Review to probability and random variables

Janne Huttunen

UEF

September 26, 2013

We start with few results related matrix calculation.

Block matrix inversion

A, B, C and D are matrices such that A and D are non-singular square matrices. Then Gauss elimination gives

$$\begin{aligned}\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} &= \begin{pmatrix} A^{-1} + A^{-1}B\Gamma_A^{-1}CA^{-1} & -A^{-1}B\Gamma_A^{-1} \\ -\Gamma_A^{-1}CA^{-1} & \Gamma_A^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \Gamma_D^{-1} & -\Gamma_D^{-1}BD^{-1} \\ -D^{-1}C\Gamma_D^{-1} & D^{-1} + D^{-1}C\Gamma_D^{-1}BD^{-1} \end{pmatrix}\end{aligned}$$

where $\Gamma_A = D - CA^{-1}B$ and $\Gamma_D = A - BD^{-1}C$ (Schur complements).

- Special case:

$$\begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix}$$

Comparing the blocks in the block matrix inversion formula gives two important matrix identities:

Matrix inversion lemma

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$$

This result is also known as Sherman–Morrison–Woodbury formula or Woodbury formula.

Matrix inversion identity

$$A^{-1}B(D - CA^{-1}B)^{-1} = (A - BD^{-1}C)^{-1}BD^{-1}$$

Short review to probability theory and random variables

- A random variable X is a function $X : \Omega \mapsto \mathbb{R}$ where Ω is a set called sample space. The elements $\omega \in \Omega$ are called samples.
- The value $X(\omega)$ for fixed ω is called as a realization of X.
- (Cumulative) distribution of X is a function $F : \mathbb{R} \mapsto [0, 1]$ such that

$$F(y) = \mathbb{P}(X \leq y), \quad y \in \mathbb{R}$$

  where $\mathbb{P}$ denotes probability: $\mathbb{P}(X \leq y)$ is probability for the event that the value of X is less or equal to y.
- The variance of X is

$$\sigma_X^2 = \mathrm{var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \geq 0$$

- The standard deviations of X is $\sigma_X = \sqrt{\mathrm{var}(X)}$

- A random variable X is called a continuos random variable if there is a function $p : \mathbb{R} \to \mathbb{R}$ such that

$$F(y) = \int_{-\infty}^{y} p(x)\mathrm{d}x$$

  The function p is called as the probability density of X.

- Note: $F(\infty) = \int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$.
- The expectations is $\mathbb{E}\left[X\right] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x$.
- More generally

$$\mathbb{E}\left[f(X)\right] = \int_{-\infty}^{\infty} f(x)p(x)\mathrm{d}x$$

  for functions f for which the integral is defined.

- During the lectures, we usually assume that random variables are continuous (especially if we are using the probability density p)

Multidimensional random variables

- A n–dimensional random variable (or random vector) is a function $X : \Omega \mapsto \mathbb{R}^n$.
- Can also be tough as a vector of random variables

$$X = (X_1, \ldots, X_n)^{\mathrm{T}}$$

where $X_1, \ldots, X_n$ are random variables.
- The expectation of X: $\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n])^{\mathrm{T}}$
- The covariance of X:

$$\begin{aligned}
\mathrm{cov} X &= \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^{\mathrm{T}}\right] \\
&= \begin{pmatrix} \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_1 - \mathbb{E}[X_1])] & \cdots & \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_n - \mathbb{E}[X_n])] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(X_n - \mathbb{E}[X_n])(X_1 - \mathbb{E}[X_1])] & \cdots & \mathbb{E}[(X_n - \mathbb{E}[X_n])(X_n - \mathbb{E}[X_n])] \end{pmatrix}
\end{aligned}$$

- A symmetric matrix K is called positive definite if $x^{\mathrm{T}}Kx > 0$ for all vectors $x \neq 0$ (or equivalently, all eigenvalues are positive).
- A symmetric matrix K is called positive semi-definite if $x^{\mathrm{T}}Kx \geq 0$ for all x (or all eigenvalues are non-negative).
- Positive semi-definite matrix is also positive definite, if it is not singular (i.e. $\det(K) \neq 0$ or $\exists K^{-1}$).
- Covariance matrices are symmetric and positive-semidefinite:

$$
\begin{aligned}
\mathrm{cov}(X)^{\mathrm{T}} &= \mathbb{E}\left[[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^{\mathrm{T}}]^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^{\mathrm{T}}]\right] = \mathrm{cov}(X) \\
x^{\mathrm{T}}\mathrm{cov}(X)x &= \mathbb{E}\left[x^{\mathrm{T}}(X - \mathbb{E}[X])(X - \mathbb{E}[X])^{\mathrm{T}}x\right] \\
&= \mathbb{E}\Big[\underbrace{[(X - \mathbb{E}[X])^{\mathrm{T}}x]}_{\in\mathbb{R}}{}^{\mathrm{T}} \underbrace{(X - \mathbb{E}[X])^{\mathrm{T}}x}_{\in\mathbb{R}}\Big] = \mathbb{E}\left[|(X - \mathbb{E}[X])^{\mathrm{T}}x|^2\right] \geq 0
\end{aligned}
$$

