RC22244 (W0111-030) November 8, 2001
Computer Science

IBM Research Report

Multivariate Density Estimation from Lower Dimensional Slices

Sankar Basu
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Multivariate Density Estimation from Lower Dimensional Slices

Mohammad Saif Ullah Khan* Peder Olsen† Sankar Basu‡
Charles A. Micchelli§

Abstract

We introduce a new method of statistical density estimation based on projections of higher dimensional data on lower dimensional subspaces. The problem of estimating densities is expected to be more well-posed in lower dimensional subspace due to the fact that in lower dimensions data points can be viewed as relatively less sparse (thus, alleviating the problems arising from curse of dimensionality). While the problem of reconstruction of the higher dimensional density function from low dimensional densities is reminiscent of tomographic reconstruction problem, the reconstruction turns out to be nonunique unless additional constraints are imposed. One such constraint that we consider in the present paper is the maximum entropy criterion. Alternatively maximum likelihood estimated could also be used. Different data models for the projected data can be assumed within this framework. Among other models, we consider the gaussian mixture model for this purpose, and show that an Expectation Maximization strategy for parameter update can be used to solve the problem. While updates of means and covariances can be obtained more or less in a manner similar to the standard EM algorithm ala e.g., Dempster-Schafer[8], computation of best directions of projection, when built into the EM algorithm, essentially involves solving an additional nonlinear optimization problem that is interesting in its own right, and has recently appeared in other statistical problems [5, 17]. Interesting connections of special cases of this optimization problem with known results on the theory of stochastic matrices are pointed out in the context of our discussion. Simple numerical examples are worked out with both real and synthetic data to validate the efficacy of the proposed method.

*Graduate student, Norwegian University of Science and Technology, Trondheim, Norway. This work was performed during his visit to the IBM T. J. Watson Research Center, Yorktown Heights, New York, USA

†IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598.

‡IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598.

§State University of New York, Department of Mathematics and Statistics, Albany, New York 12222. Also, Emeritus Research Staff Member, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598.

1 Introduction

Our motivation for this paper comes from the problem of classifying acoustic vectors for speech recognition. In that context, we are given vectors $x^1, x^2, \ldots, x^N \in \mathbf{R}^d$ where d is typically 39 and N can be about $5,000$. 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. Typically, in speech recognition the probability density is modeled by a mixture model of Gaussians (or some variations thereof) whose parameters are determined by the EM algorithm using a maximum likelihood criterion [1, 2, 8]. Acoustic vectors are important instances of high dimensional statistical data for which feature recognition can enhance the performance of density estimation in classification and training. The approach taken here to this issue is based on the fact that many methods are 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 criterion. The present paper discusses this approach to high dimensional density estimation and feature extraction in the case that 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. We can then address both analytically and computationally the task of finding the optimal directions. We also consider at this stage the possibility of using, instead of entropy, likelihood as a selection criterion for optimal feature recognition. In the first section, we follow information theoretic methods and identify the multivariate probability density on $\mathbf{R}^d$ with d given marginals with largest entropy. In subsequent sections we address the problem of choosing the directions either by further increasing the entropy or alternatively, the likelihood. The essential computational problem seems to boil down to solving a nontrivial optimization problem that also appears in other contexts e.g. gaussianization of data [5]. We examine this specific issue in some detail in later sections of the paper, and in the concluding section we provide numerical examples to demonstrate the efficacy of the proposed method.

2 Maximum entropy from marginals

In this section we compute the maximum of the entropy of a probability density function on $\mathbf{R}^d$ with prescribed marginals in d linearly independent (column) vectors, $y^1, y^2, \ldots, y^d \in \mathbf{R}^d$. Let Y be the $d \times d$ matrix whose columns are these vectors. Thus, we are given univariate probability density functions, $p_1, p_2, \ldots, p_d$ and consider all multivariate probability density functions $P : \mathbf{R}^d \to \mathbf{R}$ such that for all measurable functions f on $\mathbf{R}$ with $f \cdot p_1, f \cdot$

$p_2, \ldots, f \cdot p_d \in L(\mathbf{R})$ (the space of absolutely integrable functions on $\mathbf{R}$) we have that

$$\int_{\mathbf{R}^d} f(\langle y^i, x\rangle) P(x) dx = \int_{\mathbf{R}} f(t) p_i(t) dt, \qquad i = 1, 2, \ldots, d. \tag{2.1}$$

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

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

Our first observation computes the maximum of $H(P)$ for all $P \in \mathcal{C}(p)$,

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

$$\max\{H(P) \ : \ P \in \mathcal{C}(p)\} = -\frac{1}{2} \log \det G(Y) + \sum_{i=1}^{d} H(p_i)\,. \tag{2.2}$$

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,

$$P^*(x) = \sqrt{\det G(Y)} \prod_{i=1}^{d} p_i\left(\langle y^i, x\rangle\right), \qquad x \in \mathbf{R}^d, \tag{2.3}$$

is the unique probability density function in $\mathcal{C}(p)$ at which the maximum in (2.2) is achieved. □

Proof: First let us confirm that P^* defined by (2.3) is in $\mathcal{C}(p)$.

To this end, we compute for every f such that $f \cdot p_i \in L(\mathbf{R})$ for $i = 1, 2, \ldots, d$

$$\begin{aligned}
\int_{\mathbf{R}^d} f(\langle y^i, x\rangle) P^*(x)\, dx &= \sqrt{\det G(Y)} \int_{\mathbf{R}^d} f(\langle y^i, x\rangle) p_i(\langle y^i, x\rangle) \prod_{j \neq i} p_j(\langle y^i, x\rangle)\, dx \\
&= \int_{\mathbf{R}^d} f(t_i) p_i(t_i) \prod_{j \neq i} p_j(t_j)\, dt \\
&= \int_{\mathbf{R}} f(t) p_i(t)\, dt
\end{aligned}$$

which confirms that $P^* \in \mathcal{C}(p)$.

Next, we let P be *any* function in $\mathcal{C}(p)$. Then, making use of (2.1), we conclude that

$$\begin{aligned} -\sum_{i=1}^{d} H(p_i) &= \int_{\mathbf{R}^d} \sum_{i=1}^{d} \log p_i(\langle y^i, x\rangle) P(x)\, dx \\ &= \int_{\mathbf{R}^d} (\log P^*(x) - \frac{1}{2}\log\det G(Y)) P(x)\, dx \\ &= -\frac{1}{2}\log\det G(Y) + \int_{\mathbf{R}^d} P(x)\log P^*(x)\, dx. \end{aligned} \tag{2.4}$$

Choosing $P = P^*$ in the above equation we conclude that

$$H(P^*) = -\frac{1}{2}\log\det G(Y) + \sum_{i=1}^{d} H(p_i). \tag{2.5}$$

We recall, for any probability densities P_1 and P_2, that

$$\int_{\mathbf{R}^d} P_2(x)\log P_1(x)\, dx \le \int_{\mathbf{R}^d} P_2(x)\log P_2(x)\, dx,$$

where equality holds if and only if $P_1 = P_2$. Specializing this result to the choice $P_1 = P^*$, $P_2 = P$, and combining it with equations (2.4) and (2.5) yields the result. ∎

We let $m = (m_1, m_2, \ldots, m_d)^T$ be the vectors of means associated with probability density functions $p_1, p_2, \ldots, p_d$, that is,

$$m_j = \int_{\mathbf{R}} t p_j(t) dt, \qquad j = 1, 2, \ldots, d,$$

and $D = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)$ for the $d \times d$ diagonal matrices associated with variances

$$\sigma_j = \int_{\mathbf{R}} (t - m_j)^2 p_j(t) dt, \qquad j = 1, 2, \ldots, d.$$

Also, we use $M = (M_1, M_2, \ldots, M_d)^T$ to be the mean of a probability density P on $\mathbf{R}^d$ and Σ to be its $d \times d$ covariance matrix, that is,

$$M_i = \int_{\mathbf{R}^d} x_i\, P(x)\, dx, \qquad i = 1, 2, \ldots, d,$$

and

$$\Sigma_{ij} = \int_{\mathbf{R}^d} (x_i - M_i)\,(x_j - M_j)\, P(x)\, dx, \qquad i, j = 1, 2, \ldots, d.$$

It follows for any $P \in \mathcal{C}(p)$ that $M = Y^{-T} m$ and

$$\left(Y^T \Sigma Y\right)_{ii} = \sigma_i, \qquad i = 1, 2, \ldots, d.$$

Moreover, for the covariance matrix Σ^* of the multivariate probability density P^* given in Theorem 1 we also have that $\Sigma^* = Y^{-T} D Y^{-1}$.

