Multivariate Gaussians

Kevin P. Murphy

Last updated September 28, 2007

1 Multivariate Gaussians

The **multivariate Gaussian** or **multivariate normal** (MVN) distribution is defined by

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) \stackrel{\text{def}}{=} \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})] \tag{1}$$

where $\boldsymbol{\mu}$ is a $p \times 1$ vector, Σ is a $p \times p$ **symmetric positive definite (pd)** matrix, and p is the dimensionality of $\mathbf{x}$. It can be shown that $E[X] = \boldsymbol{\mu}$ and $\text{Cov}[X] = \Sigma$ (see e.g., [Bis06, p82]). (Note that in the 1D case, σ is the standard deviation, whereas in the multivariate case, Σ is the covariance matrix.)

The **quadratic form** $\Delta = (\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})$ in the exponent is called the **Mahalanobis distance** between $\mathbf{x}$ and $\boldsymbol{\mu}$. The equation $\Delta = \text{const}$ defines an ellipsoid, which are the level sets of constant probability density: see Figure 1. Often we just drawn the elliptical contour that contains 95% of the probability mass.

2 Bivariate Gaussians

In the 2D case, define the **correlation coefficient** between X and Y as

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{2}$$

Hence the covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \tag{3}$$

and the pdf (for the zero mean case) is given below

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho xy}{(\sigma_x\sigma_y)}\right)\right) \tag{4}$$

It should be clear from this example that when doing multivariate analysis, using matrices and vectors is easier than working with scalar variables.

3 Parsimonious covariance matrices

A full covariance matrix has $p(p+1)/2$ parameters. Hence it may be hard to estimate from data. We can restrict Σ to be diagonal; this has p parameters. Or we can use a **spherical (isotropic)** covariance, $\Sigma = \sigma^2 I$. See Figure 2 for a visualization of these different assumptions. We will consider other **parsimonious representations** for high dimensional Gaussian distributions later in the book. The problem of estimating a structured covariance matrix is called **covariance selection**.

4 Linear functions of Gaussian random variables

Linear combinations of MVN are MVN:

$$A \sim N(\mu, \Sigma) \quad \Rightarrow \quad AX \sim N(A\mu, A\Sigma A') \tag{5}$$

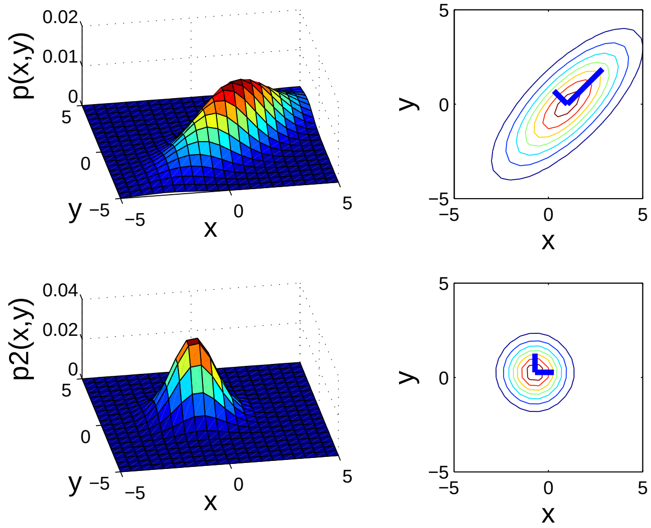

Figure 1: Visualization of a 2 dimensional Gaussian density. This figure was produced by `gaussPlot2dDemo`.

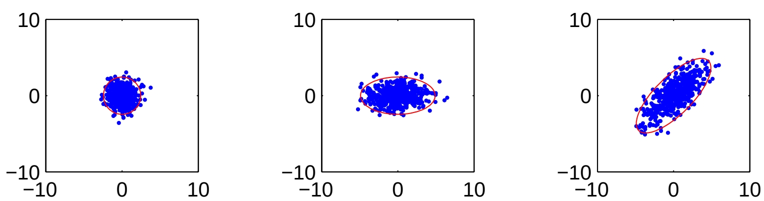

Figure 2: Samples from a spherical, diagonal and full covariance Gaussian, with 95% confidence ellipsoid superimposed. This figure was generated using `gaussSampleDemo`.

