

# **Semi-Tied Covariance Matrices for Hidden Markov Models**

Mark J. F. Gales

Presented by Winston Lee

# Abstract

---

- Covariance matrices used with CDHMMs.
    - Diagonal:  $O(n)$
    - Full:  $O(n^2)$
    - Block-diagonal:  $O(n^2/k)$
- } A dramatic increase in the number of parameters per Gaussian component!
- This paper introduces a new form of covariance matrix which
    - allows a few “**full**” covariance matrices to be shared over many distributions and
    - each distribution maintains its own “**diagonal**” covariance matrix.
  - This technique fits within the standard maximum-likelihood criterion used for training HMM’s.

# Introduction

---

- Using diagonal covariance matrices:
  - To model non-Gaussian distributions
  - To model correlations
- It is preferable to decorrelate the feature vector as far as possible. (??)
  - **Feature-space schemes:** e.g. the discrete cosine transform, LDA, KL transform
    - Hard to find a single transform which decorrelates all elements of the feature vector **for all states!**
  - **Model-space schemes:**
    - A different transform is selected depending on which component the observation was hypothesized to be generated from.
    - In the limit a transform may be used for each component, which is equivalent to a **full covariance matrix system.**

## Introduction (cont.)

---

- Semi-tied covariance matrices:
  - A natural extension of the **state-specific rotation** scheme
  - The transform is estimated in a maximum-likelihood (ML) fashion given the current model parameters.
  - The optimization is performed using a simple iterative scheme, which is guaranteed to increase the likelihood of the training data.
  - An alternative approach to solve the optimization problem of DHLDA.

## State-Specific Rotation

---

- The full covariance matrix can model inter feature-vector element correlation.
  - The number of parameters:  $n(n+3)/2$
  - **Diagonal covariance matrices** are commonly used in large-vocabulary speech recognition.
  - The data associated with each state is modeled by **multiple Gaussian components**.
  - By using multiple components, any strong correlations may be **implicitly** modeled, in addition to the possible non-Gaussian nature of the data.
- If the correlations in the data could be **explicitly** modeled
  - the number of Gaussian components per state could be reduced
  - reducing the size of the model sets
  - allowing the Gaussian components to model the non-Gaussian nature of the data, rather than the correlations

## State-Specific Rotation (cont.)

---

- **State-specific rotation** is proposed for modeling the correlations in the feature vector.
  - A full covariance matrix is calculated for each state in the system.
  - This is decomposed into its eigenvectors and eigenvalues.
  - All data from that state is then decorrelated using the eigenvectors calculated.
  - Multiple diagonal covariance matrix Gaussian components are then trained.

"...if correlations exist between the observation vector parameters, they are **implicitly** captured, and their presence (or absence) is not accounted for **explicitly**." (Ljolje, 1994)

...simulations showed that implicit models of parameter correlations (Gaussian mixtures) can **adequately** capture parameter correlations. (Rabiner, Juang, Levinson & Sondhi, 1985)

## State-Specific Rotation (cont.)

- The covariance matrix associated with each state,  $s$ , is decomposed as

$$\Sigma_{full}^{(s)} = U^{(s)} \Lambda^{(s)} U^{(s)T} \quad (\text{spectral decomposition})$$

$\downarrow$  eigenmatrix       $\searrow$  diagonal

where

$$\Sigma_{full}^{(s)} = \frac{\sum_{\tau} \gamma_s(\tau) \left( o(\tau) - \overset{\text{state mean}}{\mu^{(s)}} \right) \left( o(\tau) - \mu^{(s)} \right)^T}{\sum_{\tau=1}^T \gamma_s(\tau)}$$

$$\gamma_s(\tau) = p(q_s(\tau) | M, O_T)$$

$\searrow$  state  $s$  at time  $\tau$

## State-Specific Rotation (cont.)

---

- When training, instead of using the standard observation vector,  $o(\tau)$ , a state specific observation vector,  $o^{(s)}(\tau)$ , is used where  $o^{(s)}(\tau) = U^{(s)T} o(\tau)$

