RACSAM
Rev. R. Acad. Cien. Serie A. Mat.
VOL. **96** (2), 2002, pp. 139–156
Matemática Aplicada / Applied Mathematics

On an optimization problem arising from probability density estimation

Sankar Basu, Mohammad Saif Ullah Khan, C. A. Micchelli and Peder A. Olsen

Abstract. We consider a class of optimization problems arising from statistical density estimation of high dimensional data from projections on lower dimensional subspaces. Two criteria are used for optimal model selection, namely, maximum entropy and maximum likelihood estimation. In each case, our approach requires *univariate* density estimators and in this regard we explore the use of mixture models of gaussian densities as well as Parzen estimators, for the projected data. An expectation maximization strategy is used to update means and covariances as described in Dempster et al. [7]. However, the computation of best directions leads to a challenging class of nonlinear optimization problems which is the focus of our study here. Special cases of this optimization problem are studied analytically and an algorithm to numerically solve the general case is proposed. We provide numerical evidence, on data coming from speech recognition as well as on synthetically generated data, that validates the efficacy of the proposed method.

Sobre un problema de optimización surgido en estimación de la densidad de probabilidad

Resumen. Consideramos una clase de problemas de optimización que surgen en estimaciones de la densidad de datos en dimensión elevada a partir de proyecciones en subespacios de dimensión más baja. Los criterios que se usan para la selección óptima del modelo son máxima entropía y máxima verosimilitud.En cada caso nuestro planteamiento requiere estimadores de la densidad univariados y a este respecto exploramos el uso de modelos mezcla de densidades gaussianas y de estimadores de Parzen para los datos proyectados. Se usa una estrategia de maximización de la esperanza para actualizar medias y covarianzas como en Dempster et al. [7]. Sin embargo el cálculo de las direcciones óptimas conduce a interesantes problemas de optimización no lineal que son el núcleo del presente trabajo. Se estudian analíticamente algunos casos particulares de este problema de optimización y se propone un algoritmo para resolver numéricamente el caso general. Se presenta evidencia numérica, sobre datos procedentes de reconocimiento del lenguaje y sobre datos generados sintéticamente, que avala la eficacia del método propuesto.

1. Introduction

Our motivation comes from the problem of classifying high dimensional feature vectors arising in a number of statistical machine learning problems in the domains of speech, image and video processing. The

Presentado por Mariano Gasca.
Recibido: 6 de Junio de 2002. *Aceptado*: 13 de Septiembre de 2002.
Palabras clave / Keywords: Optimization, Density Estimation, Pattern Recognition
Mathematics Subject Classifications: 46N30,78M50

problems of automatic machine recognition of speech, or a specific object in an image, or an event in a video sequence often involve estimation of probability density of feature vectors in the feature space. The dimension of these feature spaces can be a hundred or more. We attempt to deal with the problem resulting from this curse of dimensionality in estimation of probability densities, which is further aggravated in some of the above situations by paucity of available data.

In more concrete terms, we are given vectors $x^1, x^2, \ldots, x^N \in \mathbb{R}^d$ where d is large. From this data we must compute a probability density function which represents a random variable from which the samples $x^1, x^2, \ldots, x^N$ are drawn. The approach taken here is based on the fact that there are *many* methods available for estimating *univariate* probability densities. Thus, we propose to project the multivariate data on chosen directions, reconstruct the univariate probability density from this projected data for each choice of directions and then use these univariate probability densities to reconstruct the unknown multivariate density by a maximum entropy or maximum likelihood estimation criterion. We only treat here the case of this important problem when the number of directions is the *same* as the dimension of the samples. In this case, we can identify explicitly the multivariate density which maximizes entropy subject to d marginal constraints and can then address both analytically and computationally the task of finding the optimal directions. We choose directions by either further increasing entropy or alternatively, the likelihood. In this regard, the essential computational problem is a class of numerically challenging nonlinear optimization problems for which we provide algorithms to solve numerically. Examples are provided to demonstrate the efficacy of our method. In this section, we shall review some of the observations made in [5] and [3].

2. Maximum entropy from marginals

In this section we briefly review the method proposed in [5], see also [3] for high dimensional density estimation. Let (column) vectors, $y^1, y^2, \ldots, y^d \in \mathbb{R}^d$ be given and denote by Y the $d \times d$ matrix whose columns are these vectors. We are given *univariate* probability density functions, $p_1, p_2, \ldots, p_d$ and consider all multivariate probability density functions $P : \mathbb{R}^d \to \mathbb{R}$ such that for all continuous functions f of compact support on $\mathbb{R}$ there holds the equations

$$\int_{\mathbb{R}^d} f(\langle y^i, x\rangle) P(x) dx = \int_{\mathbb{R}} f(t) p_i(t) dt, \qquad i = 1, 2, \ldots, d. \tag{1}$$

We denote the class of all densities P satisfying the above equations by $\mathcal{C}(p)$. Recall that the entropy of P is given by

$$H(P) := -\int_{\mathbb{R}^d} P(x) \log P(x) dx.$$

Theorem 1 *Given any probability density functions $p_1, p_2, \ldots, p_d$ on $\mathbb{R}$ such that for $i = 1, 2, \ldots, d$, $p_i \log p_i \in L(\mathbb{R})$ we have that*

$$\max\{H(P) \ : \ P \in \mathcal{C}(p)\} = -\frac{1}{2} \log \det G(Y) + \sum_{i=1}^{d} H(p_i) \ . \tag{2}$$

where $G := Y^T Y$ is the Gram matrix

$$G(Y) = \left[\langle y^i, y^j\rangle : i, j = 1, 2, \ldots, d\right].$$

Moreover,

$$P^*(x) = \sqrt{\det G(Y)} \prod_{i=1}^{d} p_i\left(\langle y^i, x\rangle\right), \qquad x \in \mathbb{R}^d, \tag{3}$$

is the unique probability density function in $\mathcal{C}(p)$ at which the maximum in (2) is achieved. □

The proof of this result appears in [5] and it can be used to isolate desirable features of multivariate data $x^1, x^2, \ldots, x^N$ in $\mathbb{R}^d$ in the following manner. Suppose a *univariate* family $p(\cdot; x)$, $x \in \mathbb{R}^N$ of probability densities on $\mathbb{R}$, which provide a good estimator for a random variable from which the components of x are drawn, is specified. For a given $y \in \mathbb{R}^d$, we form the vector $X^T y \in \mathbb{R}^N$ where X is the $d \times N$ matrix whose columns are the multivariate data $x^1, x^2, \ldots, x^N$ and use $p(\cdot; X^T y)$ as an estimate for the marginal probability density function of P in the direction y. Given d directions $y^1, y^2, \ldots, y^d$, Theorem 1 states that the maximum entropy estimator is

$$P^*(x) = \sqrt{\det G(Y)} \prod_{j=1}^{d} p\left(\langle y^j, x\rangle; X^T y^j\right), \qquad x \in \mathbb{R}^d, \tag{4}$$

and

$$\mathcal{J}(Y) := H(P^*) = -\frac{1}{2} \log \det G(Y) + \sum_{j=1}^{d} \mathcal{H}(y^j), \tag{5}$$

where we define

$$\mathcal{H}(y) := H\left(p(\cdot; X^T y)\right), \ y \in \mathbb{R}^d.$$

We desire to make the entropy (5) as large as possible by varying our choice of vectors over an *orthonormal basis*. In this case, (5) becomes

$$\mathcal{J}(Y) = \sum_{j=1}^{d} \mathcal{H}(y^j).$$

where Y is an orthogonal matrix.