- Cumulative distribution of a random vector X is $F : \mathbb{R}^n \to [0, 1]$ such that

$$F(y) = \mathbb{P}(X_1 \leq y_1, \ldots, X_n \leq y_n), \quad y \in \mathbb{R}^n$$

- Continuous random vectors: there is a probability density function $p : \mathbb{R}^n \to \mathbb{R}$ such that

$$F(y) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_n} p(x_1, \ldots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

- Then

$$\mathbb{E}\left[f(X)\right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x) p(x_1, \ldots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

E.g.

$$\mathbb{E}\left[X_i\right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i p(x_1, \ldots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

- Let X and Y be random variables. Then the vector (X, Y) forms a new random vector.
- The joint probability density of X and Y is the probability density function $p(x, y)$ of (X, Y)
- The covariance of (X, Y) is

$$\begin{pmatrix} \Gamma_x & \Gamma_{xy} \\ \Gamma_{yx} & \Gamma_y \end{pmatrix}$$

  where Γ_x and Γ_y are covariances of X and Y and Γ_{xy} and Γ_{yx} are the cross-covariances:

$$\begin{aligned} \Gamma_{xy} &= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^{\mathrm{T}}\right], \\ \Gamma_{yx} &= \mathbb{E}\left[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])^{\mathrm{T}}\right] = \Gamma_{xy}^{\mathrm{T}} \end{aligned}$$

- If X and Y independent, X and Y are uncorrelated ($\Gamma_{xy} = 0$, $\Gamma_{yx} = 0$). The opposite is not always true.

- Marginal densities:

$$p(x) = \int_{-\infty}^{\infty} p(x,y)\mathrm{d}y \quad \text{and} \quad p(y) = \int_{-\infty}^{\infty} p(x,y)\mathrm{d}x$$

- The conditional probability density for X given Y:

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

- If X and Y are independent: $p(x,y) = p(x)p(y)$. Also $p(x|y) = p(x)$.
- Identity $p(x,y) = p(x|y)p(y) = p(y|x)p(x)$ gives important results:

Bayes rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Normal distributions

A random variable X is called normal or Gaussian if the probability density is of the form

$$p(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Gamma_x)}} \exp\left\{-\frac{1}{2}(x-\bar{x})^{\mathrm{T}}\Gamma_x^{-1}(x-\bar{x})\right\}$$

where $\bar{x}$ and Γ_x are the expectation and covariance of X. Normal distributions are denoted as $\mathcal{N}(\bar{x}, \Gamma_x)$.

- Note that normal distributions are completely determined by its expectation and covariance.
- Thus a common approach is to check that distribution is Gaussian and then calculate the expectation and covariance

- The definition can be extended also for singular covariances Γ_x using characteristic functions $\phi(\xi) = \mathbb{E}\left[e^{i\xi^{\mathrm{T}}X}\right]$, $\xi \in \mathbb{R}^n$ where $i = \sqrt{-1}$.
- If X is continuos, $\phi(\xi) = \int e^{i\xi^{\mathrm{T}}x} p(x)\mathrm{d}x$ (the Fourier transform of p).

Normal distributions (extended definition)

A random variable X is normal if its characteristic function ϕ is of the form

$$\phi(\xi) = e^{i\xi^{\mathrm{T}}\bar{x} - \frac{1}{2}\xi^{\mathrm{T}}\Gamma_x\xi}, \quad \xi \in \mathbb{R}^n,$$

where $\bar{x}$ and Γ_x are the expectation and covariance of X.

- The Fourier transform of Gaussian density $p(x)$ is also of this form.
- The following results can be easily proved using charasteristic functions. Let $X \sim \mathcal{N}(\bar{x}, \Gamma_x)$ and $X \sim \mathcal{N}(\bar{x}, \Gamma_x)$. Then

$$X \text{ and } Y \text{ independent } \Rightarrow X + Y \sim \mathcal{N}(\bar{x} + \bar{y}, \Gamma_x + \Gamma_y).$$
$$AX \sim \mathcal{N}(A\bar{x}, A\Gamma_x A^{\mathrm{T}}) \text{ for any matrix } A.$$

- Especially components of Gaussian random vectors are also Gaussian

Example: drawing samples from Gaussian distributions

Problem: We want to draw samples from the Gaussian distribution $\mathcal{N}(\mu, \Gamma)$. How to do that?

