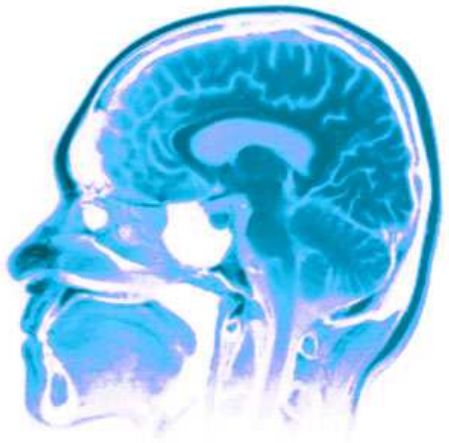
CPSC540



Multivariate Gaussian Models

Nando de Freitas

2011

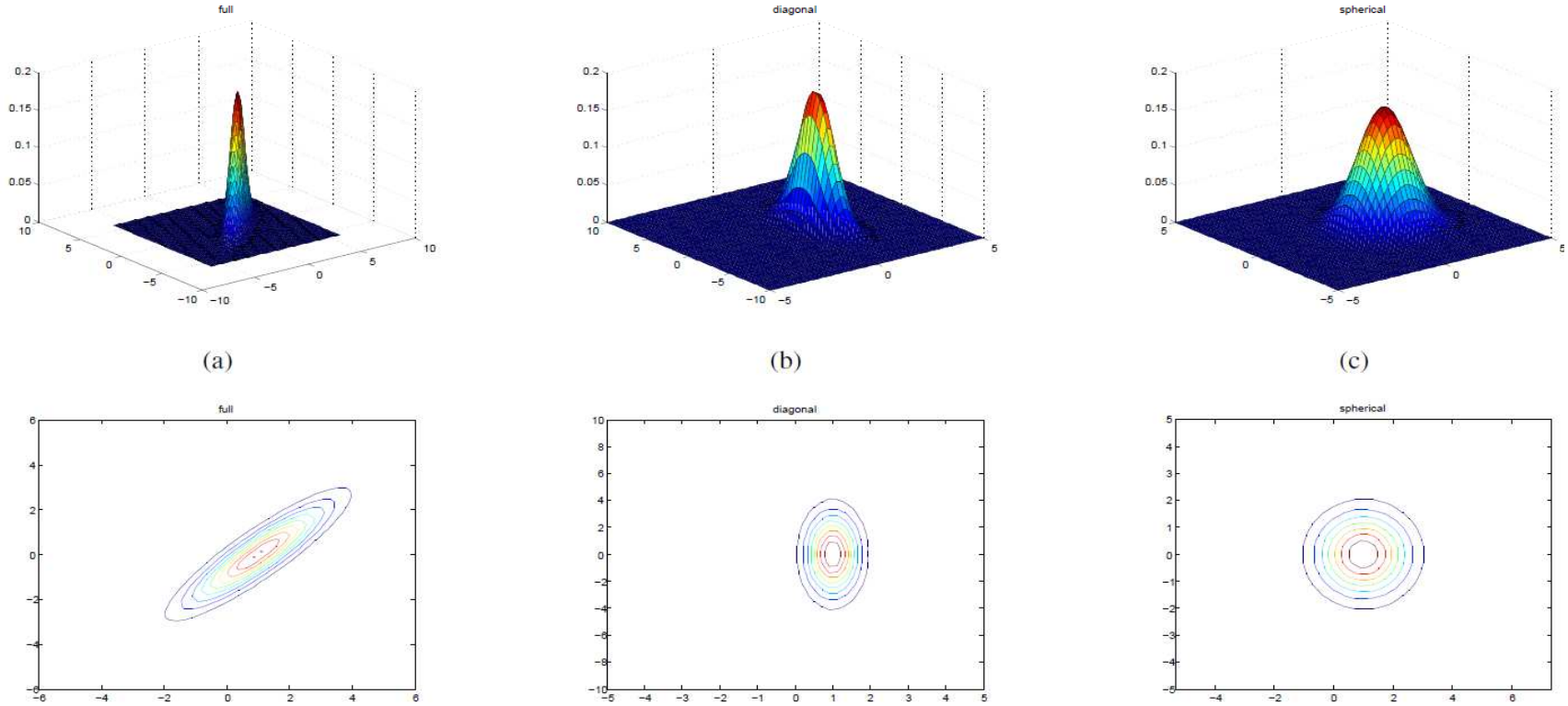
KPM Book Sections: *5 and 31.2*

- The pdf of the MVN is defined as

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad := \quad \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \times \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

  where D is the dimensionality of $\mathbf{x}$, $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ is the mean, and $\boldsymbol{\Sigma} = \text{cov}[\mathbf{X}]$ is the **covariance matrix**.

