RC23645 (W0506-120) June 29, 2005
Mathematics

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 multivariant 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 $c_{ij} = z_{ij}\frac{\delta}{\delta z_{ij}}F(z)$.
I. Discrete probabilities:

$$\hat{z}_{ij} = \frac{c_{ij} + z_{ij}C}{\sum_i c_{ij} + C} \tag{1}$$

where $z \in D = \{z_{ij} \geq 0, \sum_j z_{ij} = \sum_{j=1}^{j=m_i} z_{ij} = 1\}$

II. Gaussian mixture densities:

$$\hat{\mu}_j = \hat{\mu}_j(C) = \frac{\sum_{i\in I} c_{ij}y_i + C\mu_j}{\sum_{i\in I} c_{ij} + C} \tag{2}$$

$$\hat{\sigma}_j^2 = \hat{\sigma}_j(C)^2 = \frac{\sum_{i\in I} c_{ij}y_i^2 + C(\mu_j^2 + {\sigma_j}^2)}{\sum_{i\in I} c_{ij} + C} - \hat{\mu}_j^2 \tag{3}$$

where

$$z_{ij} = \frac{1}{(2\pi)^{1/2}\sigma_j} e^{-(y_i-\mu_j)^2/2\sigma_j^2} \tag{4}$$

and y_i is a sample of training data.

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

III. Multidemensional multivariate Gaussian mixture densities:

$$\hat{\mu}_j = \hat{\mu}_j(C) = \frac{\sum_{i\in I} c_{ij}y_i + C\mu_j}{\sum_{i\in I} c_{ij} + C} \tag{5}$$

$$\hat{\Sigma}_j = \hat{\Sigma}_j(C) = \frac{\sum_{i\in I} c_{ij}y_i y_i^T + C(\mu_j\mu_j^T + \Sigma_j)}{\sum_{i\in I} c_{ij} + C} - \hat{\mu}_j\hat{\mu}_j^T \tag{6}$$

where

$$z_{ij} = \frac{|\Sigma_j|^{-1/2}}{(2\pi)^{n/2}} e^{-1/2(y_i-\mu_j)^T\Sigma_j^{-1}(y_i-\mu_j)} \tag{7}$$

and $y_i^T = (y_{i1}, ...y_{im})$ 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 [6] 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*

$$F(z) = \tilde{F}(\{u_j\}) = \tilde{F}(\{g_j(z)\}) = \tilde{F} \circ g(z) \tag{8}$$

where $u_j = g_j(z), j = 1, ..m$ *and* z *varies in some real vector space* R^n *of dimension* n. *Let* $g_j(z)$ *for all* $j = 1, ...m$

and $F(z)$ be differentiable at z. Let, also, $\frac{\delta \tilde{F}(\{u_j\})}{\delta u_j}$ exist at $u_j = g_j(z)$ for all $j = 1, ...m$. Let, further, $L(z\prime) \equiv \nabla\tilde{F}\Big|_{g(z)} \cdot g(z\prime)$, $z\prime \in R^n$. Let T_C be a family of transformations $R^n \rightarrow R^n$ such that for some $l = (l_1...l_n) \in R^n$ $T_C(z) - z = l/C + o(1/C)$ 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

$$\nabla L\big|_z \cdot l = 0 \tag{9}$$

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 $\frac{\delta F(z)}{\delta z_k} = \sum_j \frac{\delta \tilde{F}(\{u_j\})}{\delta u_j}\frac{\delta g_j(z)}{\delta z_k} = \frac{\delta L(z)}{\delta z_k}$
Next, for $z\prime = T_C(z)$ and sufficiently large C we have: $F(z\prime) - F(z) = \sum_i \frac{\delta F(z)}{\delta z_i}(z_i\prime - z_i) + o(1/C) = \sum_i \frac{\delta F(z)}{\delta z_i} l_i/C + o(1/C) = \sum_i \frac{\delta L(z)}{\delta z_i} l_i/C + o(1/C) = \sum_i \frac{\delta L(z)}{\delta z_i}(z_i\prime - z_i) + o(1/C) = L(z\prime) - L(z) + o(1/C)$. Therefore for sufficiently large C $F(z\prime) - F(z) > 0$ iff $L(z\prime) - L(z) > 0$.