Solution (with Matlab): Compute Cholesky factor L of Γ ($L^{\mathrm{T}} L = \Gamma$) using **chol** in Matlab. Then draw samples for $X \sim \mathcal{N}(0, I)$ using **randn** and compute $Y = L^{\mathrm{T}} X + \mu$.

Practical note: Cholesky factor can only be computed for symmetric positive-definite matrices. In some cases the covariance Γ can be numerically singular (all eiqenvalues positive but some very small) and **chol** might fail. In this case the covariance can be numerically stabilized by adding a small number to the diagonal elements (i.e., compute the cholesky for, say, $\Gamma + 10^{-10} I$). Another option is to use the eigenvalue decomposition of Γ (**eig**) which works also with symmetric positive-semidefinite matrices.

Jointly Gaussian variables and conditioning

- Assume that X and Y are jointly Gaussian which means that the random vector $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$ is Gaussian.
- For example, if X and Y are Gaussian and independent, X and Y are also jointly Gaussian (easy to see using charasteristic functions)
- Note also that, if X and Y are jointly Gaussian, both X and Y have to be also Gaussian ($X = (I\ 0)Z$, $Y = (0\ I)Z$).
- Joint probability density function of X and Y is (if the covariance of Z invertible)

$$p(x, y) = p(z) \propto \exp\left\{ -\frac{1}{2} \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \Gamma_x & \Gamma_{xy} \\ \Gamma_{yx} & \Gamma_y \end{bmatrix}^{-1} \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix} \right\}$$

- It is easy to see that, if X and Y are uncorrelated ($\Gamma_{xy} = \Gamma_{yx}^{\mathrm{T}} = 0$), X and Y are independent.
  Thus for Gaussian random variables: independent $\Leftrightarrow$ uncorrelated.
  (Holds also if covariances are singular)
- Conditional distribution of jointly Gaussian random variables X and Y: the conditional distribution of X given Y is $\mathcal{N}(\hat{x}, \hat{\Gamma}_{x|y})$ where the conditional expectation $\hat{x}$ and covariance $\hat{\Gamma}_{x|y}$ are

$$
\begin{aligned}
\hat{x} &= \bar{x} + \Gamma_{xy}\Gamma_y^{-1}(y - \bar{y}) \\
\hat{\Gamma}_{x|y} &= \Gamma_x - \Gamma_{xy}\Gamma_y^{-1}\Gamma_{yx}
\end{aligned}
$$

- This results can be derived for continuous Gaussian random variables by applying the block matrix inversion equation to $p(x, y)$ as follows.

- The block inversion formula gives

$$\begin{bmatrix} \Gamma_x & \Gamma_{xy} \\ \Gamma_{yx} & \Gamma_y \end{bmatrix}^{-1} = \begin{bmatrix} \hat{\Gamma}^{-1} & -\hat{\Gamma}^{-1}\Gamma_{xy}\Gamma_y^{-1} \\ -\Gamma_y^{-1}\Gamma_{yx}\hat{\Gamma}^{-1} & \Sigma \end{bmatrix}$$

where $\hat{\Gamma} = \Gamma_x - \Gamma_{xy}\Gamma_y^{-1}\Gamma_{yx}$ (Schur complement) and $\Sigma = \Gamma_y^{-1} + \Gamma_y^{-1}\Gamma_{yx}\hat{\Gamma}^{-1}\Gamma_{xy}\Gamma_y^{-1}$.

- Then

$$\begin{aligned}
&\begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \Gamma_x & \Gamma_{xy} \\ \Gamma_{yx} & \Gamma_y \end{bmatrix}^{-1} \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix} \\
&= (x-\bar{x})^{\mathrm{T}}\hat{\Gamma}^{-1}(x-\bar{x}) - (x-\bar{x})^{\mathrm{T}}\hat{\Gamma}^{-1}\Gamma_{xy}\Gamma_y^{-1}(y-\bar{y}) \\
&\quad -(y-\bar{y})^{\mathrm{T}}\Gamma_y^{-1}\Gamma_{yx}\hat{\Gamma}^{-1}(x-\bar{x}) + (y-\bar{y})^{\mathrm{T}}\Sigma(y-\bar{y}) \\
&= (x-\bar{x})^{\mathrm{T}}\hat{\Gamma}^{-1}(x-\bar{x}) - 2(x-\bar{x})^{\mathrm{T}}\hat{\Gamma}^{-1}\Gamma_{xy}\Gamma_y^{-1}(y-\bar{y}) \\
&\quad +(y-\bar{y})^{\mathrm{T}}\Sigma(y-\bar{y})
\end{aligned}$$

since the third term is the transpose of the second ($\hat{\Gamma}$ and Γ_y are symmetric) are therefore equal (the terms are scalars)