- The normalization constant $1/(|2\pi\boldsymbol{\Sigma}|^{1/2})$ ensures that the pdf integrates to 1.



Bivariate Gaussian

- If $D = 2$, the MVN becomes the **bivariate Gaussian**. In this case, the covariance matrix can be written in the form

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

  where ρ is the **correlation coefficient**.

- The pdf of the bivariate Gaussian is as follows:

$$\begin{aligned} p(x_1, x_2) \quad = \quad & \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right.\right. \\ & \left.\left. -2\rho\frac{(x_1-\mu_1)}{\sigma_1}\frac{(x_2-\mu_2)}{\sigma_2}\right)\right) \end{aligned}$$

Cholesky decomposition

- Since $\mathbf{\Sigma}$ is symmetric positive definite, we can compute its **Cholesky decomposition**, $\mathbf{R}^T\mathbf{R} = \mathbf{\Sigma}$, where $\mathbf{R}$ is upper triangular.
- Suppose we know how to sample from a standard normal. Then we can easily sample from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. To sample from $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, first sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Now let $\mathbf{x} = \boldsymbol{\mu} + \mathbf{R}^T\mathbf{z}$. It is easy to see that $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$. This is a widely used trick.
- To evaluate the pdf of a Gaussian, we need to compute $\mathbf{\Sigma}^{-1}$ and $|\mathbf{\Sigma}|$ efficiently. If we have already computed the Cholesky decomposition, we can write $\mathbf{\Sigma}^{-1} = \mathbf{R}^{-1}\mathbf{R}^{-T}$; this is efficient to evaluate since $\mathbf{R}$ is upper triangular. For the determinant, we can use the following result:

$$|\mathbf{\Sigma}| = \prod_{j=1}^{D} R_{jj}^2$$

Information parameterization

- Suppose $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. One can show that $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ is the mean vector, and $\text{cov}[\mathbf{x}] = \boldsymbol{\Sigma}$ is the covariance matrix. These are called the **moment parameters** of the distribution. However, it is sometimes useful to use the **canonical parameters** or **natural parameters**, defined as

$$\boldsymbol{\Lambda} := \boldsymbol{\Sigma}^{-1}, \quad \boldsymbol{\xi} := \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$

- We can convert back to the moment parameters using

$$\boldsymbol{\mu} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}, \quad \boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}$$

- Using the canonical parameters, we can write the MVN in **information form** (i.e., in **exponential family** form:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\xi}, \boldsymbol{\Lambda}) = (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{x}^T\boldsymbol{\Lambda}\mathbf{x} + \boldsymbol{\xi}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi} - 2\mathbf{x}^T\boldsymbol{\xi})]$$

Inference

- **Probabilistic inference** refers to deriving unknown quantities from known quantities under uncertainty.
- Suppose we have a vector of correlated random variables with joint distribution $p(\mathbf{x}_{1:D}|\boldsymbol{\theta})$.
- Let us partition this vector into the **visible variables** $\mathbf{x}_v$, which are observed, and the **hidden variables**, $\mathbf{x}_h$, which are unobserved.
- Inference refers to computing the posterior distribution of the unknowns given the knowns:

$$p(\mathbf{x}_h|\mathbf{x}_v, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_h, \mathbf{x}_v|\boldsymbol{\theta})}{p(\mathbf{x}_v|\boldsymbol{\theta})} = \frac{p(\mathbf{x}_h, \mathbf{x}_v|\boldsymbol{\theta})}{\int p(\mathbf{x}'_h, \mathbf{x}_v|\boldsymbol{\theta}) d\mathbf{x}'_h}$$

