Generalization of the Results of Section 9.4.1

It is suggested in Section 9.4.2 that the results of Section 9.4.1 for 2-dimensional observations extend easily to $d$-dimensions. We will work through the details in this note. Specifically, we will consider an HMM with output densities attached to arcs. Let $S$ denote the set of states, let the arcs be indexed by $t \in T$, and let the outputs or emissions take values in $\mathbb{R}^d$ for some finite $d > 0$. Let $L(t)$ and $R(t)$ respectively denote the origin- and destination-states of the arc $t$, and let $p_t$ denote the probability of taking the arc $t$ when the underlying Markov chain is in the state $L(t)$. Clearly,

$$\sum_{t : L(t) = s} p_t = 1, \quad \forall s \in S. \tag{1}$$

For each non-null arc $t$, let the corresponding output density be a multivariate Gaussian,

$$\mathcal{N}_t(y) = \frac{1}{(2\pi)^{d/2} \sqrt{|U_t|}} \exp \left\{-\frac{1}{2} (y - m_t)^T U_t^{-1} (y - m_t) \right\}, \quad \forall y = \begin{bmatrix} y_1 \\ \vdots \\ y_d \end{bmatrix} \in \mathbb{R}^d, \tag{2}$$

where $m_t$ is the mean vector and $U_t$ the covariance matrix of the emitted random vector. Note that $y$ and $m_t$ are column vectors here, while they are row-vectors in the textbook, and the arc-dependence of $m_t$ and $U_t$ is denoted via a subscript here instead of writing $m(t)$ and $U(t)$. The free parameters of the HMM are $\theta = \{\theta_t, t \in T\}$, where $\theta_t = \{p_t, m_t, U_t\}$, the $p_t$’s satisfy the sum-to-one condition (1), and the $U_t$’s are symmetric and positive-definite.

Given an $n$-length observation $Y = \langle y_1, y_2, \ldots, y_n \rangle$ from this HMM, the EM auxiliary function may be constructed as

$$Q(\theta', \theta) = \sum_{\text{All paths } t} P_{\theta'}(t|Y) \log P_{\theta}(t, Y) = \sum_{\text{All paths } t} P_{\theta'}(t|Y) \log \left[ \prod_{l=1}^{k(t)} p_{t_l} \mathcal{N}_{t_l}(y_{t_l}) \right], \tag{3}$$

where $t = \langle t_1, t_2, \ldots, t_{k(t)} \rangle$ denotes any valid path through the HMM, and $k(t)$ denotes its length. While it is not made precise in the textbook, it is to be understood in (3) that
• \( P_{\theta'}(t|\mathbf{Y}) > 0 \) only for paths \( t \) of length \( k(t) \geq n \) that contain exactly \( n \) non-null arcs and \( k(t) - n \) null arcs, and hence other paths need not be considered in the sum over all \( t \);

• the reference to the \( l \)-th output symbol \( y_l \) is valid only after reindexing \( Y = (y_1, y_2, \ldots, y_n) \) and (re)assigning \( y_1 \) to the first non-null arc of \( t \), \( y_2 \) to the second non-null arc of \( t \), and so on, until \( y_n \) to the last non-null arc of \( t \), while no symbols are assigned to its null arcs;

• \( N_t(y_l) \) is computed via (2) for non-null arcs \( t_l \) in \( t \), but \( N_t(\cdot) = 1 \) for all null arcs in \( t \).

Next, given a \( \theta' \), we try to maximize \( Q(\theta', \theta) \) as a function of \( \theta \). To this end, we form the Lagrangian

\[
L(\theta) = Q(\theta', \theta) - \sum_{s \in S} \lambda_s \sum_{i: L(i) = s} p_i. \tag{4}
\]

**Updating the Transition Probabilities** \( p_t \)

Note that for every arc \( t \in T \),

\[
\frac{\partial}{\partial p_t} L(\theta) = \sum_{\text{All paths } t} P_{\theta'}(t|\mathbf{Y}) \frac{\partial}{\partial p_t} \log P_\theta(t, \mathbf{Y}) - \frac{\partial}{\partial p_t} \sum_{s \in S} \lambda_s \sum_{i: L(i) = s} p_i
\]

\[
= \sum_{\text{All paths } t} P_{\theta'}(t|\mathbf{Y}) \frac{\partial}{\partial p_t} \log \left( \prod_{l=1}^{k(t)} p_{t_l} N_{t_l}(y_l) \right) - \lambda_{L(t)}
\]

\[
= \sum_{\text{All paths } t} P_{\theta'}(t|\mathbf{Y}) \left[ \sum_{l: t_l = t} \frac{\partial}{\partial p_t} \log p_{t_l} + \sum_{l=1}^{k(t)} \log N_{t_l}(y_l) \right] - \lambda_{L(t)}
\]

\[
= \sum_{\text{All paths } t} P_{\theta'}(t|\mathbf{Y}) \sum_{l: t_l = t} \frac{1}{p_{t_l}} - \lambda_{L(t)}.
\]

