IBM Research Report

Extended Baum Transformations for General Functions, II

Dimitri Kanevsky
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
EXTENDED BAUM TRANSFORMATIONS FOR GENERAL FUNCTIONS, II

Dimitri Kanevsky*

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
{kanevsky@us.ibm.com}

ABSTRACT

The discrimination technique for estimating the parameters of Gaussian mixtures that is based on the Extended Baum transformations (EB) has had significant impact on the speech recognition community. The proof that definitively shows that these transformations increase the value of an objective function with iteration (i.e., so-called "growth transformations") was presented by the author two years ago for a diagonal Gaussian mixture densities. In this paper this proof is extended to a multidimensional multivariate Gaussian mixtures. The proof presented in the current paper is based on the linearization process and the explicit growth estimate for linear forms of Gaussian mixtures.

1. INTRODUCTION

The EB procedure involves two types of transformations that can be described as follows. Let \( F(z) = F(z_{ij}) \) be some function in variables \( z = (z_{ij}) \) and \( c_{ij} = z_{ij} \delta_{ij} F(z) \).

1.1. Discrete probabilities:

\[ z_{ij} = c_{ij} + z_{ij} C \sum_{i} c_{ij} + C \]  

(1)

where \( z \in D = \{ z_{ij} \geq 0, \sum_j z_{ij} = \sum_{j=1}^{m} z_{ij} = 1 \} \)

1.2. Gaussian mixture densities:

\[ \hat{\mu}_j = \hat{\mu}_j(C) = \frac{\sum_{i \in I} c_{ij} y_{i} + C \mu_j}{\sum_{i \in I} c_{ij} + C} \]  

(2)

\[ \hat{\Sigma}_j = \hat{\Sigma}_j(C) = \frac{\sum_{i \in I} c_{ij} y_{ij}^T + C(\mu_j \mu_j^T + \Sigma_j)}{\sum_{i \in I} c_{ij} + C} - \hat{\mu}_j \hat{\mu}_j^T \]  

(3)

where

\[ z_{ij} = \frac{1}{(2\pi)^{n/2} C_{\Sigma j}} e^{-\frac{1}{2}(y_i - \mu_j)^T \Sigma_j^{-1}(y_i - \mu_j)} \]  

(4)

and \( y_i \) is a sample of training data.

The work is partly supported by the DARPA Babylon Speech-to-Speech Translation Program.

III. Multidimensional multivariate Gaussian mixture densities:

\[ \hat{\mu}_j = \hat{\mu}_j(C) = \frac{\sum_{i \in I} c_{ij} y_{i} + C \mu_j}{\sum_{i \in I} c_{ij} + C} \]  

(5)

\[ \hat{\Sigma}_j = \hat{\Sigma}_j(C) = \frac{\sum_{i \in I} c_{ij} y_{i} y_{i}^T + C(\mu_j \mu_j^T + \Sigma_j)}{\sum_{i \in I} c_{ij} + C} - \hat{\mu}_j \hat{\mu}_j^T \]  

(6)

where

\[ z_{ij} = \frac{1}{(2\pi)^{n/2} C_{\Sigma j}} e^{-\frac{1}{2}(y_i - \mu_j)^T \Sigma_j^{-1}(y_i - \mu_j)} \]  

(7)

and \( y_{i} = (y_{i1}, \ldots, y_{im}) \) is a sample of training data.

It was shown in [4] that (1) are growth transformations for sufficiently large \( C \) when \( F \) is a rational function. Updated formulae (5, 6) for rational functions \( F \) were obtained through discrete probability approximation of Gaussian densities [7] and have been widely used as an alternative to direct gradient-based optimization approaches [9], [8]. Using the linearization technique that was originally presented in our IBM Research Report [5] and in [6] for diagonal Gaussian mixtures, we demonstrate in this paper that (5, 6) are growth transformations for sufficiently large \( C \) if functions \( F \) obey certain smoothness constraints. Axelrod [1] has recently proposed another proof of existence of a constant \( C \) that ensures validity of the MMIE auxiliary function as formulated by Gunawardana et al. [3]. We also replicate in this paper from [5] the proofs that transformations for diagonal Gaussian mixtures (5) and for discrete probabilities (1) are growth.

2. LINEARIZATION

This principle is needed to reduce proofs of growth transformation for general functions to linear forms.