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 $D = \{z_{ij} \geq 0, \sum z_{ij} = 1\}$. 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 $F(z) = l(z) = \sum a_{ij}z_{ij}$ is a linear form. Than the transformation formula for $l(x)$ is the following:

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

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

$$\sum_{j=1}^{j=n} a_{ij}\hat{z}_{ij} \geq \sum_{j=1}^{j=n} a_{ij}z_{ij}$$

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 $l(z) = \sum a_j z_j$, where $z = \{z_j\}$, $z_j \geq 0$ and $\sum z_j = 1$). We have: $l(\hat{z}) = \frac{l_2(z) + Cl(z)}{l(z)+C}$, where $l_2(z) := \sum_j a_j^2 z_j$. The linear case of Theorem 1 will follow from next two lemmas.

Lemma 2

$$l_2(z) \geq l(z)^2 \tag{11}$$

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

$$\sum_{j=1}^{j=n-1}[a_j^2 z_j + a_n^2(1 - z\prime)] \geq \sum_{j=1}^{j=n-1}(a_j - a_n)^2 z_j^2 +$$

$$2\sum_{j=1}^{j=n-1}(a_j - a_n)a_n z_j + a_n^2 \tag{12}$$

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 $(a_j - a_n)z_j \neq 0$ then the above inequality is equivalent to $a_j - a_n \geq (a_j - a_n)z_j$ 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.
$l_2(z) + Cl(z) \geq l(z)^2 + Cl(z)$,
$l(\hat{z}) = \frac{l_2(z)+Cl(z)}{l(z)+C} \geq \frac{l(z)^2+Cl(z)}{l(z)+C}$ if $l(z) + C > 0$
and $l(\hat{z}) = \frac{l_2(z)+Cl(z)}{l(z)+C} \leq \frac{l(z)^2+Cl(z)}{l(z)+C}$ if $l(z) + C < 0$.
The general case of Theorem 1 follows immediately from the observation that (9) is equivalent to $l_2(z) - l(z)^2 = 0$ for large C.

4. EB FOR GAUSSIAN DENSITIES

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

Theorem 2 *Let $F(\{z_i\})$, $i = 1...m$, be differentiable at μ, σ and $\frac{\delta F(\{z_i\})}{\delta z_i}$ exist at z_i. Let either $\hat{\mu} \neq \mu$ or $\hat{\sigma} \neq \sigma$. Then for sufficiently large C*

$$F(\{\hat{z}_i\}) - F(\{z_i\}) = T/C + o(1/C) \tag{13}$$

Where

$$T = \frac{1}{\sigma^2}\{\frac{\{\sum c_j[(y_j - \mu)^2 - \sigma^2]\}^2}{2\sigma^2} + [\sum c_j(y_j - \mu)]^2\} > 0 \tag{14}$$

In other words, $F(\{\hat{z}_i\})$ grows proportionally to $1/C$ for sufficiently large C.

Proof First, we assume that $F(\{z_i\}) = l(\mu, \sigma) := l(\{z_i\}) := \sum_{i=1}^{i=m} a_i z_i$. Let us set $l(\hat{\mu}, \hat{\sigma}) := l(\{\hat{z}_i\}) := \sum_{i=1}^{i=m} a_i \hat{z}_i$. Then $c_j = a_j z_j$ 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 $1/C^2$.

$$\hat{\mu} = \hat{\mu}(C) = \frac{\sum_{j=1}^{j=m} c_j y_j + C\mu}{\sum_{j=1}^{j=m} c_j + C} = \frac{\frac{1}{C}\sum_{j=1}^{j=m} c_j y_j + \mu}{\frac{1}{C}\sum_{j=1}^{j=m} c_j + 1} \sim$$

$$\sim (\frac{1}{C}\sum_j c_j y_j + \mu)(1 - \frac{\sum_j c_j}{C}) \sim \mu + \frac{1}{C}(\sum_j c_j y_j - \mu\sum_j c_j) \quad (15)$$

$$\hat{\mu} \sim \mu + \frac{\sum_j [c_j(y_j - \mu)]}{C} \quad (16)$$

Next, we have

$$\hat{\sigma}^2 = \hat{\sigma}(C)^2 = \frac{\sum_j c_j y_j^2 + C(\mu^2 + \sigma^2)}{\sum_j c_j + C} - \hat{\mu}^2 \quad (17)$$