This implies that marginals of a MVN are also Gaussian. To see this, suppose that $X \in \mathbb{R}^3$ and we want to compute $p(X_1, X_2)$: we can just use the projection matrix

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{6}$$

Let $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $Y = AX = (X_1, X_2)$. Then

$$E\, Y = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \tag{7}$$

and

$$\text{Cov}Y = A\text{Cov}XA^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \tag{8}$$

So to marginalize, we just select out the corresponding rows and columns of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

5 Marginals and conditionals of a MVN

Suppose $x = (x_1, x_2)$ is jointly Gaussian with parameters

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}, \tag{9}$$

In Section 9.2, we will show that we can factorize the joint as

$$p(x_1, x_2) = p(x_2)p(x_1|x_2) \tag{10}$$
$$= \mathcal{N}(x_2|\mu_2, \Sigma_{22})\mathcal{N}(x_1|\mu_{1|2}, \Sigma_{1|2}) \tag{11}$$

where the marginal parameters for $p(x_2)$ are just gotten by extracting rows and columns for x_2, and the conditional parameters for $p(x_1|x_2)$ are given by

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \tag{12}$$
$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{13}$$

Note that the new mean is a linear function of x_2, and the new covariance is independent of x_2. Note that both the marginal and conditional distributions are themselves Gaussian: see Figure 3.

5.1 Worked example

Let us consider a 2d example. The covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{14}$$

so the conditional becomes

$$p(x_1|x_2) = \mathcal{N}\left(x_1|\mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(x_2 - \mu_2),\ \sigma_1^2 - \frac{(\rho\sigma_1\sigma_2)^2}{\sigma_2^2}\right) \tag{15}$$

We see that x_1 is a linear function of x_2. If $\sigma_1 = \sigma_2 = \sigma$, we get

$$p(x_1|x_2) = \mathcal{N}\left(x_1|\mu_1 + \rho(x_2 - \mu_2),\ \sigma^2(1 - \rho^2)\right) \tag{16}$$

If $\rho = 0$, we get

$$p(x_1|x_2) = \mathcal{N}\left(x_1|\mu_1,\ \sigma_1^2\right) \tag{17}$$

since x_2 conveys no information about x_1.

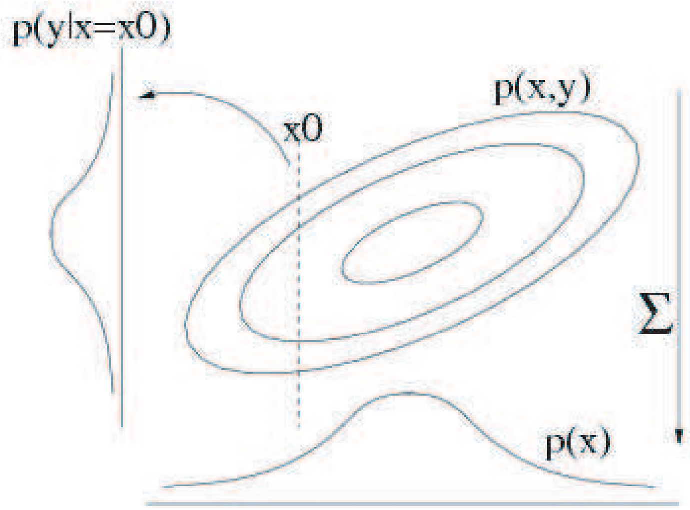

Figure 3: Marginalizing and conditionalizing a 2D Gaussian results in a 1D Gaussian. Source: Sam Roweis.