Inference

- It is clear that the key operations we need to be able to implement are
  1. **conditioning** on data, i.e., going from $p(\mathbf{x}_h, \mathbf{x}_v)$ to $p(\mathbf{x}_h|\mathbf{x}_v)$, which essentially "clamps" the visible variables to their observed values, $\mathbf{x}_v$.
  2. **marginalizing out**, i.e., going from $p(\mathbf{x}_h, \mathbf{x}_v)$ to $p(\mathbf{x}_v)$.
- Sometimes only some of the hidden variables are of interest to us. So let us partition the hidden variables into **query variables**, $\mathbf{x}_q$, whose value we wish to know, and the remaining **nuisance variables**, $\mathbf{x}_r$, which we are not interested in.
- We can compute what we are interested in by marginalizing out the nuisance variables:

$$p(\mathbf{x}_q|\mathbf{x}_v, \boldsymbol{\theta}) = \int p(\mathbf{x}_q, \mathbf{x}_r|\mathbf{x}_v, \boldsymbol{\theta}) d\mathbf{x}_r$$

Gaussian inference: Main result

- Suppose $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ is jointly Gaussian with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

  We will show that the marginal distribution for $p(\mathbf{x}_2)$ is obtained by extracting the rows and columns from $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ corresponding to $\mathbf{x}_2$:

$$p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

  We will also show that the conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$ is given by

$$\begin{aligned} p(\mathbf{x}_1|\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{aligned}$$

Schur complement

Theorem 0.1. *Consider a general partitioned matrix*

$$\mathbf{M} = \begin{pmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{G} & \mathbf{H} \end{pmatrix}$$

where we assume $\mathbf{E}$ *and* $\mathbf{H}$ *are invertible. We have*

$$\begin{aligned} \mathbf{M}^{-1} &= \begin{pmatrix} (\mathbf{M}/\mathbf{H})^{-1} & -(\mathbf{M}/\mathbf{H})^{-1}\mathbf{F}\mathbf{H}^{-1} \\ -\mathbf{H}^{-1}\mathbf{G}(\mathbf{M}/\mathbf{H})^{-1} & \mathbf{H}^{-1} + \mathbf{H}^{-1}\mathbf{G}(\mathbf{M}/\mathbf{H})^{-1}\mathbf{F}\mathbf{H}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{F}(\mathbf{M}/\mathbf{E})^{-1}\mathbf{G}\mathbf{E}^{-1} & -\mathbf{E}^{-1}\mathbf{F}(\mathbf{M}/\mathbf{E})^{-1} \\ -(\mathbf{M}/\mathbf{E})^{-1}\mathbf{G}\mathbf{E}^{-1} & (\mathbf{M}/\mathbf{E})^{-1} \end{pmatrix} \end{aligned}$$

where

$$\begin{aligned} \mathbf{M}/\mathbf{H} &:= \mathbf{E} - \mathbf{F}\mathbf{H}^{-1}\mathbf{G} \\ \mathbf{M}/\mathbf{E} &:= \mathbf{H} - \mathbf{G}\mathbf{E}^{-1}\mathbf{F} \end{aligned}$$

We say that $\mathbf{M}/\mathbf{H}$ *is the* **Schur complement** *of* $\mathbf{M}$ *wrt* $\mathbf{H}$.

Schur complement

- The proof is left as an exercise. An important intermmediate result is:

$$\begin{pmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{G} & \mathbf{H} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{H}^{-1}\mathbf{G} & \mathbf{I} \end{pmatrix} \begin{pmatrix} (\mathbf{M}/\mathbf{H})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\mathbf{F}\mathbf{H}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$$

Sherman-Morrison-Woodbury

- Equating matrix terms, we get:

$$(\mathbf{M}/\mathbf{H})^{-1} = \mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{F}(\mathbf{M}/\mathbf{E})^{-1}\mathbf{G}\mathbf{E}^{-1}$$

  Plugging in the definition of Schur complement leads to the widely used **matrix inversion lemma** or the **Sherman-Morrison-Woodbury formula**:

$$(\mathbf{E} - \mathbf{F}\mathbf{H}^{-1}\mathbf{G})^{-1} = \mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{F}(\mathbf{H} - \mathbf{G}\mathbf{E}^{-1}\mathbf{F})^{-1}\mathbf{G}\mathbf{E}^{-1}$$

- In the special case that $H = -1$ (a scalar), $\mathbf{F} = \mathbf{u}$ (a column vector), and $\mathbf{G} = \mathbf{v}^T$ (a row vector),k on we get the following formula for the **rank one update** of an inverse matrix

$$\begin{aligned}(\mathbf{E} + \mathbf{u}\mathbf{v}^T)^{-1} &= \mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{u}(-1 - \mathbf{v}^T\mathbf{E}^{-1}\mathbf{u})^{-1}\mathbf{v}^T\mathbf{E}^{-1} \\ &= \mathbf{E}^{-1} - \frac{\mathbf{E}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{E}^{-1}}{1 + \mathbf{v}^T\mathbf{E}^{-1}\mathbf{u}}\end{aligned}$$

