Bayes' Theorem for Gaussians

Chris Bracegirdle*

September 2010

The family of Gaussian-distributed variables is, generally speaking, well-behaved under Bayesian manipulation of linear combinations. This document sets out the derivations of several utility results, most of which are well-known results for inference with Gaussian variables. The aim is to present the results in a coherent, clear, and pragmatic manner.

Fact 1. *Marginal* $p(x) = \int_y p(x, y)$

If

- $\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{xy}^\top & \mathbf{\Sigma}_{yy} \end{pmatrix}\right)$

then

- $\mathbf{x} \sim \mathcal{N}(\mu_x, \mathbf{\Sigma}_{xx})$

Proof. The derivation is rather long-winded, and requires calculation of the Schur complement as well as completing the square of the Gaussian p.d.f. to integrate out the variable. For a full work-through, see *Bishop* (2007, Section 2.3.2). □

Fact 2. *Joint* $p(x, y) = p(y|x)\, p(x)$

If

- $\mathbf{x} \sim \mathcal{N}(\mu_x, \mathbf{\Sigma}_x)$ and
- $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{M}\mathbf{x} + \mu_y, \mathbf{\Sigma}_y)$

then

*Department of Computer Science, University College London

- $\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mathbf{M}\mu_x + \mu_y \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_x & \mathbf{\Sigma}_x\mathbf{M}^\top \\ \mathbf{M}\mathbf{\Sigma}_x^\top & \mathbf{M}\mathbf{\Sigma}_x\mathbf{M}^\top + \mathbf{\Sigma}_y \end{pmatrix}\right)$

Proof. We can write $\mathbf{y} = \mathbf{M}\mathbf{x}+\epsilon$, $\epsilon \sim \mathcal{N}(\mu_y, \mathbf{\Sigma}_y)$. Then we have covariance[I,II] $\langle \Delta\mathbf{x}\Delta\mathbf{y}^\top \rangle = \left\langle \Delta\mathbf{x}\left(\mathbf{M}\Delta\mathbf{x} + \Delta\epsilon\right)^\top \right\rangle = \langle \Delta\mathbf{x}\Delta\mathbf{x}^\top \rangle \mathbf{M}^\top + \langle \Delta\mathbf{x}\Delta\epsilon^\top \rangle$. Since $\langle \Delta\mathbf{x}\Delta\epsilon^\top \rangle = 0$ we therefore have $\langle \Delta\mathbf{x}\Delta\mathbf{y}^\top \rangle = \mathbf{\Sigma}_x\mathbf{M}^\top$. Similarly, $\langle \Delta\mathbf{y}\Delta\mathbf{y}^\top \rangle = \left\langle \left(\mathbf{M}\Delta\mathbf{x} + \Delta\epsilon\right)\left(\mathbf{M}\Delta\mathbf{x} + \Delta\epsilon\right)^\top \right\rangle = \mathbf{M}\langle \Delta\mathbf{x}\Delta\mathbf{x}^\top \rangle \mathbf{M}^\top + \langle \Delta\epsilon\Delta\epsilon^\top \rangle = \mathbf{M}\mathbf{\Sigma}_x\mathbf{M}^\top + \mathbf{\Sigma}_y$. The result follows. □

Corollary 3. *Marginal* $p(y) = \int_x p(y|x)\,p(x)$ *(linear transform of a Gaussian)*

If

- $\mathbf{x} \sim \mathcal{N}(\mu_x, \mathbf{\Sigma}_x)$ and
- $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{M}\mathbf{x} + \mu_y, \mathbf{\Sigma}_y)$

then

- $p(\mathbf{y}) = \int_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})\,p(\mathbf{x}) = \mathcal{N}\left(\mathbf{M}\mu_x + \mu_y, \mathbf{M}\mathbf{\Sigma}_x\mathbf{M}^\top + \mathbf{\Sigma}_y\right)$.

Proof. Immediate from 1 and 2. □

Fact 4. *Conditioning* $p(x|y) \propto p(x, y)$

If

- $\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{xy}^\top & \mathbf{\Sigma}_{yy} \end{pmatrix}\right)$

then

- $\mathbf{x}|\mathbf{y} \sim \mathcal{N}\left(\mu_x + \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}(\mathbf{y} - \mu_y), \mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{xy}^\top\right)$

Proof. Again the derivation is long-winded, and appeals to the Schur complement. See *Bishop* (2007, Section 2.3.1). □

Corollary 5. *Conditioning* $p(x|y) \propto p(y|x)\,p(x)$ *(inverse linear transform, dynamics reversal)*

If

- $\mathbf{x} \sim \mathcal{N}(\mu_x, \mathbf{\Sigma}_x)$ and
- $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{M}\mathbf{x} + \mu_y, \mathbf{\Sigma}_y)$

then

I *Displacement* of a variable $\mathbf{x}$ is given by $\Delta\mathbf{x} = \mathbf{x} - \langle \mathbf{x} \rangle$.