6 Bayes rule for linear Gaussian systems

Consider representing the joint distribution on X and Y in **linear Gaussian** form:

$$p(x) = \mathcal{N}(x|\mu, \Lambda^{-1}) \tag{18}$$
$$p(y|x) = \mathcal{N}(y|Ax + b, L^{-1}) \tag{19}$$

where Λ and L are precision matrices.

In Section 9.3, we show that we can invert this model as follows

$$p(y) = \mathcal{N}(y|A\mu + b, L^{-1} + A\Lambda^{-1}A^T) \tag{20}$$
$$p(x|y) = \mathcal{N}(x|\Sigma[A^T L(y - b) + \Lambda\mu], \Sigma) \tag{21}$$
$$\Sigma = (\Lambda + A^T LA)^{-1} \tag{22}$$

6.1 Worked example

Consider the following 1D example, where we try to estimate x from a noisy observation y:

$$p(x) = \mathcal{N}(x|\mu_0, \sigma_0^2) \tag{23}$$
$$p(y|x) = \mathcal{N}(y|x, \sigma^2) \tag{24}$$

Using

$$A = 1, b = 0, \Lambda^{-1} = \sigma_0^2, L^{-1} = \sigma^2 \tag{25}$$

the posterior on x is given by

$$p(x|y) = \mathcal{N}(x|\mu_n, \sigma_n^2) \tag{26}$$
$$\sigma_n^2 = \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}\right)^{-1} \tag{27}$$
$$\mu_n = \sigma_n^2 \left(\frac{y}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) \tag{28}$$

which matches our earlier result for deriving the posterior of a Gaussian mean (if we think of x as the unknown parameter μ). Also, from Equation 21, the posterior predictive density is

$$p(y) = \mathcal{N}(\mu_0, \sigma^2 + \sigma_0^2) \tag{29}$$

again matching our earlier result.

6.2 Worked example

Now suppose we have two noisy measurements of x, call them y_1 and y_2, with variances v_1 and v_2. Let the prior be $p(x) = \mathcal{N}(x|\mu_0, \sigma_0^2)$ where $\sigma_0^2 = \infty$ (an improper flat prior). We have

$$\mu = \mu_0, \Lambda^{-1} = \sigma_0^2, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, A = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, L^{-1} = \begin{pmatrix} v_1 & 0 \\ 0 & v_2 \end{pmatrix} \tag{30}$$

Applying the above formulae, and using the fact that $\Lambda = 0$, the posterior is

$$p(x|y_1, y_2) = \mathcal{N}(\mu_{x|y}, \sigma_{x|y}^2) \tag{31}$$

$$\sigma_{x|y}^2 = \Sigma = \left(\frac{1}{\sigma_0^2} + \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} v_1 & 0 \\ 0 & v_2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)^{-1} \tag{32}$$

$$= = \left(0 + (\frac{1}{v_1} + \frac{1}{v_2}) \right)^{-1} \tag{33}$$

$$\mu_{x|y} = \sigma_{x|y}^2 \left[\begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} v_1 & 0 \\ 0 & v_2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \frac{1}{\sigma_0^2} \mu \right] = \sigma_{x|y}^2 (\frac{y_1}{v_1} + \frac{y_2}{v_2}) \tag{34}$$

which matches the results we derived in HW3 by sequential updating (modulo the substitutions $y_1 = n_x \overline{x}$ and $y_2 = n_y \overline{y}$).

7 Maximum likelihood estimation

Given N iid datapoints $\mathbf{x}_i$ stored in rows of X, the log-likelihood is

$$\log p(X|\mu, \Sigma) = -\frac{Np}{2} \log(2\pi) - \frac{N}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu) \tag{35}$$

Below we drop the first term since it is a constant. Also, using the fact that

$$-\log|\Sigma| = \log|\Sigma^{-1}| \tag{36}$$

we can rewrite this as

$$\log p(X|\mu, \Sigma) = -\frac{Np}{2} \log(2\pi) \frac{N}{2} \log|\Lambda| - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \mu)^T \Lambda (\mathbf{x}_i - \mu) \tag{37}$$

where $\Lambda = \Sigma^{-1}$ is called the **precision matrix**.