- We denote: $p(x|y) \propto g(x,y) \Leftrightarrow p(x|y) = C_y g(x,y)$ for some constant C_y which may depend on y. In probability context, such constant is often called as a normalization constant: $C_y = (\int g(x,y)\mathrm{d}x)^{-1}$ (remember that $\int p(x|y)\mathrm{d}x = 1$).
- The conditional density $p(x|y) = p(x,y)/p(y)$ is:

$$
\begin{aligned}
p(x|y) &\propto \exp\left\{-\frac{1}{2}\left[(x-\bar{x})^{\mathrm{T}}\hat{\Gamma}^{-1}(x-\bar{x}) - 2(x-\bar{x})^{\mathrm{T}}\hat{\Gamma}^{-1}\Gamma_{xy}\Gamma_y^{-1}(y-\bar{y})\right.\right.\\
&\qquad \left.\left. +(y-\bar{y})^{\mathrm{T}}\Sigma(y-\bar{y})\right] + \frac{1}{2}(y-\bar{y})^{\mathrm{T}}\Sigma_y(y-\bar{y})\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[(x-\bar{x})^{\mathrm{T}}\hat{\Gamma}^{-1}(x-\bar{x}) - 2(x-\bar{x})^{\mathrm{T}}\hat{\Gamma}^{-1}\Gamma_{xy}\Gamma_y^{-1}(y-\bar{y})\right]\right\}\\
&= \exp\left\{-\frac{1}{2}\left[x^{\mathrm{T}}\hat{\Gamma}^{-1}x - 2x^{\mathrm{T}}\hat{\Gamma}^{-1}\bar{x} + \bar{x}^{\mathrm{T}}\hat{\Gamma}^{-1}\bar{x}\right.\right.\\
&\qquad \left.\left. -2x^{\mathrm{T}}\hat{\Gamma}^{-1}\Gamma_{xy}\Gamma_y^{-1}(y-\bar{y}) + \bar{x}^{\mathrm{T}}\hat{\Gamma}^{-1}\Gamma_{xy}\Gamma_y^{-1}(y-\bar{y})\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[x^{\mathrm{T}}\hat{\Gamma}^{-1}x - 2x^{\mathrm{T}}\hat{\Gamma}^{-1}\bar{x} - 2x^{\mathrm{T}}\hat{\Gamma}^{-1}\Gamma_{xy}\Gamma_y^{-1}(y-\bar{y})\right]\right\}
\end{aligned}
$$

where the terms that do not depend on x are included to normalization constants. We only need terms which determine the form of $p(x|y)$ as a function of x, the normalization constant is not important.

- If $p(x|y) = \mathcal{N}(\hat{x}, \hat{\Gamma}_{x|y})$, we should be able to write

$$\begin{aligned} p(x|y) &\propto \exp\left\{-\frac{1}{2}(x-\hat{x})^{\mathrm{T}}\hat{\Gamma}_{x|y}^{-1}(x-\hat{x})\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(x^{\mathrm{T}}\hat{\Gamma}_{x|y}^{-1}x - 2x^{\mathrm{T}}\hat{\Gamma}_{x}^{-1}\hat{x} + \hat{x}^{\mathrm{T}}\hat{\Gamma}_{x}^{-1}\hat{x}\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(x^{\mathrm{T}}\hat{\Gamma}_{x|y}^{-1}x - 2x^{\mathrm{T}}\hat{\Gamma}_{x|y}^{-1}\hat{x}\right)\right\} \end{aligned}$$

- We can actually see that $p(x|y)$ can be written in this form when

$$\begin{aligned} \hat{\Gamma}_{x|y} &= \hat{\Gamma} = \Gamma_x - \Gamma_{xy}\Gamma_y^{-1}\Gamma_{yx} \\ \hat{\Gamma}_{x|y}^{-1}\hat{x} &= \hat{\Gamma}^{-1}\bar{x} + \hat{\Gamma}^{-1}\Gamma_{xy}\Gamma_y^{-1}(y-\bar{y}) \end{aligned}$$

  which gives the results.
- Note that the above derivation is only valid if the covariance matrices (Γ_x, Γ_y, the joint covariance Γ_z, $\hat{\Gamma}_{x|y}$) are invertible. However the result also holds if some or all of the covariances are singular (Γ_y^{-1} in the formula is replaced with the pseudo inverse if Γ_y is singular).