II $\mathbf{y} = \mathbf{M}\mathbf{x} + \epsilon \Longrightarrow \Delta\mathbf{y} = \mathbf{y} - \langle \mathbf{y} \rangle = \mathbf{M}\mathbf{x} + \epsilon - \langle \mathbf{M}\mathbf{x} + \epsilon \rangle = \mathbf{M}\mathbf{x} + \epsilon - \mathbf{M}\langle \mathbf{x} \rangle - \langle \epsilon \rangle = \mathbf{M}\Delta\mathbf{x} + \Delta\epsilon$

- $\mathbf{x}|\mathbf{y} \sim \mathcal{N}\left(\mathbf{R}\left(\mathbf{y} - \mathbf{M}\mu_x - \mu_y\right) + \mu_x, \mathbf{\Sigma}_x - \mathbf{R}\mathbf{M}\mathbf{\Sigma}_x^\top\right)$ where
- $\mathbf{R} = \mathbf{\Sigma}_x M^\top \left(\mathbf{M}\mathbf{\Sigma}_x\mathbf{M}^\top + \mathbf{\Sigma}_y\right)^{-1}$

Equivalently, we have $\mathbf{x} = \mathbf{R}\mathbf{y} + \epsilon$, $\epsilon \sim \mathcal{N}\left(\mu_x - \mathbf{R}\left(\mathbf{M}\mu_x + \mu_y\right), \mathbf{\Sigma}_x - \mathbf{R}\mathbf{M}\mathbf{\Sigma}_x^\top\right)$.

Proof. From 2, we have

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mathbf{M}\mu_x + \mu_y \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_x & \mathbf{\Sigma}_x\mathbf{M}^\top \\ \mathbf{M}\mathbf{\Sigma}_x^\top & \mathbf{M}\mathbf{\Sigma}_x\mathbf{M}^\top + \mathbf{\Sigma}_y \end{pmatrix}\right) \tag{1}$$

and by 4,

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}\left(\mu_x + \mathbf{\Sigma}_x\mathbf{M}^\top\left(\mathbf{M}\mathbf{\Sigma}_x\mathbf{M}^\top + \mathbf{\Sigma}_y\right)^{-1}\left(\mathbf{y} - \mathbf{M}\mu_x - \mu_y\right), \mathbf{\Sigma}_x - \mathbf{\Sigma}_x\mathbf{M}^\top\left(\mathbf{M}\mathbf{\Sigma}_x\mathbf{M}^\top + \mathbf{\Sigma}_y\right)^{-1}\mathbf{M}\mathbf{\Sigma}_x^\top\right)$$

from which the result follows.

Alternative derivation (originally by and based on *Barber* (2011, Chapter 8)) follows equation (1) by aiming to write $\mathbf{x} = \mathbf{R}\mathbf{y} + \epsilon$ for Gaussian ϵ with $\left\langle \Delta\epsilon\Delta\mathbf{y}^\top\right\rangle = 0$. Consider the covariance

$$\begin{aligned} \left\langle \Delta\mathbf{x}\Delta\mathbf{y}^\top\right\rangle &= \left\langle \left(\mathbf{R}\Delta\mathbf{y} + \Delta\epsilon\right)\Delta\mathbf{y}^\top\right\rangle \\ &= \mathbf{R}\left\langle \Delta\mathbf{y}\Delta\mathbf{y}^\top\right\rangle + \cancel{\left\langle \Delta\epsilon\Delta\mathbf{y}^\top\right\rangle} \\ \mathbf{R} &= \left\langle \Delta\mathbf{x}\Delta\mathbf{y}^\top\right\rangle\left\langle \Delta\mathbf{y}\Delta\mathbf{y}^\top\right\rangle^{-1} \end{aligned}$$

from equation (1) we have $\left\langle \Delta\mathbf{x}\Delta\mathbf{y}^\top\right\rangle = \mathbf{\Sigma}_x\mathbf{M}^\top$ and $\left\langle \Delta\mathbf{y}\Delta\mathbf{y}^\top\right\rangle = \mathbf{M}\mathbf{\Sigma}_x\mathbf{M}^\top + \mathbf{\Sigma}_y$. The desired mean and covariance are therefore obtained from

$$\begin{aligned} \langle\epsilon\rangle &= \langle\mathbf{x}\rangle - \mathbf{R}\langle\mathbf{y}\rangle = \mu_x - \mathbf{R}\left(\mathbf{M}\mu_x + \mu_y\right) \\ \left\langle \Delta\epsilon\Delta\epsilon^\top\right\rangle &= \left\langle \Delta\mathbf{x}\Delta\mathbf{x}^\top\right\rangle - \mathbf{R}\left\langle \Delta\mathbf{y}\Delta\mathbf{y}^\top\right\rangle\mathbf{R}^\top = \mathbf{\Sigma}_x - \mathbf{R}\mathbf{M}\mathbf{\Sigma}_x^\top \end{aligned}$$

□

References

Barber, D., *Bayesian Reasoning and Machine Learning*, Cambridge University Press, in press, 2011.

Bishop, C. M., *Pattern Recognition and Machine Learning*, 738 pp., Springer, 2007.