This philosophy is similar to projectional pursuit where even more than d dimensions can be considered [8]. Nonparametric versions of same ideas appear in [14, 17].

Gaussian model for projections: We recall the special case of this problem connsidered in [5]. Let $x = (x_1, x_2, \ldots, x_N)^T \in \mathbb{R}^N$ be the vector of univariate[1] data obtained by considering the projections of the multivariate data $x^1, x^2, \ldots, x^N$. The univariate data are modeled by the gaussian density $N(\cdot; \mu(x), \sigma(x))$, where

$$N(t; \mu, \sigma) := \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}, \ t \in \mathbb{R},$$

with

$$\mu(x) := \frac{1}{N} \sum_{j=1}^{N} x_j \tag{6}$$

and

$$\sigma^2(x) := \frac{1}{N} \sum_{j=1}^{N} (x_j - \mu(x))^2 . \tag{7}$$

The function to be optimized in this case is

$$\mathcal{J}_N(Y) = -\frac{1}{2} \log \det N(Y) + \frac{d}{2} \log(2\pi e) + \frac{1}{2} \sum_{j=1}^{d} \log \left\langle y^j, V y^j \right\rangle, \tag{8}$$

where V is the $d \times d$ matrix $V = UU^T$ and U is the $d \times N$ matrix, whose columns are the vectors

$$u^j = \frac{1}{\sqrt{N}} \left(x^j - \bar{x}\right), \qquad j = 1, 2, \ldots, N,$$

[1] Superscripted variables denote data in d dimensional space, whereas subscripted variables denote data projected in one-dimension

where

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x^i.$$

Let $\lambda_1, \lambda_2, \ldots, \lambda_d$ be the eigenvalues of the matrix V. It was shown in [5] that

$$\min\{\mathcal{J}_N(Y) \; : \; Y \text{ is orthogonal}\} = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\sum_{j=1}^{d}\log\lambda_j.$$

where equality occurs above if and only if

$$\langle y^i, Vy^j\rangle = 0, \qquad i \neq j,\; i,j = 1,2,\ldots,d. \tag{9}$$

which, in turn, is satisfied when y^is are orthonormal eigenvectors of V. Moreover, we have that

$$\max\{\mathcal{J}_N(Y) \; : \; Y \text{ is orthogonal}\} = \frac{d}{2}\log\frac{2\pi e}{d} + \frac{d}{2}\log\operatorname{trace} V.$$

where equality occurs if and only if for all $j = 1,2,\ldots,d$,

$$\langle y^j, Vy^j\rangle = \frac{1}{d}\sum_{i=1}^{d}\lambda_i. \tag{10}$$

Gaussian mixture model for projections: In practice it is not desirable to estimate univariate densities by a *single* gaussian. As in [5] we consider the possibility of modeling univariate density by using a *mixture* model of m gaussians which has the form

$$p\,(t\,;\,\mu,\omega,\sigma) = \sum_{j=1}^{m}\omega_j\, N\,(t;\mu_j,\sigma_j) \qquad t \in \mathbb{R}, \tag{11}$$

where the mixture weights satisfy the constraints

$$\sum_{j=1}^{m}\omega_j = 1, \quad \omega_i \geq 0, \qquad i = 1,2,\ldots,m. \tag{12}$$

For any choice of parameters μ, ω and σ, we recommend computing the entropy of the probability density in (11) by a Monte Carlo method using random samples drawn from this density.

To choose the parameters of the probability density in (11) as a function of the data vector $(x_1, x_2, \ldots, x_N)^T \in \mathbb{R}^N$ we use a *fixed* number of iterations of the update formulas for the EM algorithm, see Dempster et. al. [7]

$$\hat{\omega_j} = \frac{1}{N}\sum_{i=1}^{N}P_{ij}, \qquad j = 1,2,\ldots,m$$

$$\hat{\mu_j} = \frac{\sum_{i=1}^{N}x_iP_{ij}}{\sum_{i=1}^{N}P_{ij}}, \qquad j = 1,2,\ldots,m$$

$$\hat{\sigma_j}^2 = \frac{\sum_{i=1}^{N}P_{ij}(x_i-\hat{\mu_j})^2}{\sum_{i=1}^{N}P_{ij}}, \qquad j = 1,2,\ldots,m,$$

where for $j = 1,2,\ldots,m$, $i = 1,2,\ldots,N$ we define

$$P_{ij} = \frac{\omega_j N\,(x_i\,;\,\mu_j,\sigma_j)}{p\,(x_i\,;\,\mu,\omega,\sigma)}.$$

These formulas enhance likelihood and set the values ω, μ and σ as a function of the vector x. We denote by $p(\cdot;x)$ the resulting univariate estimate for the data vector $x \in \mathbb{R}^N$ by mixture models of gaussians.

We iteratively update the orthogonal matrix Y by using *planar rotations*. In other words, we choose *any* two consecutive vectors y^j, y^{j+1} and form the new vectors

$$\begin{aligned} y^\ell(\theta) &= y^\ell, \quad \ell \neq j, j+1, \\ y^j(\theta) &= (\cos\theta)\ y^j - (\sin\theta)\ y^{j+1} \\ y^{j+1}(\theta) &= (\sin\theta)\ y^j + (\cos\theta)\ y^{j+1}. \end{aligned}$$

We choose $\hat{\theta} \in [0, 2\pi]$ to the maximize the univariate function

$$F(\theta) := \mathcal{H}(y^j(\theta)) + \mathcal{H}(y^{j+1}(\theta))$$

and replace the vectors $y^1, y^2, \ldots y^d$ by $y^1(\hat{\theta}), y^2(\hat{\theta}), \ldots y^d(\hat{\theta})$. The procedure is repeated cycling through the columns of Y until a desired outcome is achieved.

Maximum Likelihood criterion: For the remainder of this section we use a *maximum likelihood* criterion to select our model parameters. We start by considering a Parzen estimator for the date vector $x = (x_1, x_2, \ldots, x_N)^T$ of the form

$$p(t;x) = \frac{1}{Nd}\sum_{j=1}^{N} K\left(\frac{t - x_j}{h}\right), \qquad t \in \mathbb{R}.$$

where the kernel is chosen to be $K = N(\cdot; 0, 1)$. The choice of the bin width h is discussed in [16].