3 Feature extraction using maximum entropy

Our goal in this section is to use Theorem 1 as a means to isolate desirable features of multivariate data. Thus our point of view here is to take observed data $x^1, x^2, \ldots, x^N$ in $\mathbf{R}^d$ considered to be samples of some unknown probability density function P. We project this data onto d directions, estimate the probability density function of the *univariate* data, find the maximum entropy of all multivariate probability density functions in $\mathbf{R}^d$ with marginals in these directions prescribed by the estimated univariate probability density functions and then *maximize* the entropy further over the choice of these directions. This strategy for optimal feature selection depends on the method used to estimate *univariate* data.

Thus, for univariate data $x = (x_1, x_2, \ldots, x_N)^T \in \mathbf{R}^N$, we suppose $p(\cdot; x)$ is a probability density function on $\mathbf{R}$, which is a good estimator for the random variable from which x is drawn. For a given $y \in \mathbf{R}^d$, we form the vector of our univariate data $X^T y \in \mathbf{R}^N$ 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 $y^1, y^2, \ldots, y^d$ following Theorem 1, our maximum entropy estimator for the unknown probability density function is

$$P^*(x) = \sqrt{\det G(Y)} \prod_{j=1}^{d} p\left(\langle y^j, x\rangle; X^T y^j\right), \qquad x \in \mathbf{R}^d, \tag{3.6}$$

and

$$\mathcal{J}(Y) := H(P^*) = -\frac{1}{2} \log \det G(Y) + \sum_{j=1}^{d} \mathcal{H}(y^j)\ , \tag{3.7}$$

where we define

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

We desire to make the entropy (3.7) as large as possible by varying our choice of vectors $y^1, y^2, \ldots, y^d$. This is not fruitful without additional constraints on the vectors $y^1, y^2, \ldots, y^d$. Indeed, we can have independent vectors which are nearly dependent so that their Gram matrix is nearly singular and so the the first term in (3.7) will become large. Our preference is to choose an *orthonormal basis* so that the corresponding Gram matrix is the identity matrix. We turn to an example of the above which is instructive. Specifically, we have in mind as our univariate estimate for the data vector $x = (x_1, x_2, \ldots, x_N)^T \in \mathbf{R}^N$ the Gaussian density $G(\cdot; \mu(x), \sigma(x))$, where

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

with

$$\mu(x) = \frac{1}{N}\sum_{j=1}^{N} x_j \tag{3.8}$$

and

$$\sigma(x) = \frac{1}{N}\sum_{j=1}^{N} \left(x_j - \mu(x)\right)^2 \ . \tag{3.9}$$

Since

$$H\left(G(\cdot;\mu,\sigma)\right) = \frac{1}{2}\log(2\pi e\sigma),$$

the function to be optimized in this case is

$$\mathcal{J}_G(Y) = -\frac{1}{2}\log\det G(Y) + \frac{d}{2}\log(2\pi e) + \frac{1}{2}\sum_{j=1}^{d}\log\left\langle y^j, Vy^j\right\rangle, \tag{3.10}$$

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

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

where

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

First, we consider the *minimum* $\mathcal{J}_G(Y)$ over *all* $d\times d$ matrices Y. To this end, we introduce the $d\times d$ matrix $W = Y^TVY$, whose elements are given as

$$w_{ij} := \left\langle y^i, Vy^j\right\rangle, \qquad i,j = 1,2,\ldots,d.$$

Since $\det W = \det G\det V$, we conclude that

$$\mathcal{J}_G(Y) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det V + \frac{1}{2}\log\left(\prod_{j=1}^{d} w_{jj}\right) - \frac{1}{2}\log\det W.$$

Recall from Hadamard's inequality that $\det W \le \prod_{j=1}^{d} w_{jj}$ where equality occurs if and only if $w_{ij} = 0$ for all $i\neq j$, $i,j = 1,2,\ldots,d$, see for instance, [9, p.252]. Thus, we obtain that that

$$\mathcal{J}_G(Y) \ge \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det V.$$

Moreover, equality occurs if and only if

$$\left\langle y^i, Vy^j\right\rangle = 0, \qquad i\neq j,\ i,j = 1,2,\ldots,d. \tag{3.11}$$

We suppose that $v^1, v^2, \ldots, v^d$ are orthonormal eigenvectors corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$ with $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$ for V. We let $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ and S be the orthogonal $d \times d$ matrix whose columns are $v^1, v^2, \ldots, v^d$, i.e., $V = S\Lambda S^T$. Then (3.11) says that $Y^T S \Lambda S^T Y$ is a diagonal matrix. The simplest case when this condition holds is the choice $Y = S$. In other words, we have that

$$\min\{\mathcal{J}(Y) \ : \ Y\} = \mathcal{J}_G(S) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\sum_{j=1}^{d}\log\lambda_j \ .$$

Note that we have demonstrated that there exist *orthonormal* vectors at which the minimum is attained. We take a short digression from our study of the function $\mathcal{J}_G$ and elaborate on the connection of the above observation to maximum likelihood estimation.

We start with the multivariate Gaussian

$$G(x;\mu,\Sigma) = (2\pi)^{-d/2}(\det\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right), \qquad x \in \mathbf{R}^d,$$

with mean $\mu \in \mathbf{R}^d$ and covariance Σ (a positive definite $d \times d$ matrix). The likelihood function of the data is

$$L(\mu,\Sigma) = \prod_{i=1}^{N} G\left(x^i;\mu,\Sigma\right)$$

and its maximum $L = \max\{L(\mu,\Sigma) : \mu,\Sigma\}$ is achieved for $\hat{\mu} = \bar{x}$ and $\hat{\Sigma} = V$. Moreover, we have that

$$\log L = -\frac{dN}{2}\log(2\pi e) + \frac{N}{2}\log\det V. \tag{3.12}$$

Following [16], we group the data $x^1, x^2, \ldots, x^N \in \mathbf{R}^d$ into J groups where $1 \leq J \leq N$ using a function $g : \{1, 2, \ldots, N\} \to \{1, 2, \ldots, J\}$. Specifically, the j-th group, $1 \leq j \leq J$ is given by

$$G_j = \{x^i : g(i) = j\}$$

and $|G_j| = |\{x^i : g(i) = j\}| := N_j$. Each group G_j is thought of as samples from a Gaussian with mean μ_j and covariance Σ_j. We transform the data $x^1, x^2, \ldots, x^N$ into data $Y^T x^1, Y^T x^2, \ldots, Y^T x^N$ where Y is a $d \times d$ matrix with $\det Y \neq 0$ so that the first p components of the new data contains all the class information. This means that the means are restricted to have the form

$$\mu_j = \begin{pmatrix} \nu_j \\ \tau \end{pmatrix}, \qquad j = 1, 2, \ldots, J$$

where $\nu_j \in \mathbf{R}^p$ and $\tau \in \mathbf{R}^{(d-p)}$ and the covariance is of the form

$$\Sigma_j = \begin{pmatrix} \Omega_j & 0 \\ 0 & \Lambda \end{pmatrix}, \quad j = 1, 2, \ldots, J$$

where Λ is a $(d-p) \times (d-p)$ positive definite matrix and Ω_j is a $p \times p$ positive definite matrix for $j = 1, 2, \ldots J$. We call this set of parameters $\mathcal{F}$ and choose Y, μ_j, Σ_j, $j = 1, 2, \ldots, J$, to maximize the likelihood function

$$\prod_{j=1}^{N} \left(|\det Y| \; G\left(y^i; \mu_{g(i)}, \Sigma_{g(i)}\right)\right).$$

A judicious use of formula (3.12) computes the maximum likelihood over the means and the covariances. We call this quantity $L(Y)$. To identify it, we introduce the *class means*

$$\bar{x}^j = \frac{1}{N_j} \sum_{i \in G_j} x^i,$$

the *total mean*

$$\bar{x} = \sum_{j=1}^{J} \alpha_j \bar{x}^j,$$

where

$$\alpha_j = \frac{N_j}{N}, \qquad j = 1, 2, \ldots, J$$

the *class covariance*

$$V_j = \frac{1}{N_j} \sum_{i \in G_j} (x^i - \bar{x}^j)(x^i - \bar{x}^j)^T$$

and the *total covariance*

$$V = \sum_{j=1}^{J} \alpha_j V_j.$$

We have that

$$e^d (2\pi)^d L^{\frac{2}{N}}(Y) = \frac{\det G(Y)}{a \prod_{k=1}^{J} b_k^{\alpha_k}},$$

where

$$a := \det \left[\left\langle y^i, V y^j \right\rangle \; : \; i, j = p+1, p+2, \ldots, d \right]$$

and

$$b_k := \det \left[\left\langle y^i, V_k y^j \right\rangle \; : \; i, j = 1, 2, \ldots, p \right].$$