Lemma 1 Let

\[ F(z) = \tilde{F}(u_j) = \tilde{F}((g_j(z))) = \tilde{F} \circ g(z) \]  

(8)

where \( u_j = g_j(z), j = 1, \ldots, n \) and \( z \) varies in some real vector space \( R^n \) of dimension \( n \). Let \( g_j(z) \) for all \( j = 1, \ldots, n \)
and $F(z)$ be differentiable at $z$. Let, also, $\frac{\delta F(u_j)}{\delta u_j}$ exist at $u_j = g_j(z)$ for all $j = 1, \ldots, m$. Let, further, $L(z) \equiv \nabla F|_{g(z)} \cdot g(z)$, $z \in R^n$. Let $T_C$ be a family of transformations $R^n \rightarrow R^n$ such that for some $l = (l_1, \ldots, l_n) \in R^n$

$T_C(z) - z = l/C + o(1/C)$ if $C \rightarrow \infty$. (Here $o(\epsilon)$ means that $o(\epsilon)/\epsilon \rightarrow 0$ if $\epsilon \rightarrow 0$).

Let, further, $T_C(z) = z$ if

$$\nabla L|_z \cdot l = 0 \quad (9)$$

Then for sufficiently large $C$ $T_C$ is growth for $F$ at $z$ iff $T_C$ is growth for $L$ at $z$.

**Proof** First, from the definition of $L$ we have

$$\frac{\delta F(z)}{\delta z_i} = \sum_j \frac{\delta F(u_j)}{\delta u_j} \cdot \frac{\delta u_j}{\delta z_i}$$

Next, for $z = T_C(z)$ and sufficiently large $C$ we have:

$$F(z_l) - F(z) = \sum_j \frac{\delta F(z)}{\delta z_i} (z_l - z_i) + o(1/C) = \sum_j \frac{\delta F(z)}{\delta z_i} L_i/C + o(1/C)$$

Therefore for sufficiently large $C$ $F(z_l) - F(z) > 0$ iff $L(z_l) - L(z) > 0$.

### 3. EB FOR DISCRETE PROBABILITIES

The following theorem is a generalization of [4].

**Theorem 1** Let $F(z)$ be a function that is defined over $D = \{z_j \geq 0, \sum z_j = 1\}$. Let $F$ be differentiable at $z \in D$ and let $\hat{z} \neq z$ be defined as in (1). Then $F(\hat{z}) > F(z)$ for sufficiently large positive $C$ and $F(\hat{z}) < F(z)$ for sufficiently small negative $C$.

**Proof** Following the linearization principle, we first assume that $F(z) = l(z) = \sum a_{ij} z_i$ is a linear form. Then the transformation formula for $l(\hat{z})$ is the following:

$$\hat{z}_{ij} = \frac{a_{ij} z_i + C z_{ij}}{l(z) + C} \quad (10)$$

We need to show that $l(\hat{z}) \geq l(z)$. It is sufficient to prove this inequality for each linear sub component associated with $i$

$$\sum_{j=1}^{j=n} \frac{a_{ij} z_i + C z_{ij}}{l(z) + C}$$

Therefore without loss of generality we can assume that $i$ is fixed and drop subscript $i$ in the forthcoming proof (i.e. we assume that $l(z) = \sum_{j=1}^{j=n} a_{ij} z_j$, where $z = \{z_j\}$, $z_j \geq 0$ and $\sum z_j = 1$). We have:

$$l(\hat{z}) = \frac{l(z) + C(\hat{z})}{l(z) + C}$$

The linear case of Theorem 1 will follow from next two lemmas.

**Lemma 2**

$$l_2(z) \geq l(z^2) \quad (11)$$

**Proof** Let as assume that $a_j \geq a_{j+1}$ and substituting $z_l = \sum_{j=1}^{j=n} z_j$ we need to prove:

$$\sum_{j=1}^{j=n-1} (a_j^2 z_j + a_n^2 (1-z_l)) \geq \sum_{j=1}^{j=n} (a_j - a_n)^2 z_j^2 + 2 a_j - a_n \sum_{j=1}^{j=n} z_j$$

We will prove the above formula by proving for every fixed $j (a_j^2 - a_n^2) z_j \geq (a_j - a_n)^2 z_j^2 + 2 (a_j - a_n) a_n z_j$. If $(a_j - a_n) z_j \not= 0$ then the above inequality is equivalent to $a_j \geq (a_j - a_n) z_j$ and is obviously holds since $0 \leq z_j \leq 1$.

**Lemma 3** For sufficiently large $|C|$ the following holds:

$l(\hat{z}) > l(z)$ if $C$ is positive and $l(\hat{z}) < l(z)$ if $C$ is negative.

