Multivariate Gaussian Cheat Sheet

Geoff Pleiss

Definition (Multivariate Gaussian)**.** *Let $\boldsymbol{y}$ be a d-dimensional vector-valued random variable.* $\boldsymbol{y}$ *is* multivariate Gaussian *if and only all linear combination of its entries are univariate Gaussian; i.e. for all $\boldsymbol{c} \in \mathbb{R}^d$, we have that $p(\boldsymbol{c}^\top \boldsymbol{y} = a) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(a-\mu)^2\right)$ for some $\mu, \sigma \in \mathbb{R}$.*

---

1) Multivariate Gaussian Density

Let $\boldsymbol{y}$ be a multivariate Gaussian random variable with mean $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{\mu}$ and covariance $\mathbb{E}[(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})^\top] = \boldsymbol{K}$. The probability density of $\boldsymbol{y}$ is given by:

$$p(\boldsymbol{y} = \boldsymbol{a}) = \mathcal{N}(\boldsymbol{a}; \boldsymbol{\mu}, \boldsymbol{K}) := |2\pi\boldsymbol{K}|^{-1/2} \exp\left(-\tfrac{1}{2}(\boldsymbol{a}-\boldsymbol{\mu})^\top \boldsymbol{K}^{-1}(\boldsymbol{a}-\boldsymbol{\mu})\right). \tag{1}$$

We will use the notation $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K})$, $p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K})$, and $p(\boldsymbol{y} = \boldsymbol{a}) = \mathcal{N}(\boldsymbol{a}; \boldsymbol{\mu}, \boldsymbol{K})$ interchangeably. All should be read as "$\boldsymbol{y}$ is a multivariate Gaussian random variable with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{K}$."

---

2) Important Multivariate Gaussian Closures

Many important operations on multivariate Gaussians preserve Gaussianity.

1. **Closure under affine transformation.** Let $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K})$. Given matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$, we have:

   $$(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}) \sim \mathcal{N}\left(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{K}\boldsymbol{A}^\top\right). \tag{2}$$

   (For Gaussian processes with appropriate regularity conditions, this property can be generalized to *closure under arbitrary linear operations.*)

2. **Closure under linear combination.** Now let $\boldsymbol{y}' \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{K}'')$. If $\boldsymbol{y} \perp \boldsymbol{y}'$ (read: $\boldsymbol{y}$ and $\boldsymbol{y}'$ are independent random variables), then

   $$(\boldsymbol{y} + \boldsymbol{y}') \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\mu}', \boldsymbol{K} + \boldsymbol{K}''). \tag{3}$$

   (This property is analogously extended to Gaussian processes.)

3. **Closure under marginalization.** Given the following block multivariate Gaussian random variable, we have:

   $$\begin{bmatrix}\boldsymbol{y} \\ \boldsymbol{y}'\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{0} \\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{K} & \boldsymbol{K}' \\ \boldsymbol{K}'^\top & \boldsymbol{K}''\end{bmatrix}\right) \quad \implies \quad p(\boldsymbol{y}) = \int p\left(\begin{bmatrix}\boldsymbol{y} \\ \boldsymbol{y}'\end{bmatrix}\right) \mathrm{d}\boldsymbol{y}' = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K}). \tag{4}$$

   (This property naturally applies to Gaussian processes by definition: any finite subset of Gaussian process evaluations are multivariate Gaussian distributed.)

4. **Closure under conditioning.** With the same block random variable, we have:

   $$(\boldsymbol{y}' \mid \boldsymbol{y} = \boldsymbol{a}) = \mathcal{N}\left(\boldsymbol{K}'^\top \boldsymbol{K}^{-1}\boldsymbol{a},\ \boldsymbol{K}'' - \boldsymbol{K}'^\top \boldsymbol{K}^{-1}\boldsymbol{K}\right). \tag{5}$$

   We will often drop the $= \boldsymbol{a}$ and simply write $\boldsymbol{y}' \mid \boldsymbol{y}$.