For each arc \( t \in T \), equating the derivative to 0 yields

\[
\frac{\partial}{\partial p_t} L(\theta) = 0 \implies \sum_{\text{All paths } t} P_{\theta'}(t|\mathbf{Y}) \sum_{l: t_l = t} \frac{1}{p_{t_l}} = \lambda_{L(t)}
\]

\[
\frac{1}{P_{\theta'}(\mathbf{Y})} \sum_{\text{All paths } t: t_1 = t} \sum_{l: t_l = t} P_{\theta'}(t, \mathbf{Y}) \frac{1}{p_{t_l}} = \lambda_{L(t)}
\]

\[
\frac{1}{\lambda_{L(t)} P_{\theta'}(\mathbf{Y})} \sum_{\text{All paths } t: t_1 = t} \sum_{l: t_l = t} P_{\theta'}(t, \mathbf{Y}) = p_t,
\]

The brute-force way to compute the double sum \( \psi_t \) is to
1. exhaustively enumerate all paths \( t \),

2. traverse each path \( t = (t_1, \ldots, t_{k(t)}) \), and

3. every time \( t_l = t \), i.e. the arc \( t \) is traversed, add \( P_\theta(t, Y) \) to an accumulator for \( \psi_t \).

Once the \( \psi_t \) are accumulated for all \( t \in T \), the role of \( \lambda_s \) for every state \( s \in S \), is to ensure that the probabilities of arcs leaving \( s \) sum to unity. Therefore

\[
\sum_{t : L(t) = s} p_t = 1 \Rightarrow \sum_{t : L(t) = s} \psi_t \frac{\lambda_s P_\theta(Y)}{\alpha'_{l-1}(L(t))} = 1 \Rightarrow \frac{1}{\lambda_s P_\theta(Y)} = \frac{1}{\sum_{t : L(t) = s} \psi_t}.
\]

Now, a \( n \)-stage trellis captures all paths \( t \) capable of producing \( Y \): all paths \( t \) with \( P_\theta(t, Y) > 0 \). Furthermore, if \( t \) is a non-null arc, then it appears exactly \( n \) times in the trellis, once in each trellis stage, and every time a path \( t \) traverses the \( l \)-th copy of \( t \), \( l = 1, \ldots, n \), an output \( y_l \) is produced. Therefore, the contribution of the \( l \)-th copy of \( t \) to \( \psi_t \) is the sum of the probabilities of all the paths \( t \) that pass through \( t \) in the \( l \)-th stage of the trellis, namely

\[
P_\theta(y_1, \ldots, y_{l-1}, s = L(t)) \times p'_t \times N'_t(y_l) \times P_\theta(y_{l+1}, \ldots, y_n | s = R(t)) \times \beta'_l(R(t)).
\]

Therefore the total contribution from all stages of the trellis for a non-null arc \( t \) is

\[
\psi_t = \sum_{\text{All paths } t : t_1 = t} \sum_{l=1}^{\infty} \alpha'_{l-1}(L(t)) \times p'_t \times N'_t(y_l) \times \beta'_l(R(t)).
\]

Similarly, a null arc \( t \) may appear on a path \( t \) within in each “column” of a vertically aligned set of states. If the \( l \)-th copy of \( t \) in the trellis is designated as the one traversed before producing \( y_l \) (i.e. \( \text{between} \) producing \( y_{l-1} \) and \( y_l \), \( l = 1, \ldots, n \)), then its contribution to \( \psi_t \) from all paths \( t \) is

\[
P_\theta(y_1, \ldots, y_{l-1}, s = L(t)) \times p'_t \times P_\theta(y_{l+1}, \ldots, y_n | s = R(t)) \times \beta'_{l-1}(R(t)), \quad l = 1, \ldots, n.
\]