**Proof** From (11) we have the following inequalities:

$$l_2(z) + Cl(z) \geq l(z^2) + Cl(z)$$

$$l(\hat{z}) = \frac{l_2(z) + Cl(z)}{l(z) + C}$$

Therefore for sufficiently large $C$ $F(\hat{z}) - F(z) > 0$ if $l(\hat{z}) - l(z) > 0$.

### 4. EB FOR GAUSSIAN DENSITIES

For simplicity of the notation we consider the transformation (5), (6), only for a single pair of variables $\mu, \sigma$, i.e. we drop subscript $j$ everywhere in (5, 6), (7) and also set $\hat{z}_i = \frac{1}{(2\pi)^{1/2} \sigma} e^{-(y_i - \mu)^2/2\sigma^2}$

**Theorem 2** Let $F(\{z_j\})$, \(i = 1, \ldots, m\), be differentiable at $\mu, \sigma$ and $\frac{\delta F(\{z_j\})}{\delta z_{ij}}$ exist at $z_i$. Let either $\mu \not\equiv \mu$ or $\sigma \not\equiv \sigma$.

Then for sufficiently large $C$

$$F(\{z_{\hat{i}}\}) - F(\{z_{ij}\}) = \frac{T}{C} + o(1/C) \quad (13)$$

Where

$$T = \frac{1}{\sigma^2} \left\{ \sum \frac{c_j (y_j - \mu)^2 - \sigma^2}{2\sigma^2} \right\} > 0 \quad (14)$$

In other words, $F(\{z_{\hat{i}}\})$ grows proportionally to $1/C$ for sufficiently large $C$. For sufficiently large $C$ $l(\mu, \sigma) > l(\mu, \sigma)$. This inequality is sufficient to prove with the precision $1/C^2$.

**Proof** First, we assume that $F(\{z_j\}) = l(\mu, \sigma) := \sum_{j=1}^{j=n} a_j z_j$. Let us set $l(\mu, \sigma) := l(\{z_{\hat{i}}\}) := \sum_{i=1}^{i=m} a_i z_{\hat{i}}$. Then $c_j = a_j z_j$ in (5), (6). We want to prove that for sufficiently large $C l(\mu, \sigma) \geq l(\mu, \sigma)$. This inequality is sufficient to prove with the precision $1/C^2$. 


\[ \hat{\mu} = \hat{\mu}(C) = \frac{\sum_{j=1}^{m} c_j y_j + C \mu}{\sum_{j=1}^{m} c_j + C} = \frac{1}{\sigma} \sum_{j=1}^{m} c_j y_j + \mu \sim \frac{1}{C} \sum_{j=1}^{m} c_j y_j + \mu \sim (\frac{1}{C}) \sum_{j} c_j y_j + \mu (1 - \frac{1}{C}) \sum_{j} c_j \sim \mu + \frac{1}{C} \sum_{j} c_j y_j - \mu \sum_{j} c_j \]

(15)

Next, we have

\[ \hat{\sigma} = \hat{\sigma}(C)^2 = \frac{\sum_{j} c_j y_j^2 + C(\mu^2 + \sigma^2)}{\sum_{j} c_j + C} \sim \mu^2 + \frac{1}{C} \sum_{j} c_j y_j^2 - (\mu^2 + \sigma^2) \sum_{j} c_j \]

(18)

\[ \hat{\mu}^2 \sim \mu^2 + 2\frac{\mu}{C} \sum_{j=1}^{m} c_j (y_j - \mu) \]

(19)

This gives

\[ \hat{\sigma}^2 \sim \mu^2 + \sigma^2 + \frac{1}{C} \sum_{j} c_j y_j^2 - (\mu^2 + \sigma^2) \sum_{j} c_j - \left[ \mu^2 + \frac{2\mu}{C} \sum_{j} c_j (y_j - \mu) \right] \]

(20)

And finally

\[ \hat{\sigma}^2 \sim \sigma^2 + \sum_{j} [\frac{(y_j - \mu)^2 - \sigma^2 c_j}{C}] \]

(21)

\[ \frac{(y_i - \mu)^2}{\sigma^2} \sim \frac{1}{\sigma^2} \sum_{j} \frac{(y_j - \mu)^2}{C} - 2\frac{(y_i - \mu)}{\sigma^2} \sum_{j} \frac{c_j (y_j - \mu)}{C} \times \left\{ 1 - \sum_{j} \frac{c_j (y_j - \mu)^2 - \sigma^2}{\sigma^2 C} \right\} \sim \]

\[ \frac{(y_i - \mu)^2}{\sigma^2} - \frac{1}{C^2 \sigma^2} \sum_{j} (y_j - \mu)^2 - \sigma^2 c_j + 2(y_i - \mu) \sum_{j} (y_j - \mu) c_j \]