- Each component,  $m$ , associated with that particular state,  $s$ , is then trained using

$$\mu^{(sm)} = \frac{\sum_{\tau} \gamma_m(\tau) o^{(s)}(\tau)}{\sum_{\tau} \gamma_m(\tau)} \quad \left( \mu^{(sm)} = U^{(s)T} \mu^{(m)} \right)$$

$$\Sigma_{diag}^{(m)} = \text{diag} \left( \frac{\sum_{\tau} \gamma_m(\tau) \left( o^{(s)}(\tau) - \mu^{(sm)} \right) \left( o^{(s)}(\tau) - \mu^{(sm)} \right)^T}{\sum_{\tau} \gamma_m(\tau)} \right)$$



## State-Specific Rotation (cont.)

---

- Question:

$$\frac{\sum_{\tau} \gamma_m(\tau) \left( o^{(s)}(\tau) - \mu^{(sm)} \right) \left( o^{(s)}(\tau) - \mu^{(sm)} \right)^T}{\sum_{\tau} \gamma_m(\tau)} \text{ is a diagonal matrix?}$$

- Hint: if the correlations in a state are removed, the correlations in its components are also removed. Is it true?

## State-Specific Rotation (cont.)

---

- The covariance matrix associated with each component is

$$\Sigma^{(m)} = U^{(s)} \Sigma_{diag}^{(m)} U^{(s)T}$$

- During recognition and training the likelihood used for component of state is

$$L(o(\tau); \mu^{(m)}, \Sigma^{(m)}, U^{(s)}) = N(o^{(s)}(\tau); \mu^{(sm)}, \Sigma_{diag}^{(m)})$$

- Computationally, this is relatively efficient, as it is only necessary to perform one rotation per state, in contrast to standard full covariance matrices, which require the equivalent of one rotation per component.

## State-Specific Rotation (cont.)

---

- Drawbacks:
  - It does not fit within the standard ML estimation framework for training HMM's.
  - The transforms are not related to the multiple-component models being used to model the data.
- Solutions:
  - Using the average within-component covariance per state, as opposed to the global state covariance.

$$\Sigma_{full}^{(s)} = \frac{\sum_{m \in M^{(s)}, \tau} \gamma_m(\tau) \left( o(\tau) - \mu^{(m)} \right) \left( o(\tau) - \mu^{(m)} \right)^T}{\sum_{m \in M^{(s)}, \tau} \gamma_m(\tau)}$$

This still does not yield a transform that is guaranteed to increase the likelihood!

## State-Specific Rotation (cont.)

---

- In (Gales & Woodland, 1996), the component-specific variance may be written as

$$\Sigma^{(m)} = L_{diag}^{(m)} \Sigma_{full}^{(s)'} L_{diag}^{(m)T}$$

where

$$\Sigma_{full}^{(s)'} = \frac{\sum_{m \in M^{(s)}} L_{diag}^{(m)-1} \left( \sum_{\tau} \gamma_m(\tau) \left( o(\tau) - \mu^{(m)} \right) \left( o(\tau) - \mu^{(m)} \right)^T \right) \left( L_{diag}^{(m)-1} \right)^T}{\sum_{m \in M^{(s)}, \tau} \gamma_m(\tau)}$$

and

$$\Sigma_{diag}^{(m)} = L_{diag}^{(m)} L_{diag}^{(m)T} \quad \text{(Cholesky factorization)}$$

There is a significant increase  
in the computational load  
during recognition!

## Semi-Tied Covariance Matrices

---

- Instead of having a distinct covariance matrix for every component in the recognizer, each covariance matrix consists of two elements:
  - A component specific diagonal covariance element  $\Sigma_{diag}^{(m)}$
  - A semi-tied class-dependent, nondiagonal matrix  $H^{(r)}$
  - The form of the covariance matrix is then

$$\Sigma^{(m)} = H^{(r)} \Sigma_{diag}^{(m)} H^{(r)T}$$

where  $H^{(r)}$  may be tied over a set of components

- It is very complex to optimize these parameters directly so an **expectation-maximization approach** is adopted.
  - Parameters for each component  $m$ :

$$c^{(m)} \quad \mu^{(m)} \quad \Sigma_{diag}^{(m)} \quad H^{(r)}$$

## Semi-Tied Covariance Matrices (cont.)

---

- The EM approach:
  - First, rather than dealing with  $H^{(r)}$ , it is simpler to deal with its inverse,  $A^{(r)}$ , thus  $A^{(r)} = H^{(r)-1}$ .
  - The auxiliary function:

$$Q(M, \hat{M}) = \sum_{m \in M(r), \tau} \gamma_m(\tau) \left( \log \left( \frac{|\hat{A}^{(r)}|^2}{|\hat{\Sigma}_{diag}^{(m)}|} \right) - \left( o(\tau) - \hat{\mu}^{(m)} \right)^T \hat{A}^{(r)T} \hat{\Sigma}_{diag}^{(m)-1} \hat{A}^{(r)} \left( o(\tau) - \hat{\mu}^{(m)} \right) \right)$$

