Gaussians and Continuous-Density HMMs

Mark Hasegawa-Johnson

ECE 417: Multimedia Signal Processing

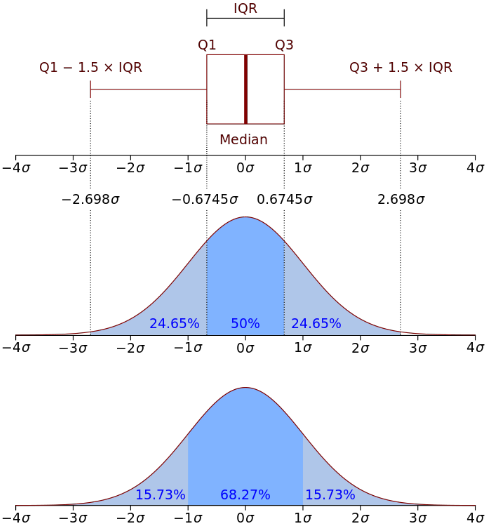
Outline

Gaussian (Normal) pdf

- Gauss considered this problem: under what circumstances does it make sense to estimate the mean of a distribution, μ, by taking the average of the experimental values, $m = \frac{1}{n}\sum_{i=1}^{n} x_i$?
- He demonstrated that m is the maximum likelihood estimate of μ if (not only if!) X is distributed with the following probability density:

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Gaussian pdf



https://commons.wikimedia.org/wiki/File:Boxplot_vs_PDF.svg

Unit Normal pdf

Suppose that X is normal with mean μ and standard deviation σ (variance σ^2):

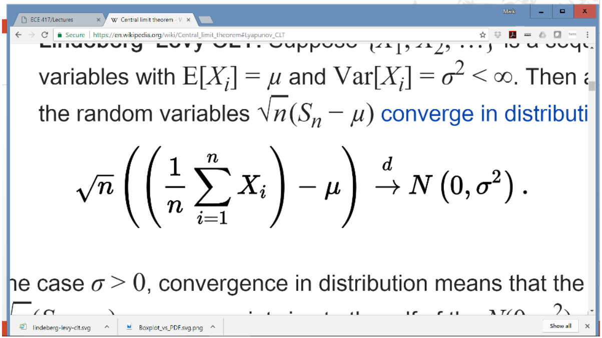
$$p_X(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Then $U = \left(\frac{X-\mu}{\sigma}\right)$ is normal with mean 0 and standard deviation 1:

$$p_U(u) = \mathcal{N}(u; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

Central Limit Theorem

The Gaussian pdf is important because of the Central Limit Theorem. Suppose X_i are i.i.d. (independent and identically distributed), each having mean μ and variance σ^2. Then


ECE 417/Lectures
Central limit theorem - W
Mark
Secure | https://en.wikipedia.org/wiki/Central_limit_theorem#Lyapunov_CLT

variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Then a
the random variables $\sqrt{n}(S_n - \mu)$ converge in distributi

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right).$$

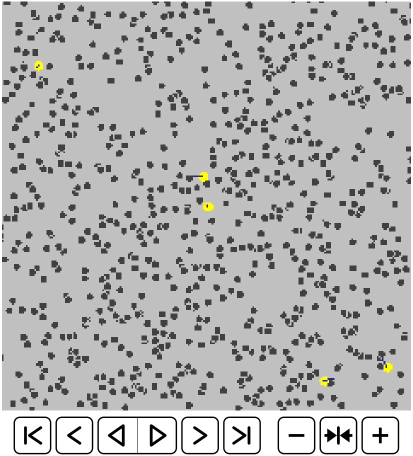
he case $\sigma > 0$, convergence in distribution means that the

lindeberg-levy-clt.svg
Boxplot_vs_PDF.svg.png
Show all

Brownian motion

The Central Limit Theorem matters because Einstein showed that the movement of molecules, in a liquid or gas, is the sum of n i.i.d. molecular collisions.

In other words, the position after t seconds is Gaussian, with mean 0, and with a variance of Dt, where D is some constant.