Let us compute $\hat{\sigma}^2$ using (38)

$$\frac{\sum_j c_j y_j^2 + C(\mu^2 + \sigma^2)}{\sum_j c_j + C} \sim$$

$$\sim (\frac{\sum_j c_j y_j^2}{C} + \mu^2 + \sigma^2)(1 - \frac{\sum_j c_j}{C}) \sim$$

$$\sim \mu^2 + \sigma^2 + \frac{1}{C}[\sum_j c_j y_j^2 - (\mu^2 + \sigma^2)\sum_j c_j] \quad (18)$$

$$\hat{\mu}^2 \sim \mu^2 + \frac{2\mu}{C}\sum_{j=1}^{j=m} c_j(y_j - \mu) \quad (19)$$

This gives

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

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

$$= \sigma^2 + \frac{1}{C}[\sum_j c_j y_j^2 - (\mu^2 + \sigma^2)\sum_j c_j - 2\mu\sum_j c_j(y_j - \mu)] \quad (20)$$

And finally

$$\hat{\sigma}^2 \sim \sigma^2 + \frac{\sum_j [(y_j - \mu)^2 - \sigma^2] c_j}{C} \quad (21)$$

$$(y_i - \hat{\mu})^2/\hat{\sigma}^2 \sim \frac{1}{\sigma^2}[(y_i - \mu)^2 -$$

$$- \frac{2(y_i - \mu)\sum_j c_j(y_j - \mu)}{C}] \times$$

$$\times \{1 - \frac{\sum_j c_j[(y_j - \mu)^2 - \sigma^2]}{\sigma^2 C}\} \sim$$

$$\sim \frac{(y_i - \mu)^2}{\sigma^2} - \frac{1}{C\sigma^2}\{\frac{(y_i - \mu)^2}{\sigma^2}\sum_j [(y_j - \mu)^2 - \sigma^2] c_j +$$

$$+ 2(y_i - \mu)\sum_j (y_j - \mu) c_j\} \quad (22)$$

$$\hat{z}_i \sim \frac{1}{(2\pi)^{1/2}\hat{\sigma}} e^{\frac{-(y_i - \mu)^2}{2\sigma^2} + \frac{A_i}{C\sigma^2}} \quad (23)$$

Where

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

Continue this we have

$$\hat{z}_i \sim K e^{\frac{-(y_i - \mu)^2}{2\sigma^2}}(1 + \frac{A_i}{C\sigma^2}) \quad (24)$$

Where

$$K = \frac{1}{(2\pi)^{1/2}\hat{\sigma}}$$

$$1/\hat{\sigma} \sim \frac{1}{\sigma}\{1 - \frac{\sum_j c_j[(y_j - \mu)^2 - \sigma^2]}{2\sigma^2 C}\} \quad (25)$$

$$(1 + \frac{A_i}{C\sigma^2})\{1 - \frac{\sum_j c_j[(y_j - \mu)^2 - \sigma^2]}{2\sigma^2 C}\} \sim$$

$$\sim 1 + \frac{1}{C\sigma^2}\{\frac{(y_i - \mu)^2}{2\sigma^2}\sum_j [(y_j - \mu)^2 - \sigma^2] c_j +$$

$$+ (y_i - \mu)\sum_j (y_j - \mu) c_j - 1/2\sum_j c_j[(y_j - \mu)^2 - \sigma^2]\} \sim$$

$$\sim 1 + \frac{B_i}{C\sigma^2} \quad (26)$$

Where $B_i = [\frac{(y_i - \mu)^2}{2\sigma^2} - 1/2]\sum_j [(y_j - \mu)^2 - \sigma^2] c_j + (y_i - \mu)\sum_j (y_j - \mu) c_j$

Using the last equalities we get

$$\hat{z}_i = z_i + \frac{B_i}{C\sigma^2} z_i \quad (27)$$

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

$$l(\{\hat{z}_i\}) = l(\{z_i\}) + \frac{l(\{B_i z_i\})}{C\sigma^2} \quad (28)$$

and

$$l(\{B_i z_i\}) = \sum_i a_i z_i \{[\frac{(y_i - \mu)^2}{2\sigma^2} - 1/2] \times$$

$$\times \sum_j c_j[(y_j - \mu)^2 - \sigma^2] + (y_i - \mu)\sum_j c_j(y_j - \mu)\} =$$

$$= \sum_i c_i \{[\frac{(y_i - \mu)^2}{2\sigma^2} - 1/2]\sum_j c_j[(y_j - \mu)^2 - \sigma^2] +$$

