boldsymbol{s}-\boldsymbol{\mu}_s)^T\boldsymbol{\Sigma}_s^{-1}(\boldsymbol{s}-\boldsymbol{\mu}_s)
\end{aligned}
$$

where $\boldsymbol{v}_i$ is an $Nd\times 1$ vector and D_i^{-1} is an $Nd\times Nd$ matrix given by

$$
\boldsymbol{v}_i=\begin{pmatrix}\boldsymbol{x}_i\\ \boldsymbol{x}_i\\ \vdots\\ \boldsymbol{x}_i\end{pmatrix}
$$

and

$$
\mathsf{D}_i^{-1}=\begin{pmatrix}\Sigma_1^{-1}p(c_1|\,\boldsymbol{x}_i,\boldsymbol{s}^k) & \mathbf{0} & \dots & \mathbf{0}\\ \mathbf{0} & \Sigma_2^{-1}p(c_2|\,\boldsymbol{x}_i,\boldsymbol{s}^k)\dots & & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \dots\Sigma_N^{-1}p(c_N|\,\boldsymbol{x}_i,\boldsymbol{s}^k)\end{pmatrix}
$$

Note that D_i^{-1} depends on the current estimate $\boldsymbol{s}^k$ of the style parameter.

To maximize $Q'(\Theta|\,\Theta^k)$, we set its derivative with respect to $\boldsymbol{s}$ equal to zero, to obtain

$$
\begin{aligned}
&\sum_{i=1}^{L}\mathsf{D}_i^{-1}(\boldsymbol{s}-\boldsymbol{v}_i)+\boldsymbol{\Sigma}_s^{-1}(\boldsymbol{s}-\boldsymbol{\mu}_s)=0\\
&\Rightarrow\left(\sum_{i=1}^{L}\mathsf{D}_i^{-1}+\boldsymbol{\Sigma}_s^{-1}\right)\boldsymbol{s}=\sum_{i=1}^{L}\mathsf{D}_i^{-1}\boldsymbol{v}_i+\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\mu}_s
\end{aligned}
$$

Thus the Bayesian EM update formula for the MAP estimation of $\boldsymbol{s}$ is

$$
\begin{aligned}
\boldsymbol{s}^{k+1}&=\left(\sum_{i=1}^{L}\mathsf{D}_i^{-1}+\boldsymbol{\Sigma}_s^{-1}\right)^{-1}\left(\sum_{i=1}^{L}\mathsf{D}_i^{-1}\boldsymbol{v}_i+\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\mu}_s\right)\\
&=\left(\boldsymbol{\Sigma}_s\sum_{i=1}^{L}\mathsf{D}_i^{-1}+\mathrm{I}\right)^{-1}\left(\boldsymbol{\Sigma}_s\sum_{i=1}^{L}\mathsf{D}_i^{-1}\boldsymbol{v}_i+\boldsymbol{\mu}_s\right)\quad(12)
\end{aligned}
$$

Note that when $L=0$, $\boldsymbol{s}^{k+1}=\boldsymbol{\mu}_s$ and when $L\to\infty$, the update formulae for the means become decoupled to be

$$
\begin{aligned}
\boldsymbol{m}_j^{k+1}&=\left(\sum_{i=1}^{L}\Sigma_j^{-1}p(c_j|\,\boldsymbol{x}_i,\boldsymbol{s}^k)\right)^{-1}\left(\sum_{i=1}^{L}\boldsymbol{x}_i\Sigma_j^{-1}p(c_j|\,\boldsymbol{x}_i,\boldsymbol{s}^k)\right)\\
&=\frac{\sum_{i=1}^{L}\boldsymbol{x}_i p(c_j|\,\boldsymbol{x}_i,\boldsymbol{s}^k)}{\sum_{i=1}^{L}p(c_j|\,\boldsymbol{x}_i,\boldsymbol{s}^k)}
\end{aligned}
$$

which is identical to the EM update formula for ML estimation. That is, the weight we assign to prior knowledge decreases with increasing sample size.

Example 1 continued

We construct and study the following classifiers for the multisource classification problem presented in the example.

- SQDF: For any field length, the features are field-class-conditionally normally distributed. We can therefore construct an optimal (for field error rate) quadratic discriminant field classifier.
- SOPT: As indicated in Sect. 3, the SQDF classifier achieves the minimum field error rate if the sources (identified by their class means) are normally distributed according to the above model. When it is desirable to minimize the character (singlet) error rate instead of the field error rate, we can construct a field classifier optimized for singlet error rate by replacing the zero-one loss function with a loss function based on the Hamming distance between field classes. Such a classifier, called *singlet error optimized classifier* (denoted SOPT), does not yield quadratic discriminant functions for our problem.

For both the ML and MAP adaptive schemes, we assume that the source-specific intraclass variance (σ^2) is known and adapt only the class means.[1] To classify the field $y=(x_1,\dots,x_L)^T$, we constructed four different adaptive classifiers as follows.

