Working with Gaussians, Linear Gaussian Models

Piyush Rai

Probabilistic Machine Learning (CS772A)

Aug 26, 2017

Gaussian Distribution

- The (multivariate) Gaussian with mean $\boldsymbol{\mu}$ and cov. matrix $\boldsymbol{\Sigma}$

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \quad = \quad \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

Gaussian Distribution

- The (multivariate) Gaussian with mean $\boldsymbol{\mu}$ and cov. matrix $\boldsymbol{\Sigma}$

$$
\begin{aligned}
\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right\} \\
&= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right] \right\} \qquad \text{where } \mathbf{S} = (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top
\end{aligned}
$$

Gaussian Distribution

- The (multivariate) Gaussian with mean $\boldsymbol{\mu}$ and cov. matrix $\boldsymbol{\Sigma}$

$$
\begin{aligned}
\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\
&= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\right\} \qquad \text{where } \mathbf{S} = (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top
\end{aligned}
$$

- An alternate representation: The "information form"

$$
\mathcal{N}_c(\boldsymbol{x}|\boldsymbol{\xi},\boldsymbol{\Lambda}) = (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^\top\boldsymbol{\Lambda}\boldsymbol{x}+\boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}-2\boldsymbol{x}^\top\boldsymbol{\xi}\right)\right\}
$$

Gaussian Distribution

- The (multivariate) Gaussian with mean $\boldsymbol{\mu}$ and cov. matrix $\boldsymbol{\Sigma}$

$$\begin{aligned}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\\ &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\right\} \qquad \text{where } \mathbf{S}=(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\end{aligned}$$

- An alternate representation: The "information form"

$$\mathcal{N}_c(\boldsymbol{x}|\boldsymbol{\xi},\boldsymbol{\Lambda}) = (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^\top\boldsymbol{\Lambda}\boldsymbol{x}+\boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}-2\boldsymbol{x}^\top\boldsymbol{\xi}\right)\right\}$$

where $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\xi}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ are the "natural parameters" (recall exp. family).

Gaussian Distribution

- The (multivariate) Gaussian with mean $\boldsymbol{\mu}$ and cov. matrix $\boldsymbol{\Sigma}$

$$\begin{aligned}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\\ &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\right\} \qquad \text{where } \mathbf{S}=(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\end{aligned}$$

- An alternate representation: The "information form"

$$\mathcal{N}_c(\boldsymbol{x}|\boldsymbol{\xi},\boldsymbol{\Lambda}) = (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^\top\boldsymbol{\Lambda}\boldsymbol{x}+\boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}-2\boldsymbol{x}^\top\boldsymbol{\xi}\right)\right\}$$

  where $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\xi}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ are the "natural parameters" (recall exp. family).

- Note that there is a term quadratic in $\boldsymbol{x}$ (involves $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$) and linear in $\boldsymbol{x}$ (involves $\boldsymbol{\xi}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$)

Gaussian Distribution

- The (multivariate) Gaussian with mean $\boldsymbol{\mu}$ and cov. matrix $\boldsymbol{\Sigma}$

$$\begin{aligned}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\right\} \qquad \text{where } \mathbf{S} = (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\end{aligned}$$

- An alternate representation: The "information form"

$$\mathcal{N}_c(\boldsymbol{x}|\boldsymbol{\xi},\boldsymbol{\Lambda}) = (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^\top\boldsymbol{\Lambda}\boldsymbol{x}+\boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}-2\boldsymbol{x}^\top\boldsymbol{\xi}\right)\right\}$$

  where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\xi} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ are the "natural parameters" (recall exp. family).

- Note that there is a term quadratic in $\boldsymbol{x}$ (involves $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$) and linear in $\boldsymbol{x}$ (involves $\boldsymbol{\xi} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$)
- Information form can help recognize $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of a Gaussian when doing algebraic manipulations

Estimating Parameters of Gaussian: MLE

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Goal: Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Goal: Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Simple to do MLE for this task

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Goal: Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Simple to do MLE for this task

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Goal: Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Simple to do MLE for this task

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Goal: Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Simple to do MLE for this task

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Goal: Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Simple to do MLE for this task

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Plugging in $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\}$

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Goal: Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Simple to do MLE for this task

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Plugging in $\mathcal{N}(x|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-\mathbf{1}}(\boldsymbol{x} - \boldsymbol{\mu})\}$ and ignoring the constants

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad = \quad \frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-\mathbf{1}} (\boldsymbol{x}_n - \boldsymbol{\mu})$$

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Goal: Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Simple to do MLE for this task

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Plugging in $\mathcal{N}(x|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-\mathbf{1}}(\boldsymbol{x}-\boldsymbol{\mu})\}$ and ignoring the constants

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-\mathbf{1}} (\boldsymbol{x}_n - \boldsymbol{\mu}) \\
&= \frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} \text{trace}[\boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top]
\end{aligned}$$

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Goal: Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Simple to do MLE for this task

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Plugging in $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\}$ and ignoring the constants

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \\
&= \frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} \text{trace}[\boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top] \\
&= \frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \text{trace}[\boldsymbol{\Sigma}^{-1} \mathbf{S}_\mu]
\end{aligned}$$

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Goal: Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Simple to do MLE for this task

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Plugging in $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\}$ and ignoring the constants

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \\
&= \frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} \text{trace}[\boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top] \\
&= \frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \text{trace}[\boldsymbol{\Sigma}^{-1} \mathbf{S}_\mu] \qquad \left[\text{where} \quad \mathbf{S}_\mu = \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top\right]
\end{aligned}$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}} \left[\frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right]$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}}\left[\frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right] = -\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n - \boldsymbol{\mu})$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}} \left[\frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right] = -\frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n - \boldsymbol{\mu}) = 0$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}} \left[\frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right] = -\frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n - \boldsymbol{\mu}) = 0$$

which gives the following MLE solution for the multivariate Gaussian's mean

$$\boxed{\boldsymbol{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n}$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}} \left[\frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right] = -\frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n - \boldsymbol{\mu}) = 0$$

which gives the following MLE solution for the multivariate Gaussian's mean

$$\boxed{\boldsymbol{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n}$$

- Taking derivatives w.r.t. $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ (instead of $\boldsymbol{\Sigma}$; leads to simpler derivatives) and setting to zero