The equation includes red dashed circles and arrows highlighting specific terms: a red dashed circle around  $\frac{|\hat{A}^{(r)}|^2}{|\hat{\Sigma}_{diag}^{(m)}|}$  with an arrow pointing to  $|\Sigma^{(m)}|^{-1}$ , and another red dashed circle around  $\hat{A}^{(r)T} \hat{\Sigma}_{diag}^{(m)-1} \hat{A}^{(r)}$  with an arrow pointing to  $\Sigma^{(m)-1}$ .

## Semi-Tied Covariance Matrices (cont.)

---

- If all the model parameters are to be simultaneously optimized then the Q-function may be rewritten as

$$Q(M, \hat{M}) = \sum_{m \in M(r), \tau} \gamma_m(\tau) \log \left( \frac{|\hat{A}^{(r)}|^2}{|\text{diag}(\hat{A}^{(r)} W^{(m)} \hat{A}^{(r)T})|} \right) - n\beta$$

where

$$W^{(m)} = \frac{\sum_{\tau} \gamma_m(\tau) (o(\tau) - \hat{\mu}^{(m)}) (o(\tau) - \hat{\mu}^{(m)})^T}{\sum_{\tau} \gamma_m(\tau)} \quad \hat{\mu}^{(m)} = \frac{\sum_{\tau} \gamma_m(\tau) o(\tau)}{\sum_{\tau} \gamma_m(\tau)}$$
$$\beta = \sum_{m \in M(r), \tau} \gamma_m(\tau)$$

## Appendix A

- Why 
$$\sum_{m \in M^{(r)}, \tau} \gamma_m(\tau) \left( o(\tau) - \hat{\mu}^{(m)} \right)^T \hat{A}^{(r)T} \hat{\Sigma}_{diag}^{(m)-1} \hat{A}^{(r)} \left( o(\tau) - \hat{\mu}^{(m)} \right)$$
  

$$= n \sum_{m \in M^{(r)}, \tau} \gamma_m(\tau)?$$

$$\begin{aligned} & \sum_{m \in M^{(r)}, \tau} \gamma_m(\tau) \left( o(\tau) - \hat{\mu}^{(m)} \right)^T \hat{A}^{(r)T} \hat{\Sigma}_{diag}^{(m)-1} \hat{A}^{(r)} \left( o(\tau) - \hat{\mu}^{(m)} \right) \\ &= \sum_{m \in M^{(r)}, \tau} \gamma_m(\tau) \text{trace} \left[ \left( o(\tau) - \hat{\mu}^{(m)} \right)^T \hat{A}^{(r)T} \hat{\Sigma}_{diag}^{(m)-1} \hat{A}^{(r)} \left( o(\tau) - \hat{\mu}^{(m)} \right) \right] \\ &= \sum_{m \in M^{(r)}, \tau} \gamma_m(\tau) \text{trace} \left[ \hat{A}^{(r)T} \hat{\Sigma}_{diag}^{(m)-1} \hat{A}^{(r)} \left( o(\tau) - \hat{\mu}^{(m)} \right) \left( o(\tau) - \hat{\mu}^{(m)} \right)^T \right] \end{aligned}$$

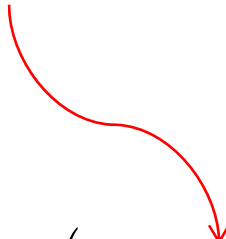
$$\because \text{trace}(AB) = \text{trace}(BA)$$