Therefore, for a null arc \( t \), the total contribution from all stages of the trellis is

\[
\psi_t = \sum_{\text{All paths } t : t_1 = t} \sum_{l=1}^{\infty} \alpha'_{l-1}(L(t)) \times p'_t \times \beta'_{l-1}(R(t)).
\]

As noted above, for every state \( s \),

\[
K_s = \sum_{t : L(t) = s} \psi_t = \sum_{l=1}^{\infty} \sum_{t : L(t) = s} \alpha'_{l-1}(s) \left[ p'_t \times N'_t(y_l) \times \beta'_l(R(t)) \text{ for non-null arcs } t \right. \\
\left. + p'_t \times \beta'_{l-1}(R(t)) \text{ for null arcs } t \right]
\]

\[
= \sum_{l=1}^{\infty} \alpha'_{l-1}(s) \sum_{t : L(t) = s} \left[ p'_t \times N'_t(y_l) \times \beta'_l(R(t)) \text{ for non-null arcs } t \right. \\
\left. + p'_t \times \beta'_{l-1}(R(t)) \text{ for null arcs } t \right] = \sum_{l=1}^{\infty} \alpha'_{l-1}(s) \beta'_{l-1}(s).
\]
Updating the Mean Vectors $m_t$

Next, note that for every non-null arc $t \in T$, if we let $m_t = [m_{t,1} \ldots m_{t,d}]^T$, then

$$\frac{\partial}{\partial m_{t,i}} L(\theta) = \sum_{\text{All paths } t} P_{0}^{t}(t|Y) \frac{\partial}{\partial m_{t,i}} \log P_{0}(t, Y) - \frac{\partial}{\partial m_{t,i}} \sum_{s \in S} \lambda_s \sum_{i : L(i)=s} p_t$$

$$= \sum_{\text{All paths } t} P_{0}^{t}(t|Y) \frac{\partial}{\partial m_{t,i}} \left[ \prod_{l=1}^{k(t)} p_i \mathcal{N}_t(y_i) \right] - 0$$

$$= \sum_{\text{All paths } t} P_{0}^{t}(t|Y) \frac{\partial}{\partial m_{t,i}} \left[ \sum_{l=1}^{k(t)} \log p_i + \sum_{l=1}^{k(t)} \log \mathcal{N}_t(y_i) \right]$$

$$= \sum_{\text{All paths } t} P_{0}^{t}(t|Y) \left[ 0 + \sum_{l : t_l = t} \frac{\partial}{\partial m_{t,i}} \log \mathcal{N}_t(y_i) \right]$$

$$= \sum_{\text{All paths } t} P_{0}^{t}(t|Y) \sum_{l : t_l = t} \frac{\partial}{\partial m_{t,i}} \left\{ -\frac{1}{2} (y_l - m_t)^T \mathcal{U}^{-1}_t (y_l - m_t) - \log (2\pi)^{d/2} \sqrt{\mathcal{U}_t} \right\}$$

$$= -\frac{1}{2} \sum_{\text{All paths } t} P_{0}^{t}(t|Y) \sum_{l : t_l = t} \frac{\partial}{\partial m_{t,i}} \left\{ (y_l - m_t)^T \mathcal{U}^{-1}_t (y_l - m_t) + 0 \right\} .$$

The partial derivatives of $L(\theta)$ with respect to the components of the mean vector $m_t$ may therefore be compactly written as the vector

$$\begin{bmatrix}
\frac{\partial}{\partial m_{t,1}} L(\theta) \\
\vdots \\
\frac{\partial}{\partial m_{t,d}} L(\theta)
\end{bmatrix} = -\frac{1}{2} \sum_{\text{All paths } t} P_{0}^{t}(t|Y) \sum_{l : t_l = t} \begin{bmatrix}
\frac{\partial}{\partial m_{t,i}} \left\{ (y_l - m_t)^T \mathcal{U}^{-1}_t (y_l - m_t) \right\} \\
\vdots \\
\frac{\partial}{\partial m_{t,d}} \left\{ (y_l - m_t)^T \mathcal{U}^{-1}_t (y_l - m_t) \right\}
\end{bmatrix} . \tag{5}$$