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}}\left[\frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right] = -\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n - \boldsymbol{\mu}) = 0$$

which gives the following MLE solution for the multivariate Gaussian's mean

$$\boxed{\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n}$$

- Taking derivatives w.r.t. $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ (instead of $\boldsymbol{\Sigma}$; leads to simpler derivatives) and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\Lambda}}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{\partial}{\partial \boldsymbol{\Lambda}}\left[\frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\text{trace}[\boldsymbol{\Lambda}\mathbf{S}_{\mu}]\right]$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}} \left[\frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right] = -\frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n - \boldsymbol{\mu}) = 0$$

which gives the following MLE solution for the multivariate Gaussian's mean

$$\boxed{\boldsymbol{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n}$$

- Taking derivatives w.r.t. $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ (instead of $\boldsymbol{\Sigma}$; leads to simpler derivatives) and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\Lambda}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{\partial}{\partial \boldsymbol{\Lambda}} \left[\frac{N}{2} \log |\boldsymbol{\Lambda}| - \frac{1}{2} \text{trace}[\boldsymbol{\Lambda} \mathbf{S}_{\mu}] \right] = \frac{N}{2} \boldsymbol{\Lambda}^{-\top} - \frac{1}{2} \mathbf{S}_{\mu}^\top$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}}\left[\frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right] = -\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n - \boldsymbol{\mu}) = 0$$

which gives the following MLE solution for the multivariate Gaussian's mean

$$\boxed{\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n}$$

- Taking derivatives w.r.t. $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ (instead of $\boldsymbol{\Sigma}$; leads to simpler derivatives) and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\Lambda}}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{\partial}{\partial \boldsymbol{\Lambda}}\left[\frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\text{trace}[\boldsymbol{\Lambda}\mathbf{S}_\mu]\right] = \frac{N}{2}\boldsymbol{\Lambda}^{-\top} - \frac{1}{2}\mathbf{S}_\mu^\top = \frac{N}{2}\boldsymbol{\Lambda}^{-1} - \frac{1}{2}\mathbf{S}_\mu$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}}\mathcal{L}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}}\left[\frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})\right] = -\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n-\boldsymbol{\mu}) = 0$$

which gives the following MLE solution for the multivariate Gaussian's mean

$$\boxed{\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n}$$

- Taking derivatives w.r.t. $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ (instead of $\boldsymbol{\Sigma}$; leads to simpler derivatives) and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\Lambda}}\mathcal{L}(\boldsymbol{\mu},\boldsymbol{\Lambda}) = \frac{\partial}{\partial \boldsymbol{\Lambda}}\left[\frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\text{trace}[\boldsymbol{\Lambda}\mathbf{S}_\mu]\right] = \frac{N}{2}\boldsymbol{\Lambda}^{-\top} - \frac{1}{2}\mathbf{S}_\mu^\top = \frac{N}{2}\boldsymbol{\Lambda}^{-1} - \frac{1}{2}\mathbf{S}_\mu = \frac{N}{2}\boldsymbol{\Sigma} - \frac{1}{2}\mathbf{S}_\mu$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}}\mathcal{L}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}}\left[\frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})\right] = -\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n-\boldsymbol{\mu}) = 0$$

which gives the following MLE solution for the multivariate Gaussian's mean

$$\boxed{\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n}$$

- Taking derivatives w.r.t. $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ (instead of $\boldsymbol{\Sigma}$; leads to simpler derivatives) and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\Lambda}}\mathcal{L}(\boldsymbol{\mu},\boldsymbol{\Lambda}) = \frac{\partial}{\partial \boldsymbol{\Lambda}}\left[\frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\text{trace}[\boldsymbol{\Lambda}\mathbf{S}_\mu]\right] = \frac{N}{2}\boldsymbol{\Lambda}^{-\top} - \frac{1}{2}\mathbf{S}_\mu^\top = \frac{N}{2}\boldsymbol{\Lambda}^{-1} - \frac{1}{2}\mathbf{S}_\mu = \frac{N}{2}\boldsymbol{\Sigma} - \frac{1}{2}\mathbf{S}_\mu = 0$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}}\left[\frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right] = -\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n - \boldsymbol{\mu}) = 0$$

which gives the following MLE solution for the multivariate Gaussian's mean

$$\boxed{\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n}$$

- Taking derivatives w.r.t. $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ (instead of $\boldsymbol{\Sigma}$; leads to simpler derivatives) and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\Lambda}}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{\partial}{\partial \boldsymbol{\Lambda}}\left[\frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\text{trace}[\boldsymbol{\Lambda}\mathbf{S}_\mu]\right] = \frac{N}{2}\boldsymbol{\Lambda}^{-\top} - \frac{1}{2}\mathbf{S}_\mu^\top = \frac{N}{2}\boldsymbol{\Lambda}^{-1} - \frac{1}{2}\mathbf{S}_\mu = \frac{N}{2}\boldsymbol{\Sigma} - \frac{1}{2}\mathbf{S}_\mu = 0$$

which gives the following MLE solution for the multivariate Gaussian's covariance matrix

$$\boxed{\boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^\top}$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}}\left[\frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right] = -\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n - \boldsymbol{\mu}) = 0$$

which gives the following MLE solution for the multivariate Gaussian's mean

$$\boxed{\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n}$$

- Taking derivatives w.r.t. $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ (instead of $\boldsymbol{\Sigma}$; leads to simpler derivatives) and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\Lambda}}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{\partial}{\partial \boldsymbol{\Lambda}}\left[\frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\text{trace}[\boldsymbol{\Lambda}\mathbf{S}_\mu]\right] = \frac{N}{2}\boldsymbol{\Lambda}^{-\top} - \frac{1}{2}\mathbf{S}_\mu^\top = \frac{N}{2}\boldsymbol{\Lambda}^{-1} - \frac{1}{2}\mathbf{S}_\mu = \frac{N}{2}\boldsymbol{\Sigma} - \frac{1}{2}\mathbf{S}_\mu = 0$$

which gives the following MLE solution for the multivariate Gaussian's covariance matrix

$$\boxed{\boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^\top}$$

- Note: The parameter estimate equations apply to univariate Gaussians too ($D = 1$)

Bayesian inference for the Gaussian's parameters?

Bayesian inference for the Gaussian's parameters?

Easy to do when priors on the parameters are conjugate

Bayesian inference for the Gaussian's parameters?

Easy to do when priors on the parameters are conjugate

Luckily such conjugate priors exist!

Bayesian Inference for Gaussian: Unknown Mean

- For simplicity, we'll look at the case of univariate Gaussian $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Bayesian Inference for Gaussian: Unknown Mean

- For simplicity, we'll look at the case of univariate Gaussian $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Goal: Estimate the mean μ of the Gaussian. Suppose (for now) that variance σ^2 is known

Bayesian Inference for Gaussian: Unknown Mean

- For simplicity, we'll look at the case of univariate Gaussian $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Goal: Estimate the mean μ of the Gaussian. Suppose (for now) that variance σ^2 is known
- Given N i.i.d. points $\mathcal{D} = \{x_1, \ldots, x_N\}$ from $\mathcal{N}(x|\mu,\sigma^2)$, the likelihood $p(\boldsymbol{x}|\mu)$ is given by

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.$$

Bayesian Inference for Gaussian: Unknown Mean

- For simplicity, we'll look at the case of univariate Gaussian $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Goal: Estimate the mean μ of the Gaussian. Suppose (for now) that variance σ^2 is known
- Given N i.i.d. points $\mathcal{D} = \{x_1, \ldots, x_N\}$ from $\mathcal{N}(x|\mu,\sigma^2)$, the likelihood $p(\boldsymbol{x}|\mu)$ is given by

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.$$

- We will assume a Gaussian prior on μ (note: Gaussian is conjugate to itself)

$$p(\mu) = \mathcal{N}(\mu|\mu_0,\sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