## Appendix A (cont.)

---

$$\begin{aligned}
 &= \text{trace} \left[ \sum_{m \in M^{(r)}} \hat{A}^{(r)T} \hat{\Sigma}_{diag}^{(m)-1} \hat{A}^{(r)} \sum_{\tau} \gamma_m(\tau) \left( o(\tau) - \hat{\mu}^{(m)} \right) \left( o(\tau) - \hat{\mu}^{(m)} \right)^T \right] \\
 &= \text{trace} \left[ \sum_{m \in M^{(r)}} \hat{A}^{(r)T} \hat{\Sigma}_{diag}^{(m)-1} \hat{A}^{(r)} \left( \hat{A}^{(r)T} \hat{\Sigma}_{diag}^{(m)-1} \hat{A}^{(r)} \right)^{-1} \sum_{\tau} \gamma_m(\tau) \right] \\
 &= \text{trace} \left[ \sum_{m \in M^{(r)}, \tau} I_n \gamma_m(\tau) \right] \\
 &= \sum_{m \in M^{(r)}, \tau} \gamma_m(\tau) * \text{trace}(I_n) \\
 &= n\beta
 \end{aligned}$$



$$\begin{aligned}
 &\therefore \frac{\sum_{\tau} \gamma_m(\tau) \left( o(\tau) - \hat{\mu}^{(m)} \right) \left( o(\tau) - \hat{\mu}^{(m)} \right)^T}{\sum_{\tau} \gamma_m(\tau)} \\
 &= \hat{\Sigma}^{(m)} = \left( \hat{A}^{(r)T} \hat{\Sigma}_{diag}^{(m)-1} \hat{A}^{(r)} \right)^{-1}
 \end{aligned}$$

## Semi-Tied Covariance Matrices (cont.)

---

- The ML estimate of the diagonal element of the covariance matrix is given by

$$\hat{\Sigma}_{diag}^{(m)} = \text{diag}\left(\hat{A}^{(r)} W^{(m)} \hat{A}^{(r)T}\right)$$

- The reestimation formulae for the component weights and transition probabilities are identical to the standard HMM cases. (Rabiner, 1989)
- Unfortunately, optimizing the new Q-function is nontrivial and more complicated, so an alternative approach is proposed next.

## Semi-Tied Covariance Matrices (cont.)

---

- $\hat{A}^{(r)}$  is initialized either with the current estimate of the semi-tied transform or an identity matrix.
  - (1) Estimate the mean, which is independent of the other model parameters.
  - (2) Use the current estimate of the semi-tied transform  $\hat{A}^{(r)}$ , and estimate the set of component specific diagonal variances. This set of parameters will be denoted as  $\{\hat{\Sigma}_{diag}^{(r)}\} = \{\hat{\Sigma}_{diag}^{(r)}, m \in M^{(r)}\}$
  - (3) Estimate the semi-tied transform  $\hat{A}^{(r)}$  using the current set  $\{\hat{\Sigma}_{diag}^{(r)}\}$
  - (4) Go to (2) until convergence, or appropriate criterion satisfied.
- However, optimizing the semi-tied transform requires an iterative estimation scheme even after fixing all other model parameters.

## Semi-Tied Covariance Matrices (cont.)

- Selecting a particular row of  $\hat{A}^{(r)}$ ,  $\hat{a}_i^{(r)}$ , and rewriting the former Q-function using the current set  $\{\hat{\Sigma}_{diag}^{(r)}\}$

$$Q\left(M, \hat{M}; \{\hat{\Sigma}_{diag}^{(r)}\}\right) = \sum_{m \in M^{(r)}, \tau} \gamma_m(\tau) \left( \log \left( \frac{|\hat{A}^{(r)}|}{a_i^{(r)} c_i^T} \right)^2 - \log |\hat{\Sigma}_{diag}^{(m)}| - \sum_j \frac{\left( \hat{a}_j^{(r)} \hat{o}^{(m)}(\tau) \right)^2}{\sigma_{diag_j}^{(m)2}} \right)$$

The  $i$ th row vector of the cofactors
element  $i$  of the leading diagonal

- In the Appendix C, it is shown that the ML estimate for the  $i$ th row of the semi-tied transform,  $\hat{a}_i^{(r)}$ , is given by

$$\hat{a}_i^{(r)} = c_i G^{(ri)-1} \sqrt{\frac{\beta}{c_i G^{(ri)-1} c_i^T}}$$

$$G^{(ri)} = \sum_{m \in M^{(r)}} \frac{1}{\hat{\sigma}_{diag_i}^{(m)2}} W^{(m)} \sum_{\tau} \gamma_m(\tau)$$