Given d orthonormal vectors $y^1, y^2, \ldots, y^d$ and corresponding marginals $p\left(\cdot; X^T y\right)$, the maximum entropy estimator given by Theorem 1 is

$$P^*(x) = \prod_{j=1}^{d} p\left(\langle y^j, x\rangle\ ;\ X^T y^j\right), \qquad x \in \mathbb{R}^d,$$

and the *log–likelihood* function of the data has the form

$$L(Y) := \sum_{i=1}^{N}\sum_{j=1}^{d} \log\left(\frac{1}{Nd}\sum_{k=1}^{N} K(\langle y^j, g^{ik}\rangle\right)$$

where

$$g^{ik} := \frac{x^i - x^k}{h}, \quad i, k = 1, 2, \ldots, N.$$

We resort to an EM strategy to develop an iterative update formula that increases the function L. To this end, we let the matrix $\hat{Y}$ be our *initial* guess for the desired orthogonal matrix Y and seek another orthogonal matrix Y such that $L(Y) \geq L(\hat{Y})$. To find a suitable Y, we set

$$c_{kij} := \frac{K(\langle \hat{y}^j, g^{ik}h\rangle)}{\sum_{m=1}^{N} K(\langle \hat{y}^j, g^{im}h\rangle)}\ . \tag{13}$$

and recall by the concavity of the log function and the form of the kernel K there holds the inequality $L(Y) - L(\hat{Y}) \geq Q(\hat{Y}) - Q(Y)$ where

$$Q(Y) := \sum_{j=1}^{d} \langle y^j, A_j y^j\rangle \tag{14}$$

and for $j = 1, 2, \dots, d$

$$A_j := \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} c_{kij} (x^i - x^k)(x^i - x^k)^T.$$

To update Y we diminish the function Q by considering the function

$$F(\theta) = \langle y^1(\theta), A_1 y^1(\theta) \rangle + \langle y^2(\theta), A_2 y^2(\theta) \rangle$$

which can be expressed as

$$F(\theta) = \alpha \sin 2\theta + \beta \cos 2\theta + \gamma$$

where

$$B = \frac{1}{2}(A_1 - A_2), \qquad C = \frac{1}{2}(A_1 + A_2)$$

and

$$\beta = \langle y^1, B y^1 \rangle - \langle y^2, B y^2 \rangle, \qquad \alpha = -2\langle y^1, B y^2 \rangle, \qquad \gamma = \langle y^1, C y^1 \rangle + \langle y^2, C y^2 \rangle.$$

The minimun of the function F occurs at $\hat{\theta} = k\pi \pm \frac{\psi}{2}$ where k is an integer and

$$\cos \psi = -\frac{\beta}{\sqrt{\alpha^2 + (\beta)^2}}.$$

We now use planar rotations as before to diminish the function Q and hence enhance L.

We consider improving the MLE estimator introduced above for our multivariate data $x^1, x^2, \dots, x^N \in \mathbb{R}^d$ by using a univariate mixture model of gaussians mixtures for the marginals. Therefore, given d orthonormal vectors $y^1, y^2, \dots, y^d \in \mathbb{R}^d$, the log–likelihood function of interest to us here is

$$L(Y, \omega, \mu, \sigma) = \sum_{i=1}^{N} \sum_{j=1}^{d} \log \left(\sum_{\ell=1}^{m} \omega_{\ell j} N \left(\langle y^j, x^i \rangle \ ; \ \mu_{\ell j}, \sigma_{\ell j} \right) \right), \tag{15}$$

where the mixture weights satisfy

$$w_{\ell j} > 0, \quad \sum_{\ell=1}^{m} w_{\ell j} = 1, \qquad j = 1, 2, \dots, d.$$

Our goal is now to describe a strategy to iteratively increase L. As before, we suppose we start with an orthogonal matrix $\hat{Y}$ and first explain how we update ω, μ and σ to increase L. We let $\hat{\omega}$, $\hat{\mu}$ and $\hat{\sigma}$ be our initial guess for these parameters and shall identify ω, μ and σ such that $L(\hat{Y}, \omega, \mu, \sigma) \geq L(\hat{Y}, \hat{\omega}, \hat{\mu}, \hat{\sigma})$. To this end, we observe by the concavity of the log function that

$$L(\hat{Y}, \omega, \mu, \sigma) - L(\hat{Y}, \hat{\omega}, \hat{\mu}, \hat{\sigma}) \geq Q(\hat{Y}, \omega, \mu, \sigma) - Q(\hat{Y}, \hat{\omega}, \hat{\mu}, \hat{\sigma}),$$

where

$$\alpha_{ij\ell} := \frac{\hat{\omega}_{\ell j} G \left(\langle \hat{y}^j, x^i \rangle \ ; \ \hat{\mu}_{\ell j}, \hat{\sigma}_{\ell j} \right)}{\sum_{k=1}^{m} \hat{\omega}_{kj} G \left(\langle \hat{y}^j, x^i \rangle \ ; \ \hat{\mu}_{kj}, \hat{\sigma}_{kj} \right)},$$

and

$$Q(Y, \omega, \mu, \sigma) := \sum_{i=1}^{N} \sum_{j=1}^{d} \sum_{\ell=1}^{m} \alpha_{ij\ell} \log \left(\omega_{\ell j} G \left(\langle y^j, x^i \rangle \ ; \ \mu_{\ell j}, \sigma_{\ell j} \right) \right). \tag{16}$$

The maximum of the function $Q(Y, \omega, \mu, \sigma)$ with respect to ω, μ, σ occurs when

$$\omega_{\ell j} = \frac{1}{N} \sum_{i=1}^{N} \alpha_{ij\ell},$$

$$\mu_{\ell j} = \frac{\sum\limits_{i=1}^{N} \alpha_{ij\ell} \left\langle \hat{y}^j, x^i \right\rangle}{\sum\limits_{i=1}^{N} \alpha_{ij\ell}}$$

and

$$\sigma_{\ell j} = \frac{\sum_{i=1}^{N} \alpha_{ij\ell}(\langle \hat{y}^j, x^i \rangle - \mu_{\ell j})^2}{\sum_{i=1}^{N} \alpha_{ij\ell}} \ .$$

These formulas update $\hat{\omega}$, $\hat{\mu}$ and $\hat{\sigma}$.

We shall now update the orthogonal matrix $\hat{Y}$. To this end, we use the new parameters ω, μ and σ for this step and reformulate the log–likelihood function in terms of unique vectors μ^ℓ, $\ell = 1, 2, \ldots, m$, defined by the equation $\mu_{j\ell} = \langle y^j, \mu^\ell \rangle$

$$L\left(Y, \omega, \mu, \sigma\right) = \sum_{i=1}^{N} \sum_{j=1}^{d} \log \left(\sum_{\ell=1}^{m} \omega_{\ell j} N\left(\left\langle y^j, x^i \right\rangle \ ; \ \left\langle y^j, \mu^\ell \right\rangle, \sigma_{\ell j}\right) \right).$$

Suppose that $\hat{Y}$ is our initial choice. We seek an orthogonal matrix Y such that $L(Y, \omega, \mu, \sigma) \geq L(\hat{Y}, \omega, \mu, \sigma)$. Similar to the argument used before we have that $L(Y, \omega, \mu, \sigma) - L(\hat{Y}, \omega, \mu, \sigma) \geq V(Y) - V(\hat{Y})$, where $V := u - Q$,

$$\bar{\alpha}_{ij\ell} = \frac{\omega_{\ell j} N\left(\left\langle \hat{y}^j, x^i \right\rangle \ ; \ \left\langle \hat{y}^j, \mu^\ell \right\rangle, \sigma_{\ell j}\right)}{\sum_{k=1}^{m} \omega_{kj} N\left(\left\langle \hat{y}^j, x^i \right\rangle \ ; \ \left\langle \hat{y}^j, \mu^k \right\rangle, \sigma_{kj}\right)},$$

$$Q(Y) := \sum_{j=1}^{d} \langle y^j, A_j y^j \rangle \tag{17}$$

$$A_j := \frac{1}{2} \sum_{i=1}^{N} \sum_{\ell=1}^{m} \frac{\bar{\alpha}_{ij\ell}}{\sigma_{\ell j}} (x_i - \mu_\ell)(x_i - \mu_\ell)^T$$

and

$$u := \sum_{i=1}^{N} \sum_{j=1}^{d} \sum_{\ell=1}^{m} \bar{\alpha}_{ij\ell} \log \left(\frac{\omega_{\ell j}}{\sqrt{2\pi \sigma_{\ell j}}} \right).$$

Since the vector u is *independent* of Y, we can use the methods described earlier to diminish the function Q and thereby enhance L.