Matrix determinant lemma

- Equating matrix terms, we also get:

$$(\mathbf{E} - \mathbf{F}\mathbf{H}^{-1}\mathbf{G})^{-1}\mathbf{F}\mathbf{H}^{-1} \quad = \quad \mathbf{E}^{-1}\mathbf{F}(\mathbf{H} - \mathbf{G}\mathbf{E}^{-1}\mathbf{F})^{-1}$$

- From the proof of the Schur complement, one can derive the following **matrix determinant lemma**

$$|\mathbf{E} - \mathbf{F}\mathbf{H}^{-1}\mathbf{G}| = |\mathbf{H} - \mathbf{G}\mathbf{E}^{-1}\mathbf{F}||\mathbf{H}^{-1}||\mathbf{E}|$$

Gaussian inference: Proof

- We factor the joint $p(\mathbf{x}_1, \mathbf{x}_2)$ as $p(\mathbf{x}_2)p(\mathbf{x}_1|\mathbf{x}_2)$ using the Schur complement.
  First, we deal with the exponent:

$$\exp\left\{-\frac{1}{2}\begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}^T\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}\right\}$$

$$= \exp\left\{-\frac{1}{2}\begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}^T\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{I}\end{pmatrix}\begin{pmatrix}(\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22})^{-1} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}\right\}$$

$$= \exp\left\{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2))^T(\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22})^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2))\right\}$$

$$\times \exp\left\{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right\}$$

  This is of the form

$$\exp(\text{quadratic form in } \mathbf{x}_1, \mathbf{x}_2) \times \exp(\text{quadratic form in } \mathbf{x}_2)$$

Gaussian inference: Proof

- Using the matrix determinant lemma, we can also split up the normalization constants

$$\begin{aligned}(2\pi)^{(d_1+d_2)/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}} &= (2\pi)^{(d_1+d_2)/2}(|\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22}||\boldsymbol{\Sigma}_{22}|)^{\frac{1}{2}} \\ &= (2\pi)^{d_1/2}|\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}}\ (2\pi)^{d_2/2}|\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}}\end{aligned}$$

  where $d_1 = |\mathbf{x}_1|$ and $d_2 = |\mathbf{x}_2|$.

- Hence we have successfully factorized the joint as

$$\begin{aligned}p(\mathbf{x}_1, \mathbf{x}_2) &= p(\mathbf{x}_2)p(\mathbf{x}_1|\mathbf{x}_2) \\ &= \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})\mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})\end{aligned}$$

  where the parameters of the conditional distribution can be read off from the above equations using

$$\begin{aligned}\boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ \boldsymbol{\Sigma}_{1|2} &= (\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22}) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\end{aligned}$$

Gaussian inference: Information form

- Recall

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\xi},\boldsymbol{\Lambda}) \quad = \quad (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{\frac{1}{2}} \ \exp[-\frac{1}{2}(\mathbf{x}^T\boldsymbol{\Lambda}\mathbf{x}+\boldsymbol{\xi}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}-2\mathbf{x}^T\boldsymbol{\xi})]$$

- It is also possible to derive the marginalization and conditioning formulas in information form, yielding:

$$\begin{aligned} p(\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_2|\boldsymbol{\xi}_2-\boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\xi}_1, \boldsymbol{\Lambda}_{22}-\boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}) \\ p(\mathbf{x}_2|\mathbf{x}_1) &= \mathcal{N}(\mathbf{x}_2|\boldsymbol{\xi}_1-\boldsymbol{\Lambda}_{12}\mathbf{x}_2, \boldsymbol{\Lambda}_{11}) \end{aligned}$$

- We see that marginalization is easier in moment form, and conditioning is easier in information form.

Linear-Gaussian systems: Prior & observation model

- Suppose we have two variables, $\mathbf{x}$ and $\mathbf{y}$.
- Let $\mathbf{x} \in \mathbb{R}^{D_x}$ be a hidden variable, and $\mathbf{y} \in \mathbb{R}^{D_y}$ be a noisy observation of $\mathbf{x}$.
- Let us assume we hav the following prior and likelihood:

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_y) \end{aligned}$$

  where $\mathbf{A}$ is a matrix of size $D_y \times D_x$.
- This is an example of a **linear Gaussian system**. Understanding these models will enable us to easily derive Gaussian processes for nonlinear regression, classification and dimensionaly reduction; Kalman filtering and factor analysis.