$$+(y_i-\mu)\sum_j c_j(y_j-\mu)\} =$$

$$= \frac{\{\sum_j c_j[(y_j-\mu)^2-\sigma^2]\}^2}{2\sigma^2} + [\sum_j c_j(y_j-\mu)]^2 \quad (29)$$

$$l(\{\hat{z}_i\}) - l(\{z_i\}) \sim \frac{T}{C}$$

Since by assumption either $\hat{\mu} \neq \mu$ or $\hat{\sigma} \neq \sigma$ $T \neq 0$. Applicability of the lineriazation 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 μ, Σ, i.e. we drop subscript j everywhere in (5, 6), (7) and also set $\hat{z}_i = \frac{|\hat{\Sigma}|^{-1/2}}{(2\pi)^{n/2}} e^{-1/2(y_i-\hat{\mu})^T \Sigma^{-1}(y_i-\hat{\mu})}$

Theorem 3 *Let $F(\{z_i\})$, $i = 1...m$, be differentiable at μ, Σ and $\frac{\delta F(\{z_i\})}{\delta z_i}$ exist at z_i. Let either $\hat{\mu} \neq \mu$ or $\hat{\Sigma} \neq \Sigma$. Then for sufficiently large C*

$$F(\{\hat{z}_i\}) - F(\{z_i\}) = T/C + o(1/C) \quad (30)$$

where $T > 0$. i.e. $F(\{\hat{z}_i\})$ grows proportionally to $1/C$ for sufficiently large C. If Σ represented as a diagonal matrix

$$\Sigma^{-1} = diag[\lambda_1, ..., \lambda_n] \quad (31)$$

then one can write T explicitly as follows:

$$T = T_1 + T_2 + T_3 \quad (32)$$

$$T_1 = \frac{1}{2}\sum_{k\neq l}(\lambda_k^2+\lambda_l^2)(\sum_i c_i a_{ki} a_{li})^2 \quad (33)$$

$$T_2 = \frac{1}{2}\sum_{k=1}^{n}(\lambda_k \sum_i c_i a_{ki}^2 - \sum c_i)^2 \quad (34)$$

$$T_3 = \sum_{k=1}^{n} \lambda_k (\sum_i c_i a_{ki})^2 \quad (35)$$

Proof Our proof will be split in several steps.
Step1: Linerarization
First, we assume that $F(\{z_i\}) = l(\mu, \Sigma) := l(\{z_i\}) := \sum_{i=1}^{i=m} a_i z_i$. Let us set $l(\hat{\mu}, \hat{\Sigma}) := l(\{\hat{z}_i\}) := \sum_{i=1}^{i=m} a_i \hat{z_i}$. Then $c_j = a_j z_j$ 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

$$\hat{\mu} = \hat{\mu}(C) = \frac{\sum_{j=1}^{j=m} c_j y_j + C\mu}{\sum_{j=1}^{j=m} c_j + C} = \frac{\frac{1}{C}\sum_{j=1}^{j=m} c_j y_j + \mu}{\frac{1}{C}\sum_{j=1}^{j=m} c_j + 1} \sim$$

$$\sim (\frac{1}{C}\sum_j c_j y_j + \mu)(1 - \frac{\sum_j c_j}{C}) \sim \mu + \frac{1}{C}(\sum_j c_j y_j - \mu \sum_j c_j) \quad (36)$$

$$\hat{\mu} \sim \mu + \frac{\sum_j [c_j(y_j-\mu)]}{C} \quad (37)$$

Next, we have

$$\hat{\Sigma} = \hat{\Sigma}(C) = \frac{\sum_j c_j y_j y_j^T + C(\mu\mu^T + \Sigma)}{\sum_j c_j + C} - \hat{\mu}\hat{\mu}^T \quad (38)$$

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

$$\frac{\sum_j c_j y_j y_j^T + C(\mu\mu^T+\Sigma)}{\sum_j c_j + C} \sim$$

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

$$\sim \mu\mu^T + \Sigma + \frac{1}{C}[\sum_j c_j y_j y_j^T - (\mu\mu^T + \Sigma)\sum_j c_j] \quad (39)$$

$$\hat{\mu}\hat{\mu}^T \sim \mu\mu^T + \frac{2\mu}{C}\sum_{j=1}^{j=m} c_j(y_j-\mu)^T \quad (40)$$