All of the methods we have described above start out from the perspective of *nonparametric* density estimation. For parameter estimation techniques based on *nongaussian* mixture components, see [1, 4, 2].

3. Optimizing the Function Q

In view of the discussion of the previous section (see equations (14) and (17)), an essential computational problem for the selection of model parameters in the methods described above is to minimize the function

$$Q(Y) := \sum_{i=1}^{d} \left\langle y^i, A_i y^i \right\rangle$$

where $A_1, A_2, \ldots, A_d$ are real symmetric matrices and $Y = [y^1, y^2, \ldots, y^d] \in \mathbb{R}^{d \times d}$ is an orthogonal matrix, that is, $Y^T Y = I$. In the present section we gather together some observations about this problem including two algorithms for minimizing Q and our computational experience with them.We first start with a result pointed out to us by Alan Hoffman that identifies an orthogonal matrix which minimizes the function Q under some special conditions on the matrices $A_i, i = 1, 2, \ldots, d$. We state this interesting fact next.

Proposition 1 *If there exist an orthogonal matrix $M \in \mathbb{R}^{d\times d}$ and diagonal matrices $D_i \in \mathbb{R}^{d\times d}$, $i = 1,2,\ldots,d$ such that $A_i = MD_iM^T$, $i = 1,2,\ldots,d$ then there is a permutation matrix $P \in \mathbb{R}^{d\times d}$ with the property that for any orthogonal matrix $Y \in \mathbb{R}^{d\times d}$, $Q(T) \leq Q(Y)$ where $T := M^TP$.*

PROOF. For any orthogonal matrix $Y = [y^1, y^2, \ldots, y^d]$ we let $z^i = M^Ty^i$, $i = 1,2,\ldots,d$ and observe that these vectors are also mutually orthogonal and satisfy for $i = 1,2,\ldots,d$ the equations $\langle y^i, A_iy^i\rangle = \langle z^i, D_iz^i\rangle$. Hence the proof boils down to showing that the function $F(Y) := \text{trace}\{SA\}$ where S is the matrix whose elements are defined by the equations $S_{ij} := (y_i^j)^2$ $i,j = 1,2,\ldots,d$ and the matrix A is built from the diagonal matrices $D_i = \text{diag}\,[a_{i1}, a_{i2}, \ldots a_{id}]$, $i,j = 1,2,\ldots,d$ is bounded below by $F(P)$ for some permutation matrix P.

Recall that a $d \times d$ matrix, $W = [w_{ij}]$,$i,j = 1,2,\ldots,d$ is called *doubly stochastic* if $\sum_{k=1}^d w_{kj} = \sum_k^d w_{jk} = 1$ and $w_{ij} \geq 0$, $i,j = 1,2,\ldots,d.$ Note that since Y is an orthogonal matrix the matrix S defined above is doubly stochastic. Moreover, since the function F is *linear* in S we can appeal to a theorem attributed to Birkhoff [6, 11] which states: *The set $\mathcal{S}$ of all $n \times n$ doubly stochastic matrices is the convex hull of the set $\mathcal{P}$ of $n \times n$ permutation matrices* to prove the result. ■

Another fact of some utility is a simple lower bound on the function Q. We state this result next.

Proposition 2 *If $A_1, A_2, \ldots, A_d \in \mathbb{R}^{d\times d}$ are real symmetric matrices and Y is a $\mathbb{R}^{d\times d}$ orthogonal matrix then*

$$Q(Y) \geq \sum_{i=1}^d \lambda_i,$$

where λ_i is the minimum eigenvalue of A_i.

PROOF. Since the matrices A_i's are real symmetric we have for any vector y

$$\langle y, A_iy\rangle \geq \lambda_i \langle y, y\rangle$$

We use this inequality to bound each summand that appears in the definition of Q to prove the result. ■

We now turn to the the description of two algorithms which we have used to minimize the function Q. The first was described earlier and uses planar rotations. We note that all orthogonal matrices of size d can be written as a product of $d(d-1)/2$ planar rotations (see [18, section 14.6]), each of which is in turn parameterized by a rotation angle, say θ. Our strategy in this section is to successively use these planar rotations in such a way that Q is minimized at each step as a function of the rotation angle θ. We achieve this goal by using the following algorithm:

Algorithm 1

- For each vector pair y^j and y^k, form vectors:
- $\begin{cases} y^j(\theta) = (\cos\theta)y^j - (\sin\theta)y^k \\ y^k(\theta) = (\sin\theta)y^j + (\cos\theta)y^k \end{cases}$
- Find $\hat{\theta}$ s. t. $F(\hat{\theta}) = \min\left\{F(\theta) = \langle y^j(\theta), A_jy^j(\theta)\rangle + \langle y^k(\theta), A_ky^k(\theta)\rangle : \theta \in [0, 2\pi]\right\}$
- Replace θ with $\hat{\theta}$ in $y^j(\theta)$ and $y^k(\theta)$.
- Repeat until convergence occurs.

Note that the orthogonality of the vectors $y^j(\theta)$ and $y^k(\theta)$ in the algorithm above is preserved, since these merely represent y^j and y^k in a rotated coordinate system. Furthermore, note that this algorithm in fact performs $d(d-1)/2$ planar rotations in each loop. We can simplify $F(\theta)$ by writing

$$F(\theta) = \alpha\sin(2\theta) + \beta\cos(2\theta) + \gamma,$$

where

$$\alpha = -2\left\langle y^j, By^k \right\rangle, \qquad \beta = \left\langle y^j, By^j \right\rangle - \left\langle y^k, By^k \right\rangle, \qquad \gamma = \left\langle y^j, Cy^j \right\rangle + \left\langle y^k, Cy^k \right\rangle$$

and

$$B = \frac{1}{2}(A_j - A_k), \qquad C = \frac{1}{2}(A_j + A_k).$$

We locate the stationary points of $F(\theta)$ with the formulas

$$\hat{\theta} = \frac{1}{2}(k\pi + \psi); \qquad \psi = \arctan\left(\frac{\alpha}{\beta}\right), \quad k \in \mathcal{Z}.$$

and are interested in those which minimize of F. For this purpose, we observe that F can be rewritten as

$$F(\theta) = \left[\ \beta, \alpha\ \right] \left[\begin{array}{c} \cos 2\theta \\ \sin 2\theta \end{array} \right]$$

so that the minimum of F is $-(\beta^2 + \alpha^2)^{\frac{1}{2}}$ and is achieved when $\left[\begin{array}{c} \cos 2\theta \\ \sin 2\theta \end{array} \right]$ points in the opposite direction of $\left[\ \beta, \alpha\ \right]$. That is, if $\beta + i\alpha = \exp(i\psi_0)$ then $\theta = -\psi_0/2 + k\pi$ is a minimizes F.