3) Other Useful Properties

1. **Uncorrelation implies independence.** Two random variables $\boldsymbol{y}$ and $\boldsymbol{y}'$ are uncorrelated if $\mathbb{E}[(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])(\boldsymbol{y}' - \mathbb{E}[\boldsymbol{y}'])^\top] = \boldsymbol{0}$. For arbitrary random variables, uncorrelation does not imply independence. However, for multivariate Gaussians:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{K} & \boldsymbol{K}' \\ \boldsymbol{K}'^\top & \boldsymbol{K}'' \end{bmatrix} \right) : \qquad \boldsymbol{K}' = \boldsymbol{0} \quad \Leftrightarrow \quad \boldsymbol{y} \perp \boldsymbol{y}' \tag{6}$$

2. **Conditional variances from the Cholesky factorization.** The *Cholesky factor* of the (positive-definite) covariance matrix $\boldsymbol{K}$ is the unique lower triangular matrix $\boldsymbol{L}$ such that 1) $\boldsymbol{L}\boldsymbol{L}^\top = \boldsymbol{K}$ and 2) $L_{ii} > 0$ for all i. Consider the N-dimensional multivariate Gaussian random variable $\begin{bmatrix} y_1 & \cdots & y_N \end{bmatrix} \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$. We can write the density of $\boldsymbol{y}$ in *autoregressive form*:

$$p(\boldsymbol{y}) \;=\; p(\underbrace{y_1}_{:=z_1}) \times p(\underbrace{y_2 \mid y_1}_{:=z_2}) \times p(\underbrace{y_3 \mid y_1, y_2}_{:=z_3}) \times \ldots \times p(\underbrace{y_N \mid y_1, \ldots, y_{N-1}}_{:=z_N}). \tag{7}$$

   We can interpret Eq. (7) as sequential Bayesian inference. We first consider the random variable y_1. After observing y_1, we then consider y_2 (conditioned on our observation). After observing y_2, we then consider y_3 (conditioning on our observations). And so on.

   By Eq. (4), z_1, ..., z_N are univariate Gaussians. The Cholesky factor gives us a convenient way to automatically compute the variance of these conditional random variables:

$$\mathbb{V}[z_i] = \mathbb{V}[y_i \mid y_1, \ldots, y_{i-1}] = L_{ii}^2. \tag{8}$$

3. **Sampling.** To draw a sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K})$, we often use the following computational routine:

   - Use e.g. `np.random.randn` to draw a sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
   - Compute $\boldsymbol{L}\boldsymbol{\epsilon} + \boldsymbol{\mu}$, where $\boldsymbol{L}$ is the Cholesky factor of $\boldsymbol{K}$.

   We do not specifically need to use the matrix $\boldsymbol{L}$ in the second step; we could instead use any matrix $\boldsymbol{A}$ such that $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{K}$. (Note that all matrices $\boldsymbol{A}$ that satisfy this condition are equivalent to $\boldsymbol{L}$ up to an orthogonal rotation. In other words, there exists some orthogonal matrix $\boldsymbol{Q}$ so that $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{L}$.)

4. **Sequences.** Consider a sequence of multivariate Gaussian variables $\{\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{K}_i)\}$, where $\{\boldsymbol{\mu}_i\}$ and $\{\boldsymbol{K}_i\}$ represent a sequence of means and covariances, respectively. (Weak) convergence of $\boldsymbol{y}_i$ is uniquely determined by convergence of the mean/covariance sequences:

$$\{\boldsymbol{\mu}_i\} \to \boldsymbol{\mu},\ \{\boldsymbol{K}_i\} \to \boldsymbol{K} \qquad \Longrightarrow \qquad \{p(\boldsymbol{y}_i)\} \xrightarrow{\text{dist.}} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K}) \tag{9}$$