- ML1: Maximum-likelihood estimation of the style parameter s (i.e., the shift of the mean of class A from the origin). The means of A and B are given by $\hat{m}_A^k=\hat{s}^k$ and $\hat{m}_B^k=d_c+\hat{s}^k$, where $\hat{s}^k$ is the estimate of s at the k^{th} iteration of the EM algorithm. The EM update formula for ML estimation of s is

$$
\begin{aligned}
\hat{s}^{k+1}&=\frac{1}{L}\Big(\sum_{i=1}^{L}x_i-\mu_A\sum_{i=1}^{L}p(A|\,x_i,\hat{m}_A^k,\hat{m}_B^k)\\
&\quad-\mu_B\sum_{i=1}^{L}p(B|\,x_i,\hat{m}_A^k,\hat{m}_B^k)\Big)\\
&=\frac{1}{L}\Big(\sum_{i=1}^{L}x_i-d_c\sum_{i=1}^{L}p(B|\,x_i,\hat{m}_A^k,\hat{m}_B^k)\Big)
\end{aligned}
$$

- MAP1: Maximum a posteriori estimation of the style parameter s. Assuming that s is distributed as described by Eq. 4, the EM update formula for MAP

[1] The prefixes ML and MAP distinguish the types of estimates of the parameters from the test data and not the classification scheme.

estimation of s is

$$\hat{s}^{k+1} = \frac{1}{L + 4\sigma^2/d_s^2}\Big(\sum_{i=1}^{L} x_i - \mu_A \sum_{i=1}^{L} p(A|\, x_i, \hat{m}_A^k, \hat{m}_B^k) - \mu_B \sum_{i=1}^{L} p(B|\, x_i, \hat{m}_A^k, \hat{m}_B^k)\Big) = \frac{\sum_{i=1}^{L} x_i - d_c \sum_{i=1}^{L} p(B|\, x_i, \hat{m}_A^k, \hat{m}_B^k)}{L + 4\sigma^2/d_s^2}$$

At each iteration of the EM algorithm, new estimates of the means of A and B are computed from the current estimate of s.[2]

Now we present the ML2 and MAP2 adaptive schemes, where we attempt to estimate the means of A and B, ignoring the constraint that they are separated by d_c. That is, the correlation between the means of A and B induced by commonality of source is disregarded. However, the MAP2 scheme assumes the prior distributions of the means to be normal with variance $d_s^2/4$, centered at $\mu_A = 0$ for A and at $\mu_B = d_c$ for B.

– ML2: The EM update formulae for the maximum-likelihood estimation of the means are

$$\hat{m}_A^{k+1} = \frac{\sum_{i=1}^{L} x_i p(A|\, x_i, \hat{m}_A^k, \hat{m}_B^k)}{\sum_{i=1}^{L} p(A|\, x_i, \hat{m}_A^k, \hat{m}_B^k)}$$

$$\hat{m}_B^{k+1} = \frac{\sum_{i=1}^{L} x_i p(B|\, x_i, \hat{m}_A^k, \hat{m}_B^k)}{\sum_{i=1}^{L} p(B|\, x_i, \hat{m}_A^k, \hat{m}_B^k)}$$

– MAP2: The EM update formulae for the maximum a posteriori estimation of the means are

$$\hat{m}_A^{k+1} = \frac{\sum_{i=1}^{L} x_i p(A|\, x_i, \hat{m}_A^k, \hat{m}_B^k) + 4\mu_A\sigma^2/ds^2}{\sum_{i=1}^{L} p(A|\, x_i, \hat{m}_A^k, \hat{m}_B^k) + 4\sigma^2/ds^2} = \frac{\sum_{i=1}^{L} x_i p(A|\, x_i, \hat{m}_A^k, \hat{m}_B^k)}{\sum_{i=1}^{L} p(A|\, x_i, \hat{m}_A^k, \hat{m}_B^k) + 4\sigma^2/ds^2}$$

$$\hat{m}_B^{k+1} = \frac{\sum_{i=1}^{L} x_i p(B|\, x_i, \hat{m}_A^k, \hat{m}_B^k) + 4\mu_B\sigma^2/ds^2}{\sum_{i=1}^{L} p(B|\, x_i, \hat{m}_A^k, \hat{m}_B^k) + 4\sigma^2/ds^2} = \frac{\sum_{i=1}^{L} x_i p(B|\, x_i, \hat{m}_A^k, \hat{m}_B^k) + 4d_c\sigma^2/ds^2}{\sum_{i=1}^{L} p(B|\, x_i, \hat{m}_A^k, \hat{m}_B^k) + 4\sigma^2/ds^2}$$

Error rates of the classifiers. We simulated random test fields of length $L = 2$ generated according to the above continuously distributed source model. For the adaptive classifiers, in order to reduce the incidence of convergence to local maxima, we use four different initializations and choose the one yielding the maximum likelihood. The four initial values for the class means used are $\{(0, d_c),\ (\bar{\mathcal{X}} - d_c/2, \bar{\mathcal{X}} + d_c/2),\ (\bar{\mathcal{X}} - d_c, \bar{\mathcal{X}}),\ (\bar{\mathcal{X}}, \bar{\mathcal{X}} + d_c)\}$, where $\bar{\mathcal{X}}$ is the sample mean of the test set (field).

[2] The EM update equations for estimation of the two means in this case can be directly obtained using Eq. 12

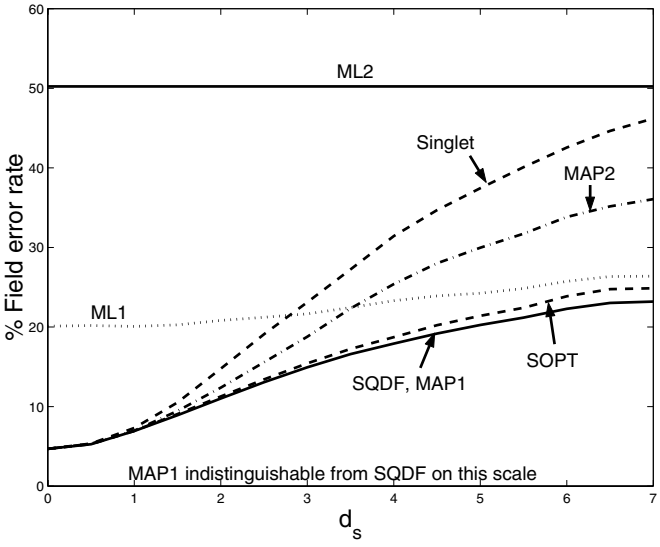


Fig. 2. Field error rates as a function of d_s – continuously distributed sources