As an alternative to the strategy of using planar rotations discussed, we consider the possibility of using conjugate gradient to search for local maxima of the function Q. We attack this problem by parameterizing orthogonal matrices by skew-symmetric matrices. This frees the variables of any constraint equations, thus enabling the use of gradient ascent methods.

Recall that a complex matrix is skew hermitian symmetric whenever $S = -S^{*T}$ and we have the following fact.

Lemma 1 *A matrix $Y \in \mathbb{C}^{m\times m}$ with $\det Y = 1$ is unitary if and only if there exists a skew-Hermitian matrix S such that*

$$Y = e^S.$$

Furthermore, if $Y \in \mathbb{R}^{m\times m}$ then $S \in \mathbb{R}^{m\times m}$.

Note that the representation of Y in terms of S is not unique.Indeed for any $k \in \mathbb{Z}$ and

$$S = \left[\begin{array}{cc} 2\pi k & 0 \\ 0 & 2\pi k \end{array} \right].$$

we have that $\exp S = I$. Let us briefly present the proof of this result for the convenience of the reader.
PROOF. We write Y in the form

$$Y = P\Lambda P^*;$$

where P is a unitary matrix and $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ is the diagonal matrix of eigenvalues of Y. Certainly, for $i = 1, 2, \ldots, d$, $\lambda_i = \exp(\imath\theta_i)$ for some $\theta_i \in \mathbb{R}$. Thus, we obtain the desired representation $Y = e^S$, where $S := P\text{diag}[\theta_1, \theta_2, \cdots, \theta_m]P^*$.

Additional effort is needed to show that S can be chosen to be real when $Y \in \mathbb{R}^{m\times m}$. In this case, complex eigenvalues (and the associated eigenvectors) of Y occur in conjugate pairs. Thus, we may choose an ordering of the eigenvalues (and the corresponding eigenvectors in P) such that $\lambda_i = \bar{\lambda}_{i+1}$ for $i = 1, 3, \ldots, (2n-1)$, $\lambda_i = -1$ for $i = 2n+1, 2n+2\ldots, 2\ell$, and $\lambda_i = 1$ for $i = 2\ell+1, 2\ell+2\ldots, m$. Note that since $\det Y = 1$, the number of λ_i's that equals -1 is even and therefore we can write the matrix Λ in block form

$$\Lambda = \exp\Gamma$$

where

$$\Gamma := \operatorname{diag}\left[\Gamma_1^c, \ldots, \Gamma_n^c, \Gamma_1^\pi, \ldots, \Gamma_s^\pi, \Gamma_1^p, \ldots, \Gamma_r^p\right],$$

with

$$\Gamma_i^c := \begin{bmatrix} \imath\theta_i & 0 \\ 0 & -\imath\theta_i \end{bmatrix}, \quad \Gamma_i^\pi = \begin{bmatrix} 0 & \pi \\ -\pi & 0 \end{bmatrix}, \quad \text{and} \quad \Gamma_i^p = [0].$$

We set $S = P\Gamma P^*$ so that we have

$$Y = Pe^{\tilde{\Lambda}}P^*.$$

Moreover, for $i = 1, 2, \ldots, n$ we introduce matrices

$$\Gamma_i^c = J\Gamma_i^r J^*$$

where $J = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -\imath \\ 1 & \imath \end{bmatrix}$ and $\Gamma_i^r := \begin{bmatrix} 0 & \theta_i \\ -\theta_i & 0 \end{bmatrix}$. Thus, we may write

$$S = Q\Gamma^r Q^*,$$

where

$$\Gamma^r := \operatorname{diag}\left[\Gamma_1^r, \ldots, \Gamma_n^r, \Gamma_1^\pi, \ldots, \Gamma_s^\pi, \Gamma_1^p, \ldots, \Gamma_r^p,\right]$$

$$Q = P\bar{J}$$

and

$$\bar{J} = \operatorname{diag}[\underbrace{J, \ldots, J,}_{n \text{ copies}} \quad \underbrace{1, \ldots, 1}_{(2s+r) \text{ copies}} \quad]$$

Since $\Gamma^r \in \mathbb{R}^{m\times m}$, the proof that $S \in \mathbb{R}^{m\times m}$ will be complete when we show that $Q \in \mathbb{R}^{m\times m}$. For this purpose,let the eigenvectors of Y be

$$P = [p_1, p_2, \ldots, p_m]$$

and by choice we have that $p_i = \bar{p}_{i+1}$ for $i = 1, 3, \ldots, (2n-1)$, and p_i's are real for $i = 2n+1, 2n+2\ldots, m$. Invoking this fact we conclude that

$$Q = [2\Re(p_1), 2\Im(p_1), \ldots, 2\Re(p_{2n-1}), 2\Im(p_{2n-1}), p_{2n+1}, \ldots, p_m] \in \mathbb{R}^{m\times m},$$

Now since $Q \in \mathbb{R}^{m\times m}$, we have $S^T = Q(\tilde{\Lambda^r})^T Q^T = -Q\tilde{\Lambda}^r Q^T = -S$ which shows that S is skew-symmetric, thus, completing the proof. ■

For matrices of small size e^S can be explicitly computed by using the Cayley-Hamilton theorem. For example, when $m = 2, 3, 4$ we have that

$$\begin{aligned} e^{S_2} &= \cos\theta_2 I_2 + \frac{\sin\theta_2}{\theta_2}S_2 \\ e^{S_3} &= I_3 + \frac{\sin\theta_3}{\theta_3}S_3 + \frac{\cos\theta_3 - 1}{\theta_3^2}S_3^2 \\ e^{S_4} &= \alpha_0 I_4 + \alpha_1 S_4 + \alpha_2 S_4^2 + \alpha_3 S_4^3, \end{aligned}$$

where

$$S_2 = \begin{bmatrix} 0 & a_{12} \\ -a_{12} & 0 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 0 & b_{12} & b_{13} \\ -b_{12} & 0 & b_{23} \\ -b_{13} & -b_{23} & 0 \end{bmatrix}, \quad S_4 = \begin{bmatrix} 0 & c_{12} & c_{13} & c_{14} \\ -c_{12} & 0 & c_{23} & c_{24} \\ -c_{13} & -c_{23} & 0 & c_{34} \\ -c_{14} & -c_{24} & -c_{34} & 0 \end{bmatrix},$$