(22)

\[ \hat{z}_i \sim \frac{1}{(2\pi)^{1/2} \sigma} e^{-\frac{(y_i - \mu)^2}{2\sigma^2} - \frac{A_1}{C^2 \sigma^2}} \]

(23)

Where

\[ A_i = \frac{2\sigma^2}{2\sigma^2} \sum_{j} (y_j - \mu)^2 - \sigma^2 c_j + (y_i - \mu) \sum_{j} (y_j - \mu) c_j \]

(24)

Continue this we have

\[ \hat{z}_i \sim K e^{-\frac{(y_i - \mu)^2}{2\sigma^2}} \left( 1 + \frac{A_i}{C^2 \sigma^2} \right) \]

(25)

Where

\[ K = \frac{1}{(2\pi)^{1/2} \sigma} \]

\[ 1/\hat{\sigma} \sim \frac{1}{\sigma} \left( 1 - \frac{\sum_{j} c_j [(y_j - \mu)^2 - \sigma^2]}{2\sigma^2 C} \right) \]

\[ (1 + \frac{A_i}{C^2 \sigma^2}) \left( 1 - \frac{\sum_{j} c_j [(y_j - \mu)^2 - \sigma^2]}{2\sigma^2 C} \right) \sim \]

\[ 1 + \frac{B_i}{C^2 \sigma^2} \]

(26)

Where \[ B_i = \frac{2\sigma^2}{2\sigma^2} \sum_{j} (y_j - \mu)^2 - \sigma^2 c_j + (y_i - \mu) \sum_{j} (y_j - \mu) c_j \]

Using the last equalities we get

\[ \hat{z}_i = z_i + \frac{B_i}{C^2 \sigma^2} \hat{z}_i \]

(27)

Since \( l(\hat{\mu}, \hat{\sigma}) \) is a linear form in the \( z_i \), we have

\[ l(\{\hat{z}_i\}) = l(\{z_i\}) + l(\{B_i z_i\}) \]

(28)

\[ l(\{B_i z_i\}) = \sum_{i} a_i z_i \left[ \frac{(y_i - \mu)^2}{2\sigma^2} - 1/2 \right] \times \sum_{j} c_j [(y_j - \mu)^2 - \sigma^2] + (y_i - \mu) \sum_{j} c_j (y_j - \mu) = \]

\[ \sum_{i} c_i \left[ \frac{(y_i - \mu)^2}{2\sigma^2} - 1/2 \right] \sum_{j} c_j [(y_j - \mu)^2 - \sigma^2] + \]
\[+(y_i - \mu) \sum_j c_j(y_j - \mu) =\]
\[= \left[\sum_j c_j[(y_j - \mu)^2 - \sigma^2]\right]^2 + \left[\sum_j c_j(y_j - \mu)^2\right]^2\]

\[l(\{z_i\}) = \frac{T}{C}\]

Since by assumption either \(\hat{\mu} \neq \nu\) or \(\hat{\sigma} \neq \sigma\) \(T \neq 0\). Applicability of the linearization principle follows from the fact that if (14) holds then the left part in the equation (9) is not equal to zero. Q.E.D.

5. EB FOR MULTIDIMENSIONAL MULTIVARIATE GAUSSIAN DENSITIES

For simplicity of the notation we consider the transformation (5), (6), only for a single pair of variables \(\mu, \Sigma\), i.e. we drop subscript \(j\) everywhere in (5, 6), (7) and also set \(z_i = \frac{1}{(2\pi)^{n/2}} e^{-1/2(y_i-\mu)^T \Sigma^{-1}(y_i-\mu)}\)

Theorem 3 Let \(F\) \((\{z_i\})\), \(i = 1...m\), be differentiable at \(\mu, \Sigma\) and \(SF(\{z_i\})\) exist at \(z_i\). Let either \(\hat{\mu} \neq \mu\) or \(\hat{\Sigma} \neq \Sigma\).