Next, note that $x^T b = \sum_{m=1}^{d} x_m b_m$, and thus

$$\frac{\partial}{\partial x_i} x^T b = \frac{\partial}{\partial x_i} \sum_{m=1}^{d} x_m b_m = b_i,$$

from which it follows that for any $d \times 1$ vectors $x = [x_1 \ldots x_d]^T$ and $b = [b_1 \ldots b_d]^T$

$$\begin{bmatrix}
\frac{\partial}{\partial x_1} x^T b \\
\vdots \\
\frac{\partial}{\partial x_d} x^T b
\end{bmatrix} = b.$$

Similarly $x^T A x = \sum_{m=1}^{d} \sum_{n=1}^{d} x_m a_{mn} x_n$, and therefore

$$\frac{\partial}{\partial x_i} x^T A x = \frac{\partial}{\partial x_i} \sum_{m=1}^{d} \sum_{n=1}^{d} x_m a_{mn} x_n = \sum_{m=1, m \neq i}^{d} x_m a_{mi} + \sum_{n=1, n \neq i}^{d} a_{in} x_n + 2a_{ii} x_i = \sum_{m=1}^{d} [a_{mi} + a_{im}] x_m.$$
Therefore, for a symmetric $d \times d$ matrix $A$,

$$\begin{bmatrix}
\frac{\partial}{\partial x_1} x^T A x \\
\vdots \\
\frac{\partial}{\partial x_d} x^T A x
\end{bmatrix} = A^T x + A x = 2 A x.$$ 

Set $A = U_t^{-1}$ and $x = (y_t - m_t)$ and note that

$$\begin{bmatrix}
\frac{\partial}{\partial m_{t,1}} (y_l - m_t)^T U_t^{-1} (y_l - m_t) \\
\vdots \\
\frac{\partial}{\partial m_{t,d}} (y_l - m_t)^T U_t^{-1} (y_l - m_t)
\end{bmatrix} = - \begin{bmatrix}
\frac{\partial}{\partial (y_l, m_{t,1})} (y_l - m_t)^T U_t^{-1} (y_l - m_t) \\
\vdots \\
\frac{\partial}{\partial (y_l, m_{t,d})} (y_l - m_t)^T U_t^{-1} (y_l - m_t)
\end{bmatrix},$$

where the negative sign comes about due to the fact that

$$\begin{bmatrix}
\frac{\partial}{\partial m_{t,1}} (y_l, m_{t,1}) \\
\vdots \\
\frac{\partial}{\partial m_{t,d}} (y_l, m_{t,d})
\end{bmatrix} = \begin{bmatrix} 1 \\
\vdots \\
1 \end{bmatrix}.$$ 

This easily provides the partial derivatives (5) of $L(\theta)$ with respect to the components of $m_t$. To obtain the update equation for $m_t$ we must set it to zero. i.e. Set

$$\begin{bmatrix}
\frac{\partial}{\partial m_{t,1}} L(\theta) \\
\vdots \\
\frac{\partial}{\partial m_{t,d}} L(\theta)
\end{bmatrix} = 0 \Rightarrow U_t \times \sum_{\text{All paths } t : t_l = t} P_{\theta'}(t|\mathbf{Y}) \sum_{l : t_l = t} U_t^{-1} (y_l - m_t) = U_t \times 0$$

$$\Rightarrow \sum_{\text{All paths } t : t_l = t} \sum_{l : t_l = t} P_{\theta'}(t|\mathbf{Y})(y_l - m_t) = 0$$

$$\Rightarrow \sum_{\text{All paths } t : t_l = t} \sum_{l : t_l = t} P_{\theta'}(t|\mathbf{Y})y_l = \sum_{\text{All paths } t : t_l = t} \sum_{l : t_l = t} P_{\theta'}(t|\mathbf{Y})m_t$$

$$\Rightarrow \sum_{\text{All paths } t} \frac{1}{\sum_{l : t_l = t} P_{\theta'}(t, \mathbf{Y})} \sum_{l : t_l = t} \sum_{\text{All paths } t : t_l = t} P_{\theta'}(t, \mathbf{Y})y_l = m_t$$

$$\Rightarrow \sum_{l=1}^n \left[ \alpha_{l-1}'(L(t)) \frac{p_t' N_t'(y_l) \beta_t'(R(t))}{P_{\theta'}(Y)} \right] y_l = m_t,$$

where the last step, once again, follows from the argument that the overall double sum, similar to $\psi_t$ above, may be obtained by separately computing the contribution of all paths $t$ to the arc $t$ in a particular ($l$-th) stage of the trellis, and accumulating such contributions for $l = 1, \ldots, n$.