and $\theta_2 = a_{12}$, $\theta_3 = \sqrt{b_{12}^2 + b_{13}^2 + b_{23}^2}$,

$$\begin{aligned}
\alpha_0 &= \frac{1}{r_- - r_+}\left[r_- \cos(\sqrt{-r_+}) - r_+ \cos(\sqrt{-r_-})\right] \\
\alpha_1 &= \frac{1}{r_- - r_+}\left[\frac{r_-}{\sqrt{-r_+}}\sin(\sqrt{-r_+}) - \frac{r_+}{\sqrt{-r_-}}\sin(\sqrt{-r_-})\right] \\
\alpha_2 &= \frac{1}{r_- - r_+}\left[\cos(\sqrt{-r_-}) - \cos(\sqrt{-r_+})\right] \\
\alpha_3 &= \frac{1}{r_- - r_+}\left[\frac{\sin(\sqrt{-r_-})}{\sqrt{-r_-}} - \frac{\sin(\sqrt{-r_+})}{\sqrt{-r_+}}\right],
\end{aligned}$$

where $r_\pm = (-\beta^2 \pm \sqrt{\delta})/2$, $\beta = \sqrt{c_{12}^2 + c_{13}^2 + c_{14}^2 + c_{23}^2 + c_{24}^2 + c_{34}^2}$ and

$$\delta = \left((c_{12}+c_{34})^2 + (c_{13}-c_{24})^2 + (c_{14}+c_{23})^2\right) \times \left((c_{12}-c_{34})^2 + (c_{13}+c_{24})^2 + (c_{14}-c_{23})^2\right).$$

These formulas allow for the direct use of numerical optimization packages, however, for $m \geq 5$ this procedure impractical. As an alternative, we discuss numerical procedures for computing e^S and its derivatives by using the following result, see [13, 12, 15]:

Proposition 3 *If $S = [s_{ij}]$, $i, j = 1, 2, \ldots, d$ is a skew-symmetric matrix and $p = 1, 2, \ldots$ then*

$$\frac{\partial S^p}{\partial s_{ij}} = S\frac{\partial S^{p-1}}{\partial s_{ij}} + (E_{ij} - E_{ji})S^{p-1}$$

where all the elements of the matrix E_{ij} are zero except for the i-th, j-th element, which is equal to one.

PROOF. The proposition follows as a simple application of the product rule for matrix differentiation [4, 13, 15, 12]. ■

If the eigenvalues of S are small the Taylor series for e^S converges rapidly and this can be exploited to compute approximations to e^S and its derivatives. When the eigenvalues of S are large, or simply in order to accelerate convergence, we use the relation

$$e^S = \left(e^{\frac{1}{\alpha}S}\right)^\alpha$$

where α is chosen to be some power of two for rapid computation of e^S and for its derivatives use the product rule for matrix differentiation.

We associate with every skew-symmetric $S = [s_{ij}]$, $i, j = 1, 2, \ldots, d$ a vector $v \in \mathbb{R}^{d(d-1)/2}$ (and visa - versa) by the formula

$$v = \left[s_{12}, \cdots, s_{1d} | s_{23}, \cdots, s_{2d} | \cdots | s_{(d-1)d}\right]$$

and transform Q into the unconstrained optimization problem of finding v^* so that

$$U(v^*) = \min\{U(v) : v \in \mathbb{R}^{d(d-1)/2}\}$$

where $U(v) := Q(\exp S)$.

By using results of Proposition 3 we may now compute the gradient of the function U follow the algorithm:

Algorithm 2

- Initialize $S = \log Y$ and form the vector v.
- Compute the desired derivatives of $\exp S$ and therefore obtain ∇U.
- Find the optimal v^* by using a conjugate gradient update strategy such as: $v = \hat{v} - \rho \nabla_{\hat{v}} U(\hat{v})$, where the constant ρ is chosen appropriately.
- Finally, rearrange elements of v^* to get S^* and compute $Y^* = e^{S^*}$.

Other optimization routines can be used which are better suited for large scale problems. This algorithm was implemented in Matlab using the function 'fminunc' from the Optimization Toolbox, which uses a method involving preconditioned conjugate gradients [9].

4. Numerical Results

The two algorithms presented in the previous section for minimizing the function Q were implemented in Matlab, using the $d \times d$ identity matrix as initial value for Y and utilizing the Matlab function 'fminunc' in the second algorithm. We consider their performance on the following three examples. First we consider an example where the algorithms should reach the lower bound presented in Proposition 2.

Example 1 We consider the following 3 orthogonal matrices:

$$Q_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad Q_2 = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} \end{bmatrix} \quad \text{and} \quad Q_3 = \begin{bmatrix} 0 & \frac{1}{\sqrt{3}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

and the diagonal matrices (with ascending values on the diagonal):

$$D_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 5 \end{bmatrix} \quad \text{and} \quad D_3 = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 8 \end{bmatrix}.$$

Next, we construct matrices $A_i = Q_i D_i Q_i^T$ for $i = 1, 2, 3$. Note that this corresponds to the eigenvalue-decomposition of the matrices A_1, A_2 and A_3. Moreover, the eigenvectors corresponding to the minimal eigenvalue for each A_i are chosen so that they are mutually orthogonal. Hence, the global minimum of Q is achieved by the orthogonal matrix

$$Y^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}.$$

Algorithm 1 yielded the following solution after twelve iterations:

$$\hat{Y}_1^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.7071 & -0.7071 \\ 0 & 0.7071 & 0.7071 \end{bmatrix},$$

whereas Algorithm 2 converged after six iterations yielding the solution:

$$\hat{Y}_2^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.7072 & -0.7071 \\ 0 & 0.7071 & 0.7072 \end{bmatrix}.$$

The small inaccuracy in the last solution can be attributed to round-off errors resulting from the computation of matrix logarithms and exponentials. Also, it is easily verified that $Q(\hat{Y}_1^*) = Q(\hat{Y}_2^*) = 6$ (within a precision of 10^{-3}), thus, confirming that both algorithms achieve the global minimum in this case. The next example concerns the diagonal case discussed in Proposition 1.

Example 2 Consider the 5×5 diagonal matrices consisting of random numbers:

$$\begin{aligned} D_1 &= \text{diag}\{[0.0178, 0.2477, 0.3662, 0.0510, 0.5587]\} \\ D_2 &= \text{diag}\{[0.5223, 0.1489, 0.2709, 0.3607, 0.0850]\} \\ D_3 &= \text{diag}\{[0.6148, 0.8254, 0.9150, 0.1290, 0.2233]\} \\ D_4 &= \text{diag}\{[0.8930, 0.2746, 0.6009, 0.4612, 0.6609]\} \\ D_5 &= \text{diag}\{[0.7574, 0.2787, 0.9980, 0.7245, 0.7056]\} \end{aligned}$$

Algorithm 1 converges after two iterations for this example and yields the following minimizer:

$$\hat{Y}_1^* = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \end{bmatrix}.$$

However, the unconstrained optimization method of Algorithm 2 converged to the identity matrix $\hat{Y}_2^*$,in one iteration this being the case because the identity matrix is a local extremum, which we chose as an initial starting point. Since $F(\hat{Y}_1^*) < F(\hat{Y}_2^*)$, in this case, $\hat{Y}_2^*$ is a local minimizer. Interestingly, we note that both of the Y's are permutation matrices. This is consistent with Proposition 1 which asserts that in the case when all A_i's are diagonal the local minima of Q are permutation matrices.