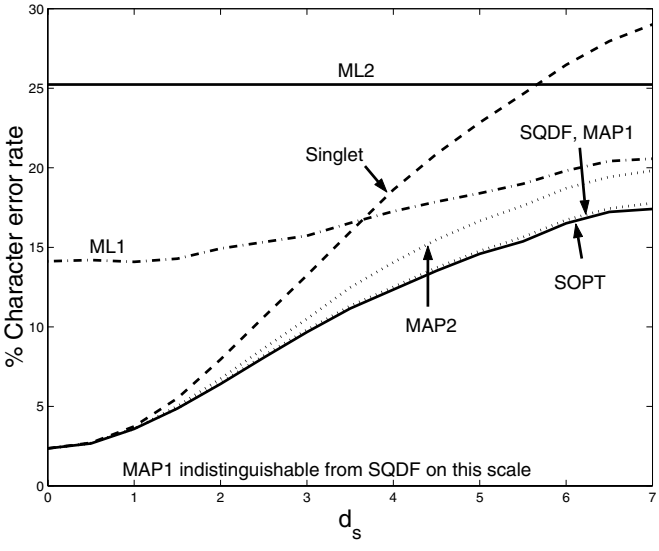
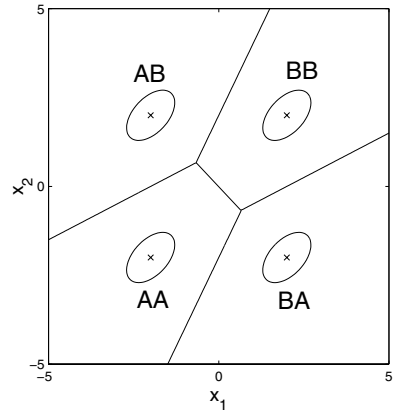


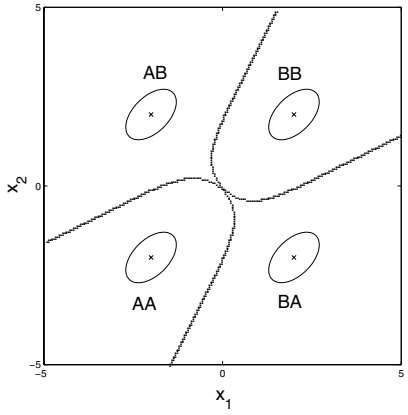
Fig. 3. Character error rates as a function of d_s – continuously distributed sources

The field error and character error rates for various field classifiers are plotted against increasing d_s with the interclass distance fixed at $d_c = 4$, in Figs. 2 and 3, respectively.

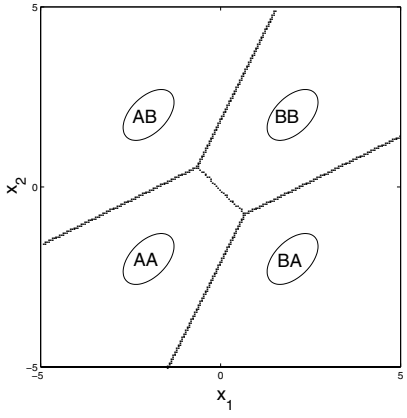
The SQDF and the SOPT classifiers are significantly better than the singlet classifier. As expected, the field-optimal SQDF classifier achieves the lowest field error rate and the singlet-optimal SOPT classifier achieves the lowest character error rate. For this example, the MAP1 classifier is almost as good as the optimal SQDF classifier. The ML2 and MAP2 adaptive classifiers have a considerably higher error rate than the ML1 and MAP1 classifiers because they disregard the fact that the two class means are maximally correlated. The reason for ML2 yielding a field error rate of approximately 50% will become evident



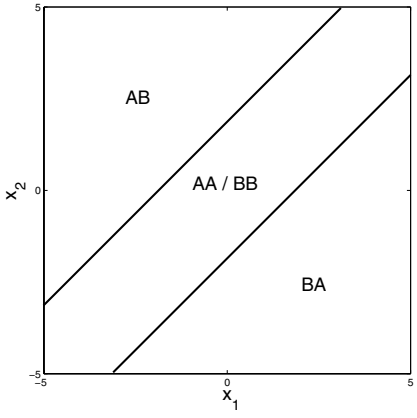
SQDF field classifier
Field error rate = 11.0
Character error rate = 6.5



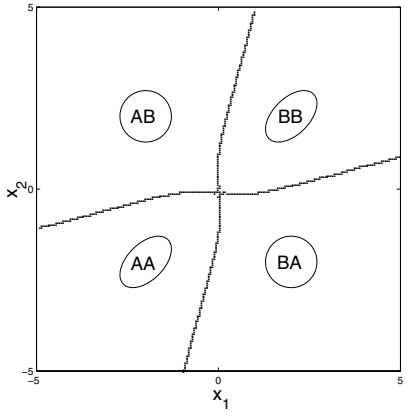
SOPT field classifier
Field error rate = 11.3
Character error rate = 6.4