## Appendix B

---

- If  $A$  is a square matrix, then the minor entry of  $a_{ij}$  is denoted by  $M_{ij}$  and is defined to be the determinant of the submatrix that remains after the  $i$ -th row and the  $j$ -th column are deleted from  $A$ . The number  $(-1)^{i+j}M_{ij}$  is denoted by  $c_{ij}$  and is called the **cofactor** of  $a_{ij}$ .
- Given the matrix

$$B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

$$c_{23} = (-1)^{2+3}(M_{23})$$

$$c_{33} = (-1)^{3+3}(M_{33})$$

$$M_{23} = \begin{bmatrix} b_{11} & b_{12} & \times \\ \times & \times & \times \\ b_{31} & b_{32} & \times \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ b_{31} & b_{32} \end{bmatrix} = b_{11}b_{32} - b_{12}b_{31}$$

## Appendix B (cont.)

---

- Given the  $n$  by  $n$  matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

- The determinant of  $A$  can be written as the sum of its cofactors multiplied by the entries that generated them.

$$\det(A) = a_{1j}c_{1j} + a_{2j}c_{2j} + a_{3j}c_{3j} + \cdots + a_{nj}c_{nj} = A^{(j)T} c^{(j)}$$

(cofactor expansion along the  **$j$ th** column)

$$\det(A) = a_{i1}c_{i1} + a_{i2}c_{i2} + a_{i3}c_{i3} + \cdots + a_{in}c_{in} = A_{(i)} c_{(i)}^T$$

(cofactor expansion along the  **$i$ th** row)

## Appendix C

- The objective is to maximize the following expression with respect to  $\hat{A}^{(r)}$  :

$$\begin{aligned}
 & Q\left(M, \hat{M}; \left\{ \hat{\Sigma}_{diag}^{(r)} \right\} \right) \\
 &= \sum_{m \in M^{(r)}, \tau} \gamma_m(\tau) \left( \log \left( \hat{a}_i^{(r)} c_i^T \right)^2 - \log \left| \hat{\Sigma}_{diag}^{(m)} \right| - \sum_j \frac{\left( \hat{a}_j^{(r)} \hat{\sigma}^{(m)}(\tau) \right)^2}{\hat{\sigma}_{diag_j}^{(m)2}} \right) \\
 &= \beta \log \left( c_i \hat{a}_i^{(r)T} \right)^2 - \sum_j \left( \hat{a}_j^{(r)} G^{(rj)} \hat{a}_j^{(r)T} \right) + K
 \end{aligned}$$

All terms independent of  $\hat{A}^{(r)}$

$\sum_{m \in M^{(r)}, \tau} \gamma_m(\tau)$

$\sum_{m \in M^{(r)}} \frac{1}{\hat{\sigma}_{diag_j}^{(m)2}} W^{(m)} \sum_{\tau} \gamma_m(\tau)$

## Appendix C (cont.)

- Differentiating with respect to  $\hat{a}_i^{(r)T}$  and equating to zero:

$$\frac{\partial \left( \beta \log \left( c_i \hat{a}_i^{(r)T} \right)^2 - \sum_j \left( \hat{a}_j^{(r)} G^{(rj)} \hat{a}_j^{(r)T} \right) + K \right)}{\partial \left( \hat{a}_i^{(r)T} \right)} = 0$$

$$\boxed{\frac{\partial (x^T A x)}{\partial x} = (A^T + A)x}$$

$$\Rightarrow \frac{\beta}{\left( c_i \hat{a}_i^{(r)T} \right)^2} \times 2 \times c_i \hat{a}_i^{(r)T} \times c_i^T - 2G^{(ri)} \hat{a}_i^{(r)T} = 0$$

$$\Rightarrow \frac{\beta}{c_i \hat{a}_i^{(r)T}} \times c_i^T = G^{(ri)} \hat{a}_i^{(r)T}$$

$$\Rightarrow \frac{\beta}{c_i \hat{a}_i^{(r)T}} \times c_i = \hat{a}_i^{(r)} G^{(ri)}$$

$$\Rightarrow \beta c_i G^{(ri)-1} = \underbrace{c_i \hat{a}_i^{(r)T}}_{\text{scalar}} \hat{a}_i^{(r)}$$