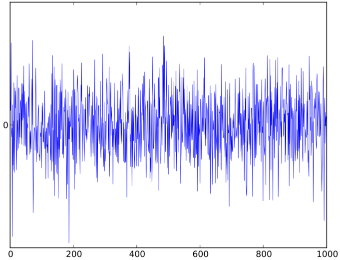
https://commons.wikimedia.org/wiki/File:Brownianmotion5particles150fra gif

White Noise

- Sound = air pressure fluctuations caused by velocity of air molecules
- Velocity of warm air molecules without any external sound source = Gaussian

Therefore:

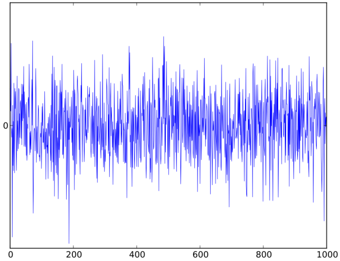
- Sound produced by warm air molecules without any external sound source = Gaussian noise
- Electrical signals: same.



https://commons.wikimedia.org/wiki/File:White_noise.svg

White Noise

- White Noise = noise in which each sample of the signal, x_n, is i.i.d.
- Why "white"? Because the Fourier transform, $X(\omega)$, is a zero-mean random variable whose variance is independent of frequency ("white")
- Gaussian White Noise: $x[n]$ are i.i.d. and Gaussian



https://commons.wikimedia.org/wiki/File:White_noise.svg

Outline

1. Gaussians, Brownian motion, and white noise
2. Gaussian Random Vector
3. HMM with Gaussian Observation Probabilities
4. Summary

Vector of Independent Gaussian Variables

Suppose we have a frame containing D samples from a Gaussian white noise process, $x_1, \ldots, x_D$. Let's stack them up to make a vector:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix}$$

This whole frame is random. In fact, we could say that $\mathbf{x}$ is a sample value for a Gaussian random vector called X, whose elements are $X_1, \ldots, X_D$:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_D \end{bmatrix}$$

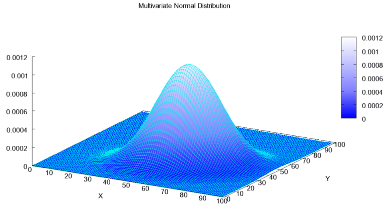
Vector of Independent Gaussian Variables

Suppose that the N samples are i.i.d., each one has the same mean, μ, and the same variance, σ^2. Then the pdf of this random vector is

$$p_X(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \sigma^2 \mathbf{I}) = \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

Vector of Independent Gaussian Variables

Here's an example from Wikipedia with a mean of about 50 and a standard deviation of about 12.



https://commons.wikimedia.org/wiki/File:Multivariate_Gaussian.png

Independent Gaussians that aren't identically distributed

Suppose that the N samples are independent Gaussians that aren't identically distributed, i.e., X_i has mean μ_i and variance σ_i^2. Then the pdf of this random vector is

$$p_X(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi\sigma_d^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and covariance matrix:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_D \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots \\ 0 & \sigma_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Independent Gaussians that aren't identically distributed

Anpther useful form is:

$$p_X(\mathbf{x}) = \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi\sigma_d^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2}$$
$$= \frac{1}{(2\pi)^{D/2}\prod_{i=1}^{D}\sigma_d} e^{-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_d-\mu_d}{\sigma_d}\right)^2}$$

Example

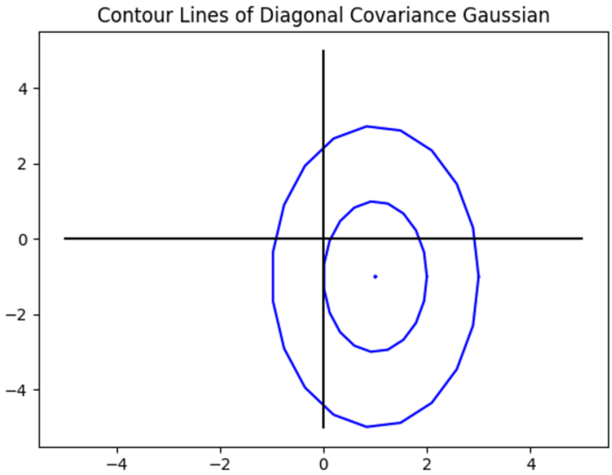
Suppose that $\mu_1 = 1$, $\mu_2 = -1$, $\sigma_1^2 = 1$, and $\sigma_2^2 = 4$. Then