To interpret the mean update equation (6) qualitatively, recall that

$$\frac{\alpha_{l-1}'(L(t)) p_t' N_t'(y_l) \beta_t'(R(t))}{P_{\theta'}(Y)} = P_{\theta'}(t_l = t | \mathbf{Y}) = P_{\theta'}(y_l \text{ was emitted by arc } t | \mathbf{Y}).$$

The value of $m_t$ in (6) that maximizes $L(\theta)$ may thus be seen as a “sample mean,” where each observation $y_l$ in the sample has a fractional count — the probability that it came from arc $t$ — and the sample size is the expected number of times the arc $t$ was traversed, or as a “weighted mean,” where the weight of the sample $y_l$ is the probability that it was emitted from $t$. 


Updating the Covariance Matrices $U_t$

To find the $U_t$ that maximizes $L(\theta)$, let $V_t = U_t^{-1}$, and $v_{t,ij}$ denote\textsuperscript{1} the $ij$-th element of $V_t$.

\[
\frac{\partial}{\partial v_{t,ij}} L(\theta) = \sum_{\text{All paths } t} P_{\theta}(t|Y) \frac{\partial}{\partial v_{t,ij}} \log P_{\theta}(t, Y) - \frac{\partial}{\partial v_{t,ij}} \sum_{s \in S} \lambda_s \sum_{t : L(\tilde{t}) = s} p_t = 0
\]

\[
\sum_{\text{All paths } t} P_{\theta}(t|Y) \frac{\partial}{\partial v_{t,ij}} \left[ \sum_{l=1}^{k(t)} \log p_{t_l} + \sum_{l=1}^{k(t)} \log N_{t_l}(y_l) \right] = 0
\]

\[
\sum_{\text{All paths } t} P_{\theta}(t|Y) \left[ 0 + \sum_{l : t_l = t} \frac{\partial}{\partial v_{t,ij}} \log N_{t_l}(y_l) \right] = 0
\]

\[
\sum_{\text{All paths } t} P_{\theta}(t|Y) \left[ \frac{1}{2} (y_l - m_t)^T U_t^{-1} (y_l - m_t) - \log (2\pi)^\frac{d}{2} \sqrt{|U_t|} \right]
\]

\[
= -\frac{1}{2} \sum_{\text{All paths } t} P_{\theta}(t|Y) \left[ (y_l - m_t)^T U_t^{-1} (y_l - m_t) + \log |U_t| \right]
\]

\[
= -\frac{1}{2} \sum_{\text{All paths } t} P_{\theta}(t|Y) \left[ (y_l - m_t)^T V_t (y_l - m_t) - \log |V_t| \right]. \tag{7}
\]

Next, for any $d \times 1$ vector $x = [x_1 \ldots x_d]^T$ and symmetric positive-definite $d \times d$ matrix $A$, the partial derivatives of the scalar $x^T Ax$ with respect to the components of the matrix $A$ are

\[
\begin{bmatrix}
\frac{\partial}{\partial a_{11}} x^T Ax & \cdots & \frac{\partial}{\partial a_{1d}} x^T Ax \\
\vdots & \ddots & \vdots \\
\frac{\partial}{\partial a_{d1}} x^T Ax & \cdots & \frac{\partial}{\partial a_{dd}} x^T Ax
\end{bmatrix} = xx^T,
\]

and the partial derivatives of the scalar $\log |A|$ with respect to the components of $A$ are

\[
\begin{bmatrix}
\frac{\partial}{\partial a_{11}} \log |A| & \cdots & \frac{\partial}{\partial a_{1d}} \log |A| \\
\vdots & \ddots & \vdots \\
\frac{\partial}{\partial a_{d1}} \log |A| & \cdots & \frac{\partial}{\partial a_{dd}} \log |A|
\end{bmatrix} = A^{-1}.
\]

The derivatives $\frac{\partial}{\partial v_{t,ij}} L(\theta)$ are obtained by setting $x = (y_l - m_t)$ and $A = U_t^{-1}$ in the formulae above:

\[
\begin{bmatrix}
\frac{\partial}{\partial v_{t,11}} L(\theta) & \cdots & \frac{\partial}{\partial v_{t,1d}} L(\theta) \\
\vdots & \ddots & \vdots \\
\frac{\partial}{\partial v_{t,d1}} L(\theta) & \cdots & \frac{\partial}{\partial v_{t,dd}} L(\theta)
\end{bmatrix} = -\frac{1}{2} \sum_{\text{All paths } t} P_{\theta}(t|Y) \sum_{l : t_l = t} \{(y_l - m_t)(y_l - m_t)^T - V_t^{-1}\},
\]