## Appendix C (cont.)

$$\beta c_i G^{(ri)-1} = c_i \hat{a}_i^{(r)T} \hat{a}_i^{(r)} \rightarrow \text{scalar}$$

$$(\hat{a}_i^{(r)} = \alpha c_i G^{(ri)-1}) \quad \hat{a}_i^{(r)} \text{ must be in the direction of } c_i G^{(ri)-1}$$

$$\Rightarrow \beta c_i G^{(ri)-1} = \alpha^2 c_i G^{(ri)-1} c_i^T c_i G^{(ri)-1}$$

$$\Rightarrow \alpha = \pm \sqrt{\frac{\beta}{c_i G^{(ri)-1} c_i^T}} \quad \text{only the positive root is considered}$$

$$\Rightarrow \hat{a}_i^{(r)} = c_i G^{(ri)-1} \sqrt{\frac{\beta}{c_i G^{(ri)-1} c_i^T}}$$

## Semi-Tied Covariance Matrices (cont.)

---

- It can be shown that

$$Q(M, \hat{M}) \geq Q(M, \hat{M}; \{\hat{\Sigma}_{diag}^{(r)}\})$$

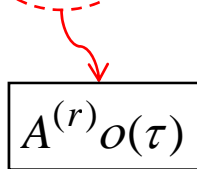
with equality when diagonal elements of the covariance matrix are given by

$$\hat{\Sigma}_{diag}^{(m)} = \text{diag}(\hat{A}^{(r)} W^{(m)} \hat{A}^{(r)T})$$

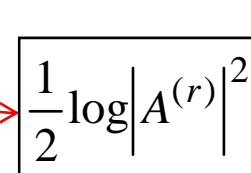
- During recognition the log-likelihood is based on

$$\log(L(o(\tau); \mu^{(m)}, \Sigma^{(m)}, A^{(r)}))$$

$$= \log(N(o^{(r)}(\tau); A^{(r)} \mu^{(m)}, \Sigma_{diag}^{(m)})) + \log|A^{(r)}|$$



$$A^{(r)} o(\tau)$$



$$\frac{1}{2} \log|A^{(r)}|^2$$

## Appendix D

---

$$\begin{aligned}
& \log\left(L\left(o(\tau); \mu^{(m)}, \Sigma^{(m)}, A^{(r)}\right)\right) \\
&= -\frac{1}{2} \log|\Sigma^{(m)}| - \frac{1}{2} \left(o(\tau) - \mu^{(m)}\right)^T \Sigma^{(m)-1} \left(o(\tau) - \mu^{(m)}\right) \\
&= -\frac{1}{2} \log\left|A^{(r)-1} \Sigma_{diag}^{(m)} A^{(r)T-1}\right| - \dots \\
& \quad \frac{1}{2} \left(A^{(r)-1} o^{(r)}(\tau) - A^{(r)-1} \mu^{(rm)}\right)^T A^{(r)T} \Sigma_{diag}^{(m)-1} A^{(r)} \left(A^{(r)-1} o^{(r)}(\tau) - A^{(r)-1} \mu^{(rm)}\right) \\
&= -\frac{1}{2} \log|\Sigma_{diag}^{(m)}| - \frac{1}{2} \log\left|A^{(r)T-1} A^{(r)-1}\right| - \dots \quad \boxed{|AB| = |BA| = |B||A|} \\
& \quad \frac{1}{2} \left(o^{(r)}(\tau) - \mu^{(rm)}\right)^T A^{(r)T-1} A^{(r)T} \Sigma_{diag}^{(m)-1} A^{(r)} A^{(r)-1} \left(o^{(r)}(\tau) - \mu^{(rm)}\right) \\
&= -\frac{1}{2} \log|\Sigma_{diag}^{(m)}| - \frac{1}{2} \left(o^{(r)}(\tau) - A^{(r)} \mu^{(m)}\right)^T \Sigma_{diag}^{(m)-1} \left(o^{(r)}(\tau) - A^{(r)} \mu^{(m)}\right) - \frac{1}{2} \log\left|A^{(r)-1}\right|^2 \\
&= \log\left(N\left(o^{(r)}(\tau); A^{(r)} \mu^{(m)}, \Sigma_{diag}^{(m)}\right)\right) + \frac{1}{2} \log\left|A^{(r)}\right|^2 \quad \boxed{|A^{-1}| = |A|^{-1}}
\end{aligned}$$

## Semi-Tied Covariance Matrices (cont.)

---

- The difference between semi-tied covariance matrices and state-specific rotations:
  - The semi-tied covariance matrices are trained in an ML sense on the training data given the current model set.
  - It would only be possible to train a state-based rotation in an ML sense when all the values of  $\Sigma_{diag}^{(m)}$  associated with a particular transform are **the same**.