Then for sufficiently large \(C\)

\[F(\{\hat{z}_i\}) = F(\{z_i\}) = T/C + o(1/C)\]

where \(T > 0\), i.e. \(F(\{z_i\})\) grows proportionally to \(1/C\) for sufficiently large \(C\). If \(\Sigma\) represented as a diagonal matrix

\[\Sigma^{-1} = \text{diag}[\lambda_1, ..., \lambda_n]\]

then one can write \(T\) explicitly as follows:

\[T = T_1 + T_2 + T_3\]

\[T_1 = \frac{1}{2} \sum_{k=1}^{m} (\lambda_k^2 + \lambda_i^2) \sum_i c_k a_k a_i^T \]

\[T_2 = \frac{1}{2} \sum_{k=1}^{n} (\lambda_k \sum_i c_k a_k a_i^T - \sum_i c_i^T)\]

\[T_3 = \frac{1}{2} \sum_{k=1}^{n} \lambda_k \sum_i c_k a_k a_i^T\]

\[\text{Proof:}\]

Our proof will be split in several steps.

Step 1: Linerarization

First, we assume that \(F(\{z_i\}) = l(\mu, \Sigma) := l(\{z_i\}) := \sum_i a_i z_i\). Let us set \(l(\hat{\mu}, \hat{\Sigma}) := l(\{\hat{z}_i\}) := \sum_i a_i \hat{z}_i\).

Then \(c_j = a_j \hat{z}_j\) in (5), (6). We want to prove that for sufficiently large \(C\) \(l(\hat{\mu}, \hat{\Sigma}) \geq l(\mu, \Sigma)\). This inequality is sufficiently to prove with the precision \(1/C^2\).

Step 2: Computation of \(T\)

\[\hat{\mu} = \hat{\mu}(C) = \frac{\sum_j c_j y_j + C \mu}{\sum_j c_j + C} \approx \frac{C}{\nu} \sum_j c_j y_j + \mu + \frac{1}{C} \sum_j c_j y_j - \mu \sigma_j\]

\[\sim \left[\frac{1}{C} \sum_j c_j y_j + \mu\right] (1 - \frac{1}{C} \sum_j c_j) \sim \mu^T + \frac{1}{C} \sum_j c_j y_j - \mu \sigma_j\]

\[\hat{\Sigma} = \frac{\sum_j c_j y_j y_j^T + C(\mu \mu^T + \Sigma)}{\sum_j c_j + C} - \frac{\hat{\mu} \hat{\mu}^T}{C}\]

Let us compute \(\hat{\Sigma}\) using (38)

\[\sim \left[\frac{1}{C} \sum_j c_j y_j y_j^T + C(\mu \mu^T + \Sigma)\right] - \mu \mu^T + \frac{2}{C} \sum_{j=1}^{m} c_j (y_j - \mu)^T\]

This gives

\[\hat{\Sigma} \sim \mu \mu^T + \Sigma + \frac{1}{C} \sum_j c_j y_j y_j^T - (\mu \mu^T + \Sigma) \sum_j c_j -\]

\[= \Sigma + \frac{1}{C} \sum_j c_j y_j y_j^T - (\mu \mu^T + \Sigma) \sum_j c_j - 2 \mu \sum_j c_j (y_j - \mu)^T\]

And finally

\[\hat{\Sigma} \sim \Sigma + \sum_j [(y_j - \mu)(y_j - \mu)^T - \Sigma] c_j\]

\[\hat{\Sigma}^{-1} \sim \Sigma^{-1} - \frac{\sum_j [(y_j - \mu)(y_j - \mu)^T - \Sigma] c_j}{C}\]

\[1/2(v_i - \hat{v}_i)(yi - \hat{v}_i) \sim 1/2 [(yi - \mu) - \sum_j c_j (y_j - \mu)]^T \times\]
\[
\begin{align*}
[\Sigma^{-1} - \frac{\Sigma^{-2}}{C} (\sum_j [ (y_j - \mu)(y_j - \mu)^T - \Sigma] C_j)] \\
\times [ (y_i - \mu) - \frac{\sum_j C_j (y_j - \mu)}{\Sigma}] \sim \\
\sim \frac{1}{2} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) - \frac{A_1}{C}
\end{align*}
\]

(44)

\[A_{11} = A_{11} + A_{i2}\]

\[A_{22} = 1/2 \sum_j C_j [ (y_j - \mu)^T \Sigma^{-1} (y_j - \mu) + (y_i - \mu)^T \Sigma^{-1} (y_j - \mu) ] \] 

(45)

\[\tilde{z}_i \sim K e^{-\frac{1}{2} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)} (1 + \frac{A_1}{C}) \] 

(48)

Where

\[K = \frac{|\hat{\Sigma}|^{-1/2}}{(2\pi)^{n/2}} \]

\[|\Sigma|^{-1/2} \sim |\hat{\Sigma}|^{-1/2} \{ 1 + \frac{\sum_j C_j [ n - Tr \Sigma^{-1}(y_j - \mu)^T(y_j - \mu)^T ] }{C} \} \] 