Bayesian Inference for Gaussian: Unknown Mean

- For simplicity, we'll look at the case of univariate Gaussian $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Goal: Estimate the mean μ of the Gaussian. Suppose (for now) that variance σ^2 is known
- Given N i.i.d. points $\mathcal{D} = \{x_1, \dots, x_N\}$ from $\mathcal{N}(x|\mu,\sigma^2)$, the likelihood $p(\mathbf{x}|\mu)$ is given by

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.$$

- We will assume a Gaussian prior on μ (note: Gaussian is conjugate to itself)

$$p(\mu) = \mathcal{N}(\mu|\mu_0,\sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

- The posterior over the mean μ will be

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

Bayesian Inference for Gaussian: Unknown Mean

- For simplicity, we'll look at the case of univariate Gaussian $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Goal: Estimate the mean μ of the Gaussian. Suppose (for now) that variance σ^2 is known
- Given N i.i.d. points $\mathcal{D} = \{x_1, \ldots, x_N\}$ from $\mathcal{N}(x|\mu,\sigma^2)$, the likelihood $p(\mathbf{x}|\mu)$ is given by

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.$$

- We will assume a Gaussian prior on μ (note: Gaussian is conjugate to itself)

$$p(\mu) = \mathcal{N}(\mu|\mu_0,\sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

- The posterior over the mean μ will be

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

- Collecting terms quadratic and linear in μ, we get $p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N,\sigma_N^2)$

Bayesian Inference for Gaussian: Unknown Mean

- For simplicity, we'll look at the case of univariate Gaussian $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Goal: Estimate the mean μ of the Gaussian. Suppose (for now) that variance σ^2 is known
- Given N i.i.d. points $\mathcal{D} = \{x_1, \ldots, x_N\}$ from $\mathcal{N}(x|\mu,\sigma^2)$, the likelihood $p(\mathbf{x}|\mu)$ is given by

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.$$

- We will assume a Gaussian prior on μ (note: Gaussian is conjugate to itself)

$$p(\mu) = \mathcal{N}(\mu|\mu_0,\sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

- The posterior over the mean μ will be

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

- Collecting terms quadratic and linear in μ, we get $p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N,\sigma_N^2)$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

Bayesian Inference for Gaussian: Unknown Variance

- Now suppose the variance σ^2 is unknown but mean μ is known

Bayesian Inference for Gaussian: Unknown Variance

- Now suppose the variance σ^2 is unknown but mean μ is known
- Suppose $\lambda = 1/\sigma^2$ (precision). We can rewrite the likelihood $p(\boldsymbol{x}|\lambda)$ as follows

$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

Bayesian Inference for Gaussian: Unknown Variance

- Now suppose the variance σ^2 is unknown but mean μ is known
- Suppose $\lambda = 1/\sigma^2$ (precision). We can rewrite the likelihood $p(\boldsymbol{x}|\lambda)$ as follows

$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

- This representation of Gaussian (in terms of λ) has the same functional form as a gamma distrib.

Bayesian Inference for Gaussian: Unknown Variance

- Now suppose the variance σ^2 is unknown but mean μ is known
- Suppose $\lambda = 1/\sigma^2$ (precision). We can rewrite the likelihood $p(\boldsymbol{x}|\lambda)$ as follows

$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

- This representation of Gaussian (in terms of λ) has the same functional form as a gamma distrib.
- So let's assume a gamma prior on $\lambda = 1/\sigma^2$ (gamma is conjugate to Gaussian in this case)

$$p(\lambda) = \text{Gamma}(\lambda|a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0-1} \exp(-b_0\lambda)$$

Bayesian Inference for Gaussian: Unknown Variance

- Now suppose the variance σ^2 is unknown but mean μ is known
- Suppose $\lambda = 1/\sigma^2$ (precision). We can rewrite the likelihood $p(\boldsymbol{x}|\lambda)$ as follows

$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

- This representation of Gaussian (in terms of λ) has the same functional form as a gamma distrib.
- So let's assume a gamma prior on $\lambda = 1/\sigma^2$ (gamma is conjugate to Gaussian in this case)

$$p(\lambda) = \text{Gamma}(\lambda|a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0 - 1} \exp(-b_0\lambda)$$

- The posterior over λ will be $\propto p(\lambda)p(\boldsymbol{x}|\lambda)$. Therefore

$$p(\lambda|\mathbf{x}) \propto \lambda^{a_0-1}\lambda^{N/2} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}$$

Bayesian Inference for Gaussian: Unknown Variance

- Now suppose the variance σ^2 is unknown but mean μ is known
- Suppose $\lambda = 1/\sigma^2$ (precision). We can rewrite the likelihood $p(\boldsymbol{x}|\lambda)$ as follows

$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

- This representation of Gaussian (in terms of λ) has the same functional form as a gamma distrib.
- So let's assume a gamma prior on $\lambda = 1/\sigma^2$ (gamma is conjugate to Gaussian in this case)

$$p(\lambda) = \text{Gamma}(\lambda|a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0 - 1} \exp(-b_0 \lambda)$$

- The posterior over λ will be $\propto p(\lambda)p(\boldsymbol{x}|\lambda)$. Therefore

$$p(\lambda|\mathbf{x}) \propto \lambda^{a_0 - 1}\lambda^{N/2} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}$$

- Thus the posterior $p(\lambda|\boldsymbol{x}) = \text{Gamma}(\lambda|a_N, b_N)$ with $a_N = a_0 + \frac{N}{2}$ and $b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^2$

Bayesian Inference for Gaussian: Unknown Mean and Variance

- Need priors on μ and $\lambda = 1/\sigma^2$.

Bayesian Inference for Gaussian: Unknown Mean and Variance

- Need priors on μ and $\lambda = 1/\sigma^2$. Can't use independent Gaussian and gamma priors we used before

Bayesian Inference for Gaussian: Unknown Mean and Variance

- Need priors on μ and $\lambda = 1/\sigma^2$. Can't use independent Gaussian and gamma priors we used before
- Reason: μ and λ are coupled in the likelihood term (also note that we assume both are unknown)

$$p(\mathbf{x}|\mu,\lambda) = \prod_{n=1}^{N} \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\}$$

$$\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}.$$

Bayesian Inference for Gaussian: Unknown Mean and Variance

- Need priors on μ and $\lambda = 1/\sigma^2$. Can't use independent Gaussian and gamma priors we used before
- Reason: μ and λ are coupled in the likelihood term (also note that we assume both are unknown)

$$
\begin{aligned}
p(\mathbf{x}|\mu,\lambda) &= \prod_{n=1}^{N}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\} \\
&\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{N}\exp\left\{\lambda\mu\sum_{n=1}^{N}x_n-\frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}.
\end{aligned}
$$

- Thus we need a joint prior on μ and λ that has the same functional form as above

$$
\begin{aligned}
p(\mu,\lambda) &= \mathcal{N}(\mu|\mu_0,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b) \\
&\propto \exp\left\{-\frac{\beta\lambda}{2}(\mu-\mu_0)^2\right\}\lambda^{a-1}\exp\left\{-b\lambda\right\}
\end{aligned}
$$