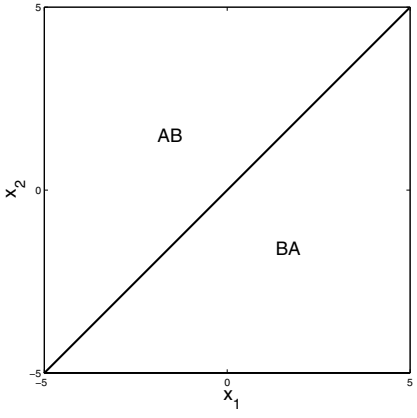
MAP1 adaptive classifier
Field error rate = 11.0



ML1 adaptive classifier
Field error rate = 33.0



MAP2 adaptive classifier
Field error rate = 12.4

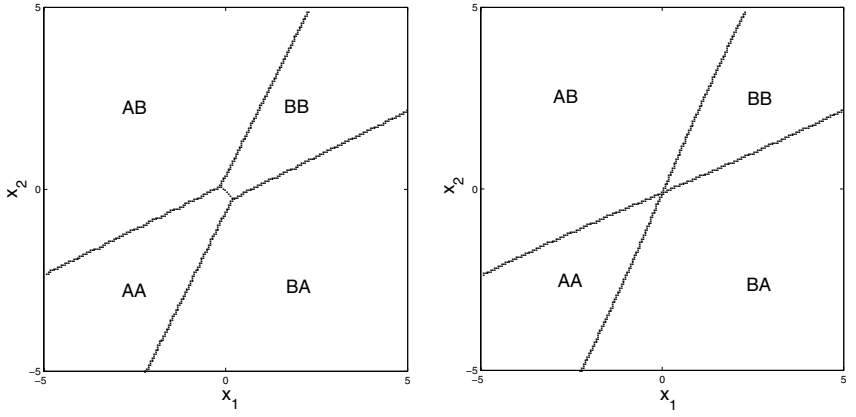


ML2 adaptive classifier
Field error rate = 50.1

Fig. 4. Continuously distributed styles – field classification boundaries for different classifiers, with $d_s = 2$, $d_c = 4$. In all subfigures, the regions of the feature space classified into the four field classes are shown. The x marks in the SQDF and SOPT subfigures indicate the field-class means. The *circles* and *ellipses* indicate the equiprobability contours of the assumed field-class distributions. There are no such contours for the ML classifiers because they are blind to the prior distribution of the styles

when we study the decision boundaries in the field-feature space.

Table 1 shows the character error rates obtained by the style-conscious classifiers operating on fields of increasing length. The SQDF classifier has a slightly higher character error rate than that of the SOPT and MAP1 classifiers because the SQDF classifier is designed to optimize the field error rate. As expected from the EM update formulae, the ML1 classifier approaches the MAP1 classifier as the field length increases, although it is worse than the MAP1 classifier on short fields. We observe that with increasing L the adaptive classifiers approach the least achievable character error rate for this problem (which is the intrasource error rate $= 100 \times Q(\frac{d_c}{2\sigma}) = 2.3\%$,



SQDF field classifier MAP1 field classifier

Fig. 5. Continuously distributed styles – field classification boundaries for SQDF and MAP1 classifiers, with $d_s = 2$, $d_c = 1$

Table 1. Character error rates in percentage for various classifiers with increasing field length L for $d_c = 4$, $d_s = 2$

L	SQDF	SOPT	ML1	MAP1	ML2	MAP2
2	6.5	6.3	14.8	6.5	25.2	6.6
4	4.7	4.6	10.4	4.6	10.2	5.8
6	3.9	3.8	5.7	3.8	6.2	4.3
10	–	–	3.4	3.1	3.9	3.4
20	–	–	2.6	2.6	2.8	2.7
100	–	–	2.3	2.3	2.3	2.3

where $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})\ dx$). We expect that the SOPT classifier would perform likewise, but the required computation is excessive for $L > 6$.

Decision boundaries. We plotted the decision boundaries in the field-feature space for the above-described classifiers for $L = 2$. The decision boundaries highlight the distinctions and similarities between the various classifiers.

Figure 4 shows the decision boundaries for various classifiers for $d_s = 2$ and $d_c = 4$. The SOPT field classifier has slightly different boundaries from the SQDF classifier. Note that the confusions between "AA" and "BB" are reduced by the SOPT classifier, consequently reducing the character error rate at the expense of increasing the field error rate.

For these values of $d_c = 4$ and $d_s = 2$, the MAP1 classifier is almost identical to the SQDF classifier. Figure 5 shows the decision boundaries for the SQDF and MAP1 classifiers for $d_c = 1$ and $d_s = 2$, where the difference between the two classifiers is more apparent.

The ML1 classifier classifies every pair of patterns sufficiently close to each other to be from the same class because it operates under the assumption that the mean of A and B for a particular style must be d_c apart. However, since the likelihood of the samples being from either of the singlet classes is equal, the banded region in the middle has the label AA/BB. When the patterns are sufficiently separated, the pattern with the lower value is labeled as an A and the pattern with the higher value is called a B.

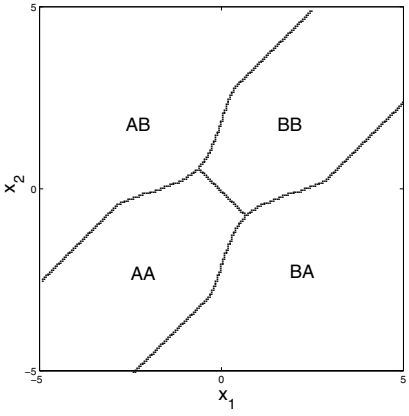


Fig. 6. ML1 decision boundaries for $d_c = 4$, $d_s = 2$ with only one initialization

The ML2 adaptive classifier attempts to fit two normal distributions to the two patterns in the field because the correlation between the means A and B is disregarded. Thus unless the two patterns coincide, it is assumed that they are from different classes. Also, the likelihood obtained by labeling the two modes as AB and BA is equal. However, in Fig. 4, the region $x_1 > x_2$ is labeled as BA and the region $x_1 < x_2$ is labeled as AB. This is an artifact of our initialization of the EM algorithm (all four initializations for mean of A are lower than the corresponding initialization for the mean of B). Thus the field error rate obtained is at least 50% (all AAs and BBs are misclassified). In addition, there are some confusions between ABs and BAs.