(49)

\[(1 + \frac{A_1}{C}) [ 1 + \frac{\sum_j C_j [ n - Tr \Sigma^{-1}(y_j - \mu)^T(y_j - \mu)^T ] }{C} ] \sim \\
\sim 1 + \frac{A_1}{C} \{ A_1 + 1/2 \sum_j C_j [ n - Tr \Sigma^{-1}(y_j - \mu)^T(y_j - \mu)^T ] \} \\
\sim 1 + \frac{B_1}{C}
\]

(50)

Where

\[B_1 = A_1 + D\]

Here we use

\[D = 1/2 \sum_j C_j [ n - (y_j - \mu)^T \Sigma^{-1}(y_j - \mu) ] \]

and

\[Tr \Sigma^{-1} (y_j - \mu)^T(y_j - \mu)^T = (y_j - \mu)^T \Sigma^{-1}(y_j - \mu) \]

Using the last equalities we get

\[\tilde{z}_i \sim z_i + \frac{B_1}{C} z_i \] 

(51)

Since \(l(\mu, \Sigma)\) is a linear form in the \(z_i\) we have

\[l(\{\tilde{z}_i\}) \sim l(\{z_i\}) + \frac{1}{C} B_1 z_i \]

(52)

and

\[T = l(\{B_1 z_i\}) = \\
= \sum_i c_i A_1 + 1/2(\sum_i c_i) \sum_j C_j [ n - (y_j - \mu)^T \Sigma^{-1}(y_j - \mu) ] \\
= \sum_i c_i A_1 = \sum_i c_i A_{11} + \sum_i c_i A_{i2} \]

(53)

\[\tilde{A}_1 = \sum_i c_i A_{11} = \]

\[= 1/2 \sum_{ij} c_i c_j (y_i - \mu)^T \Sigma^{-2} [(y_j - \mu)(y_j - \mu)^T - \Sigma] (y_i - \mu) \] 

(54)

\[\tilde{A}_2 = \sum_i c_i A_{i2} = \]

\[= 1/2 \sum_{ij} c_i c_j [ (y_i - \mu)^T \Sigma^{-1} (y_j - \mu) + (y_i - \mu)^T \Sigma^{-1} (y_j - \mu) ] = \\
= \sum_j c_j (y_j - \mu)^T \Sigma^{-1} \sum_i c_i (y_i - \mu) \]

(55)

\[l(\{\tilde{z}_i\}) - l(\{z_i\}) \sim \frac{T}{C} \]

**Step 3: Reduction to a diagonal case**

Since \(\Sigma\) is a symmetric matrix there exists an orthogonal matrix \(O\) such that \(O \Sigma O^{-1}\) is a diagonal matrix. It is easy to see that \(T\) is invariant under such orthogonal change of coordinates. For example, the component \(\tilde{A}_1\) of \(T\) is invariant under orthogonal change of coordinate as one can see from the following computations:

\[\tilde{A}_1 = \]

\[= 1/2 \sum_{ij} c_i c_j (y_i - \mu)^T O^T O \Sigma^{-2} O^T x \times \\
\times [O(y_j - \mu)(y_j - \mu)^T O^T O \Sigma O^T] O(y_i - \mu) \] 

(56)

**Step 4: special case - 2-dimensional Gaussians**

We will perform computations for simplicity for 2-dimensional case.

Without loss of generality we can assume the \(\Sigma^{-1} = \text{diag}[\lambda_1, \lambda_2]\) is a diagonal \(2 \times 2\) - matrix with diagonal elements \(\lambda_1\) and \(\lambda_2\).

Let compute

\[A'_1 = \]

\[\frac{1}{2} \sum_{ij} c_i c_j (y_i - \mu)^T \Sigma^{-2} [(y_j - \mu)(y_j - \mu)^T (y_i - \mu)] \] 

(57)

Let set

\[(y_i - \mu)^T = (a_{11}, a_{2i}) \]

(58)