This gives

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

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

$$= \Sigma + \frac{1}{C}[\sum_j c_j y_j y_j^T - (\mu\mu^T+\Sigma)\sum_j c_j - 2\mu\sum_j c_j(y_j-\mu)^T] \quad (41)$$

And finally

$$\hat{\Sigma} \sim \Sigma + \frac{\sum_j[(y_j-\mu)(y_j-\mu)^T - \Sigma]c_j}{C} \quad (42)$$

$$\hat{\Sigma}^{-1} \sim \Sigma^{-1} - \frac{\Sigma^{-2}\{\sum_j[(y_j-\mu)(y_j-\mu)^T - \Sigma]c_j\}}{C} \quad (43)$$

$$1/2(y_i-\hat{\mu})^T \hat{\Sigma}^{-1}(y_i-\hat{\mu}) \sim 1/2[(y_i-\mu) - \frac{\sum_j c_j(y_j-\mu)}{C}]^T \times$$

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

$$\times[(y_i-\mu) - \frac{\sum_j c_j(y_j-\mu)}{C}] \sim$$

$$\sim 1/2(y_i-\mu)^T\Sigma^{-1}(y_i-\mu) - \frac{A_i}{C} \quad (44)$$

$$A_i = A_{i1} + A_{i2}$$

$$A_{i1} = 1/2\sum_j c_j(y_i-\mu)^T\Sigma^{-2}[(y_j-\mu)(y_j-\mu)^T - \Sigma](y_i-\mu) \quad (45)$$

$$A_{i2} = 1/2\sum_j c_j[(y_j-\mu)^T\Sigma^{-1}(y_i-\mu)+(y_i-\mu)^T\Sigma^{-1}(y_j-\mu)] \quad (46)$$

$$\hat{z}_i \sim \frac{|\hat{\Sigma}|^{-1/2}}{(2\pi)^{n/2}}e^{\frac{-1}{2}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)+\frac{A_i}{C}} \quad (47)$$

Continue this we have

$$\hat{z}_i \sim Ke^{-\frac{1}{2}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)}(1+\frac{A_i}{C}) \quad (48)$$

Where

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

$$|\hat{\Sigma}|^{-1/2} \sim |\Sigma|^{-1/2}\{1+\frac{\sum c_j}{2C}[n - Tr\Sigma^{-1}(y_j-\mu)(y_j-\mu)^T]\} \quad (49)$$

$$(1+\frac{A_i}{C})\{1+\frac{\sum c_j}{2C}[n - Tr\Sigma^{-1}(y_j-\mu)(y_j-\mu)^T]\} \sim$$

$$\sim 1+\frac{1}{C}\{A_i + 1/2\sum_j c_j[n - Tr\Sigma^{-1}(y_j-\mu)(y_j-\mu)^T]\}$$

$$\sim 1+\frac{B_i}{C} \quad (50)$$

Where

$$B_i = A_i + D$$

Here we use

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

and

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

Using the last equalities we get

$$\hat{z}_i \sim z_i + \frac{B_i}{C}z_i \quad (51)$$

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

$$l(\{\hat{z}_i\}) \sim l(\{z_i\}) + \frac{l(\{B_i z_i\})}{C} \quad (52)$$

and

$$T = l(\{B_i z_i\}) =$$

$$= \sum_i c_iA_i + 1/2(\sum_i c_i)\sum_j c_j[n-(y_j-\mu)^T\Sigma^{-1}(y_j-\mu)]$$

$$\sum_i c_iA_i = \sum_i c_iA_{i1} + \sum_i c_iA_{i2} \quad (53)$$

$$\tilde{A}_1 = \sum_i c_iA_{i1} =$$

$$= 1/2\sum_{ij} c_ic_j(y_i-\mu)^T\Sigma^{-2}[(y_j-\mu)(y_j-\mu)^T - \Sigma](y_i-\mu) \quad (54)$$

$$\tilde{A}_2 = \sum_i c_iA_{i2} =$$

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

$$= [\sum_j c_j(y_j-\mu)]^T\Sigma^{-1}[\sum_i c_i(y_i-\mu)] \quad (55)$$

$$l(\{\hat{z}_i\}) - l(\{z_i\}) \sim \frac{T}{C}$$

Step3: Reduction to a diagonal case

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

$$\tilde{A}_1 =$$

$$= 1/2\sum_{ij} c_ic_j(y_i-\mu)^TO^TO\Sigma^{-2}O^T\times$$

$$\times[O(y_j-\mu)(y_j-\mu)^TO^T - O\Sigma O^T]O(y_i-\mu) \quad (56)$$

Step 4: special case - 2-dimensional Gaussians

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

Withot loss of generality we can assume the $\Sigma^{-1} = diag[\lambda_1,\lambda_2]$ is a diagonal 2×2 - matrix with diagonal elemens λ_1 and λ_2.