Bayesian Inference for Gaussian: Unknown Mean and Variance

- Need priors on μ and $\lambda = 1/\sigma^2$. Can't use independent Gaussian and gamma priors we used before
- Reason: μ and λ are coupled in the likelihood term (also note that we assume both are unknown)

$$p(\mathbf{x}|\mu,\lambda) = \prod_{n=1}^{N}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\}$$

$$\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}.$$

- Thus we need a joint prior on μ and λ that has the same functional form as above

$$p(\mu,\lambda) = \mathcal{N}(\mu|\mu_0,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b)$$

$$\propto \exp\left\{-\frac{\beta\lambda}{2}(\mu-\mu_0)^2\right\}\lambda^{a-1}\exp\{-b\lambda\}$$

- This is known as Gaussian-gamma prior (conjugate to a Gaussian with unknown mean and var.)
- The posterior will also be Gaussian-gamma

Bayesian Inference for Multivariate Gaussian

- Can be done similarly as the univariate case

Bayesian Inference for Multivariate Gaussian

- Can be done similarly as the univariate case
- Prior on the covariance matrix will depend on the type of covariance (spherical, diagonal, or full)

Bayesian Inference for Multivariate Gaussian

- Can be done similarly as the univariate case
- Prior on the covariance matrix will depend on the type of covariance (spherical, diagonal, or full)
- In the most general case (full covariance matrix), we need to use some prior on p.s.d. matrices

Bayesian Inference for Multivariate Gaussian

- Can be done similarly as the univariate case
- Prior on the covariance matrix will depend on the type of covariance (spherical, diagonal, or full)
- In the most general case (full covariance matrix), we need to use some prior on p.s.d. matrices
    - Wishart prior can be used as a conjugate prior on the precision matrix (inverse of covariance matrix), if mean is known. Or can also work with inverse Wishart for cov. matrix.

Bayesian Inference for Multivariate Gaussian

- Can be done similarly as the univariate case
- Prior on the covariance matrix will depend on the type of covariance (spherical, diagonal, or full)
- In the most general case (full covariance matrix), we need to use some prior on p.s.d. matrices
  - Wishart prior can be used as a conjugate prior on the precision matrix (inverse of covariance matrix), if mean is known. Or can also work with inverse Wishart for cov. matrix.
  - Note: Wishart is like a generalized form of univariate gamma distribution

Bayesian Inference for Multivariate Gaussian

- Can be done similarly as the univariate case
- Prior on the covariance matrix will depend on the type of covariance (spherical, diagonal, or full)
- In the most general case (full covariance matrix), we need to use some prior on p.s.d. matrices
  - Wishart prior can be used as a conjugate prior on the precision matrix (inverse of covariance matrix), if mean is known. Or can also work with inverse Wishart for cov. matrix.
  - Note: Wishart is like a generalized form of univariate gamma distribution
  - If both mean and covariance are unknown then Gaussian-Wishart prior can be used

Bayesian Inference for Multivariate Gaussian

- Can be done similarly as the univariate case
- Prior on the covariance matrix will depend on the type of covariance (spherical, diagonal, or full)
- In the most general case (full covariance matrix), we need to use some prior on p.s.d. matrices
  - Wishart prior can be used as a conjugate prior on the precision matrix (inverse of covariance matrix), if mean is known. Or can also work with inverse Wishart for cov. matrix.
  - Note: Wishart is like a generalized form of univariate gamma distribution
  - If both mean and covariance are unknown then Gaussian-Wishart prior can be used
- More details can be found in (MLAPP Chap. 4). Please take a look.

Predictive Density?

- Given a new point $\boldsymbol{x}_*$, what is its probability under the (inferred) Gaussian?

Predictive Density?

- Given a new point $\boldsymbol{x}_*$, what is its probability under the (inferred) Gaussian?

$$p(\boldsymbol{x}_*|\mathcal{D}) = p(\boldsymbol{x}_*|\boldsymbol{\mu}_{ML}, \boldsymbol{\Sigma}_{ML}) \text{ or } p(\boldsymbol{x}_*|\boldsymbol{\mu}_{MAP}, \boldsymbol{\Sigma}_{MAP}) \qquad \text{(when using point estimates)}$$

Predictive Density?

- Given a new point $\boldsymbol{x}_*$, what is its probability under the (inferred) Gaussian?

$$
\begin{aligned}
p(\boldsymbol{x}_*|\mathcal{D}) &= p(\boldsymbol{x}_*|\boldsymbol{\mu}_{ML}, \boldsymbol{\Sigma}_{ML}) \text{ or } p(\boldsymbol{x}_*|\boldsymbol{\mu}_{MAP}, \boldsymbol{\Sigma}_{MAP}) && \text{(when using point estimates)} \\
p(\boldsymbol{x}_*|\mathcal{D}) &= \int p(\boldsymbol{x}_*|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) d\boldsymbol{\mu} d\boldsymbol{\Sigma} && \text{(when using full posterior)}
\end{aligned}
$$

Predictive Density?

- Given a new point $\boldsymbol{x}_*$, what is its probability under the (inferred) Gaussian?

$$
\begin{aligned}
p(\boldsymbol{x}_*|\mathcal{D}) &= p(\boldsymbol{x}_*|\boldsymbol{\mu}_{ML}, \boldsymbol{\Sigma}_{ML}) \text{ or } p(\boldsymbol{x}_*|\boldsymbol{\mu}_{MAP}, \boldsymbol{\Sigma}_{MAP}) && \text{(when using point estimates)} \\
p(\boldsymbol{x}_*|\mathcal{D}) &= \int p(\boldsymbol{x}_*|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) d\boldsymbol{\mu} d\boldsymbol{\Sigma} && \text{(when using full posterior)}
\end{aligned}
$$

- Case 1 will just be a Gaussian with MLE estimates of the parameters

Predictive Density?

- Given a new point $\boldsymbol{x}_*$, what is its probability under the (inferred) Gaussian?

$$
\begin{aligned}
p(\boldsymbol{x}_*|\mathcal{D}) &= p(\boldsymbol{x}_*|\boldsymbol{\mu}_{ML}, \boldsymbol{\Sigma}_{ML}) \text{ or } p(\boldsymbol{x}_*|\boldsymbol{\mu}_{MAP}, \boldsymbol{\Sigma}_{MAP}) && \text{(when using point estimates)} \\
p(\boldsymbol{x}_*|\mathcal{D}) &= \int p(\boldsymbol{x}_*|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) d\boldsymbol{\mu} d\boldsymbol{\Sigma} && \text{(when using full posterior)}
\end{aligned}
$$

- Case 1 will just be a Gaussian with MLE estimates of the parameters
- Case 2 will be a Student-t distribution

Some Useful Properties of Gaussians

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix}$$

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}$$

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}$$

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}$$

- The marginal distribution of one block, say $\boldsymbol{x}_a$, is a Gaussian

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}$$

- The marginal distribution of one block, say $\boldsymbol{x}_a$, is a Gaussian

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

- The conditional distribution of $\boldsymbol{x}_a$ given $\boldsymbol{x}_b$, is Gaussian, i.e., $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}$$