The appearance of the determinant above makes this problem different from the maximum entropy problem considered earlier. However, if we assume that all

the group covariances are *diagonal* then the maximum of the likelihood function can be shown to be

$$e^d(2\pi)^d L^{\frac{2}{N}}(Y) = \frac{\det G(Y)}{\prod_{i=p+1}^{n} \langle y^i, Vy^i\rangle \prod_{k=1}^{J} (\prod_{i=1}^{p} \langle y^i, V_k y^i\rangle)^{\alpha_k}}.$$

In particular, when $J = 1$ and $p = d$ we get that

$$e^d(2\pi)^d L^{\frac{2}{N}}(Y) = \frac{\det G(Y)}{\prod_{i=1}^{d} \langle y^i, Vy^i\rangle}.$$

Consequently, we see that maximizing $L(Y)$ over all matrices Y is the same as minimizing $\mathcal{J}_G(Y)$.

Let us now consider the problem of *maximizing* $\mathcal{J}_G(Y)$ over all $d \times d$ matrices Y which are orthogonal. In this case, we have from (3.10) that

$$\mathcal{J}_G(Y) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\sum_{j=1}^{d} \log\langle y^j, Vy^j\rangle$$

and by the concavity of the log function obtain the inequality

$$\mathcal{J}_G(Y) \leq \frac{d}{2}\log\frac{2\pi e}{d} + \frac{d}{2}\log\sum_{j=1}^{d}\lambda_j \tag{3.13}$$

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

$$\langle y^j, Vy^j\rangle = \frac{1}{d}\sum_{i=1}^{d}\lambda_i. \tag{3.14}$$

In other words, the matrix W has *constant* diagonal elements. To ensure that the bound in (3.13) is sharp we need to establish the existence of an orthogonal matrix Y such that (3.14) is satisfied. Our first observation in this regard is the following fact. Suppose H is a $d \times d$ *Hadamard matrix*, that is, all the entries of H are either 1 or -1 and satisfies the condition that

$$HH^T = dI.$$

We set $J = d^{-\frac{1}{2}}H$ and $Y = SJ$ so that Y is orthonormal. Since $V = S\Lambda S^T$ we have, for $i = 1, 2, \ldots, d$, that

$$\langle y^i, Vy^i\rangle = \sum_{j=1}^{d} J_{ij}^2\lambda_j = \frac{1}{d}\sum_{j=1}^{d}\lambda_j,$$

which is the desired equation (3.14). Hadamard matrices *do not* exist for all orders, see for example [10, chapter 14]. When they do exist, the above remark

tells us how to construct an orthonormal matrix from the eigenvalues of V which maximizes the function $\mathcal{J}_G$.

Returning to the general case, we describe inductive proof for equation (3.14). We start with an arbitrary real symmetric matrix A and assume that when A has order $k < d$ there exist orthonormal vectors $y^1, y^2, \ldots, y^k$, such that

$$\left\langle y^i, Ay^i \right\rangle = \frac{1}{k}\text{trace } A, \qquad i = 1, 2, \ldots, k.$$

The case $k = 2$ is proved by using the Hadamard matrix

$$H = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Now, let A be of order d. By the mean value theorem, there is a vector $x \in \mathbf{R}^d$ with $\langle x, x \rangle = 1$ such that

$$\langle x, Ax \rangle = \frac{1}{d}\text{trace } A.$$

We let X be *any* orthogonal matrix whose first column is x. Hence the matrix $X^T AX$ has the property that $(X^T AX)_{11} = \frac{1}{d}\text{trace } A$. In particular, we can write it in the form

$$(X^T AX) = \begin{bmatrix} \frac{1}{d}\text{trace } A & w^T \\ w & B \end{bmatrix}$$

for some $w \in \mathbf{R}^{d-1}$, where B is a $(d-1) \times (d-1)$ symmetric matrix. Consequently, we have the following formula

$$\frac{1}{d-1}\text{trace } B = \frac{1}{d}\text{trace } A.$$

By the induction hypothesis, there exists a $(d-1)\times(d-1)$ orthonormal matrix Z such that, for $i = 1, 2, \ldots, d-1$, $(Z^T BZ)_{ii} = \frac{1}{d-1}\text{trace } B$. Then the orthogonal matrix Y defined as

$$Y = X \begin{bmatrix} 1 & 0^T \\ 0 & Z \end{bmatrix}$$

where $0 \in \mathbf{R}^{d-1}$ (the zero vector), advances the induction hypothesis.

Another way to prove this result is to recall a theorem of A. Horn [19, p. 220] that gives necessary and sufficient conditions for the existence of a symmetric matrix eigenvalues and diagonal elements. To make use of it we note that the $d\times d$ doubly stochastic matrix P all of whose elements are $\frac{1}{d}$ has the property that $P\lambda = \bar{\lambda}$, where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_d)^T$ and $\bar{\lambda} = \left(\frac{1}{d}\sum_{i=1}^d \lambda_i, \frac{1}{d}\sum_{i=1}^d \lambda_i, \ldots, \frac{1}{d}\sum_{i=1}^d \lambda_i\right)^T$. Hence, there is a symmetric matrix B such that $B_{ii} = \frac{1}{d}\text{trace } A$, $i = 1, 2, \ldots, d$ and the eigenvalues of B are $\lambda_1, \lambda_2, \ldots, \lambda_d$. We write B in the form

$$B = R^T \Lambda R$$

where R is an orthogonal matrix. Thus, the matrix $Y = SR$ is orthogonal and its columns have the desired property (3.14). The above proofs were pointed out to us by Alan J. Hoffman [13].

In summary, we have proved that

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

4 Feature selection using maximum entropy and mixtures of Gaussians

In the last section we showed how to identify an orthonormal set of directions which maximizes entropy when we use a *single* Gaussian to model univariate data. In this section we discuss this problem when we use a *mixture* model of Gaussians to model univariate data. In this case, which is of greater practical importance, we are unable to analytically solve the problem and so only discuss possible numerical algorithms for its solution. We begin by recalling that a probability density which is a mixture model of m Gaussians has the form

$$p\left(t \ ; \ \mu, \omega, \sigma\right) = \sum_{j=1}^{m} \omega_j \, G\left(t; \mu_j, \sigma_j\right) \qquad t \in \mathbf{R}, \tag{4.15}$$

where the mixture weights satisfy the constraints

$$\sum_{j=1}^{m} \omega_j = 1, \quad \omega_i \geq 0, \qquad i = 1, 2, \ldots, m. \tag{4.16}$$

For any choice of parameters μ, ω and σ, the entropy of the probability density in (4.15) can easily be computed by a Monte Carlo method using random samples drawn from this density.

To choose the parameters of the probability density in (4.15) as a function of the data vector $(x_1, x_2, \ldots, x_N)^T \in \mathbf{R}^N$ we consider the likelihood function

$$\prod_{i=1}^{N} p\left(x_i; \mu, \omega, \sigma\right).$$

It is well known that for $m \geq 2$ the likelihood function will not have a maximum, see for example [21]. Therefore, we turn to the EM algorithm to *iteratively* set the means and variances. For this purpose, we fix the number of iterations in the EM update formulas appearing below *independent* of the data vector x. The first step is the initialization of the means and variances. This is done by *clustering* the coordinates of the data vector into m groups of (approximately) equal size. On each of the m clusters we use the formulas (3.8) and (3.9) for

the mean and the variances. We then update them by the EM formulas, (see, for instance, [4]), which produces a nondecreasing sequence of likelihoods. One step of the update formulas is given by

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

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

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

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

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

With these formulas we use a *fixed* number of iterations to set the values ω, μ and σ and denote by $p(\cdot; x)$ our univariate estimate for the data vector $x \in \mathbf{R}^N$ by mixture models of Gaussians.