Linear-Gaussian systems: Joint & posterior

- Let $\mathbf{z} = (\mathbf{x}, \mathbf{y})^T$. Then,

$$\begin{aligned} p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \\ \boldsymbol{\mu}_z &= \begin{pmatrix} \boldsymbol{\mu}_x \\ \mathbf{A}\boldsymbol{\mu}_x + \mathbf{b} \end{pmatrix} \\ \boldsymbol{\Sigma}_z &= \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_x\mathbf{A}^T \\ \mathbf{A}\boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_y + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T \end{pmatrix} \end{aligned}$$

- So we see that linear Gaussian systems are just a way to create large jointly Gaussian distributions.

- The conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ is given by the following Gaussian distribution (so the Gaussian is conjugate to itself):

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \\ \boldsymbol{\Sigma}_{x|y}^{-1} &= \boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{A} \\ \boldsymbol{\mu}_{x|y} &= \boldsymbol{\Sigma}_{x|y}[\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\,(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x] \end{aligned}$$

Linear-Gaussian systems: Convolution

- The normalization constant is given by the following equation, which we will call the Gaussian marginal likelihood equation:

$$p(\mathbf{y}) = \int \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b},\boldsymbol{\Sigma}_y)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)d\mathbf{x} = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}_x+\mathbf{b},\boldsymbol{\Sigma}_y+\mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T)$$

- Another useful result is the following expression for the expected value of the log of a Gaussian:

$$\begin{aligned}\int \log \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x},\boldsymbol{\Sigma}_y)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)d\mathbf{x} &= -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_y| \\ &\quad -\frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}_y^{-1}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}_x) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_y^{-1}\mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T)\end{aligned}$$

Linear-Gaussian systems: Derivation

- Consider the linear Gaussian system

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Lambda}^{-1}) \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}) \end{aligned}$$

  Let $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ and consider the log of the joint distribution:

$$\log p(\mathbf{x}, \mathbf{y}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_x)^T\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}_x) - \frac{1}{2}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b})^T\mathbf{L}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}) + \text{const}$$

  Since this is a quadratic form, we see that $p(\mathbf{x}, \mathbf{y})$ is a Gaussian.

- Expanding out the terms involving $\mathbf{x}$ and $\mathbf{y}$ (and ignoring constants) we have

$$\begin{aligned} &-\frac{1}{2}\mathbf{x}^T(\boldsymbol{\Lambda}+\mathbf{A}^T\mathbf{L}\mathbf{A})\mathbf{x} - \frac{1}{2}\mathbf{y}^T\mathbf{L}\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{L}\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{L}\mathbf{y} \\ &\quad = -\frac{1}{2}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^T\begin{pmatrix}\boldsymbol{\Lambda}+\mathbf{A}^T\mathbf{L}\mathbf{A} & -\mathbf{A}^T\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = -\frac{1}{2}\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z} \end{aligned}$$

Linear-Gaussian systems: Derivation

- Using the Schur complement, we invert the information matrix $\mathbf{\Sigma}^{-1}$, to get the covariance:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1}\mathbf{A}^T \\ \mathbf{A}\mathbf{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T \end{pmatrix}$$

- The mean of the joint is given by

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

- Given the joint, we can easily write down the marginal $p(\mathbf{y})$ by extracting the appropriate rows and columns:

$$\begin{aligned} p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbb{E}[\mathbf{y}], \mathrm{cov}[\mathbf{y}]) \\ \mathbb{E}[\mathbf{y}] &= \mathbf{A}\boldsymbol{\mu}_x + \mathbf{b} \\ \mathrm{cov}[\mathbf{y}] &= \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T \end{aligned}$$