Although the MAP2 adaptive classifier assumes that the mean of A and the mean of B are drawn independently from a given source, the classification boundaries are not parallel to the coordinate axes owing to the correlation between patterns from the same class. The assumed field-feature probability contours are shown superimposed on the MAP2 decision boundaries in Fig. 4. That is, the MAP2 adaptive classifier uses only the intrasource same-class consistency but no source-induced correlation between classes.

We have observed that the accuracy of all four adaptive classifiers depends critically on the initialization scheme. For this simple simulation example we solved the problem by using multiple starts. However, it is unclear how this can be done for high-dimensional data. Also, for ML1 and ML2 adaptive classifiers incorrect labeling of mixture components can result from initialization methods designed to obtain parameter estimates that yield the maximum likelihood. That is, local optima close to the parameters obtained from the training data may be more desirable. For example, Fig. 6 shows the ML1 decision boundaries when the EM algorithm is initialized only with $\hat{s}^0 = 0$ (cf. Fig. 4 where four different initializations were used).

4 Experimental results

4.1 Description of the data

To test our algorithms on realistic data, we experimented with the databases SD3 and SD7, which are contained in the NIST Special Database SD19 [9]. The database contains samples of handwritten numerals labeled by writer and class. SD3 was the training data released for the First Census OCR Systems Conference and SD7 was the test data. SD3 and SD7 were obtained from different populations, and SD7 is considered to be much more difficult to recognize. There are approximately ten samples per class per writer.

We constructed four datasets, two from each of SD3 and SD7, as shown in Table 2. The writers in the Train and Test sets are disjoint, which allows us to verify our hypothesis that broad styles can be gleaned from a sufficiently large sample of writers.

Since we compute the field class-conditional covariance matrices from source-specific class-conditional matrices, we require that each writer have at least two samples for each singlet class. We therefore deleted all writers not satisfying this criterion from the training sets. In Table 2, the numbers in parentheses indicate the total number of writers from each set that remain after the deletion.

Figure 7 shows the scatterplot of the top two principal component features of the writer-specific class means of the writers in the training sets. We show only some of the classes for legibility. The writer-specific class means seem to vary in a continuous fashion.

We extracted 100 blurred directional (chain-code) features from each sample [12]. The samples of each writer in the test sets were randomly permuted, and L patterns were chosen at a time to simulate fields of length L.

In order to alleviate the adverse effect of finite sample estimation on accuracy, we smoothed the class-conditional covariance matrices by the *regularized discriminant function* (RDF) method [7]. Regularization reduces the variance of the estimator at the cost of increasing its bias. In RDF, the regularized (smoothed) covariance matrix of a class is an interpolation of the sample covariance matrix and the identity matrix

$$\hat{\mathrm{C}}_i = (1-\gamma)\mathrm{C}_i + \gamma\sigma_i^2\mathrm{I} \tag{13}$$

where $\sigma_i^2 = tr(\mathrm{C}_i)/d$ and $0 < \gamma < 1$. For our experiments we chose a value of $\gamma = 0.2$.

Table 2. Handwritten numeral datasets

	Writers	Number of samples
SD3-Train	0-399 (395)	42698
SD7-Train	2100-2199 (99)	11495
SD3-Test	400-799 (399)	42821
SD7-Test	2200-2299 (100)	11660

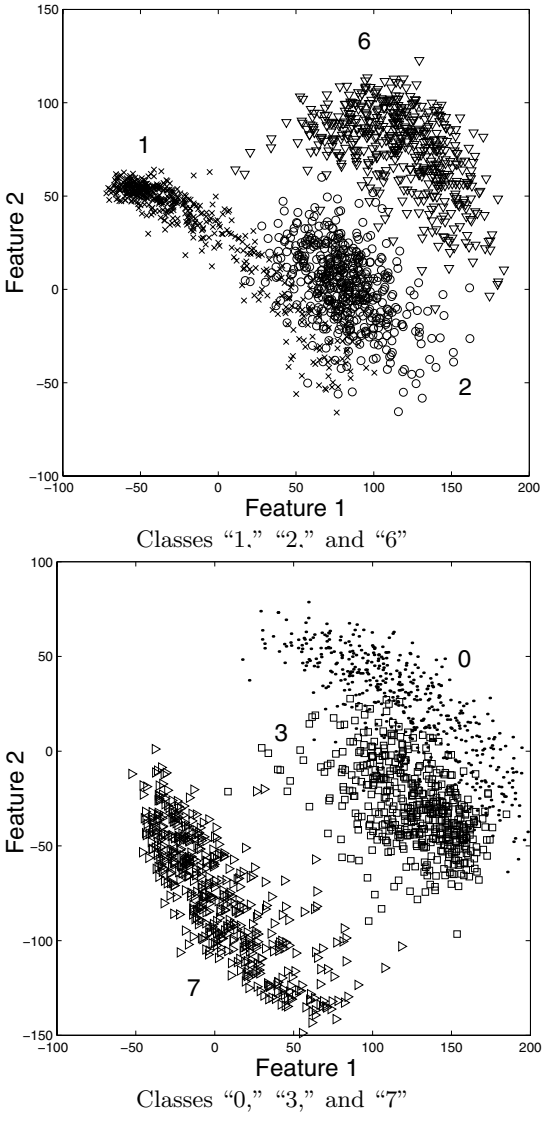


Fig. 7. Scatterplot of the top two principal component features of writer-specific class means

4.2 Results and discussion

On the handwritten data we present results on adapting the class-conditional covariance matrices as well as the means. Starting with the statistics of the training set, we adapt to the statistics of the test set (modeled as a Gaussian mixture with one Gaussian per class) via the EM algorithm. We will first test the ML adaptive classifier and then the MAP adaptive classifier.