$$p_X(\mathbf{x}) = \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi\sigma_d^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2} = \frac{1}{4\pi} e^{-\frac{1}{2}\left((x_1-1)^2+\left(\frac{x_2+1}{2}\right)^2\right)}$$

The pdf has its maximum value, $p_X(\mathbf{x}) = \frac{1}{4\pi}$, at $\mathbf{x} = \boldsymbol{\mu} = \left[\begin{array}{c} 1 \\ -1 \end{array}\right]$.

It drops to $p_X(\mathbf{x}) = \frac{1}{4\pi\sqrt{e}}$ at $\mathbf{x} = \left[\begin{array}{c} \mu_1 \pm \sigma_1 \\ \mu_2 \end{array}\right]$ and at $\mathbf{x} = \left[\begin{array}{c} \mu_1 \\ \mu_2 \pm \sigma_2 \end{array}\right]$. It drops to $p_X(\mathbf{x}) = \frac{1}{4\pi e^2}$ at $\mathbf{x} = \left[\begin{array}{c} \mu_1 \pm 2\sigma_1 \\ \mu_2 \end{array}\right]$ and at $\mathbf{x} = \left[\begin{array}{c} \mu_1 \\ \mu_2 \pm 2\sigma_2 \end{array}\right]$.

Example


Contour Lines of Diagonal Covariance Gaussian
4
2
0
−2
−4
−4
−2
0
2
4

Facts about linear algebra #1: determinant of a diagonal matrix

Suppose that $\mathbf{\Sigma}$ is a diagonal matrix, with variances on the diagonal:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots \\ 0 & \sigma_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Then its determinant is

$$|\mathbf{\Sigma}| = \prod_{i=1}^{D} \sigma_d^2$$

So we can write the Gaussian pdf as

$$p_X(\mathbf{x}) = \frac{1}{|2\pi\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_d-\mu_d}{\sigma_d}\right)^2}$$

Facts about linear algebra #2: inverse of a diagonal matrix

Suppose that $\mathbf{\Sigma}$ is a diagonal matrix, with variances on the diagonal:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots \\ 0 & \sigma_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Then its inverse is:

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots \\ 0 & \frac{1}{\sigma_2^2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Facts about linear algebra #3: weighted distance

Suppose that

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_D \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots \\ 0 & \sigma_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Then

$$\sum_{i=1}^{D} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 = [x_1 - \mu_1, x_2 - \mu_2, \ldots] \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots \\ 0 & \frac{1}{\sigma_2^2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \end{bmatrix}$$

$$= (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

Mahalanobis distance: Diagonal covariance

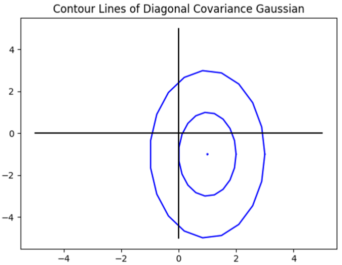
The Mahalanobis distance between vectors $\mathbf{x}$ and $\boldsymbol{\mu}$, weighted by covariance matrix $\boldsymbol{\Sigma}$, is defined to be

$$d_{\boldsymbol{\Sigma}}(\mathbf{x},\boldsymbol{\mu}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

If $\boldsymbol{\Sigma}$ is a diagonal matrix, the Mahalanobis distance is

$$d_{\boldsymbol{\Sigma}}(\mathbf{x},\boldsymbol{\mu}) = \sum_{i=1}^{D}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2$$

The contour lines of equal Mahalanobis distance are ellipses.