The next step is to iteratively update the orthogonal matrix Y. Our iteration is initialized per our previous discussion of the single Gaussian model. To update the matrix Y we use two dimensional *planar rotations*. Thus, we choose *any* two consecutive vectors y^j, y^{j+1} and form new vectors

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

where $\theta \in [0, 2\pi]$. We then optimize the univariate function

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

over $\theta \in [0, 2\pi)$ and replace the vectors $y^1, y^2, \ldots y^d$ by $y^1(\hat{\theta}), y^2(\hat{\theta}), \ldots y^d(\hat{\theta})$ where $\hat{\theta}$ is a good choice for the maximum of the function F. If $j < d-1$, we increment j by one and repeat the procedure otherwise we set $j = 1$ and again repeat the procedure continually cycling through the columns of Y as needed.

5 Feature selection using maximum likelihood and Parzen estimators

In Section 4 we considered modeling our multivariate data by using the maximal entropy estimator (3.6) with univariate marginals constructed from mixture

models of Gaussians. In this section, we study the possibility of using Parzen estimators for our univariate data and maximum likelihood as a feature selection criterion instead of maximum entropy. Starting with a bin width h and a univariate kernel K, the corresponding Parzen estimator for $x = (x_1, x_2, \ldots, x_N)^T$ is given by

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

The choice of the bin width is discussed in [22]. In what follows, we choose the kernel K to be the standard Gaussian $K = G(\cdot; 0, 1)$. 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

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

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

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

where

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

We resort to an EM strategy and develop an iterative update formula that increases $L(Y)$. 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 consider the equation

$$L(Y) - L(\hat{Y}) = \sum_{i=1}^{N} \sum_{j=1}^{d} \log\left(\sum_{k=1}^{N} c_{kij} a_k\right). \tag{5.17}$$

where

$$a_k = \frac{K(\langle y^j, g^{ik}\rangle)}{K(\langle \hat{y}^j, g^{ik}\rangle)} \tag{5.18}$$

and

$$c_{kij} = \frac{K(\langle \hat{y}^j, g^{ik} h\rangle)}{\sum_{m=1}^{N} K(\langle \hat{y}^j, g^{im} h\rangle)}. \tag{5.19}$$

Substituting (5.19) and (5.18) into (5.17) and using the concavity of the log function as well as the form of the kernel K we get

$$L(Y) - L(\hat{Y}) \quad \geq \quad \sum_{i=1}^{N} \sum_{j=1}^{d} \sum_{k=1}^{N} c_{kij} \log a_k$$

$$= \quad Q(Y) - Q(\hat{Y})$$

where

$$Q(Y) := -\frac{1}{2}\sum_{j=1}^{d}\langle y^j, A_j y^j\rangle$$

and

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

The task is now to compute the minimum of $\sum_{j=1}^{d}\langle y^j, A_j y^j\rangle$ over all orthonormal matrices Y. We approach this problem as in Section 4 by using planar rotations. In the case at hand, our angle θ can be *determined explicitly.* To this end, we consider 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 simplifies to

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

where

$$B = \frac{1}{2}\left(A_1 - A_2\right),$$

$$C = \frac{1}{2}\left(A_1 + A_2\right)$$

$$\beta = \langle y^1, By^1\rangle - \langle y^2, By^2\rangle$$

$$\alpha = -2\langle y^1, By^2\rangle$$

$$\gamma = \langle y^1, Cy^1\rangle + \langle y^2, Cy^2\rangle$$

and its minimum occurs at

$$\hat{\theta} = k\pi \pm \frac{\psi}{2}$$

where k is an integer and

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

6 Feature selection using maximum likelihood Gaussian mixture estimators

In this section, we discuss optimal feature selection for the data $x^1, x^2, \ldots, x^N \in \mathbf{R}^d$ by using (different) univariate Gaussian mixtures for the marginals in the maximum entropy estimator of Theorem 1 with a maximum likelihood criterion. Given d orthonormal vectors $y^1, y^2, \ldots, y^d \in \mathbf{R}^d$, the log–likelihood function of interest to us here is

$$L\left(Y, \omega, \mu, \sigma\right) = \sum_{i=1}^{N} \sum_{j=1}^{d} \log \left(\sum_{\ell=1}^{m} \omega_{\ell j} G\left(\left\langle y^j, x^i \right\rangle \ ; \ \mu_{\ell j}, \sigma_{\ell j} \right) \right), \tag{6.20}$$

where

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

We describe a method to iteratively increase L. First, we fix Y and point out how to update ω, μ and σ to increase L.

We let the parameters $\hat{\omega}$, $\hat{\mu}$ and $\hat{\sigma}$ be the initial guess. We shall identify ω, μ and σ such that $L(Y, \omega, \mu, \sigma) \geq L(Y, \hat{\omega}, \hat{\mu}, \hat{\sigma})$. To this end, we consider the equation

$$L(Y, \omega, \mu, \sigma) - L(Y, \hat{\omega}, \hat{\mu}, \hat{\sigma}) = \sum_{i=1}^{N} \sum_{j=1}^{d} \log \left(\sum_{\ell=1}^{m} \alpha_{ij\ell} \beta_{ij\ell} \right), \tag{6.21}$$

where

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

and

$$\beta_{ij\ell} = \frac{\omega_{\ell j} G\left(\left\langle y^j, x^i \right\rangle \ ; \ \mu_{\ell j}, \sigma_{\ell j} \right)}{\hat{\omega}_{\ell j} G\left(\left\langle y^j, x^i \right\rangle \ ; \ \hat{\mu}_{\ell j}, \hat{\sigma}_{\ell j} \right)}.$$

Note that, by definition, $\sum_{\ell=1}^{m} \alpha_{ij\ell} = 1$. Thus, using the concavity of the log function in (6.21) we obtain that

$$\begin{aligned} L(Y, \omega, \mu, \sigma) - L(Y, \hat{\omega}, \hat{\mu}, \hat{\sigma}) &\geq \sum_{i=1}^{N} \sum_{j=1}^{d} \sum_{\ell=1}^{m} \alpha_{ij\ell} \log \beta_{ij\ell} \\ &= Q(Y, \omega, \mu, \sigma) - Q(Y, \hat{\omega}, \hat{\mu}, \hat{\sigma}), \end{aligned}$$

where

$$Q(Y, \omega, \mu, \sigma) = \sum_{i=1}^{N} \sum_{j=1}^{d} \sum_{\ell=1}^{m} \alpha_{ij\ell} \log \left(\omega_{\ell j} G\left(\left\langle y^j, x^i \right\rangle \ ; \ \mu_{\ell j}, \sigma_{\ell j} \right) \right). \tag{6.22}$$

We can *explicitly* maximize $Q(Y,\omega,\mu,\sigma)$ with respect to ω, μ, σ to get the equations

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

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

and

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

These formulas provide an effective means for updating the parameters ω, μ and σ. We will now consider an update formula for 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 $\nu_{j\ell} = \langle y^j, \mu^\ell \rangle$

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

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

$$L(Y,\omega,\mu,\sigma) - L(\hat{Y},\omega,\mu,\sigma) \geq \sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{\ell=1}^{m} \bar{\alpha}_{ij\ell} \log \bar{\beta}_{ij\ell},$$

where

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

and

$$\bar{\beta}_{ij\ell} = \frac{\omega_{\ell j} G\left(\left\langle y^j, x^i\right\rangle \ ; \ \left\langle y^j, \mu^\ell\right\rangle, \sigma_{\ell j}\right)}{\omega_{\ell j} G\left(\left\langle \hat{y}^j, x^i\right\rangle \ ; \ \left\langle \hat{y}^j, \mu^\ell\right\rangle, \sigma_{\ell j}\right)} \ .$$

Hence, we conclude that $L(Y,\omega,\mu,\sigma) - L(\hat{Y},\omega,\mu,\sigma) \geq Q(Y,\omega,\mu,\sigma) - Q(\hat{Y},\omega,\mu,\sigma)$, where

$$\begin{aligned} Q(Y,\omega,\mu,\sigma) &= \sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{\ell=1}^{m} \bar{\alpha}_{ij\ell} \log\left(\omega_{\ell j} G\left(\left\langle y^j, x^i\right\rangle \ ; \ \left\langle y^j, \mu^\ell\right\rangle, \sigma_{\ell j}\right)\right) \\ &= -\frac{1}{2}\sum_{j=1}^{d} \langle y^j, A_j y^j\rangle + Q_C \end{aligned}$$

with

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

and

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

Since Q_C is independent of Y, we can use the methods described in Section 5 to increase the function Q.