ML adaptation. The maximum-likelihood estimation of covariance matrices for Gaussian mixtures with the EM algorithm with insufficient samples is beset with problems such as convergence to the boundary of the parameter space. That is, when the class-covariances are unconstrained, the EM algorithm sometimes fits the samples with a mixture containing zero-variance Gaussians. We avoid this phenomenon by constraining the class-conditional covariance matrices to be equal, for every class for a particular writer, during adaptation. Even so, the number of samples available per writer in our dataset (approximately ten samples per class) is insufficient to estimate the full 100×100 covariance matrix for 100 features.

We solve the problem of covariance adaptation as follows. We first estimate the class-conditional covariance matrices C_i, $i = 1, \ldots, N$ from the training set. Let D_i be the matrix obtained by setting all the off-diagonal terms in C_i equal to 0. We compute the class-conditional correlation matrices $\mathrm{R}_i = \mathrm{D}_i^{-1/2}\mathrm{C}_i\mathrm{D}_i^{-1/2}$. Now we find the average correlation matrix $\mathrm{R}_{avg} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{R}_i$. Note that the feature variances in C_i include the variance of means across writers.

At the first iteration of the EM algorithm we use the full class-conditional covariance matrices C_i to evaluate the posterior class probabilities of the test samples. At every succeeding iteration of the EM algorithm we update only the class-conditional feature variances for all features under the assumption that, for a particular feature, the class-conditional feature variance is equal for all classes. That is, for d-dimensional features, we estimate only d variances irrespective of the number of classes present.

Let $\hat{\mathrm{D}}^k$ be the $d \times d$ matrix with the principal diagonal set to the feature variances estimated at the k^{th} iteration of the EM algorithm and all other elements set to 0. The class-conditional covariance matrices at the k^{th} iteration are updated according to $\hat{\mathrm{C}}_i^k = \hat{\mathrm{D}}^k\mathrm{R}_{avg}\hat{\mathrm{D}}^k$. The estimates $\hat{\mathrm{C}}_i^k$ are used in the next iteration to compute the posterior probabilities of each class for each sample. We classify the test samples after every EM iteration and terminate when no classification decision is changed.

In other words, we assume that the feature-to-feature correlation in the test set is the same as in the training set and is independent of the class, but the individual feature variances, averaged over all classes, are estimated from the test set. The algorithm for ML mean and covariance adaptation is given in Fig. 8.

Tables 3 and 4 present the character error rates for various combinations of training and test sets without any adaptation, as well as adapting only the means and adapting both means and covariance matrices. For the experiments in Table 3 the covariance matrices obtained from the training set were not regularized, whereas they were regularized for the experimental results in Table 4. We do not report running times because all the classifiers were implemented in MATLAB.

We observe a reduction in the character error rate even when only the class-conditional means are adapted to those of the test set. This improvement is increased dramatically when the covariance matrices are also adapted (the total reduction in error rate is more than 50% in many cases). The final error rates after adaptation with the regularized covariance matrices are lower than for the unregularized covariance matrices, indicating the dependence of the adaptation accuracy on the training method. However, the benefit of covariance adaptation is more significant when the matrices are not regularized.

For each of the above experiments, Table 5 shows the percentage of writers in the test set with increased and decreased accuracy with adaptive classification (both class-conditional means and covariance matrices were adapted). The maximum decrease in accuracy for any writer was never higher than 2.5% of the total number of samples from the writer, while the maximum increase in accuracy was around 25%. Table 5 also lists the number of iterations before the stopping criterion was met, averaged over all the writers in each test set. We observe that the adaptive algorithm converges rapidly, requiring approximately two iterations on average. These results indicate that even when the test fields are not very large (here we have only ten samples/class), ML adaptation can be performed to improve accuracy.

MAP adaptation. According to the model of normally distributed writer-specific class means, we constructed a maximum a posteriori adaptive classifier (Sect. 3.2). The means and covariance matrices of the writer-specific means were estimated from the training data. However, we could not experiment with all 100 features because we had fewer than 500 writers in the training data (the covariance matrix of the vector writer-specific means would be singular if all 100 features were used). Therefore, we conducted experiments with only the top 25 principal component features. In Table 6, we compare the accuracy of the MAP adaptive classifier with that of the ML adaptive scheme for increasing field length. The last row shows the results when all the samples from each writer in the test set are considered as one field (i.e., approximately 100 patterns per field).

