Bayesian Inference for Gaussians (Contd)

CS698X: Topics in Probabilistic Modeling and Inference

Piyush Rai

Multivariate Gaussian

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

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

Trace notation

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

- An alternate representation: The "information form"

Quadratic in $\boldsymbol{x}$

Linear in $\boldsymbol{x}$

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

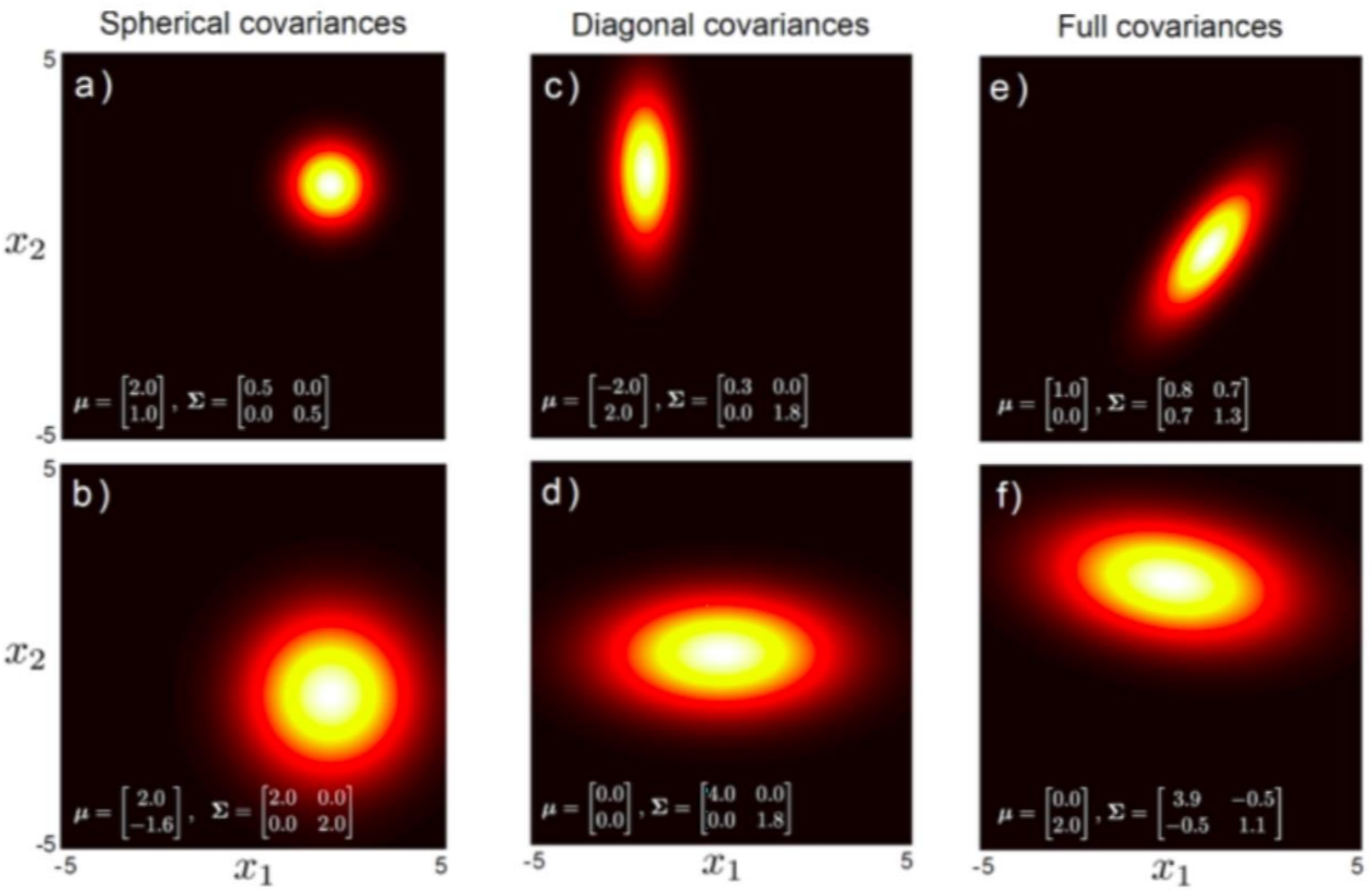
More when we discuss exp. family

where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\xi} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ are known as natural parameters of Gaussian