7 The problem of optimizing Q as a function of Y

In view of the discussion of the previous section, the essential computational problem boils down to solving the following nonlinear optimization problem (minimize Q for maximum entropy problem and maximize Q for maximum likelihood problem).

$$\min_{\{Y | Y^T Y = I\}} \{F(Y) = \sum_{i=1}^{d} \left\langle y^i, A_i y^i \right\rangle\}; \quad Y = [y^1, y^2, \ldots, y^d] \in \mathbf{R}^{d \times d}. \tag{7.23}$$

In the present section we suggest two different algorithms for solving this optimization problem.

7.1 2D planar rotations

We note that all orthogonal matrices of size d can be written as a product of $\frac{1}{2}d(d-1)$ planar rotations (see e.g., [12], [32, 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 $F(Y)$ 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 $\hat{\theta}_{\min}$ s. t. $F(\hat{\theta}_{\min}) = \min \left\{ F(\theta) = \left\langle y^j(\theta), A_j y^j(\theta) \right\rangle + \left\langle y^k(\theta), A_k y^k(\theta) \right\rangle ; \theta \in [0, 2\pi] \right\}$

- Replace θ with $\hat{\theta}_{\min}$ 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 $\frac{d(d-1)}{2}$ planar rotations in each loop. We can simplify $F(\theta)$ by writing

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

where

$$\begin{aligned} \alpha &= -2\left\langle y^j, By^k \right\rangle \\ \beta &= \left\langle y^j, By^j \right\rangle - \left\langle y^k, By^k \right\rangle \\ \gamma &= \left\langle y^j, Cy^j \right\rangle + \left\langle y^k, Cy^k \right\rangle \\ B &= \frac{1}{2}(A_j - A_k) \\ C &= \frac{1}{2}(A_j + A_k). \end{aligned}$$

By simple differentiation of $F(\theta)$ we obtain the extremal points

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

We are interested in the minimizers of $F(\theta)$. For this, we observe that $F(\theta)$ can be rewritten as

$$F(\theta) = \begin{bmatrix} \beta & \alpha \end{bmatrix} \begin{bmatrix} \cos 2\theta \\ \sin 2\theta \end{bmatrix}$$

so that the minimum of $F(\theta)$ is $-(\beta^2 + \alpha^2)^{\frac{1}{2}}$ and is achieved when $\begin{bmatrix} \cos 2\theta \\ \sin 2\theta \end{bmatrix}$ points to the opposite direction of $\begin{bmatrix} \beta & \alpha \end{bmatrix}$. That is, if $\beta + i\alpha = \exp(i\psi_0)$ then $\theta = -\psi_0/2 + k\pi$ is a minimizing solution.

7.2 Transformation to unconstrained form

As an alternative to the numerical optimization strategy discussed in the previous section, we consider using conjugate gradient strategies to directly search for the local maxima of the problem. However, the large number of constraint equations make it impractical for us to use standard optimization packages. We, therefore, consider a parameterization of orthogonal matrices via skew-symmetric matrices. This strategy exposes all the variables without any constraint equations, thus enabling use of unconstrained gradient ascent methods.

It is easy to see that if S is a skew-symmetric matrix, then e^S is an orthogonal matrix with determinant 1. A partial converse of this fact is that if Y is a real valued orthogonal matrix with determinant 1, then there is a real skew-symmetric matrix S, such that $Y = e^S$. Since this latter fact, albeit known in the theory of Lie Groups, is not so well known in statistical literature, we provide an elementary matrix theoretic proof in the Appendix. Our current intent is to exploit this fact so that constrained optimization problem (7.23) is transformed into an unconstrained one. We begin with a formal statement of the fact just stated.

Fact 1 *If Y is a unitary matrix with $\det Y = 1$, then there exists a skew-Hermitian matrix S such that*

$$Y = e^S$$

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

Note that the representation of Y in terms of S is not unique. We have e.g., $\exp \begin{bmatrix} 2\pi k & 0 \\ 0 & 2\pi k \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. For small matrices S explicitly compute e^S. This is done by using the Cayley-Hamilton theorem [12] to compute S^k recursively from lower powers and then gathering terms carefully. For $m = 2, 3, 4$ we have

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

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

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

Differentiation of e^S with respect to the exposed variables is a simple matter and can now be done analytically such that a numerical optimization package can be easily used. For $m > 5$ this procedure becomes tedious and highly impractical. As an alternative we discuss numerical procedures for computing e^S and $\frac{\partial e^S}{\partial s_{ij}}$. To compute the derivative $\frac{\partial e^S}{\partial s_{ij}}$ efficiently we use the following result [23, 31, 24]:

Proposition 1 *If $S = [s_{ij}]$ is a skew-symmetric matrix such that $s_{ii} = 0$ and $s_{ij} = -s_{ij}$ we have*

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

for $p = 1, 2, \ldots$. Furthermore, $\frac{\partial S^p}{\partial s_{ij}}$ can be efficiently computed using the recursive equation:

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

for $p = 1, 2, \ldots$.

Proof: The proposition follows as a simple application of the product rule for matrix differentiation [2]. ■

If the eigenvalues of S are sufficiently small the Taylor series for $e^S = \sum_{k=0}^{\infty} \frac{S^k}{k!}$ converges rapidly and this can be exploited to compute approximations to e^S and $\frac{\partial e^S}{\partial s_{ij}}$. For the case when the eigenvalues of S are large or simply in order to accelerate the convergence rate of the Taylor series we use the relation

$$e^S = \left(e^{\frac{1}{\alpha}S}\right)^\alpha = \left(\sum_{k=0}^{\infty} \frac{S^k}{\alpha^k k!}\right)^\alpha$$

with the choice $\alpha = 2^n$ convenient for rapid computation of e^S. For rapid computation of $\frac{\partial e^S}{\partial s_{ij}}$ we may use $e^S = M^\alpha$ for $M = e^{S/\alpha}$, and the product rule for matrix differentiation

$$\frac{\partial e^S}{\partial s_{ij}} = \sum_{k=1}^{\infty} M^{k-1}\frac{\partial M}{\partial s_{ij}}M^{\alpha-k}$$

then follows Section 7.2. Recalling the fact that since $S = [s_{ij}]$ is skew-symmetric we have $s_{ij} = -s_{ji}$, and in particular, $s_{ij} = 0$ if $i = j$, we define the vector $v(S) \in \mathbf{R}^{\frac{1}{2}d(d-1)}$ from elements of S as:

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

We may now transform (7.23) into the unconstrained optimization problem of finding v^* so that

$$F_2(v^*) = \min_v \{F_2(v(S))\}, \quad \text{where} \quad F_2(v) = F_1(S).$$

By using results of Proposition 1 we may now compute the gradient $\nabla_{v(S)} F_2(v(S))$ of F_2 with respect to $v(S)$. In view of this, the following algorithm is suggested:

Algorithm 2

- Initialize $S = \log Y$ and form $v = v(S)$.
- Compute $\frac{\partial e^S}{\partial s_{ij}}$ by using formulas (**??**) and (**??**), and subsequently compute $\nabla_{v(S)} F_2(v(S))$.
- Find the optimal v^* by using a iterative Conjugate Gradient update strategy such as: $v_{\text{update}}(S) = v(S) - \rho \nabla_{v(S)} F_2(v(S))$, where the constant ρ is chosen appropriately.
- Finally, rearrange elements of $v^*(S)$ to get S^*, and subsequently compute $Y^* = e^{S^*}$.

Alternatives to Conjugate Gradient method can be used for minimization and may be more suitable for large scale problems. The algorithm was implemented in Matlab using the function 'fminunc' from the Optimization Toolbox, which uses a method involving preconditioned conjugate gradients [11, 6, 7].