We observe that the MAP adaptive classifier is more accurate than the singlet classifier even on fields as short as ten patterns. The ML adaptive classifier performs poorly when the fields are short but improves with increasing field length. On SD7-Test we observe that ML adaptation performs better than the MAP adaptive scheme when each writer constitutes a separate field. We attribute this to the estimation error in the writer-specific

Table 3. Character error rates on handwritten data before and after adaptation with all 100 features. Training covariance matrices not regularized

		Character error rate (%)		
Training set	Test set	Before adaptation	After mean adaptation	After mean & cov adapt
SD3-Train	SD3-Test	2.2	1.9	0.8
	SD7-Test	8.0	6.6	3.0
SD7-Train	SD3-Test	3.7	2.7	1.1
	SD7-Test	3.6	3.1	1.8
SD3-Train	SD3-Test	1.7	1.5	0.6
+SD7-Train	SD7-Test	4.7	3.8	1.9

Table 4. Character error rates on handwritten data before and after adaptation with all 100 features. Training covariance matrices regularized ($\gamma = 0.2$)

		Character error rate (%)		
Training set	Test set	Before adaptation	After mean adaptation	After mean & cov adapt
SD3-Train	SD3-Test	1.1	0.7	0.6
	SD7-Test	5.0	2.6	2.2
SD7-Train	SD3-Test	1.7	0.9	0.8
	SD7-Test	2.4	1.6	1.7
SD3-Train	SD3-Test	0.9	0.6	0.6
+SD7-Train	SD7-Test	3.2	1.9	1.8

Table 5. Percentage of writers in each experiment that improved or worsened after mean and covariance adaptation. The number of EM iterations, averaged over all writers for each test set, before the stopping criterion was met is also shown

		Without regularization			With regularization		
		% of Writers		Avg	% of Writers		Avg
Training set	Test set	improved	worsened	iter	improved	worsened	iter
SD3-Train	SD3-Test	59.1	5.0	1.8	23.8	5.8	1.4
	SD7-Test	83.0	0.0	3.1	68.0	3.0	2.4
SD7-Train	SD3-Test	79.2	1.8	2.2	47.6	4.8	1.7
	SD7-Test	64.0	9.0	2.3	38.0	11.0	1.9
SD3-Train	SD3-Test	52.6	4.8	1.7	22.6	6.0	1.4
+SD7-Train	SD7-Test	70.0	4.0	2.5	50.0	6.0	2.0

covariance matrix (which is a covariance matrix for a 250-dimensional vector estimated from 500 samples).

5 Conclusions

By modeling the test fields as being drawn from a normal mixture with unknown parameters, we developed adaptive methodologies to cope with nonrepresentative training data. We suggested that multiple sources in the training data can give rise to such nonrepresentative training sets.

We proposed a Gaussian model for situations with a continuous style variation. The model identifies a style by its class means, which are assumed to be Gaussian distributed. Furthermore, intrastyle consistency induces interclass correlation between the class means, which can be estimated from training data. Under such a model we derived ML and MAP adaptive classifiers that exploit style consistency in test fields.

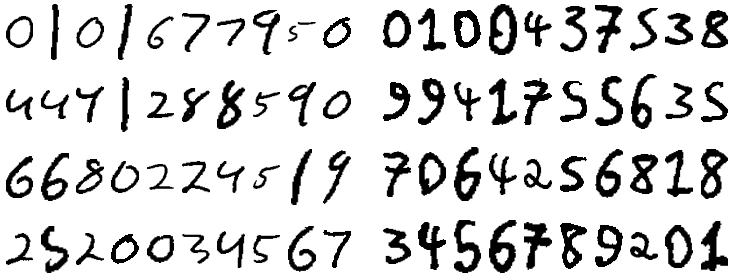
Extensive simulations showed that the computationally less expensive MAP adaptive classifier is a good approximation to the optimal SQDF classifier even on short fields, when the styles are Gaussian distributed and interclass separation is large. For long fields the even more economical ML adaptive classifier achieves accuracy comparable to that of the MAP adaptive classifier. The simulations also demonstrated that as the field length increases the accuracy of all the style-conscious classifiers approaches the average intrastyle accuracy.

On the NIST handwritten data, we experimented with mean and covariance adaptive schemes with all the samples from a single test writer considered as one field. The ML adaptive classifiers that operate on one test writer at a time (≈ 100 samples in the test field) reduce

function Training( )
  For all classes $i = 1, \ldots, N$, compute class-conditional means μ_i and
    covariance matrices C_i from training set.
  Compute D_i by setting every nondiagonal term in C_i to 0, $i = 1, \ldots, N$.
  Compute $\mathrm{R}_i = \mathrm{D}_i^{-1/2}\mathrm{C}_i\mathrm{D}_i^{-1/2}$, $i = 1, \ldots, N$.
  Compute $\mathrm{R}_{avg} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{R}_i$.

function Adaptive Classification (Test samples from one writer $\mathcal{X}$, Max Iterations)
  $\hat{\mu}_i^0 = \mu_i$ and $\hat{\mathrm{C}}_i^0 = \mathrm{C}_i$, for $i = 1, \ldots, N$.
  Classify Samples($\mathcal{X}$ with $\hat{\mu}_i^0, \hat{\mathrm{C}}_i^0$).
  for k=1 to Max Iterations
    $(\hat{\mu}_i^k, \hat{\mathrm{D}}^k)$ = EM Estimation($\mathcal{X}, \hat{\mu}_i^{k-1}, \hat{\mathrm{C}}_i^k$)
      Here estimate the N class-conditional means
        and only one diagonal covariance matrix $\hat{\mathrm{D}}^k$.
    $\hat{\mathrm{C}}_i^k = \hat{\mathrm{D}}^k\mathrm{R}_{avg}\hat{\mathrm{D}}^k$ for $i = 1, \ldots, N$.
    Classify Samples($\mathcal{X}$ with $\hat{\mu}_i^k, \hat{\mathrm{C}}_i^k$).
    If no classification result for any sample in $\mathcal{X}$
      is changed from previous result, then terminate for.
  end for

Fig. 8. Pseudocode for the ML mean and covariance matrix adaptation algorithm.