- Information form can help recognize $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of a multivariate Gaussian when doing algebraic manipulations (e.g., when computing a posterior)

Multivariate Gaussian

- The covariance matrix can be spherical, diagonal, or full



Picture courtesy: Computer vision: models, learning and inference (Simon Price)

Marginals and Conditionals from Gaussian Joint

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

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

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

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

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

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

Extremely useful results when working with Gaussian joint distributions

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

Note that $\Sigma_{a|b}$ is "smaller" than Σ_{aa} (conditioning reduces variance)

Linear Transformations of Random Variables

- Let $\boldsymbol{x} = f(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}$ be a linear function of a vector r.v. $\boldsymbol{z}$

Need not be a Gaussian random var

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

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

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

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

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

$$\text{var}[x] = \text{var}[\boldsymbol{a}^\top \boldsymbol{z} + b] = \boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}$$

$p(\boldsymbol{x})$ will also be Gaussian with mean and covariance/variance given by these expressions

Especially when $p(\boldsymbol{z})$ is Gaussian

- These properties are often helpful in obtaining the marginal distribution $p(\boldsymbol{x})$ from $p(\boldsymbol{z})$

Linear Gaussian Model

Independently added and drawn from $\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{L}^{-1})$

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

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}$$

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

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

- A Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of interest. Assuming $\boldsymbol{\Sigma} = \left(\boldsymbol{\Lambda} + \boldsymbol{A}^\top \boldsymbol{L}\boldsymbol{A}\right)^{-1}$

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

If $p(\boldsymbol{z})$ is a prior and $p(\boldsymbol{x}|\boldsymbol{z})$ is likelihood then this is the posterior

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

If $p(\boldsymbol{z})$ is a prior and $p(\boldsymbol{x}|\boldsymbol{z})$ is likelihood then this is the marginal likelihood

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

Applications of Gaussian-based Models

- Gaussians and Linear Gaussian Models widely used in probabilistic models, e.g.,
  - Probability density estimation: Given $x_1, x_2, \dots, x_N$, estimate $p(x)$ assuming Gaussian lik./noise
  - Given N sensor obs. $\{x_n\}_{n=1}^N$ with $x_n = \mu + \epsilon_n$ (zero-mean Gaussian noise ϵ_n) estimate the underlying true value μ (possibly along with the variance of the estimate of μ)
  - Estimating missing data: $p(x_n^{\text{miss}} | x_n^{\text{obs}})$ or $\mathbb{E}[x_n^{\text{miss}} | x_n^{\text{obs}}]$
  - Linear Regression with Gaussian Likelihood

$$\underbrace{\boldsymbol{y}}_{\text{Training responses}} = \underbrace{\boldsymbol{X}}_{\text{Training feat. mat}}\underbrace{\boldsymbol{w}}_{\text{The prior } p(\boldsymbol{w}) \text{ is Gaussian}} + \underbrace{\boldsymbol{\epsilon}}_{\text{i.i.d. Gaussian noise}}$$

  - Linear latent variable models (probabilistic PCA, factor analysis, Kalman filters) and their mixtures
  - Gaussian Processes (GP) extensively use Gaussian conditioning and marginalization rules

$$\boldsymbol{y} = \boldsymbol{f} + \text{noise} \qquad \text{(GP assumes } \boldsymbol{f} = [f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_N)] \text{ is jointly Gaussian)}$$

  - More complex models where parts of the model use Gaussian likelihoods/priors

Coming Up Next

- Exponential Family distributions
- Conditional Models for supervised learning