Then

\[A'_1 = \]

\[= \frac{1}{2} \sum_{ij} c_i c_j (a_{11}, a_{2i}) \times \text{diag}[\lambda_1^2, \lambda_2^2] \times \]
\[
(a_{1j}, a_{2j})^T \times (a_{1j}, a_{2j}) \times (a_{1j}, a_{2j})^T
\]
(59)

\[
= 1/2 \sum_{ij} c_i c_j (\lambda_i^2 a_{11} + \lambda_j^2 a_{22}) (a_{1j} a_{2j}) (a_{1j} a_{1j} + a_{2j} a_{2j})
\]
(60)

\[
= 1/2 \sum_{ij} c_i c_j (\lambda_i^2 a_{11} a_{1j} + \lambda_j^2 a_{22} a_{2j}) (a_{1j} a_{1j} + a_{2j} a_{2j})
\]
(61)

\[
= 1/2(\lambda_i^2 + \lambda_j^2) \sum_{ij} c_i c_j a_{11} a_{1j} a_{2j} +
\]

\[
+ 1/2 \lambda_i^3 \sum_{ij} c_i c_j a_{11}^2 a_{1j}^2
\]

\[
+ 1/2 \lambda_j^3 \sum_{ij} c_i c_j a_{22}^2 a_{2j}^2
\]

\[
= 1/2(\lambda_i^2 + \lambda_j^2) \left( \sum_{ij} c_i a_{11} a_{22} \right)^2 +
\]

\[
+ 1/2 \lambda_i^2 \left( \sum_{i} c_i a_{11}^2 \right)^2
\]

\[
+ 1/2 \lambda_j^2 \left( \sum_{i} c_i a_{22}^2 \right)^2
\]
(62)

\[
A''_i = \Lambda_i - A'_i =
\]

\[
= -1/2 \sum_{ij} c_i c_j (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)
\]
(63)

\[
-1/2 \sum_{ij} c_i c_j (a_{11} a_{2j}) \times (\lambda_1, \lambda_2) \times (a_{11}, a_{2j})^T
\]

\[
= -1/2 \left( \sum_{ij} \sum_{ij} c_i a_{11} a_{2j} \right)
\]
(64)

Next, for our 2-dimensional case we have

\[
\sum c_i D = \left( \sum c_i \right)^2 - 1/2 \left( \sum c_i \right) \left( \sum c_i \lambda_1 a_{11} + \sum c_i \lambda_2 a_{22} \right)
\]

Therefore

\[
T =
\]

\[
1/2(\lambda_i^2 + \lambda_j^2) \left( \sum_{ij} c_i a_{11} a_{22} \right)^2 +
\]

\[
+ 1/2 \lambda_i^2 \left( \sum_{i} c_i a_{11}^2 \right)^2
\]

\[
+ 1/2 \lambda_j^2 \left( \sum_{i} c_i a_{22}^2 \right)^2 -
\]

\[
- \left( \sum_{ij} c_i \lambda_1 a_{11} + \sum c_i \lambda_2 a_{22} \right)
\]

\[
+ \left( \sum c_i \right)^2
\]

\[
+ \sum c_i c_j (\lambda_1 a_{11} a_{1j} + \lambda_2 a_{2j} a_{2j})
\]
(66)

And finally

\[
T =
\]

\[
1/2(\lambda_i^2 + \lambda_j^2) \left( \sum_{ij} c_i a_{11} a_{22} \right)^2 +
\]

\[
+ 1/2 \lambda_i^2 \left( \sum_{i} c_i a_{11}^2 \right)^2
\]

\[
+ 1/2 \lambda_j^2 \left( \sum_{i} c_i a_{22}^2 \right)^2 +
\]

\[
+ \lambda_1 \left( \sum_{i} c_i a_{11} \right)^2
\]

\[
+ \lambda_2 \left( \sum_{i} c_i a_{22} \right)^2
\]
(67)

In the above equation:

\[
T_1 = 1/2(\lambda_i^2 + \lambda_j^2) \left( \sum_{ij} c_i a_{11} a_{22} \right)^2
\]
(68)

\[
T_2 = 1/2 \lambda_i \left( \sum_{i} c_i a_{11}^2 - \sum c_i \right)^2 +
\]

\[
+ 1/2 \lambda_j \left( \sum_{i} c_i a_{22}^2 - \sum c_i \right)^2
\]
(69)

\[
T_3 = \lambda_1 \left( \sum_{i} c_i a_{11} \right)^2
\]

\[
+ \lambda_2 \left( \sum_{i} c_i a_{22} \right)^2
\]
(70)

**Step 4: General case - n-dimensional Gaussians**

We will perform computations for n-dimensional case.

Without loss of generality we can assume the \( \Sigma^{-1} = diag[\lambda_1, \lambda_2, ..., \lambda_n] \) is a diagonal \( n \times n \) matrix with diagonal elements \( \lambda_1, \lambda_2, \lambda_n \).