- The marginal distribution of one block, say $\boldsymbol{x}_a$, is a Gaussian

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

- The conditional distribution of $\boldsymbol{x}_a$ given $\boldsymbol{x}_b$, is Gaussian, i.e., $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ where

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}$$

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}$$

- The marginal distribution of one block, say $\boldsymbol{x}_a$, is a Gaussian

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

- The conditional distribution of $\boldsymbol{x}_a$ given $\boldsymbol{x}_b$, is Gaussian, i.e., $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ where

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}$$

- The marginal distribution of one block, say $\boldsymbol{x}_a$, is a Gaussian

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

- The conditional distribution of $\boldsymbol{x}_a$ given $\boldsymbol{x}_b$, is Gaussian, i.e., $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ where

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$ (“smaller” than Σ_{aa}; makes sense intuitively)

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}$$

- The marginal distribution of one block, say $\boldsymbol{x}_a$, is a Gaussian

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

- The conditional distribution of $\boldsymbol{x}_a$ given $\boldsymbol{x}_b$, is Gaussian, i.e., $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ where

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$ (“smaller” than Σ_{aa}; makes sense intuitively)

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\}$$

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}$$

- The marginal distribution of one block, say $\boldsymbol{x}_a$, is a Gaussian

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

- The conditional distribution of $\boldsymbol{x}_a$ given $\boldsymbol{x}_b$, is Gaussian, i.e., $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ where

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

("smaller" than Σ_{aa}; makes sense intuitively)

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}$$

- The marginal distribution of one block, say $\boldsymbol{x}_a$, is a Gaussian

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

- The conditional distribution of $\boldsymbol{x}_a$ given $\boldsymbol{x}_b$, is Gaussian, i.e., $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ where

$$\begin{aligned} \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \qquad (\text{“smaller” than } \Sigma_{aa}\text{; makes sense intuitively}) \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \\ &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}^{-1}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

Marginals and Conditionals from Gaussian Joint Distribution

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}$$

- The marginal distribution of one block, say $\boldsymbol{x}_a$, is a Gaussian

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

- The conditional distribution of $\boldsymbol{x}_a$ given $\boldsymbol{x}_b$, is Gaussian, i.e., $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ where

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \qquad \text{(“smaller” than } \Sigma_{aa}\text{; makes sense intuitively)}$$

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \\ &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}^{-1}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

- Both results are extremely useful when working with Gaussian joint distributions

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{L}^{-1})$$

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \epsilon} \qquad \text{where} \quad p(\epsilon) = \mathcal{N}(\epsilon|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \epsilon} \qquad \text{where} \quad p(\epsilon) = \mathcal{N}(\epsilon|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \epsilon} \qquad \text{where} \quad p(\epsilon) = \mathcal{N}(\epsilon|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest.

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \epsilon} \qquad \text{where} \quad p(\epsilon) = \mathcal{N}(\epsilon|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \epsilon} \qquad \text{where} \quad p(\epsilon) = \mathcal{N}(\epsilon|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})}$$

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma})$$

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \epsilon} \qquad \text{where} \quad p(\epsilon) = \mathcal{N}(\epsilon|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) \quad = \quad \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) \quad = \quad \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})} \quad = \quad \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}) \qquad \text{(a Gaussian posterior :-))}$$

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}) \qquad \text{(a Gaussian posterior :-))}$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}) \qquad \text{(a Gaussian posterior :-))}$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top + \mathbf{L}^{-1})$$

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \epsilon} \qquad \text{where} \quad p(\epsilon) = \mathcal{N}(\epsilon|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}) \qquad \text{(a Gaussian posterior :-))}$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top + \mathbf{L}^{-1}) \qquad \text{(a Gaussian predictive/marginal :-))}$$

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}) \qquad \text{(a Gaussian posterior :-))}$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top + \mathbf{L}^{-1}) \qquad \text{(a Gaussian predictive/marginal :-))}$$

- Exercise: Prove the above two results (MLAPP Chap. 4 and PRML Chap. 2 contain the proof)

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}) \qquad \text{(a Gaussian posterior :-))}$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top + \mathbf{L}^{-1}) \qquad \text{(a Gaussian predictive/marginal :-))}$$

- Exercise: Prove the above two results (MLAPP Chap. 4 and PRML Chap. 2 contain the proof)
  - Write down joint $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ (work with logs).

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}) \qquad \text{(a Gaussian posterior :-))}$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top + \mathbf{L}^{-1}) \qquad \text{(a Gaussian predictive/marginal :-))}$$

- Exercise: Prove the above two results (MLAPP Chap. 4 and PRML Chap. 2 contain the proof)
  - Write down joint $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ (work with logs). Use information-form to note that it will be Gaussian.

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top\mathbf{L}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}) \qquad \text{(a Gaussian posterior :-))}$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top + \mathbf{L}^{-1}) \qquad \text{(a Gaussian predictive/marginal :-))}$$

- Exercise: Prove the above two results (MLAPP Chap. 4 and PRML Chap. 2 contain the proof)
  - Write down joint $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ (work with logs). Use information-form to note that it will be Gaussian. Identify mean/covar of $p(\boldsymbol{x}, \boldsymbol{z})$.

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}) \qquad \text{(a Gaussian posterior :-))}$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top + \mathbf{L}^{-1}) \qquad \text{(a Gaussian predictive/marginal :-))}$$

- Exercise: Prove the above two results (MLAPP Chap. 4 and PRML Chap. 2 contain the proof)
  - Write down joint $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ (work with logs). Use information-form to note that it will be Gaussian. Identify mean/covar of $p(\boldsymbol{x}, \boldsymbol{z})$. Finally use conditional/marginal results from previous slide.

Probabilistic Linear Regression as a Linear Gaussian Model

- Recall the linear regression model: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ and

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Recall the linear regression model: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ and

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$

- Denote by $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$: $N \times 1$ response vector, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix

Probabilistic Linear Regression as a Linear Gaussian Model

- Recall the linear regression model: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ and

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$

- Denote by $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$: $N \times 1$ response vector, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix
- Denote by $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$: $N \times 1$ noise vector.