7.3 Special types of A_i's

We first consider the case when all the matrices A_i's in (7.23) are diagonal i.e., for all $i = 1$ to d we have $A_i = \text{diag}\,[a_{i1}, a_{i2}, \cdots a_{id}]$. In this case the problem of optimizing $F(Y)$ becomes a linear programming problem and, as discussed below, its solution has interesting properties. The function $F(Y)$ in (7.23) can be written as:

$$F(Y) = \sum_{i=1}^{d} \langle y^i, A_i y^i \rangle = \sum_{i=1}^{d} \sum_{j=1}^{d} (y_j^i)^2 a_{ij}. \tag{7.24}$$

A square matrix, $S = [s_{ij}]$ of size $m \times m$, is called *doubly stochastic* if $\sum_k s_{kj} = \sum_k s_{jk} = 1$ for $i, j = 1, 2, \cdots, m$ and $s_{ij} \geq 0$. For our ensuing discussions it would be of interest to note the following fact attributed to Birkhoff [3, 25, 26, 27, 28]. *The set $\mathcal{S}$ of all $n \times n$ stochastic matrices is the convex hull of the set $\mathcal{P}$ of $n \times n$ permutation matrices.*

Proposition 2 *A minimum of $F(Y)$ is obtained when $[(y_j^i)^2]$ is a permutation matrix.*

Proof: Observe that since Y is an orthogonal matrix, $S = [s_{ij}]$, where $s_{ij} = (y_i^i)^2$, is a doubly stochastic matrix. It follows from standard results in the theory of linear programming [29, 20] that since F is linear in s_{ij}, the minimum of $F(Y)$ is attained at a vertex of the polyhedral set:

$$\{s_{ij};\ \sum_{i=1}^{d} s_{ij} = 1, \sum_{j=1}^{d} s_{ij} = 1, s_{ij} \geq 0\}. \tag{7.25}$$

To determine a vertex in a d^2 dimensional space, at least d^2 constraints are needed. There are $(2d-1)$ linearly independent constraints of the form $\sum_{i=1}^{d} s_{ij} = \sum_{j=1}^{d} s_{ij} = 1$, d^2 constraints of the form $s_{ij} \geq 0$ and since the solution is a vertex of the region defined in (7.25), we have at least $(d^2 - (2d-1))$ pairs (i,j) such that $s_{ij} = 0$. Therefore, there are at most $2d-1$ nonzero elements in $S = [s_{ij}]$. Now, there is at least one row (or column) that has at most one non-zero element, because if all of them had 2 or more elements we would then need $2d$ or more elements to be nonzero. Assume that this element is s_{kl}. Since S is a doubly stochastic matrix we have $s_{kl} = 1$, with the rest of the elements in the k-th row and l-th column equal zero.

Next, define $S^{(d)} = S$, and $S^{(d-1)}$ to be the sub-matrix obtained by deleting the k-th row and l-th column of S. We can now rewrite the $F(S)$ as:

$$F(S) = F(S^{(d)}) = F(S^{(d-1)}) - d_{kl}, \quad \text{where} \quad F(S^{(d-1)}) = -\sum_{\substack{i=1 \\ i \neq k}}^{d} \sum_{\substack{j=1 \\ j \neq l}}^{d} (y_j^i)^2 d_{ij}.$$

Since $S^{(d-1)}$ is a stochastic matrix and $F(S^{(d-1)})$ has the same structure as $F(S^{(d)})$, we can repeat the same argument with $S^{(d-1)}$ and subsequently, with $S^{(d-2)}, \ldots, S^{(1)}$. It then follows directly that the minimizer of $F(S)$, $S_{\min}$, is a permutation matrix. ∎

Proposition 2 implies that a minimum of $F(Y)$ subject to the constraint $Y^T Y = I$ is obtained when $[y_{ij}^2]$ is a permutation matrix. This elegant result, provided to us by Alan Hoffman [13, 14], also provides us with a criterion for checking the correctness of any implemented algorithm for solving the optimization problem (7.23).

We now consider the slightly generalized situation when the matrices A_i's are simultaneously diagonalizable by an orthogonal matrix M i.e., for all i, $A_i = M D_i M^T$, where D_i's are each diagonal matrices. We then have:

Proposition 3 *If for all i we have $A_i = M D_i M^T$, where $D_i \in \mathbf{R}^{d \times d}$ are diagonal matrices, $M \in \mathbf{R}^{d \times d}$ is an orthogonal matrix then the solution of (7.23) is on the form $Y = M^T Z$, where $Z = [(z_j^i)^2]$ is a permutation matrix.*

Proof: Note that $\langle y^i, A_i y^i \rangle = \langle z^i, D_i z^i \rangle$ for all i, where $z^i = M^T y^i$. Hence the proof follows directly from Proposition 2. ∎

As before, the above result gives us another way to check the correctness of the algorithm for solving (7.23).

7.4 A lower bound of $F(Y)$

Yet another approach for checking the robustness of an algorithm for minimizing $F(Y)$ is to obtain a lower bound for $\min\{F(Y)\}$, which can be used for comparison. To this end we consider a problem very close to (7.23):

$$\min_{\{Y|\|y^i\|_2^2=1\}}\{F(Y)=\sum_{i=1}^{d}\left\langle y^i, A_i y^i\right\rangle\}; \quad Y=[y^1,y^2,\ldots,y^d]\in\mathbf{R}^{d\times d}. \tag{7.26}$$

Note that in this problem Y need not be an orthogonal matrix, and the only condition is that the vectors $y^1, y^2, \ldots, y^d$ are normalized to have unit length. The following result gives us the means to find a lower bound for $F(Y)$ corresponding to (7.26):

Proposition 4 *For real symmetric matrices $A_1, A_2, \ldots, A_d \in \mathbf{R}^{d\times d}$ and unit length vectors $y^1, y^2, \ldots, y^d \in \mathbf{R}^d$ we have that:*

$$\min_{\{Y|\|y^i\|_2^2=1\}}\{F(Y)=\sum_{i=1}^{d}\left\langle y^i, A_i y^i\right\rangle\}=\sum_{i=1}^{d}\lambda_{\min}^i,$$

where $\lambda_{\min}^i$ is the minimal eigenvalue of A_i for all i. Moreover, this value is achieved for

$$Y=[\pm e_{\min}^1, \pm e_{\min}^2, \ldots, \pm e_{\min}^d],$$

where $e_{\min}^i$ is the (normalized) eigenvector of A^i corresponding to $\lambda_{\min}^i$ for all i.

Proof: Since the matrices A_i's are real symmetric we have,

$$A_i = Q_i \Lambda_i Q_i^T,$$

where Q_i's are orthogonal matrices containing the orthonormal eigenvectors of A_i, and Λ_i is a diagonal matrix containing the corresponding eigenvalues on the diagonal arranged in *descending order*. Thus, by defining unit length vectors $z^i = Q_i^T y^i$ we can write:

$$\begin{aligned}\left\langle y^i, A_i y^i\right\rangle &= \left\langle z^i, A_i z^i\right\rangle \\ &\geq \lambda_d^i \|z^i\|_2^2 \\ &= \lambda_d^i.\end{aligned}$$

We clearly see that

$$\min_{\{Y:\|y^i\|_2^2=1\}}\left\langle y^i, A_i y^i\right\rangle = \lambda_d^i = \lambda_{\min}^i,$$

It follows that $z^i_{\min} = \arg\min\{\langle z^i, A_i z^i\rangle\} = \pm[0,\ldots,0,1]^T$, which is obtained when $y^i = Q_i z^i_{\min} = \pm e^i_d = \pm e^i_{\min}$. Hence the result is proven. ■

From the discussion above we note that the lower bound given in the above result is achieved for (7.23) in the special case when $e^1_{\min}, e^2_{\min}, \ldots, e^d_{\min}$ are mutually orthonormal. When $e_{\min}$ are orthogonal this gives an exact lower bound for (7.23) .

7.5 Numerical Results

The algorithms given in Sections 7.1 and 7.2 for minimizing $F(Y)$ 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 the following three examples to check the correctness of our algorithms: First we consider an example where the algorithms should reach the lower bound discussed in Section 7.4:

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, it follows directly from the discussion in Section 7.4 that the global minimum of (7.23), for the given A_i's is achieved by the orthogonal matrix

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

The Planar Rotation Algorithm yielded the following solution after 12 iterations:

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

whereas unconstrained optimization formulation Algorithm 2 described in Section 7.2 converged after 6 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 when computing matrix logarithms and exponentials. Also, it is easily verified that $F(Y^*) = F(\hat{Y}_1^*) = 6$ (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 section 7.3:

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

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

The Planar Rotation Algorithm 1 converges after 2 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 in one iteration i.e., we have $\hat{Y}_2^* = I_{(5\times 5)}$, this being the case because the identity matrix is already a local extremum, which we chose as an initial starting point. Since we have $F(\hat{Y}_1^*) < F(\hat{Y}_2^*)$, in this case, $\hat{Y}_2^*$ is a local minimizer. Interestingly, we note that both of the minimal Y's are permutation matrices. This is consistent with the results of Section 7.3, which asserts that in the case when all A_i's are diagonal the local minima of the optimization problem (7.23) are permutation matrices.

