(Multivariate) Gaussian (Normal) Probability Densities

Carl Edward Rasmussen, José Miguel Hernández-Lobato & Richard Turner

April 20th, 2018

Gaussian Density

The probability density of a D-dimensional Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ is given by

$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) \;=\; \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) \;=\; \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\big(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\big),$$

and we also write

$$\mathbf{x}|\boldsymbol{\mu}, \Sigma \;\sim\; \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma).$$

The covariance matrix Σ must be symmetric and positive definite.

In the special (scalar) case where $D = 1$ we have

$$p(x|\mu, \sigma^2) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\tfrac{1}{2}(x-\mu)^2/\sigma^2\big),$$

where σ^2 is the variance and σ is the standard deviation.

The *standard* Gaussian has $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = I$ (the unit matrix).

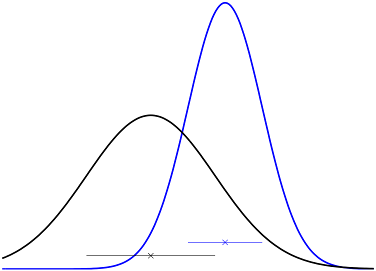
Parametrisation

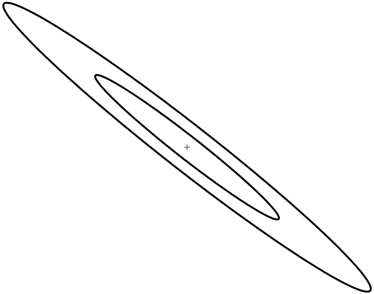
There are two commonly used parametrisations of Gaussians

- *standard* parametrisation:
  - *mean* $\boldsymbol{\mu}$ and
  - *covariance* Σ
- *natural* parametrisation:
  - *natural mean* $\boldsymbol{\nu} = \Sigma^{-1}\boldsymbol{\mu}$ and
  - *precision* matrix $\mathsf{R} = \Sigma^{-1}$.

Different operations are more convenient in either parametrisation.

Gaussian Pictures



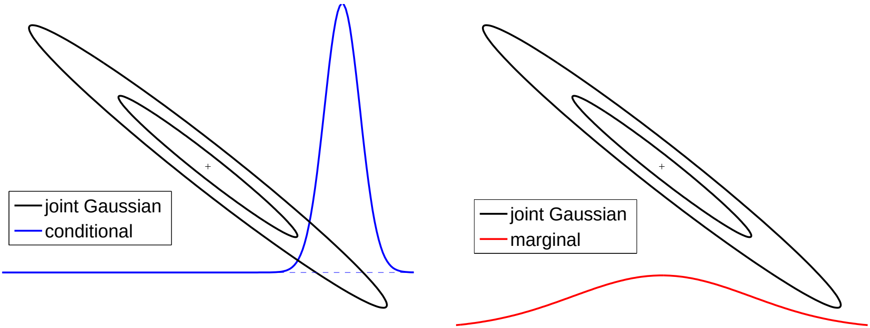


The mean corresponds to the location or center of the distribution.

In one dimension, the square root of the variance corresponds to the *width* of the distribution.

In multiple dimensions, the eigen-vectors of the covariance matrix give the principal axis of the elliptical equi-probability contours of the distribution, and the square root of the eigenvalues the width of the distribution in the corresponding directions.

Conditionals and Marginals of a Gaussian, pictorial



Both the conditionals $p(x|y)$ and the marginals $p(x)$ of a joint Gaussian $p(x, y)$ are again Gaussian.

Conditionals and Marginals of a Gaussian, algebra

If **x** and **y** are jointly Gaussian

$$p(\mathbf{x}, \mathbf{y}) \;=\; p\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}\right) \;=\; \mathcal{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}\right),$$

we get the marginal distribution of **x**, $p(\mathbf{x})$ by

$$p(\mathbf{x}, \mathbf{y}) \;=\; \mathcal{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}\right) \;\Longrightarrow\; p(\mathbf{x}) \;=\; \mathcal{N}(\mathbf{a},\; A),$$

and the conditional distribution of **x** given **y** by

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}\right) \;\Longrightarrow\; p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{a} + BC^{-1}(\mathbf{y} - \mathbf{b}),\; A - BC^{-1}B^\top),$$

where **x** and **y** can be scalars or vectors.