Probabilistic Linear Regression as a Linear Gaussian Model

- Recall the linear regression model: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ and

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$

- Denote by $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$: $N \times 1$ response vector, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix
- Denote by $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$: $N \times 1$ noise vector.
- Can write the model in the matrix-vector notation as

$$\boxed{\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \beta^{-1}\mathbf{I}_N)$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Recall the linear regression model: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ and
$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$
- Denote by $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$: $N \times 1$ response vector, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix
- Denote by $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$: $N \times 1$ noise vector.
- Can write the model in the matrix-vector notation as
$$\boxed{\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \beta^{-1}\mathbf{I}_N)$$
- This is essentially a Linear Gaussian Model ($\boldsymbol{w}$ transformed into $\boldsymbol{y}$ via $\mathbf{X}$, plus noise) with

Probabilistic Linear Regression as a Linear Gaussian Model

- Recall the linear regression model: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ and
$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$
- Denote by $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$: $N \times 1$ response vector, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix
- Denote by $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$: $N \times 1$ noise vector.
- Can write the model in the matrix-vector notation as
$$\boxed{\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \beta^{-1}\mathbf{I}_N)$$
- This is essentially a Linear Gaussian Model ($\boldsymbol{w}$ transformed into $\boldsymbol{y}$ via $\mathbf{X}$, plus noise) with
$$p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N)$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Recall the linear regression model: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ and
$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$
- Denote by $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$: $N \times 1$ response vector, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix
- Denote by $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$: $N \times 1$ noise vector.
- Can write the model in the matrix-vector notation as
$$\boxed{\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \beta^{-1}\mathbf{I}_N)$$
- This is essentially a Linear Gaussian Model ($\boldsymbol{w}$ transformed into $\boldsymbol{y}$ via $\mathbf{X}$, plus noise) with
$$\begin{aligned} p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) &= \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N) \\ p(\boldsymbol{w}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{aligned}$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Recall the linear regression model: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ and

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$

- Denote by $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$: $N \times 1$ response vector, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix
- Denote by $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$: $N \times 1$ noise vector.
- Can write the model in the matrix-vector notation as

$$\boxed{\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \beta^{-1}\mathbf{I}_N)$$

- This is essentially a Linear Gaussian Model ($\boldsymbol{w}$ transformed into $\boldsymbol{y}$ via $\mathbf{X}$, plus noise) with

$$\begin{aligned} p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) &= \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N) \\ p(\boldsymbol{w}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{aligned}$$

- Note: In our earlier discussion of prob. linear regression, we assumed $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\Sigma}_0 = \lambda^{-1}\mathbf{I}_D$

Probabilistic Linear Regression as a Linear Gaussian Model

- Recall the linear regression model: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ and
$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$
- Denote by $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$: $N \times 1$ response vector, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix
- Denote by $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$: $N \times 1$ noise vector.
- Can write the model in the matrix-vector notation as
$$\boxed{\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \beta^{-1}\mathbf{I}_N)$$
- This is essentially a Linear Gaussian Model ($\boldsymbol{w}$ transformed into $\boldsymbol{y}$ via $\mathbf{X}$, plus noise) with
$$\begin{aligned} p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) &= \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N) \\ p(\boldsymbol{w}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{aligned}$$
- Note: In our earlier discussion of prob. linear regression, we assumed $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\Sigma}_0 = \lambda^{-1}\mathbf{I}_D$
- Since it's an LGM, we can easily compute posterior $p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X})$,

Probabilistic Linear Regression as a Linear Gaussian Model

- Recall the linear regression model: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ and

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$

- Denote by $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$: $N \times 1$ response vector, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix
- Denote by $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$: $N \times 1$ noise vector.
- Can write the model in the matrix-vector notation as

$$\boxed{\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \beta^{-1}\mathbf{I}_N)$$

- This is essentially a Linear Gaussian Model ($\boldsymbol{w}$ transformed into $\boldsymbol{y}$ via $\mathbf{X}$, plus noise) with

$$\begin{aligned} p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) &= \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N) \\ p(\boldsymbol{w}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{aligned}$$

- Note: In our earlier discussion of prob. linear regression, we assumed $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\Sigma}_0 = \lambda^{-1}\mathbf{I}_D$
- Since it's an LGM, we can easily compute posterior $p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X})$, marginal $p(\boldsymbol{y}|\mathbf{X})$,

Probabilistic Linear Regression as a Linear Gaussian Model

- Recall the linear regression model: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ and
$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$
- Denote by $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$: $N \times 1$ response vector, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix
- Denote by $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$: $N \times 1$ noise vector.
- Can write the model in the matrix-vector notation as
$$\boxed{\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \beta^{-1}\mathbf{I}_N)$$
- This is essentially a Linear Gaussian Model ($\boldsymbol{w}$ transformed into $\boldsymbol{y}$ via $\mathbf{X}$, plus noise) with
$$\begin{aligned} p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) &= \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N) \\ p(\boldsymbol{w}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{aligned}$$
- Note: In our earlier discussion of prob. linear regression, we assumed $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\Sigma}_0 = \lambda^{-1}\mathbf{I}_D$
- Since it's an LGM, we can easily compute posterior $p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X})$, marginal $p(\boldsymbol{y}|\mathbf{X})$, posterior predictive $p(y_*|\boldsymbol{x}_*)$, etc,

Probabilistic Linear Regression as a Linear Gaussian Model

- Recall the linear regression model: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ and

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$

- Denote by $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$: $N \times 1$ response vector, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix
- Denote by $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$: $N \times 1$ noise vector.
- Can write the model in the matrix-vector notation as

$$\boxed{\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \beta^{-1}\mathbf{I}_N)$$

- This is essentially a Linear Gaussian Model ($\boldsymbol{w}$ transformed into $\boldsymbol{y}$ via $\mathbf{X}$, plus noise) with

$$\begin{aligned} p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) &= \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N) \\ p(\boldsymbol{w}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{aligned}$$

- Note: In our earlier discussion of prob. linear regression, we assumed $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\Sigma}_0 = \lambda^{-1}\mathbf{I}_D$
- Since it's an LGM, we can easily compute posterior $p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X})$, marginal $p(\boldsymbol{y}|\mathbf{X})$, posterior predictive $p(y_*|\boldsymbol{x}_*)$, etc, using the LGM results from the previous slide

Probabilistic Linear Regression as a Linear Gaussian Model

- Using the LGM results, the marginal $p(\boldsymbol{y}|\mathbf{X})$ will be (exercise: plug-in and verify)

$$p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Using the LGM results, the marginal $p(\boldsymbol{y}|\mathbf{X})$ will be (exercise: plug-in and verify)

$$p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{\mu_0}, \beta^{-1}\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top)$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Using the LGM results, the marginal $p(\boldsymbol{y}|\mathbf{X})$ will be (exercise: plug-in and verify)

$$p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{\mu}_{\mathbf{0}}, \beta^{-1}\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top)$$

- Using the LGM results, the posterior of $\boldsymbol{w}$ will be (exercise: plug-in and verify)

$$p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Using the LGM results, the marginal $p(\boldsymbol{y}|\mathbf{X})$ will be (exercise: plug-in and verify)

$$p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{\mu_0}, \beta^{-1}\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top)$$

- Using the LGM results, the posterior of $\boldsymbol{w}$ will be (exercise: plug-in and verify)

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \\ \boldsymbol{\Sigma}_N &= (\boldsymbol{\Sigma}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X})^{-1} \end{aligned}$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Using the LGM results, the marginal $p(\boldsymbol{y}|\mathbf{X})$ will be (exercise: plug-in and verify)

$$p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{\mu}_{\mathbf{0}}, \beta^{-1}\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top)$$

- Using the LGM results, the posterior of $\boldsymbol{w}$ will be (exercise: plug-in and verify)

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \\ \boldsymbol{\Sigma}_N &= (\boldsymbol{\Sigma}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X})^{-1} = (\boldsymbol{\Sigma}_0^{-1} + \beta\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top)^{-1} \end{aligned}$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Using the LGM results, the marginal $p(\boldsymbol{y}|\mathbf{X})$ will be (exercise: plug-in and verify)

$$p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{\mu_0}, \beta^{-1}\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top)$$