Example 3 Consider $A_i = MD_iM^T$ 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 where orthogonal matrix M is given by:

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

In this case, Planar Rotation Algorithm 1 converged to the following matrix after 4 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, $F(\hat{Y}_1^*) = 6$, which by Proposition 4, is the global minimum. Thus, Algorithm 1 gives us the global minimum in this case as well. Although Algorithm 2 used 65 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 $F(\hat{Y}_2^*) = 6.0001 \approx F(Y^*)$ one concludes that the unconstrained optimization based Algorithm 2 converges to the correct solution as well.

Comments: The above examples and other experiments conducted indicate that in general Algorithm 1 seem more efficient than Algorithm 2, in terms of compute time. Furthermore, the algorithms in some cases converged to different points, according to the chosen starting points, which indicates that the function $F(Y)$ has multiple local minima.

Experiments with maximum entropy model: In this section, we seek validation of a our maximum entropy method for estimation for the multivariate probability distribution $P(x)$ corresponding to equation (6.20). where the parameters are determined by the following two-step algorithm:

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

This algorithm was implemented using the planar rotation scheme from Section 7.1 to update Y in each step (the unconstrained optimization method via transformation to skew-symmetric matrices could also be used as an alternative, but was not pursued here). Moreover, the Gaussian Mixture Model [15] was also implemented for the sake of comparison. Then the algorithms were tested for the two dimensional cases stated below:

Case 1: In this case, 5000 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. [1] The results shown in figure **??**, illustrates that both models estimate the true distribution fairly well.

Case 2: In order to examine the behavior of our models in a 'difficult' case, the two most correlated dimensions (of 60) were found. Here, two-dimensional projections of high dimensional acoustic feature vectors obtained from the same subphonetic speech unit wee used. 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 **??** displays the plots of the densities obtained via the two methods and the histogram of the data as well. Here, an inherent weakness of the Maximum Entropy Model is revealed, namely, when dealing with spherically nonsymmetric distributions, the model smooths out the structure to depict an elliptic one. This is due to the way the model is built in the first place; 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 **??**.

8 Conclusion

We have considered a number of problems associated with estimation of probability densities from projection of data on lower dimensional subspaces. Maxi-

[1] details of mixture gaussian estimation not given e.g., how many mixture components etc.

mum entropy and maximum likelihood criteria are used to exploit the nonuniqueness of the problem of reconstruction from lower dimensional slices. 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 leading to the main problem of estimating densities are developed. Simple numerical examples are worked out to demonstrate the method.

Appendix

Proof of Fact 1: Since Y is unitary, and therefore, also a normal matrix, we can write [30]:

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

where P is a unitary matrix and $\Lambda = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ is a diagonal matrix containing the eigenvalues of Y. Since P is unitary $|\lambda_i| = 1$ i.e., $\lambda_i = \exp(\imath\theta_i)$ for all i, with $\theta_i \in \mathbf{R}$. Thus, by further exploiting the unitary property of P we may write $Y = e^S$, where $S = P\mathrm{diag}[\theta_1, \theta_2, \cdots, \theta_m]P^*$.

Additionally, if $Y \in \mathbf{R}^{m\times m}$, 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) as: $\lambda_i = \bar{\lambda}_{i+1}$ 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 $\det Y = \det \Lambda = 1$, the number of λ_i's that equals -1 are necessarily even. Thus, we can write:

$$\Lambda = \mathrm{diag}\,[\Lambda_1^c, \ldots, \Lambda_n^c, \Lambda_1^\pi, \ldots, \Lambda_s^\pi, \Lambda_1^p, \ldots, \Lambda_r^p,]$$

where for all i and with nonzero $\theta_i \in \mathbf{R}$,

$$\Lambda_i^c = \exp\begin{bmatrix} \imath\theta_i & 0 \\ 0 & -\imath\theta_i \end{bmatrix}, \quad \Lambda_i^\pi = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} = \exp\begin{bmatrix} 0 & \pi \\ -\pi & 0 \end{bmatrix}, \quad \text{and} \quad \Lambda_i^p = [1].$$

We then have

$$\Lambda = \exp\tilde{\Lambda}$$

where

$$\tilde{\Lambda} = \mathrm{diag}\left[\tilde{\Lambda}_1^c, \ldots, \tilde{\Lambda}_n^c, \tilde{\Lambda}_1^\pi, \ldots, \tilde{\Lambda}_s^\pi, \tilde{\Lambda}_1^p, \ldots, \tilde{\Lambda}_r^p\right],$$

with

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

Therefore, by invoking the unitarity of P we can write

$$Y = Pe^{\tilde{\Lambda}}P^* = e^S, \quad \text{where} \quad S = P\tilde{\Lambda}P^*$$

To this end, we note the relation

$$\begin{bmatrix} \imath\theta & 0 \\ 0 & -\imath\theta \end{bmatrix} = \mathcal{I}_c \begin{bmatrix} 0 & \theta \\ -\theta & 0 \end{bmatrix} \mathcal{I}_c^*$$

where $\mathcal{I}_c = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -\imath \\ 1 & \imath \end{bmatrix}$. Thus, we may write

$$S = Q\tilde{\Lambda}^r Q^*,$$

where

$$\tilde{\Lambda}^r = \text{diag}\left[\tilde{\Lambda}_1^r, \ldots, \tilde{\Lambda}_n^r, \tilde{\Lambda}_1^\pi, \ldots, \tilde{\Lambda}_s^\pi, \tilde{\Lambda}_1^p, \ldots, \tilde{\Lambda}_r^p\right]$$

with

$$Q = P\mathcal{I} \text{ and } \mathcal{I} = \text{diag}[\underbrace{\mathcal{I}_c, \ldots, \mathcal{I}_c}_{n \text{ copies}}, \underbrace{1, \ldots, 1}_{(2s+r) \text{ copies}}]$$

Since clearly $\tilde{\Lambda}^r \in \mathbf{R}^{m\times m}$, the proof of $S = Q\tilde{\Lambda}^r Q^* \in \mathbf{R}^{m\times m}$ will be complete if we can show that $Q \in \mathbf{R}^{m\times m}$, which we undertake next. For this purpose let us identify the eigenvectors p_i of Y as:

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

and note that due to the particular order of eigenvalues chosen we have $p_i = \bar{p}_{i+1}$ for $i = 1, 3, \ldots, (2n-1)$, and p_i's are real for $i = 2n+1, 2n+2 \ldots, m$. Invoking this fact it follows from $Q = P\mathcal{I}$ that

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

Now since $Q \in \mathbf{R}^{m\times m}$, 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. ■

References

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

[2] S. Basu, C.A. Micchelli and Peder Olsen, Power exponential densities for the training and classification of acoustic feature vectors in speech recognition, to appear Journal for Computational and Graphical Statistics, 1998.

[3] G. Birkhoff, Tres observations sobre el algebra lineal, Rev. univ. nac. Tucumán, ser. A, vol.5, pp. 147-151, 1946.

[4] C. M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, 1995.

[5] S. S. Chen and R. A. Gopinath, Gaussianization, proceeding of NIPS-2000, Denver, Colorado, Nov. 2000.

[6] Coleman, T.F. and Y. Li, An interior trust region approach for nonlinear minimization subject to bounds, SIAM Journal on Optimization, Vol. 6, pp. 418-445, 1996.

[7] Coleman, T.F. and Y. Li, On the convergence of reflective newton methods for large-scale nonlinear minimization subject to bounds, Mathematical Programming, Vol. 67, Number 2, pp. 189-224, 1994.

[8] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of Royal Statistical Soc., Ser. B, vol. 39, pp. 1-38, 1977.

[9] F. R. Gantmacher, The Theory of Matrices, Vol. 1, Chelsea Publishing Company, 1960.

[10] Marshall Hall Jr., Combinatorial Theory, Blaisdell Publishing Company, 1967.

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

[12] G. H. Golub, C. F. Van Loan, Matrix Computations, The Johns Hopkins University Press, 1989.

[13] Alan Hoffman, Private communication.

[14] A. J. Hoffman and H. W. Wieland, The variation of the spectrum of a normal matrix, Duke J. of Math, p. 37-39, 1953.

[15] Frederick Jelinek, Statistical Methods for Speech Recognition, The MIT Press, Cambridge, MA, 1997.

[16] Nagendra Kumar, Investigation of silicon auditory model and generalization of linear discriminant analysis for improved speech recognition, Ph.D dissertation, Johns Hopkins University, 1997.