https://commons.wikimedia.org/wiki/File:Multivariate_Gaussian.png

Independent Gaussians that aren't identically distributed

So if we have independent Gaussians that aren't identically distributed, we can write the pdf as

$$p_X(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}\prod_{i=1}^{D}\sigma_i} e^{-\frac{1}{2}\sum_{i=1}^{D}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2}$$

or as

$$p_X(\mathbf{x}) = \frac{1}{|2\pi\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

or as

$$p_X(\mathbf{x}) = \frac{1}{|2\pi\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}d_{\boldsymbol{\Sigma}}^2(\mathbf{x},\boldsymbol{\mu})}$$

Outline

Review: HMM with Discrete Observations

1. **Initial State Probabilities:**

$$\pi_i' = \frac{E\left[\#\text{ state sequences that start with } q_1 = i\right]}{\#\text{ state sequences in training data}}$$

2. **Transition Probabilities:**

$$\pi_i' = \frac{E\left[\#\text{ frames in which } q_{t-1} = i, q_t = j\right]}{E\left[\#\text{ frames in which } q_{t-1} = i\right]}$$

3. **Observation Probabilities:**

$$b_j'(k) = \frac{E\left[\#\text{ frames in which } q_t = j, k_t = k\right]}{E\left[\#\text{ frames in which } q_t = j\right]}$$

Baum-Welch with Gaussian Probabilities

The requirement that we vector-quantize the observations is a problem. It means that we can't model the observations very precisely.

It would be better if we could model the observation likelihood, $b_j(\mathbf{x})$, as a probability density in the space $\mathbf{x} \in \Re^D$. One way is to use a parameterized function that is guaranteed to be a properly normalized pdf. For example, a Gaussian:

$$b_i(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$$

Diagonal-Covariance Gaussian pdf

Let's assume the feature vector has D dimensions, $\mathbf{x}_t = [x_{t,1}, \ldots, x_{t,D}]$. The Gaussian pdf is

$$b_i(\mathbf{x}_t) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}_t-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_i)^T}$$

The logarithm of a Gaussian is

$$\ln b_i(\mathbf{x}_t) = -\frac{1}{2}\left((\mathbf{x}_t - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i) + \ln|\boldsymbol{\Sigma}_i| + C\right)$$

where the constant is $C = D\ln(2\pi)$.

Baum-Welch

Baum-Welch maximizes the expected log probability, i.e.,

$$E_{\mathbf{q}|\mathbf{X}}\left[\ln b_i(\mathbf{x}_t)\right] = -\frac{1}{2}\sum_{i=1}^{N}\gamma_t(i)\left((\mathbf{x}_t-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_i)+\ln|\boldsymbol{\Sigma}_i|+C\right)$$

If we include all of the frames, then we get

$$\begin{aligned} E_{\mathbf{q}|\mathbf{X}}\left[\ln p(\mathbf{X},\mathbf{q}|\Lambda)\right] &= \text{other terms} \\ &-\frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\gamma_t(i)\left((\mathbf{x}_t-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_i)+\ln|\boldsymbol{\Sigma}_i|+C\right) \end{aligned}$$

where the "other terms" are about $a_{i,j}$ and π_i, and have nothing to do with $\boldsymbol{\mu}_i$ or $\boldsymbol{\Sigma}_i$.

M-Step: optimum $\boldsymbol{\mu}$

First, let's optimize $\boldsymbol{\mu}$. We want

$$0 = \frac{\partial}{\partial \boldsymbol{\mu}_q} \sum_{t=1}^{T} \sum_{i=1}^{N} \gamma_t(i)(\mathbf{x}_t - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_i)$$

Re-arranging terms, we get

$$\boldsymbol{\mu}_q' = \frac{\sum_{t=1}^{T} \gamma_t(q)\mathbf{x}_t}{\sum_{t=1}^{T} \gamma_t(q)}$$