Linear-Gaussian systems: Derivation

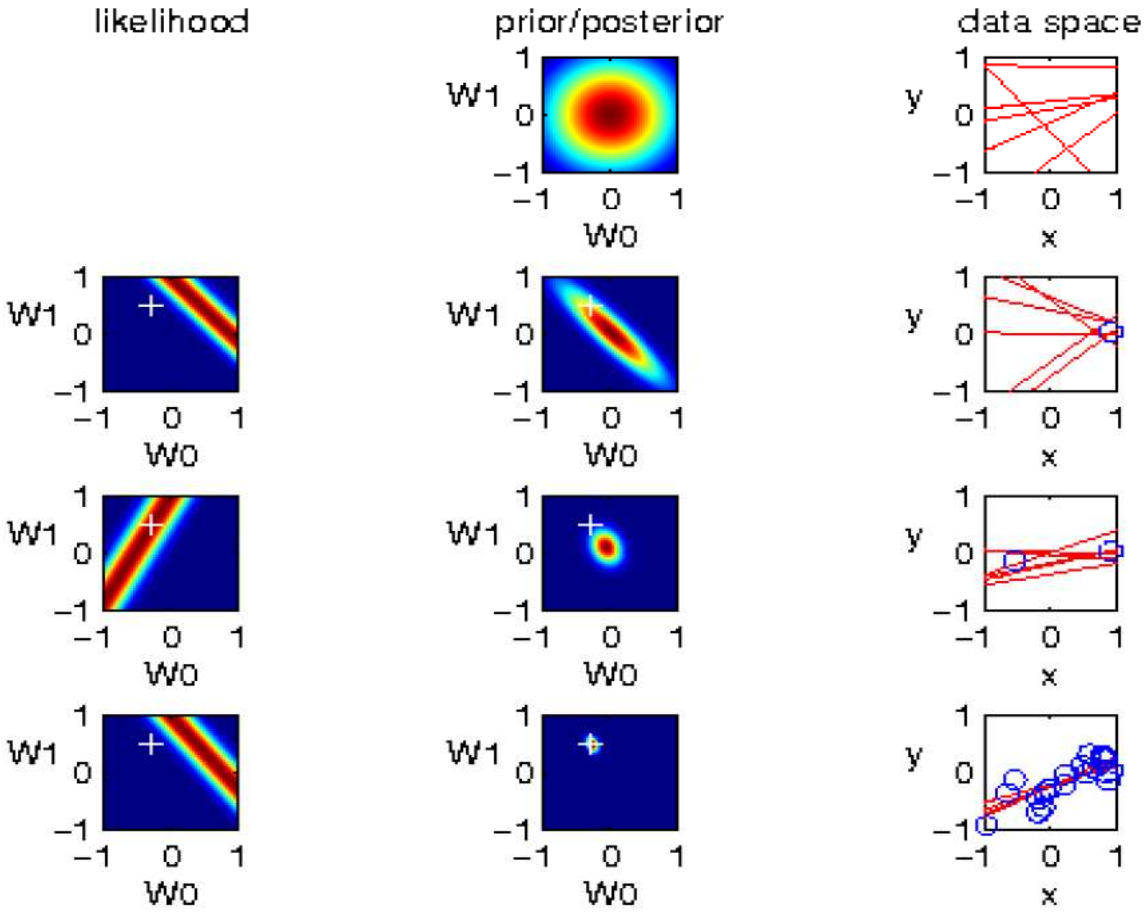
- To compute the conditional $p(\mathbf{x}|\mathbf{y})$, we use the information form results:

$$
\begin{aligned}
\boldsymbol{\Lambda}_{x|y} &= \boldsymbol{\Lambda}_{xx} = \boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A} \\
\boldsymbol{\Sigma}_{x|y} &= \boldsymbol{\Lambda}_{x|y}^{-1} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1} \\
\boldsymbol{\mu}_{x|y} &= \boldsymbol{\Lambda}_{x|y}^{-1}\boldsymbol{\eta}_{x|y} = \boldsymbol{\Sigma}_{x|y}(\boldsymbol{\eta}_x - \boldsymbol{\Lambda}_{xy}\mathbf{y}) \\
&= \boldsymbol{\Sigma}_{x|y}(\boldsymbol{\Lambda}_{xx}\boldsymbol{\mu}_x + \boldsymbol{\Lambda}_{xy}\boldsymbol{\mu}_y - \boldsymbol{\Lambda}_{xy}\mathbf{y}) \\
&= \boldsymbol{\Sigma}_{x|y}((\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})\boldsymbol{\mu}_x - \mathbf{A}^T\mathbf{L}(\mathbf{A}\boldsymbol{\mu}_x + \mathbf{b}) + \mathbf{A}^T\mathbf{L}\mathbf{y}) \\
&= \boldsymbol{\Sigma}_{x|y}(\boldsymbol{\Lambda}\boldsymbol{\mu}_x + \mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}))
\end{aligned}
$$

Bayesian linear regression

- We can use Bayes rule for Gaussians to infer the parameters of a linear regression model. For simplicity, we will assume the noise variance σ^2 is known.