Gaussian Properties

Gaussians are closed both under marginalisation and conditioning.
If $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$$

Kullback-Leibler Divergence (Relative Entropy)

The Kullback-Leibler (KL) divergence between continuous distributions is

$$\mathcal{KL}(q(x)\|p(x)) \;=\; \int q(x)\log\frac{q(x)}{p(x)}dx.$$

The KL divergence is an asymmetric measure of distance between distributions.
The KL divergence between two Gaussians is

$$\mathcal{KL}(\mathcal{N}_0\|\mathcal{N}_1) \;=\; \tfrac{1}{2}\log|\Sigma_1\Sigma_0^{-1}| + \tfrac{1}{2}\,\mathrm{tr}\left(\Sigma_1^{-1}\big((\mu_0-\mu_1)(\mu_0-\mu_1)^\top + \Sigma_0 - \Sigma_1\big)\right).$$

KL matching constrained Gaussians

It is often convenient to approximate one distribution with another, simpler one, by finding the *closest match* within a constrained family.

Minimizing KL divergence between a general Gaussian $\mathcal{N}_g$ and a factorized Gaussian $\mathcal{N}_f$ will match the means $\boldsymbol{\mu}_f = \boldsymbol{\mu}_g$ and for the covariances either:

$$\frac{\partial \mathcal{KL}(\mathcal{N}_f \| \mathcal{N}_g)}{\partial \Sigma_f} = -\tfrac{1}{2}\Sigma_f^{-1} + \tfrac{1}{2}\Sigma_g^{-1} = 0 \;\Rightarrow\; (\Sigma_f)_{ii} = 1/(\Sigma_g^{-1})_{ii},$$

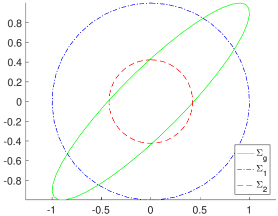
or

$$\frac{\partial \mathcal{KL}(\mathcal{N}_g \| \mathcal{N}_f)}{\partial \Sigma_f} = \tfrac{1}{2}\Sigma_f^{-1} - \tfrac{1}{2}\Sigma_f^{-1}\Sigma_g\Sigma_f^{-1} = 0 \;\Rightarrow\; (\Sigma_f)_{ii} = (\Sigma_g)_{ii}.$$

Interpretation:

- averaging wrt the *factorized* Gaussian, the fitted variance equals the *conditional* variance of Σ_g,
- averaging wrt the *general* Gaussian, the fitted variance equals the *marginal* variance of Σ_g,

with straight forward generalization to block diagonal Gaussians.



Appendix: Some useful Gaussian identities

If x is multivariate Gaussian with mean μ and covariance matrix Σ

$$p(\mathbf{x}; \mu, \Sigma) \;=\; (2\pi|\Sigma|)^{-D/2} \exp\big(-(\mathbf{x}-\mu)^\top \Sigma^{-1} (\mathbf{x}-\mu)/2\big),$$

then

$$\begin{aligned} \mathbb{E}[\mathbf{x}] &= \mu, \\ \mathbb{V}[\mathbf{x}] &= \mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}])^2] \;=\; \Sigma. \end{aligned}$$

For any matrix A, if $\mathbf{z} = A\mathbf{x} + \mathbf{b}$ then $\mathbf{z}$ is Gaussian and

$$\begin{aligned} \mathbb{E}[\mathbf{z}] &= A\mu + \mathbf{b}, \\ \mathbb{V}[\mathbf{z}] &= A\Sigma A^\top. \end{aligned}$$

Matrix and Gaussian identities cheat sheet

Matrix identities

- Matrix inversion lemma (Woodbury, Sherman & Morrison formula)

$$(Z + UWV^\top)^{-1} = Z^{-1} - Z^{-1}U(W^{-1} + V^\top Z^{-1}U)^{-1}V^\top Z^{-1}$$

- A similar equation exists for determinants

$$|Z + UWV^\top| = |Z|\,|W|\,|W^{-1} + V^\top Z^{-1}U|$$

The product of two Gaussian density functions

$$\mathcal{N}(\mathbf{x}|\mathbf{a}, A)\,\mathcal{N}(P^\top\mathbf{x}|\mathbf{b}, B) = z_c\,\mathcal{N}(\mathbf{x}|\mathbf{c}, C)$$

- is proportional to a Gaussian density function with covariance and mean

$$C = \left(A^{-1} + P\,B^{-1}P^\top\right)^{-1} \qquad\qquad c = C\,\left(A^{-1}\mathbf{a} + P\,B^{-1}\,\mathbf{b}\right)$$

- and has a normalizing constant z_c that is Gaussian both in $\mathbf{a}$ and in $\mathbf{b}$

$$z_c = (2\pi)^{-\frac{m}{2}}|B + P^\top A\,P|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{b} - P^\top\mathbf{a})^\top\left(B + P^\top A\,P\right)^{-1}(\mathbf{b} - P^\top\mathbf{a})\right)$$