[17] Te-Won Lee, Michael S. Lewicki, Terrence J. Sejnowski, ICA mixture models for unsupervised classification of non-gaussian classes and automatic context switching in blind signal separation, IEEE Trans. PAMI, vol.22, n0.10, pp.1078-1088, October 2000.

[18] Louis A. Liporace, Maximum likelihood estimation for multivariate observations of Markov Sources, IEEE Transactions of Information theory 5, pp. 729-734, 1982.

[19] Albert W. Marshall, Ingram Olkin, Inequalities: Theory of Majorization and its Applications, Academic Press 1979.

[20] J. Nocedal and S. J. Wright, Numerical Optimization, Springer Verlag, 1999.

[21] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer Verlag, 1995.

[22] Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, 1996.

[23] G. R. Rogers, Matrix Derivatives, Marcel Dekker, New York, 1980.

[24] C. R. Rao, Matrix derivatives, in Encyclopedia of Statistical Science, vol. 5, pp. 320-325, John Wiley, New York, 1985.

[25] S. Sherman, On a conjecture concerning doubly stochastic matrices, Proc. Am. Math. Soc., vol 3, pp. 511-513, 1952.

[26] S. Shreiber, On a result of S. Sherman concerning doubly stochastic matrices, Proc. Am. Math. Soc., vol 9, pp. 350-353, 1953.

[27] S. Sherman, A correction to "On a conjecture concerning doubly stochastic matrices", Proc. Am. Math. Soc., vol 5, pp. 998-999, 1954.

[28] L. Mirsky, Proofs of two theorems on doubly-stochastic matrices, Proc. Am. Math. Soc., vol 9, pp. 371-374, 1958.

[29] G. Strang, Introduction to Applied Mathematics, Wellesly-Cambridge Press, 1986.

[30] G. Strang, Linear Algebra and its Applications, Harcourt Brace Jovanovich, 1988.

[31] P. Sebastiani, On the derivatives of matrix powers, SIAM J. Matrix Anal. Appl.,vol 17, no 3, pp. 640-648, July 1996.

[32] P. P. Vaidyanathan, Multivariate Systems and Filter Banks, Prentice-Hall, 1993.

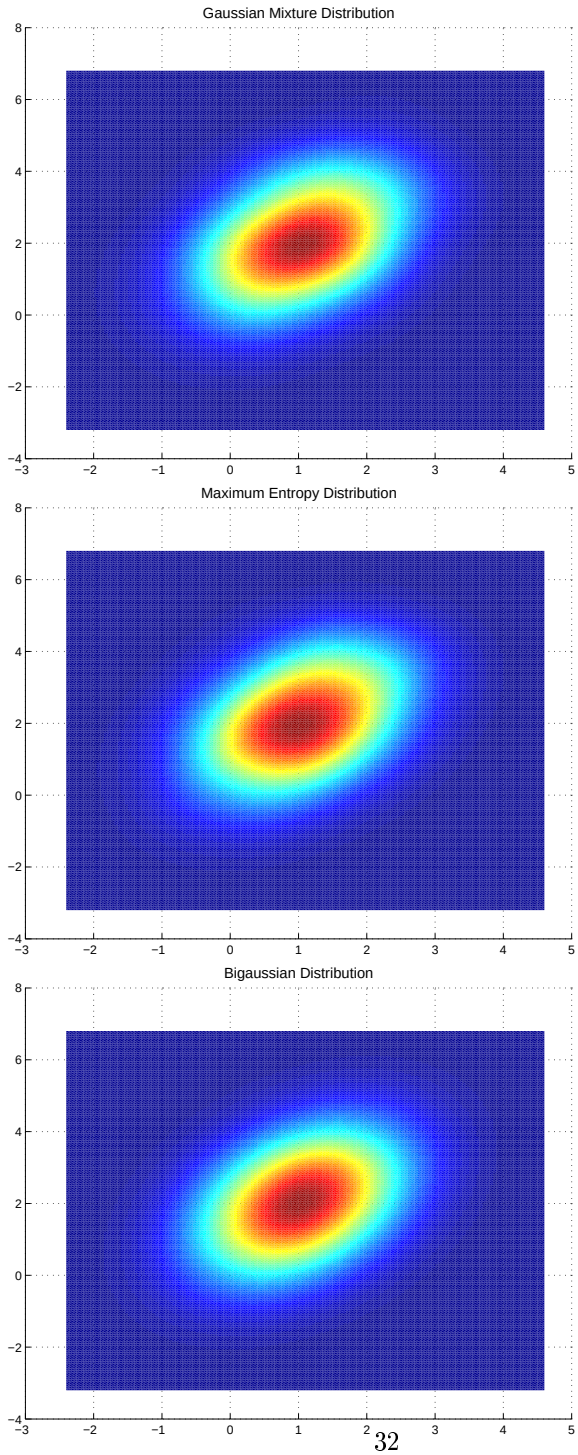

Figure 1: These two dimensional graphs illustrate the good performances of both the Gaussian Mixture Model and the Maximum Entropy Model for estimating the Bigaussian distribution shown on the bottom.

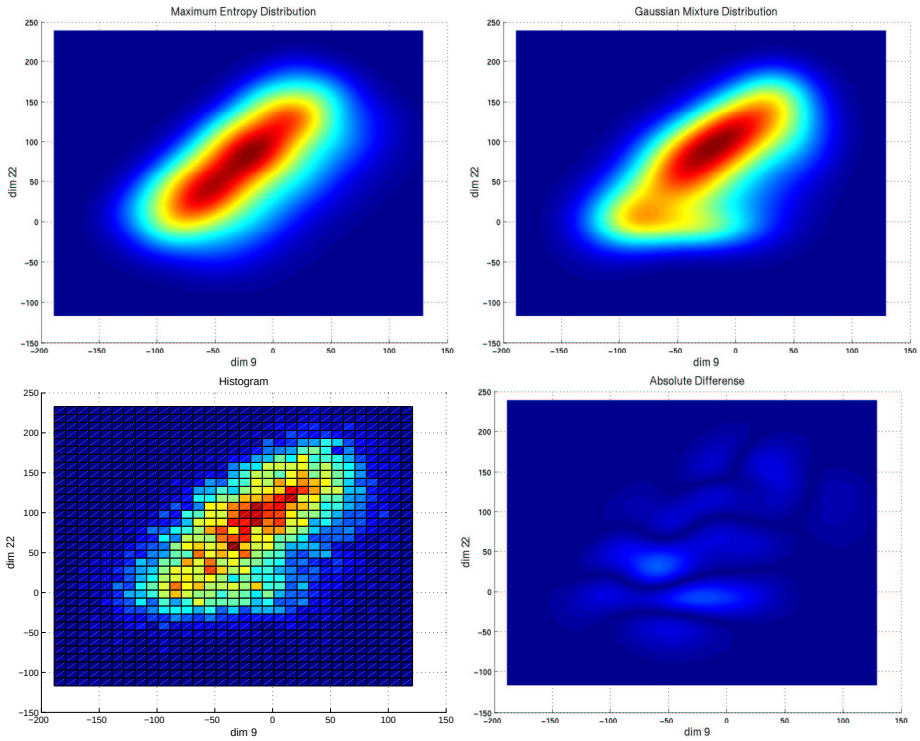

Figure 2: The Maximum Entropy and the Gaussian Mixture estimates of the distribution of the most correlated dimensions (9 and 22) of leaf 1. The bottom left graph shows the histogram of the data, and the bottom right graph illustrates the absolute difference between the two models.

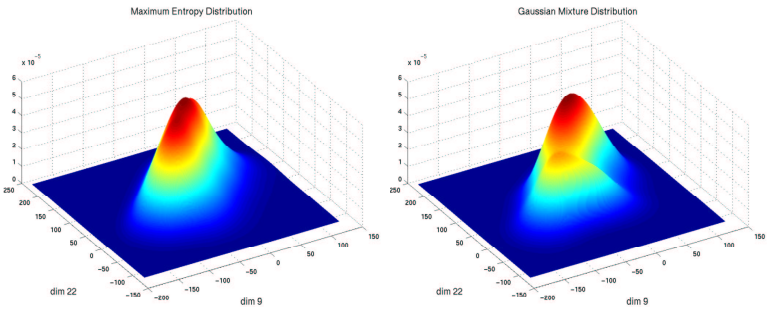

Figure 3: The estimates of the distribution of leaf 1's dimensions 9 and 22, viewed in three dimensions.