- Using the LGM results, the posterior of $\boldsymbol{w}$ will be (exercise: plug-in and verify)

$$\begin{aligned}
p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \\
\boldsymbol{\Sigma}_N &= (\boldsymbol{\Sigma}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X})^{-1} = (\boldsymbol{\Sigma}_0^{-1} + \beta\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top)^{-1} \\
\boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\mathbf{X}^\top\boldsymbol{y})
\end{aligned}$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Using the LGM results, the marginal $p(\boldsymbol{y}|\mathbf{X})$ will be (exercise: plug-in and verify)

$$p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{\mu_0}, \beta^{-1}\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top)$$

- Using the LGM results, the posterior of $\boldsymbol{w}$ will be (exercise: plug-in and verify)

$$\begin{aligned}
p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \\
\boldsymbol{\Sigma}_N &= (\boldsymbol{\Sigma}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X})^{-1} = (\boldsymbol{\Sigma}_0^{-1} + \beta\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top)^{-1} \\
\boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\mathbf{X}^\top\boldsymbol{y}) = \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\sum_{n=1}^{N}y_n\boldsymbol{x}_n)
\end{aligned}$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Using the LGM results, the marginal $p(\boldsymbol{y}|\mathbf{X})$ will be (exercise: plug-in and verify)

$$p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{\mu}_{\mathbf{0}}, \beta^{-1}\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top)$$

- Using the LGM results, the posterior of $\boldsymbol{w}$ will be (exercise: plug-in and verify)

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \\ \boldsymbol{\Sigma}_N &= (\boldsymbol{\Sigma}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X})^{-1} = (\boldsymbol{\Sigma}_0^{-1} + \beta\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top)^{-1} \\ \boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\mathbf{X}^\top\boldsymbol{y}) = \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\sum_{n=1}^{N}y_n\boldsymbol{x}_n) \end{aligned}$$

- The "brute-force" method to get the above posterior is to use the "completing the squares" trick

Probabilistic Linear Regression as a Linear Gaussian Model

- Using the LGM results, the marginal $p(\boldsymbol{y}|\mathbf{X})$ will be (exercise: plug-in and verify)

$$p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{\mu}_{\mathbf{0}}, \beta^{-1}\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top)$$

- Using the LGM results, the posterior of $\boldsymbol{w}$ will be (exercise: plug-in and verify)

$$\begin{aligned}
p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \\
\boldsymbol{\Sigma}_N &= (\boldsymbol{\Sigma}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X})^{-1} = (\boldsymbol{\Sigma}_0^{-1} + \beta\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top)^{-1} \\
\boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\mathbf{X}^\top\boldsymbol{y}) = \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\sum_{n=1}^{N}y_n\boldsymbol{x}_n)
\end{aligned}$$

- The "brute-force" method to get the above posterior is to use the "completing the squares" trick
- Using the LGM results, the posterior predictive dist. $p(y_*|\boldsymbol{x}_*)$ will be (exercise: plug-in and verify)

$$p(y_*|\boldsymbol{x}_*) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w})p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})d\boldsymbol{w}$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Using the LGM results, the marginal $p(\boldsymbol{y}|\mathbf{X})$ will be (exercise: plug-in and verify)

$$p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X},\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{\mu_0}, \beta^{-1}\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top)$$

- Using the LGM results, the posterior of $\boldsymbol{w}$ will be (exercise: plug-in and verify)

$$\begin{aligned}
p(\boldsymbol{w}|\boldsymbol{y},\mathbf{X}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \\
\boldsymbol{\Sigma}_N &= (\boldsymbol{\Sigma}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X})^{-1} = (\boldsymbol{\Sigma}_0^{-1} + \beta\sum_{n=1}^N \boldsymbol{x}_n\boldsymbol{x}_n^\top)^{-1} \\
\boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\mathbf{X}^\top\boldsymbol{y}) = \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\sum_{n=1}^N y_n\boldsymbol{x}_n)
\end{aligned}$$

- The "brute-force" method to get the above posterior is to use the "completing the squares" trick
- Using the LGM results, the posterior predictive dist. $p(y_*|\boldsymbol{x}_*)$ will be (exercise: plug-in and verify)

$$p(y_*|\boldsymbol{x}_*) = \int p(y_*|\boldsymbol{x}_*,\boldsymbol{w})p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y})d\boldsymbol{w} = \mathcal{N}(y_*|\boldsymbol{\mu}_N^\top\boldsymbol{x}_*, \beta^{-1} + \boldsymbol{x}_*^\top\boldsymbol{\Sigma}_N\boldsymbol{x}_*)$$

Probabilistic Linear Regression as a Linear Gaussian Model

- Using the LGM results, the marginal $p(\boldsymbol{y}|\mathbf{X})$ will be (exercise: plug-in and verify)

$$p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{\mu}_{\mathbf{0}}, \beta^{-1}\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top)$$

- Using the LGM results, the posterior of $\boldsymbol{w}$ will be (exercise: plug-in and verify)

$$\begin{aligned}
p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \\
\boldsymbol{\Sigma}_N &= (\boldsymbol{\Sigma}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X})^{-1} = (\boldsymbol{\Sigma}_0^{-1} + \beta\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top)^{-1} \\
\boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\mathbf{X}^\top\boldsymbol{y}) = \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\sum_{n=1}^{N}y_n\boldsymbol{x}_n)
\end{aligned}$$

- The "brute-force" method to get the above posterior is to use the "completing the squares" trick
- Using the LGM results, the posterior predictive dist. $p(y_*|\boldsymbol{x}_*)$ will be (exercise: plug-in and verify)

$$p(y_*|\boldsymbol{x}_*) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w})p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})d\boldsymbol{w} = \mathcal{N}(y_*|\boldsymbol{\mu}_N^\top\boldsymbol{x}_*, \beta^{-1} + \boldsymbol{x}_*^\top\boldsymbol{\Sigma}_N\boldsymbol{x}_*)$$

- Note: In our earlier discussion of lin. reg., we assumed $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\Sigma}_0 = \lambda^{-1}\mathbf{I}_D$

An Aside: Transformation of General Random Variables

Suppose $\boldsymbol{x} = f(\boldsymbol{z}) = \mathbf{A}\boldsymbol{z} + \mathbf{b}$ be a linear function of an r.v. $\boldsymbol{z}$ (not necessarily Gaussian)

Suppose $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{\mu}$ and $\text{cov}[\boldsymbol{z}] = \boldsymbol{\Sigma}$

An Aside: Transformation of General Random Variables

Suppose $\boldsymbol{x} = f(\boldsymbol{z}) = \mathbf{A}\boldsymbol{z} + \mathbf{b}$ be a linear function of an r.v. $\boldsymbol{z}$ (not necessarily Gaussian)

Suppose $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{\mu}$ and $\text{cov}[\boldsymbol{z}] = \boldsymbol{\Sigma}$

- Expectation of $\boldsymbol{x}$

$$\mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\mathbf{A}\boldsymbol{z} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

