# 1 Multivariate Gaussians: A review

(a) Consider a two-dimensional random variable $Z \in \mathbb{R}^2$. In order for the random variable to be jointly Gaussian, a necessary and sufficient condition is that

- $Z_1$ and $Z_2$ are each marginally Gaussian, and
- $Z_1|Z_2 = z$ is Gaussian, and $Z_2|Z_1 = z$ is Gaussian.

A second characterization of a jointly Gaussian Random Variable (RV) $Z$ is that it can be written as $Z = AX$, where $X$ is a collection of i.i.d. standard normal RVs and $A \in \mathbb{R}^{2\times 2}$ is a matrix.

Note that the probability density function of a Gaussian RV with mean vector $\mu$ and covariance matrix $\Sigma$ is:

$$f(z) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)}{\sqrt{(2\pi)^k|\Sigma|}}$$

.

Let $X_1$ and $X_2$ be i.i.d. standard normal RVs. Let $U$ denote a random variable uniformly distributed on $\{-1, 1\}$, independent of everything else. Verify if the conditions of the first characterization hold for the following random variables, and calculate the covariance matrix $\Sigma_Z$.

- $Z_1 = X_1$ and $Z_2 = X_2$.
- $Z_1 = X_1$ and $Z_2 = X_1 + X_2$. (Use the second characterization to argue joint Gaussianity.)
- $Z_1 = X_1$ and $Z_2 = -X_1$.
- $Z_1 = X_1$ and $Z_2 = UX_1$.

(b) Use the above example to show that two Gaussian random variables can be uncorrelated, but not independent. On the other hand, show that two uncorrelated, jointly Gaussian RVs are independent.

(c) With the setup above, let $Z = VX$, where $V \in \mathbb{R}^{2\times 2}$ as a fixed non-random matrix, and $Z, X \in \mathbb{R}^2$. What is the covariance matrix $\Sigma_Z$? Is this also true for a RV other than Gaussian?

(d) Use the above setup to show that $X_1 + X_2$ and $X_1 - X_2$ are independent. Give another example pair of linear combinations that are independent.

(e) Given a jointly Gaussian RV $Z \in \mathbb{R}^2$ with covariance matrix $\Sigma_Z = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12} & \Sigma_{22}\end{bmatrix}$, how would you derive the distribution of $Z_1|Z_2 = z$?

Hint: The following identity may be useful

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{b}{c} & 1 \end{bmatrix} \begin{bmatrix} \left(a - \frac{b^2}{c}\right)^{-1} & 0 \\ 0 & \frac{1}{c} \end{bmatrix} \begin{bmatrix} 1 & -\frac{b}{c} \\ 0 & 1 \end{bmatrix}.$$

## 2 Kernel Validity

For a function $k(x_i, x_j)$ to be a valid kernel, it suffices to show either of the following conditions is true:

1. $k$ has an inner product representation: $\exists\ \Phi : \mathbb{R}^d \to \mathcal{H}$, where $\mathcal{H}$ is some (possibly infinite-dimensional) inner product space such that $\forall x_i, x_j \in \mathbb{R}^d,\ k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$.
2. For every sample $x_1, x_2, \ldots, x_n \in R^d$, the kernel matrix

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & k(x_i, x_j) & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}$$

   is positive semidefinite. For the following parts you can use either condition (1) or (2) in your proofs.

(a) Show that the first condition implies the second one, i.e. if $\forall x_i, x_j \in \mathbb{R}^d,\ k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ then the kernel matrix $K$ is PSD.

(b) Given a positive semidefinite matrix $A \in \mathbb{R}^{d \times d}$, show that $k(x_i, x_j) = x_i^\top A x_j$ is a valid kernel.

(c) Show why $k(x_i, x_j) = x_i^\top(\text{rev}(x_j))$ (where $\text{rev}(x)$ reverses the order of the components in $x$) is *not* a valid kernel.

(d) Soon we will cover a regression method based on kernels called kernel ridge regression (KRR). A key intermediate step to solve KRR is the following optimization problem:

$$\text{argmin}_{\alpha \in \mathbb{R}^n} \left[ \frac{1}{2} \alpha^T (K + \lambda I) \alpha - \lambda \langle \alpha, y \rangle \right]$$

where $y \in \mathbb{R}^n$, $\lambda \geq 0$, and $K \in \mathbb{R}^{n \times n}$ is the kernel matrix computed by applying a kernel function $k$ on every sample pair: $k(x_i, x_j)$. How does the requirement that K be a kernel affect the properties of this optimization problem? You may want to consider the cases where $\lambda$ is close to zero or even $\lambda = 0$.?