Writer worsened (2 more errors than before adaptation)

Writer improved (26 fewer errors than before adaptation)

Fig. 9. Some samples from writers on whom error rate improved and worsened after adaptation

Table 6. Character error rates in % for the MAP and ML adaptive classifiers with increasing field length. The training set is SD3-Train+SD7-Train. Only the top 25 pca features were used, without covariance regularization. For comparison, the character error rate obtained by the singlet classifier with 25 features on SD3-Test is 1.5% and on SD7-Test 4.4%. The last row shows the error rates when all the samples from one writer comprise the same field

Test set ⇒	SD3-Test		SD7-Test	
Field length L	MAP	ML	MAP	ML
10	1.5	5.1	4.2	7.5
20	1.4	2.2	4.1	5.1
40	1.2	1.7	3.6	4.5
60	1.2	1.5	3.2	4.2
Writer	1.0	1.0	3.0	2.7

the error rate by more than 50%. We also implemented the MAP adaptive classifier for class mean adaptation and demonstrated that adaptive classification increases accuracy even on relatively short fields ($\approx$ 10 samples).

An important problem in the area of adaptive classification is to characterize situations when the adaptive classifier degrades accuracy. For the approaches we presented, this problem is related to the separation of styles relative to the separation between classes as well as the convergence properties of the EM algorithm. Although we currently initialize the EM algorithm to the parameters estimated from the training set, we intend to explore other initialization methods.

Acknowledgements. We would like to thank Dr. Hiromichi Fujisawa, Dr. Cheng-Lin Liu, and Dr. Prateek Sarkar for their valuable suggestions and comments. We are also grateful to anonymous reviewers for their suggestions.

References

1. Baird HS, Nagy G (1994) A self-correcting 100-font classifier. In: Vincent L, Pavlidis T (eds) Document recognition. Proc SPIE 2181:106–115
2. Casey RG, Nagy G (1968) An autonomous reading machine. IEEE Trans Comput C-17(5):492–503

3. Castelli V, Cover TM (1996) The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. IEEE Trans Informat Theory 42:2102–2117
4. Chiavaccini E, Vitetta GM (2001) MAP symbol estimation on frequency-flat Rayleigh fading channels via a Bayesian EM algorithm. IEEE Trans Commun 49(11):1869–1872
5. Dempster AP, Laird MM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39(1):1–38
6. Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York
7. Friedman H (1989) Regularized discriminant analysis. J Am Stat Assoc 84(405):166–175
8. Gauvain JL, Lee CH (1994) Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans Speech Audio Process 2(2):291–298
9. Grother P (1995) Handprinted forms and character database, NIST special database 19. Technical report and CDROM.
10. Ho TK, Nagy G OCR with no shape training. In: Proceedings of the 15th international conference on pattern recognition, Barcelona, September 2000, pp 27–30
11. Leggetter CJ, Woodland PC (1995) Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput Speech Lang 9(2):171–185
12. Liu CL, Sako H, Fujisawa H (2002) Performance evaluation of pattern classifiers for handwritten character recognition. IJDAR 4(3):191–204
13. Lucky RW (1966) Techniques for adaptive equalization of digital communication systems. Bell Sys Tech J 45:255–286
14. Mathis C, Breuel T (2002) Classification using a hierarchical Bayesian approach. In: Proceedings of the 16th international conference on pattern recognition, QUEBEC CITY, AUGUST 2002, 4:103–106
15. Nagy G, Shelton Jr GL (1966) Self-corrective character recognition system. IEEE Trans Informat Theory IT-12(2):215–222
16. Nagy G, Seth S, Einspahr K (1987) Decoding substitution ciphers by means of word matching with application to OCR. IEEE Trans Patt Analysis Mach Intell 9(5):710–715
17. Redner RA, Walker HF (1984) Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev 26(2):195–239
18. Selfridge OG, Neisser U (1960) Pattern recognition by machine. Sci Am 203:60–68
19. Shashahani BM, Landgrebe DA (1994) The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. IEEE Trans Geosci Remote Sens 32(5):1087–1095
20. Veeramachaneni S (2002) Style constrained quadratic field classifiers. PhD thesis, Rensselaer Polytechnic Institute, Troy, NY

George Nagy received his B. Eng. and M. Eng. from McGill University and his Ph.D. in electrical engineering from Cornell University in 1962 (on neural networks). For the next 10 years he studied pattern recognition at the IBM T.J. Watson Research Center in Yorktown Heights, NY. From 1972 to 1985 he was professor of computer science at the University of Nebraska, Lincoln (9 years as chair) and worked on geographic information systems, remote sensing applications, and human-computer interfaces. Since 1985 he has been professor of computer engineering at Rensselaer Polytechnic Institute, where he established ECSE DocLab. In addition to document image analysis, OCR, geographic information systems, and computational geometry, his students have engaged in solid modeling, finite-precision spatial computation, and interactive computer vision, often with a focus on systems that improve with use. He has benefited from visiting appointments at the Stanford Research Institute, Cornell, the University of Montreal, the National Scientific Research Institute of Quebec, the University of Genoa and the Italian National Research Council in Naples and Genoa, AT&T and Lucent Bell Laboratories, IBM Almaden, McGill University, the Institute for Information Science Research at the University of Nevada, and the Center for Image Analysis in Uppsala.

Sriharsha Veeramachaneni received his Ph.D. in computer engineering from the Rensselaer Polytechnic Institute, New York in 2002. He is currently a postdoctoral researcher at the Istituto per la Ricerca Scientifica e Tecnologica, Trento, Italy. His research interests include statistical pattern classification, machine learning, and information theory.