- So our goal is to compute $p(\mathbf{w}|\mathcal{D}, \sigma^2)$, where $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^N$.



Bayesian linear regression: Posterior

- The likelihood is a Gaussian, $\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}_N)$. The conjugate prior is also a Gaussian, which we will denote by $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{V}_0)$.
- Using Bayes rule for Gaussians, the posterior is given by

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) &\propto \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{V}_0)\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}_N) = \mathcal{N}(\mathbf{w}|\mathbf{w}_N, \mathbf{V}_N) \\
\mathbf{w}_N &= \mathbf{V}_N\mathbf{V}_0^{-1}\mathbf{w}_0 + \frac{1}{\sigma^2}\mathbf{V}_N\mathbf{X}^T\mathbf{y} \\
\mathbf{V}_N^{-1} &= \mathbf{V}_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}
\end{aligned}
$$

Bayesian linear regression: Ridge regression

- Consider the special case where $\mathbf{w}_0 = \mathbf{0}$ and $\mathbf{V}_0 = \tau_0^2 \mathbf{I}_d$, which is a spherical Gaussian prior. Then the posterior mean reduces to

$$\begin{aligned} \mathbf{w}_N &= \frac{1}{\sigma^2}\mathbf{V}_N \mathbf{X}^T \mathbf{y} = \frac{1}{\sigma^2}\left(\frac{1}{\tau_0^2}\mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ &= \left(\lambda \mathbf{I}_d + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

  where we have defined $\lambda := \frac{\sigma^2}{\tau_0^2}$. This is known as **ridge regression**.

Bayesian linear regression: Predicting

- The **posterior predictive distribution** at a test point $\mathbf{x}$ is:

$$\begin{aligned} p(y|\mathbf{x}, \mathcal{D}, \sigma^2) &= \int \mathcal{N}(y|\mathbf{x}^T\mathbf{w}, \sigma^2)\mathcal{N}(\mathbf{w}|\mathbf{w}_N, \mathbf{V}_N)d\mathbf{w} \\ &= \mathcal{N}(y|\mathbf{x}^T\mathbf{w}_N, \sigma_N^2(\mathbf{x})) \\ \sigma_N^2(\mathbf{x}) &= \sigma^2 + \mathbf{x}^T\mathbf{V}_N\mathbf{x} \end{aligned}$$

- The variance in this prediction, $\sigma_N^2(\mathbf{x})$, depends on two terms: the observation noise with variance σ^2, and the parameter uncertainty $\sigma_N^2(\mathbf{x})$; this latter term varies depending on how close $\mathbf{x}$ is to the training data $\mathcal{D}$. The error bars get larger as we move away from the training points, representing increased uncertainty.

- By contrast, the plugin approximation has constant sized error bars, since

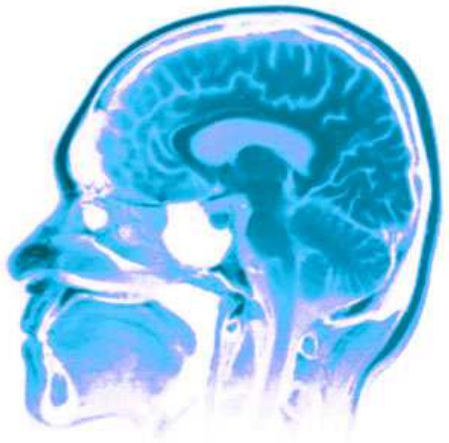
$$p(y|\mathbf{x}, \mathcal{D}, \sigma^2) \approx \int \mathcal{N}(y|\mathbf{x}^T\mathbf{w}, \sigma^2)\delta_{\hat{\mathbf{w}}}(\mathbf{w})d\mathbf{w} = p(y|\mathbf{x}^T\hat{\mathbf{w}}, \sigma^2)$$

Reading assignment

Please read on:

- MLE for the MVN
- Bayes for the MVN
- How to place Wishart priors on the covariance to derive the Bayesian posterior distribution of the covariance matrix.

Next class



Gaussian Processes

Nando de Freitas

2011

KPM Book Sections: *16*