Let compute

$$A'_1 =$$

$$1/2\sum_{ij} c_ic_j(y_i-\mu)^T\Sigma^{-2}[(y_j-\mu)(y_j-\mu)^T(y_i-\mu)] \quad (57)$$

Let set

$$(y_i-\mu)^T = (a_{1i}, a_{2i}) \quad (58)$$

Then

$$A'_1 =$$

$$1/2\sum_{ij} c_ic_j(a_{1i},a_{2i})\times diag[\lambda_1^2,\lambda_2^2]\times$$

$$(a_{1j}, a_{2,j})^T \times (a_{1j}, a_{2j}) \times (a_{1i}, a_{2i})^T \tag{59}$$

$$= 1/2 \sum_{ij} c_i c_j (\lambda_1^2 a_{1i}, \lambda_2^2 a_{2i})(a_{1j}, a_{2,j})^T (a_{1j} a_{1i} + a_{2j} a_{2i}) \tag{60}$$

$$= 1/2 \sum_{ij} c_i c_j (\lambda_1^2 a_{1i} a_{1j} + \lambda_2^2 a_{2i} a_{2j})(a_{1j} a_{1i} + a_{2j} a_{2i}) \tag{61}$$

$$\begin{gathered} = 1/2(\lambda_1^2 + \lambda_2^2) \sum_{ij} c_i c_j a_{1i} a_{1j} a_{2i} a_{2,j} + \\ +1/2\lambda_1^2 \sum_{ij} c_i c_j a_{1i}^2 a_{1j}^2 \\ +1/2\lambda_2^2 \sum_{ij} c_i c_j a_{2i}^2 a_{2j}^2 = \\ 1/2(\lambda_1^2 + \lambda_2^2)(\sum_{ij} c_i a_{1i} a_{2i})^2 + \\ +1/2\lambda_1^2 (\sum_i c_i a_{1i}^2)^2 \\ +1/2\lambda_2^2 (\sum_i c_i a_{2i}^2)^2 \end{gathered} \tag{62}$$

$$\begin{gathered} A''_1 = \tilde{A}_1 - A'_1 = \\ = -1/2 \sum_{ij} c_i c_j (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) = \end{gathered} \tag{63}$$

$$\begin{gathered} -1/2(\sum_j c_j) \sum_i c_i (a_{1i}, a_{2i}) \times (\lambda_1, \lambda_2) \times (a_{1i}, a_{2i})^T = \\ -1/2(\sum_j c_j)(\sum_i c_i \lambda_1 a_{1i}^2 + \sum_i c_i \lambda_2 a_{2i}^2) \end{gathered} \tag{64}$$

Next, for our 2-dimensional case we have

$$\sum c_i D = (\sum c_i)^2 - 1/2(\sum_j c_j)(\sum_i c_i \lambda_1 a_{1i}^2 + \sum_i c_i \lambda_2 a_{2i}^2) \tag{65}$$

Therefore

$$\begin{gathered} T = \\ 1/2(\lambda_1^2 + \lambda_2^2)(\sum_{ij} c_i a_{1i} a_{2i})^2 + \\ +1/2\lambda_1^2 (\sum_i c_i a_{1i}^2)^2 \\ +1/2\lambda_2^2 (\sum_i c_i a_{2i}^2)^2 - \\ -(\sum_j c_j)(\sum_i c_i \lambda_1 a_{1i}^2 + \sum_i c_i \lambda_2 a_{2i}^2) \\ +(\sum c_i)^2 \\ + \sum c_i c_j (\lambda_1 a_{i1} a_{j1} + \lambda_2 a_{i2} a_{j2}) \end{gathered} \tag{66}$$