Example 3 Consider the matrices $A_i = MD_iM^T$ for $i = 1, 2, 3$ with

$$D_1 = \text{diag}[1, 5, 3]; \quad D_2 = \text{diag}[2, 7, 1]; \quad D_3 = \text{diag}[6, 4, 9],$$

and the orthogonal matrix M

$$M = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{-2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \end{bmatrix}.$$

In this case, Algorithm 1 converged to the following matrix after four iterations:

$$\hat{Y}_1^* = \begin{bmatrix} -0.5774 & 0 & 0.8615 \\ -0.5774 & 0.7071 & -0.4082 \\ -0.5774 & -0.7071 & -0.4082 \end{bmatrix},$$

and we obtain

$$\hat{Z}_1^* = M^T\hat{Y}_1^* = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix},$$

which is a permutation matrix. Moreover, $Q(\hat{Y}^*) = 6$, which by Proposition 1, is the global minimum. Thus, Algorithm 1 gives us the global minimum in this case as well. Although Algorithm 2 used sixty five iterations to stop at the following solution

$$\hat{Y}_2^* = \begin{bmatrix} 0.5753 & 0.0006 & -0.8179 \\ 0.5770 & 0.7085 & 0.4063 \\ 0.5798 & -0.7057 & 0.4073 \end{bmatrix},$$

which gives

$$\hat{Z}_2^* = M^T \hat{Y}_2^* = \begin{bmatrix} 1.0000 & 0.0020 & -0.0025 \\ 0.0025 & 0.0007 & 1.0000 \\ -0.0020 & 1.0000 & -0.0006 \end{bmatrix},$$

which is a close approximation to a permutation matrix. Taking further into consideration the fact that $Q(\hat{Y}^*) = 6.0001 \approx Q(Y^*)$ one concludes that the unconstrained optimization based Algorithm 2 converges to the correct solution as well.

We remark that the above examples and other experiments we conducted indicate that in terms of computation time Algorithm 1 generally is more efficient than Algorithm 2. Furthermore, the point to which the convergence algorithm converged depended on the starting point chosen, which indicates that the function Q has multiple local (and possibly also global) minima.

We now turn our attention to density estimation using our maximum entropy method for estimation for multivariate probability densities corresponding to equation (15). where the parameters are determined by the following two-step algorithm:

- Update $w_{\ell j}, \mu_{\ell j}$ and $\sigma_{\ell j}^2$ by using the EM algorithm
- Update the orthogonal projection directions Y by minimizing $F(Y)$.

This algorithm was implemented using Algorithm 1 and compared to the gaussian mixture model [10] for the two dimensional examples described below.

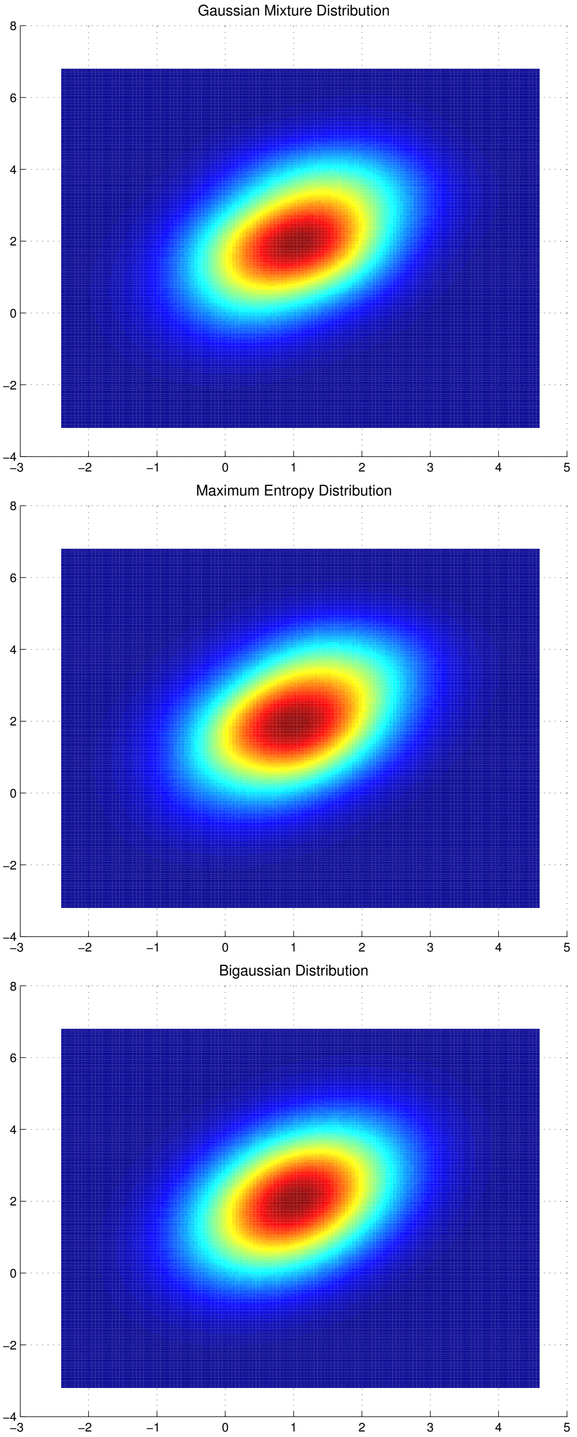
Example 4 In this case, five thousand random samples were drawn from a two-dimensional gaussian distribution, with mean and variance $\mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{bmatrix}$. The maximum entropy model and the mixture gaussian models were trained using this data, and the estimated probability densities were plotted using the same grid for comparison. The results shown in figure 1, illustrates that both models estimate the true distribution fairly well.

Example 5 In order to examine the behavior of our models in a somewhat more involved scenario, we considered data used in speech recognition experiments. We used two-dimensional projections of 60 dimensional acoustic feature vectors from a sub-phonetic speech unit. The two models under investigation were trained using a set of 10248 of two-dimensional vectors so obtained, and the estimated distributions were plotted.

Figure 2 displays the plots of the densities obtained via the two methods together with the histogram of the data. Here, an inherent weakness of the maximum entropy model is revealed, namely, when dealing with spherically non-symmetric distributions, the model smooths out the structure. This is due to the way the model is built, projecting the data into optimal orthogonal directions and estimating the distribution from the marginals in these directions. In comparison, the gaussian mixture model is more of a multidimensional kernel estimator, and has, therefore, more advantage in estimating such distributions. In particular, the histogram indicates a lump in the structure, which is captured by the gaussian mixture model, whereas it is smoothed out by the maximum entropy model. Furthermore, it seems that the former emphasizes bimodality in this case, where the latter does not. This is better observed in the three-dimensional graphs displayed in figure 3.

5. Conclusion

We have considered problems associated with estimation of probability densities from projection of data on lower dimensional subspaces. We have indicated how maximum entropy and maximum likelihood criteria



Gaussian Mixture Distribution

Maximum Entropy Distribution

Bigaussian Distribution

Figure 1. These two dimensional graphs illustrate the good performances of both the Gaussian Mixture Model and the Maximum Entropy Model for estimating the Bigaussian distribution shown on the bottom.