M-Step: optimum $\boldsymbol{\Sigma}$

Second, let's optimize $\boldsymbol{\Sigma}_i$. For this, it's easier to express the log likelihood as

$$E_{\mathbf{q}|\mathbf{X}}\left[\ln p_X(\mathbf{X},\mathbf{q})\right] = \text{other stuff} - \frac{1}{2}\sum_{t=1}^{T}\gamma_t(i)\sum_{d=1}^{D}\left(\ln\sigma_{i,d}^2 + \frac{(x_{t,d}-\mu_{i,d})^2}{\sigma_{i,d}^2}\right)$$

Its scalar derivative is

$$\frac{\partial E_{\mathbf{q}|\mathbf{X}}\left[\ln p_X(\mathbf{X},\mathbf{q})\right]}{\partial\sigma_{i,d}^2} = -\frac{1}{2}\sum_{t=1}^{T}\gamma_t(i)\left(\frac{1}{\sigma_{i,d}^2} - \frac{(x_{t,d}-\mu_{i,d})^2}{\sigma_{i,d}^4}\right)$$

Which we can solve to find

$$\sigma_{i,d}^2 = \frac{\sum_{t=1}^{T}\gamma_t(i)(x_{t,d}-\mu_{t,d})^2}{\sum_{t=1}^{T}\gamma_t(i)}$$

Minimizing the cross-entropy: optimum σ

Arranging all the scalar derivatives into a matrix, we can write

$$\boldsymbol{\Sigma}_i' = \frac{\sum_{t=1}^{T} \gamma_t(i)(\mathbf{x}_t - \boldsymbol{\mu}_i)(\mathbf{x}_t - \boldsymbol{\mu}_i)^T}{\sum_{t=1}^{T} \gamma_t(i)}$$

- Actually, the above formula holds even if the Gaussian has a non-diagonal covariance matrix, but Gaussians with non-diagonal covariance matrices work surprisingly badly in HMMs.
- For a diagonal-covariance Gaussian, we evaluate only the diagonal elements of the vector outer product $(\mathbf{x}_t - \boldsymbol{\mu}_i)(\mathbf{x}_t - \boldsymbol{\mu}_i)^T$

Summary: Gaussian Observation PDFs

So we can use Gaussians for $b_j(\mathbf{x})$:

- **E-Step:**

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i'} \alpha_t(i')\beta_t(i')}$$

- **M-Step:**

$$\boldsymbol{\mu}_i' = \frac{\sum_{t=1}^{T} \gamma_t(i)\mathbf{x}_t}{\sum_{t=1}^{T} \gamma_t(i)}$$

$$\boldsymbol{\Sigma}_i' = \frac{\sum_{t=1}^{T} \gamma_t(i)(\mathbf{x}_t - \boldsymbol{\mu}_i)(\mathbf{x}_t - \boldsymbol{\mu}_i)^T}{\sum_{t=1}^{T} \gamma_t(i)}$$

Outline

Summary: Independent Gaussians that aren't identically distributed

$$
\begin{aligned}
p_X(\mathbf{x}) &= \frac{1}{(2\pi)^{D/2}\prod_{i=1}^{D}\sigma_i} e^{-\frac{1}{2}\sum_{i=1}^{D}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2} \\
&= \frac{1}{|2\pi\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \\
&= \frac{1}{|2\pi\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}d_{\boldsymbol{\Sigma}}^2(\mathbf{x},\boldsymbol{\mu})}
\end{aligned}
$$

Summary: Gaussian Observation PDFs

So we can use Gaussians for $b_j(\mathbf{x})$:

- **E-Step:**

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i'} \alpha_t(i')\beta_t(i')}$$

- **M-Step:**

$$\boldsymbol{\mu}_i' = \frac{\sum_{t=1}^{T} \gamma_t(i)\mathbf{x}_t}{\sum_{t=1}^{T} \gamma_t(i)}$$

$$\boldsymbol{\Sigma}_i' = \frac{\sum_{t=1}^{T} \gamma_t(i)(\mathbf{x}_t - \boldsymbol{\mu}_i)(\mathbf{x}_t - \boldsymbol{\mu}_i)^T}{\sum_{t=1}^{T} \gamma_t(i)}$$