Let compute

\[
A'_1 =
\]

\[
= 1/2 \sum_{ij} c_i c_j (y_i - \mu)^T \Sigma^{-2} [(y_j - \mu)(y_j - \mu)^T (y_i - \mu)]
\]
(71)

Let set

\[
(y_i - \mu)^T = (a_{1i}, a_{2i}, ..., a_{ni})
\]
(72)

Then

\[
A'_1 =
\]

\[
1/2 \sum_{ij} c_i c_j (a_{1j}, a_{2j}, ..., a_{nj}) \times diag[\lambda_1^2, \lambda_2^2, ..., \lambda_n^2] \times (a_{1j}, a_{2j}, ..., a_{nj})^T
\]

\[
\times (a_{1j}, a_{2j}, ..., a_{nj}) \times (a_{1i}, a_{2i}, ..., a_{ni})^T
\]
(73)

\[
= 1/2 \sum_{ij} c_i c_j (\lambda_1^2 a_{11} a_{2j} + \lambda_2^2 a_{2j} a_{2j}) (a_{1j}, a_{2j}, ..., a_{nj})^T \times
\]

\[
\times (a_{1i}, a_{2i}, ..., a_{ni})
\]
\[ \begin{align*}
\times (\sum \alpha_{kj}a_{ki}) & = 1/2 \sum_{ij} c_ic_j (\sum_k \lambda_k^2 a_{ki} a_{kj}) \times \\
& = 1/2 \sum_{k,l,k\neq l} \sum_{ij} c_ic_j (\lambda_k^2 + \lambda_l^2) \times \\
& = 1/2 \sum_{k,l,k\neq l} \sum_{ij} \lambda_k^2 a_{ki} a_{kj} + \\
+1/2 \sum_{k,l,k\neq l} \sum_{ij} \lambda_k^2 a_{ki} a_{kj}^2 = \\
1/2 \sum_{k,i} (\lambda_k^2 + \lambda_i^2) (\sum_{ij} c_i a_{ki} a_{lj})^2 & + \\
+1/2 \sum_{k,i} \lambda_k c_i a_{ki}^2 & (76)
\end{align*} \]

Adding to both parts of the above equation \( C' \Sigma \) for any \( C' \) we get

\[ \begin{align*}
\Sigma \sum c_i + C' \Sigma & = \sum c_i y_i y_i^T - \sum c_i \mu \mu^T + \\
+ C'(\mu \mu^T + \Sigma) & - C' \mu \mu^T \rightarrow \\
\Sigma & = \sum c_i y_i y_i^T + C'(\mu \mu^T + \Sigma) \rightarrow \\
& - \mu \mu^T \rightarrow \\
\Sigma & = \Sigma (C')
\end{align*} \]

c) \( \mu = \mu(0) \) implies that \( T_3 = 0 \). \( \Sigma_{kk} = \Sigma_{kk}(0) \) implies that \( T_2 = 0 \). Finally, \( \Sigma_{kl} = \Sigma_{kl}(0) = 0 \) for \( k \neq l \) implies that \( \sum c_i(y_{ik} - \mu_k)(y_{il} - \mu_l) = 0 \), i.e. \( T_1 = 0 \).

We can now finish the proof of the theorem. Since by assumption either \( \mu \neq \mu \) or \( \Sigma \neq \Sigma T \neq 0 \). Applicability of the linearization principle follows from the fact that if (14) holds then the left part in the equation (9) is not equal to zero. Q.E.D.

6. NEW GROWTH TRANSFORMATIONS

One can derive new updates for means and variances applying EB algorithm of the section 3 by introducing probability constraints for means and variances as follows. Let us assume that \( 0 \leq \mu_j \leq D_j, 0 \leq \sigma_j \leq E_j \). Then we can introduce slack variables \( \mu_j', \sigma_j' \geq 0 \) such that \( \mu_j/D_j + \mu_j'/D_j = 1, \sigma_j/E_j + \sigma_j'/E_j = 1 \). Then we can compute updates as in (1), with \( c_j \) as in (5, 6).

\[ \begin{align*}
\hat{\mu_j} & = \frac{D_j \mu_j'}{\sum_j c_j \frac{(y_{ij} - \mu_j')^2}{\sigma_j^2} + D_j C} \\
\hat{\sigma_j} & = E_j \frac{\sum_j c_j \frac{(y_{ij} - \mu_j')^2}{\sigma_j^2} + C \sigma_j}{\sum_j c_j \frac{(y_{ij} - \mu_j')^2}{\sigma_j^2} + C \sigma_j}
\end{align*} \]

If some \( \mu_j < 0 \) one can make them positive by adding positive constants, compute updates for new variables in the new coordinate system and then go back to the old system of coordinates.

7. REFERENCES