And finally

$$\begin{gathered} T = \\ 1/2(\lambda_1^2 + \lambda_2^2)(\sum_{ij} c_i a_{1i} a_{2i})^2 + \\ +1/2(\lambda_1 \sum_i c_i a_{1i}^2 - \sum c_i)^2 \\ +1/2(\lambda_2 \sum_i c_i a_{2i}^2 - \sum c_i)^2 + \\ +\lambda_1 (\sum_i c_i a_{1i})^2 \\ +\lambda_2 (\sum_i c_i a_{2i})^2 \end{gathered} \tag{67}$$

In the above equation:

$$T_1 = 1/2(\lambda_1^2 + \lambda_2^2)(\sum_{ij} c_i a_{1i} a_{2i})^2 \tag{68}$$

$$\begin{gathered} T_2 = 1/2(\lambda_1 \sum_i c_i a_{1i}^2 - \sum c_i)^2 + \\ +1/2(\lambda_2 \sum_i c_i a_{2i}^2 - \sum c_i)^2 \end{gathered} \tag{69}$$

$$\begin{gathered} T_3 = \lambda_1 (\sum_i c_i a_{1i})^2 \\ +\lambda_2 (\sum_i c_i a_{2i})^2 \end{gathered} \tag{70}$$

Step 4: General case - n-dimensional Gaussians

We will perform computations for n -dimensional case.

Withot 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 elemens λ_1, λ_2 and λ_n.

Let compute

$$\begin{gathered} A'_1 = \\ 1/2 \sum_{ij} c_i c_j (y_i - \mu)^T \Sigma^{-2} [(y_j - \mu)(y_j - \mu)^T (y_i - \mu)] \end{gathered} \tag{71}$$

Let set

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

Then

$$\begin{gathered} A'_1 = \\ 1/2 \sum_{ij} c_i c_j (a_{1i}, a_{2i}, ...a_{ni}) \times \\ diag[\lambda_1^2, \lambda_2^2, ...\lambda_n^2] \times (a_{1j}, a_{2j}, ...a_{nj})^T \\ \times (a_{1j}, a_{2j}, ...a_{nj}) \times (a_{1i}, a_{2i}, ...a_{nj})^T \end{gathered} \tag{73}$$

$$= 1/2 \sum_{ij} c_i c_j (\lambda_1^2 a_{1i}, \lambda_2^2 a_{2i}, ...\lambda_2^2 a_{ni})(a_{1j}, a_{2j}, ...a_{nj})^T \times$$

$$\times(\sum_k a_{kj}a_{ki}) \quad (74)$$

$$= 1/2\sum_{ij} c_i c_j (\sum_k \lambda_k^2 a_{ki}a_{kj})\times$$

$$\times(\sum_k a_{kj}a_{ki}) \quad (75)$$

$$= 1/2\sum_{k,l,k\neq l}\sum_{ij} c_i c_j(\lambda_k^2+\lambda_l^2)\times$$

$$\times a_{ki}a_{kj}a_{li}a_{l,j}+$$

$$+1/2\sum_{ij} c_i c_j \sum_k \lambda_k^2 a_{ki}^2 a_{kj}^2 =$$

$$1/2\sum_{k,i\neq l}(\lambda_k^2+\lambda_l^2)(\sum_{ij} c_i a_{ki}a_{li})^2+$$

$$+1/2(\sum_{k,i}\lambda_k c_i a_{2i}^2)^2 \quad (76)$$

Similar (like for the 2-dimensional case) one can compute other componets in T.

Step 5: Invariant transformation points

Here we prove the following

Lemma 4 *Let Σ be a diagonal matrix. Then the following holds. a) $T = 0$ implies that $\Sigma(C) = \Sigma$ and $\mu(C) = \mu$ for $C = 0$.*

b) $\Sigma(C) = \Sigma$ and $\mu(C) = \mu$ for $C = 0$ implies that $\Sigma(C) = \Sigma$ and $\mu(C) = \mu$ for any C.

c) $\Sigma(C) = \Sigma$ and $\mu(C) = \mu$ for some $C \to T = 0$

Proof of Lemma

a) $T = 0 \to T_3 = 0 \to \mu = \frac{\sum c_i y_i}{\sum c_i} \to \mu(0) = \mu$.

Next, $T2 = 0 \to \lambda_k \sum c_i a_{ki}^2 - \sum c_i = 0 \to \lambda_k^{-1} = \frac{\sum y_{ik}^2}{\sum c_i} - \mu_k\mu_k = \Sigma(0)_{kk}$.