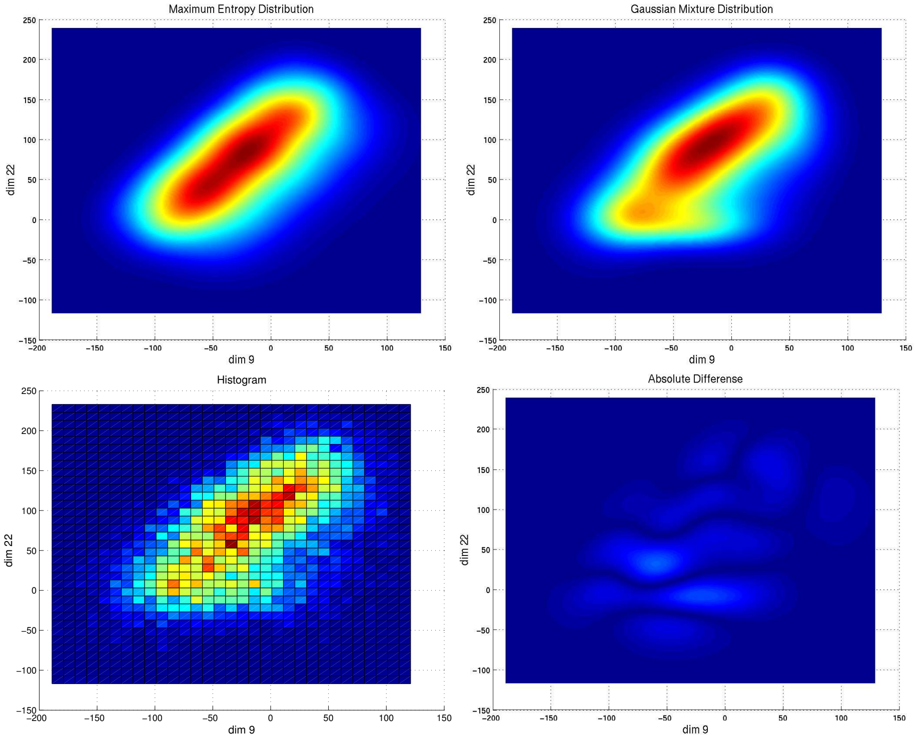


Figure 2. The Maximum Entropy and the Gaussian Mixture estimates of the distribution of the most correlated dimensions (9 and 22) of leaf 1. The bottom left graph shows the histogram of the data, and the bottom right graph illustrates the absolute difference between the two models.

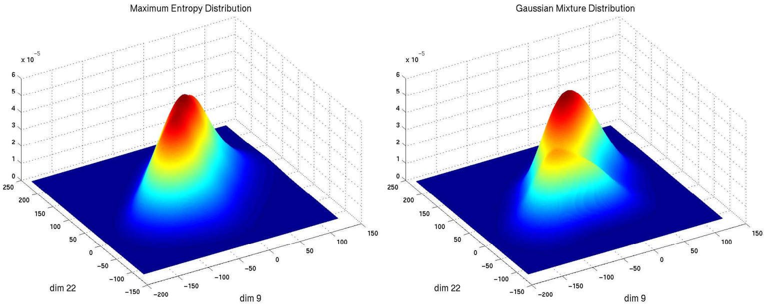


Figure 3. The estimates of the distribution of leaf 1’s dimensions 9 and 22, viewed in three dimensions.

can be used to exploit the nonuniqueness of the problem of reconstruction. Estimates of densities in the lower dimensional subspace can be obtained from a variety of models including e.g., the gaussian mixture model. A nonlinear optimization problem of specific nature appears to play a central role in our discussion. Algorithms for solving this optimization problem for estimating these densities are discussed. Simple numerical examples are worked out to demonstrate the method. The present research could be pursued in several other directions. An issue that we have not addresses would be to consider reconstruction of d-dimensional densities from projections along *more* than d directions.

Acknowledgement. Mohammad Saif Ullah Khan's work was carried out during a visit to the IBM T. J. Watson Research Center, Yorktown Heights, New York, USA. Charles A. Micchelli's work was partially supported by National Science Foundation grant DMS-9504780.

References

[1] Basu, S. and Micchelli, C. A. (1998). Parametric density estimation for the classification of acoustic feature vectors in speech recognition, in *Nonlinear Modeling: Advanced Black-Box Techniques*, J. A. K. Suykens and J. Vandewalle (eds), Kluwer Academic Publishers, Boston, 87–118.

[2] Basu, S., Micchelli, C. A. and Olsen, P. (1999). Maximum likelihood estimates for exponential type density families, *IEEE Int. Conf. on Acoustics Speech & Signal Processing*.

[3] Basu, S., Micchelli, C. A. and Olsen, P. (2000). A maximum entropy and maximum likelihood criteria for feature selection from multivariable data, *IEEE Int. Symp. on Circuits and Systems*, Geneva, Switzerland.

[4] Basu, S., C. A. Micchelli, C. A. and Olsen, P. (2001). Power exponential densities for the training and classification of acoustic feature vectors in speech recognition, *J. Comput. Graph. Statist.*, **10**, (1), 158–192.

[5] Basu, S., Micchelli, C. A. and Olsen, P. (2002). A maximum entropy criterion for feature extraction from multivariable data, in *Wavelet Analysis and Applications*, AMS/IP Studies in Advanced Mathematics, **25**.

[6] Birkhoff, G. (1946). Tres observations sobre el Álgebra lineal, *Rev. Univ. Nac. Tucumán*, ser. A, **5**, 147– 151.

[7] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc.*, Ser. B, **39**, 1–38.

[8] Friedman, J. H., Stuetzle, W. and Schreder, A. (1984). Projection pursuit density estimation, *J. Amer. Statist. Assoc.*, **79**, 599–608.

[9] *MATLAB 6.0 Reference guide: Optimization Toolbox, 'fminunc'*.

[10] Jelinek, F. (1997). *Statistical Methods for Speech Recognition*, The MIT Press, Cambridge, MA.

[11] Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and its Applications*, Academic Press.

[12] Rao, C. R. (1985). Matrix derivatives, in *Encyclopedia of Statistical Science*, John Wiley, New York, **5**, 320–325.

[13] Rogers, G. R. (1980). *Matrix Derivatives*, Marcel Dekker, New York.

[14] Scott, D. W. (1992). *Multivariate Density Estimation*, Wiley & Sons.

[15] Sebastiani, P. (1996). On the derivatives of matrix powers, *SIAM J. Matrix Anal. Appl.*, **17**, (3), 640–648.

[16] Silverman, B. W. (1996). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall.

[17] Tapia, R. A. and Thompson, J. R. (1992). *Nonparametric probability density estimation*, John Wiley New York.

[18] Vaidyanathan, P. P. (1993). *Multivariate Systems and Filter Banks*, Prentice-Hall.

Sankar Basu
IBM T.J. Watson Research Center
New York 10532
USA
sbasu@us.ibm.com

Mohammad Saif Ullah Khan
Norwegian University of Science and Technology
Trondheim
Norway

C. A. Micchelli
Department of Mathematics and Statistics
State University of New York, Albany
New York 12222
cam @math.albany.edu

Peder A. Olsen
IBM T.J. Watson Research Center
Yorktown Heights
New York 10598
pederao@watson.ibm.com