$V_t$, the inverse of the covariance matrix, is sometimes called the precision matrix, and is often of interest in multivariate statistics and factor analysis.
and the choice of $V_t$ (equivalently $U_t$) that makes $\frac{\partial}{\partial \psi_{t,i}} L(\theta) = 0$ for all $i$ and $j$ is

$$\frac{1}{2} \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \sum_{l:t_l=1} \left\{ (\mathbf{y}_l - \mathbf{m}_l)(\mathbf{y}_l - \mathbf{m}_l)^T - V_t^{-1} \right\} = 0$$

$$\frac{1}{\sum_{\text{All paths } \mathbf{t}} \sum_{l:t_l=1}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \sum_{l:t_l=1} \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) (\mathbf{y}_l - \mathbf{m}_l)(\mathbf{y}_l - \mathbf{m}_l)^T = V_t^{-1}$$

$$\frac{1}{\sum_{\text{All paths } \mathbf{t}} \sum_{l:t_l=1}} P_{\theta'}(\mathbf{t}, \mathbf{Y}) \sum_{l:t_l=1} \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}, \mathbf{Y}) (\mathbf{y}_l - \mathbf{m}_l)(\mathbf{y}_l - \mathbf{m}_l)^T = U_t,$$  

(8)

where, once again, the parameters $\theta'_t$ and $N'_t$ correspond to $\theta'$, and $\alpha_{t-1}'(\cdot)$ and $\beta_{t}'(\cdot)$ are the forward- and backward-probabilities computed using $\theta'$ on a $n$-stage trellis, with the null arcs going between vertically aligned states and only non-null arcs traversing left-to-right.

Finally, the covariance update equation (9) once again follows by observing that, similar to $\psi_t$, the double sum may be obtained by first accumulating the contribution of all paths $\mathbf{t}$ in the trellis to the $t$-th copy of an arc $t$, and then summing these contributions for $l = 1, \ldots, n$.

$$U_t = \frac{\sum_{l=1}^n \left[ \sum_{t_l=1} \alpha_{t-1}'(L(t)) \left( \alpha_{l}'(R(t)) \right) \right] (\mathbf{y}_l - \mathbf{m}_l)(\mathbf{y}_l - \mathbf{m}_l)^T}{\sum_{l=1}^n \left[ \sum_{t_l=1} \alpha_{t-1}'(L(t)) \left( \alpha_{l}'(R(t)) \right) \right]}.$$  

(9)

It remains to verify that the updated matrix $U_t$ of (9) is symmetric and positive-(semi)definite, thereby justifying why the Lagrangian of (4) did not impose any constraints on the components of the parameter set $\theta$ corresponding to the $U_t$'s.

It is straightforward to see that the updated $U_t$ of (9) satisfies

$$U_t^T = \left[ \sum_{l=1}^n P_{\theta'}(t_l = t, \mathbf{Y}) (\mathbf{y}_l - \mathbf{m}_l)(\mathbf{y}_l - \mathbf{m}_l)^T \right]^T$$

$$= \frac{\sum_{l=1}^n P_{\theta'}(t_l = t, \mathbf{Y}) [(\mathbf{y}_l - \mathbf{m}_l)(\mathbf{y}_l - \mathbf{m}_l)^T]^T}{\sum_{l=1}^n P_{\theta'}(t_l = t, \mathbf{Y})}$$

$$= U_t,$$

and that for any vector $\mathbf{x}$,

$$\mathbf{x}^T U_t \mathbf{x} = \frac{\sum_{l=1}^n P_{\theta'}(t_l = t, \mathbf{Y}) \mathbf{x}^T(\mathbf{y}_l - \mathbf{m}_l)(\mathbf{y}_l - \mathbf{m}_l)^T \mathbf{x}}{\sum_{l=1}^n P_{\theta'}(t_l = t, \mathbf{Y})}$$

$$= \frac{\sum_{l=1}^n P_{\theta'}(t_l = t, \mathbf{Y}) (\mathbf{x}^T(\mathbf{y}_l - \mathbf{m}_l))^2}{\sum_{l=1}^n P_{\theta'}(t_l = t, \mathbf{Y})}$$

$$\geq 0.$$

This guarantees that $U_t$ will always be a bona fide covariance matrix.