Finally, $T_1 = 0 \to \sum c_i(y_{ik}-\mu_k)(y_{il}-\mu_l) = 0 \to \sum c_i y_{ik}y_{il} - c_i y_{ik}\mu_l - c_i y_{il}\mu_k + c_i\mu_k\mu_l = 0 \to \sum c_i(y_{ik}y_{il} - \mu_k\mu_l) = 0 \to \Sigma_{kl}(0) = 0$. This proves a) for $C = 0$.

b) It follows from (5) that if $\mu(C) = \mu$ then

$$\mu(\sum c_i + C) = \sum c_i y_i + C\mu \to$$

$$\mu\sum c_i = \sum c_i y_i$$

Adding to both parts of the above equation $C'\mu$ for any C' we get

$$\mu(\sum c_i + C') = \sum c_i y_i + C'\mu \to \mu(C') = \mu$$

This proves b) for μ.

Similarly, from (6) and a part b) of the lemma for μ we have that $\Sigma(C) = \Sigma$ implies

$$\Sigma\sum c_i = \sum c_i y_i y_i^T - \sum c_i \mu\mu^T$$

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

$$\Sigma\sum c_i + \Sigma C' = \sum c_i y_i y_i^T - \sum c_i\mu\mu^T +$$

$$+C'(\mu\mu^T + \Sigma) - C'\mu\mu^T \to$$

$$\Sigma = \frac{\sum c_i y_i y_i^T + C'(\mu\mu^T+\Sigma)}{\sum c_i + C'} - \mu\mu^T \to$$

$$\Sigma = \Sigma(C')$$

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 ot the theorem. Since by assumption either $\hat{\mu} \neq \mu$ or $\hat{\Sigma} \neq \Sigma$ $T \neq 0$. Applicability of the lineriazation 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 $\mu_j\prime \geq 0$, $\sigma_j\prime \geq 0$ such that $\mu_j/D_j + \mu_j\prime/D_j = 1$, $\sigma_j/E_j + \sigma_j\prime/E_j = 1$. Then we can compute updates as in (1), with c_j as in (5, 6).

$$\hat{\mu}_j = D_j\mu_j\frac{\sum_i c_{ij}\frac{(y_i-\mu_j)}{\sigma_j^2}+C}{\sum_i c_{ij}\frac{(y_i-\mu_j)}{\sigma_j^2}\mu_j + D_jC}$$

$$\hat{\sigma}_j = E_j\frac{\sum_i c_{ij}[-1+\frac{(y_i-\mu_j)^2}{\sigma_j^2}]+C\sigma_j}{\sum_i c_{ij}[-1+\frac{(y_i-\mu_j)^2}{\sigma_j^2}]+E_jC}$$

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

[1] S. Axelrod, V. Goel, R. Gopinath, P. Olsen, and K. Visweswariah, "Discriminative Training of Subspace Constrained GMMs for Speech Recognition," to be submitted to IEEE Transactions on Speech and Audio Processing.

[2] L.E.Baum and J.A. Eagon, "An inequality with applications to statistical prediction for functions of Markov processes and to a model of ecology," *Bull. Amer. Math. Soc.*, vol. 73, pp.360-363, 1967.

[3] A. Gunawardana and W. Byrne, “Discriminative Speaker Adaptation with Conditional Maximum Likelihood Linear Regression,” ICASSP, 2002.

[4] P.S. Gopalakrishnan, D. Kanevsky, D. Nahamoo and A. Nadas, ”An inequality for rational functions with applications to some statistical estimation problems”, IEEE Trans. Information Theory, Vol. 37, No.1 January 1991

[5] D. Kanevsky, ”Growth Transformations for General Functions”, RC22919 (W0309-163), September 25, 2003.

[6] D. Kanevsky, ”Extended Baum transformations for general functions”, in Proc. ICASSP, 2004.

[7] Y. Normandin, ”An improved MMIE Training Algorithm for Speaker Independent, Small Vocabulary, Continuous Speech Recognition”, Proc. ICASSP’91, pp. 537-540, 1991.

[8] R. Schluter, W. Macherey, B. Muler and H. Ney, ”Comparison of discriminative training criteria and optimization methods for speech recognition”, Speech Communication, Vol. 34, pp.287-310, 2001.

[9] V. Valtchev , P.C. Woodland and S. J. Young, ”Lattice-based Discriminative Training for Large Vocabulary Speech Recognition Systems”, Speech Communication, Vol. 22, pp. 303-314, 1996.