An Aside: Transformation of General Random Variables

Suppose $\boldsymbol{x} = f(\boldsymbol{z}) = \mathbf{A}\boldsymbol{z} + \mathbf{b}$ be a linear function of an r.v. $\boldsymbol{z}$ (not necessarily Gaussian)

Suppose $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{\mu}$ and $\text{cov}[\boldsymbol{z}] = \boldsymbol{\Sigma}$

- Expectation of $\boldsymbol{x}$

$$\mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\mathbf{A}\boldsymbol{z} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

- Covariance of $\boldsymbol{x}$

$$\text{cov}[\boldsymbol{x}] = \text{cov}[\mathbf{A}\boldsymbol{z} + \mathbf{b}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

An Aside: Transformation of General Random Variables

Suppose $\boldsymbol{x} = f(\boldsymbol{z}) = \mathbf{A}\boldsymbol{z} + \mathbf{b}$ be a linear function of an r.v. $\boldsymbol{z}$ (not necessarily Gaussian)

Suppose $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{\mu}$ and $\text{cov}[\boldsymbol{z}] = \boldsymbol{\Sigma}$

- Expectation of $\boldsymbol{x}$

$$\mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\mathbf{A}\boldsymbol{z} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

- Covariance of $\boldsymbol{x}$

$$\text{cov}[\boldsymbol{x}] = \text{cov}[\mathbf{A}\boldsymbol{z} + \mathbf{b}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

Likewise if $x = f(\boldsymbol{z}) = \boldsymbol{a}^T\boldsymbol{z} + b$ is a scalar-valued linear function of an r.v. $\boldsymbol{z}$:

An Aside: Transformation of General Random Variables

Suppose $\boldsymbol{x} = f(\boldsymbol{z}) = \mathbf{A}\boldsymbol{z} + \mathbf{b}$ be a linear function of an r.v. $\boldsymbol{z}$ (not necessarily Gaussian)

Suppose $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{\mu}$ and $\text{cov}[\boldsymbol{z}] = \boldsymbol{\Sigma}$

- Expectation of $\boldsymbol{x}$

$$\mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\mathbf{A}\boldsymbol{z} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

- Covariance of $\boldsymbol{x}$

$$\text{cov}[\boldsymbol{x}] = \text{cov}[\mathbf{A}\boldsymbol{z} + \mathbf{b}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

Likewise if $x = f(\boldsymbol{z}) = \boldsymbol{a}^T\boldsymbol{z} + b$ is a scalar-valued linear function of an r.v. $\boldsymbol{z}$:

- $\mathbb{E}[x] = \mathbb{E}[\boldsymbol{a}^T\boldsymbol{z} + b] = \boldsymbol{a}^T\boldsymbol{\mu} + b$

An Aside: Transformation of General Random Variables

Suppose $\boldsymbol{x} = f(\boldsymbol{z}) = \mathbf{A}\boldsymbol{z} + \mathbf{b}$ be a linear function of an r.v. $\boldsymbol{z}$ (not necessarily Gaussian)

Suppose $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{\mu}$ and $\text{cov}[\boldsymbol{z}] = \boldsymbol{\Sigma}$

- Expectation of $\boldsymbol{x}$

$$\mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\mathbf{A}\boldsymbol{z} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

- Covariance of $\boldsymbol{x}$

$$\text{cov}[\boldsymbol{x}] = \text{cov}[\mathbf{A}\boldsymbol{z} + \mathbf{b}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

Likewise if $x = f(\boldsymbol{z}) = \boldsymbol{a}^T\boldsymbol{z} + b$ is a scalar-valued linear function of an r.v. $\boldsymbol{z}$:

- $\mathbb{E}[x] = \mathbb{E}[\boldsymbol{a}^T\boldsymbol{z} + b] = \boldsymbol{a}^T\boldsymbol{\mu} + b$
- $\text{var}[x] = \text{var}[\boldsymbol{a}^T\boldsymbol{z} + b] = \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}$

Another very useful property worth remembering

Some Thoughts

- Gaussian distributions are easy to manipulate algebraically

Some Thoughts

- Gaussian distributions are easy to manipulate algebraically
  - Eases inference when (at least some of) the distributions involved are Gaussians

Some Thoughts

- Gaussian distributions are easy to manipulate algebraically
  - Eases inference when (at least some of) the distributions involved are Gaussians
- Helpful to know tricks such as conditioning and marginalizing from a joint distribution

Some Thoughts

- Gaussian distributions are easy to manipulate algebraically
  - Eases inference when (at least some of) the distributions involved are Gaussians
- Helpful to know tricks such as conditioning and marginalizing from a joint distribution
- Linear Gaussian Models (LGM) widely used in many problems

Some Thoughts

- Gaussian distributions are easy to manipulate algebraically
  - Eases inference when (at least some of) the distributions involved are Gaussians
- Helpful to know tricks such as conditioning and marginalizing from a joint distribution
- Linear Gaussian Models (LGM) widely used in many problems
  - Probabilistic linear regression is perhaps the most straightforward example of an LGM

Some Thoughts

- Gaussian distributions are easy to manipulate algebraically
  - Eases inference when (at least some of) the distributions involved are Gaussians
- Helpful to know tricks such as conditioning and marginalizing from a joint distribution
- Linear Gaussian Models (LGM) widely used in many problems
  - Probabilistic linear regression is perhaps the most straightforward example of an LGM
- We'll come across LGM when looking at many other models, especially latent variable models

Some Thoughts

- Gaussian distributions are easy to manipulate algebraically
  - Eases inference when (at least some of) the distributions involved are Gaussians
- Helpful to know tricks such as conditioning and marginalizing from a joint distribution
- Linear Gaussian Models (LGM) widely used in many problems
  - Probabilistic linear regression is perhaps the most straightforward example of an LGM
- We'll come across LGM when looking at many other models, especially latent variable models, e.g.,
  - Probabilistic PCA and Factor Analysis (dimensionality reduction)

Some Thoughts

- Gaussian distributions are easy to manipulate algebraically
  - Eases inference when (at least some of) the distributions involved are Gaussians
- Helpful to know tricks such as conditioning and marginalizing from a joint distribution
- Linear Gaussian Models (LGM) widely used in many problems
  - Probabilistic linear regression is perhaps the most straightforward example of an LGM
- We'll come across LGM when looking at many other models, especially latent variable models, e.g.,
  - Probabilistic PCA and Factor Analysis (dimensionality reduction)
  - State-Space Models and Kalman Filtering

Some Thoughts

- Gaussian distributions are easy to manipulate algebraically
  - Eases inference when (at least some of) the distributions involved are Gaussians
- Helpful to know tricks such as conditioning and marginalizing from a joint distribution
- Linear Gaussian Models (LGM) widely used in many problems
  - Probabilistic linear regression is perhaps the most straightforward example of an LGM
- We'll come across LGM when looking at many other models, especially latent variable models, e.g.,
  - Probabilistic PCA and Factor Analysis (dimensionality reduction)
  - State-Space Models and Kalman Filtering
- Inference in such models can be performed easily using the properties we saw today