7.1 Mean

Using the following results for taking derivatives wrt vectors (where $\mathbf{a}$ is a vector and A is a matrix)

$$\frac{\partial(\mathbf{a}^T \mathbf{y})}{\partial \mathbf{y}} = \mathbf{a} \tag{38}$$

$$\frac{\partial(\mathbf{y}^T A \mathbf{y})}{\partial \mathbf{y}} = (A + A^T)\mathbf{y} \tag{39}$$

and using the substitution $\mathbf{y}_i = \mathbf{x}_i - \boldsymbol{\mu}$, we have

$$\frac{\partial}{\partial \boldsymbol{\mu}}(\mathbf{x}_i - \mu)^T \Sigma^{-1}(\mathbf{x}_i - \mu) = \frac{\partial}{\partial \mathbf{y}_i}\frac{\partial \mathbf{y}_i}{\partial \boldsymbol{\mu}}\mathbf{y}_i^T \Sigma^{-1}\mathbf{y}_i \tag{40}$$

$$= -1(\Sigma^{-1} + \Sigma^{-T})\mathbf{y}_i \tag{41}$$

Hence

$$\frac{\partial}{\partial \mu}\log p(X|\mu, \Sigma) = -\frac{1}{2}\sum_{i=1}^{N} -2\Sigma^{-1}(\mathbf{x}_i - \mu) \tag{42}$$

$$= \Sigma^{-1}\sum_{i=1}^{N}(\mathbf{x}_i - \mu) = 0 \tag{43}$$

so

$$\mu_{ML} = \frac{1}{N}\sum_i \mathbf{x}_i \tag{44}$$

which is just the empirical mean.

7.2 Covariance

To compute Σ_{ML} is a little harder We will need to take derivatives wrt a matrix of a quadratic form and a determinant. We introduce the required algebra, since we will be using multivariate Gaussians a lot.

First, recall $\text{tr}(A) = \sum_i A_{ii}$ is the **trace** of a matrix (sum of the diagonal elements). This satisfies the **cyclic permutation property**

$$\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA) \tag{45}$$

We can therefore derive the **trace trick**, which reorders the scalar inner product $x^T Ax$ as follows

$$x^T Ax = \text{tr}(x^T Ax) = \text{tr}(xx^T A) \tag{46}$$

Hence the log-likelihood becomes

$$\ell(\mathcal{D}|\Lambda, \hat{\mu}) = \frac{N}{2}\log|\Lambda| - \tfrac{1}{2}\sum_i (x_i - \mu)^T \Lambda (x_i - \mu) \tag{47}$$

$$= \frac{N}{2}\log|\Lambda| - \tfrac{1}{2}\sum_i \text{tr}[(x_i - \mu)(x_i - \mu)^T \Lambda] \tag{48}$$

$$= \frac{N}{2}\log|\Lambda| - \tfrac{1}{2}\sum_i \text{tr}[S\Lambda] \tag{49}$$

where S is the **scatter matrix**

$$S \stackrel{\text{def}}{=} \sum_i (\mathbf{x}_i - \overline{x})(\mathbf{x}_i - \overline{x})^T = (\sum_i \mathbf{x}_i \mathbf{x}_i^T) - N\overline{x}\overline{x}^T \tag{50}$$

We need to take derivatives of this expression wrt Λ. We use the following results

$$\frac{\partial}{\partial A}\text{tr}(BA) = B^T \tag{51}$$

$$\frac{\partial}{\partial A}\log|A| = A^{-T} \tag{52}$$

Hence

$$\frac{\partial \ell(\mathcal{D}|\Sigma)}{\partial \Lambda} = \frac{N}{2}\Lambda^{-T} - \frac{1}{2}S^T = 0 \tag{53}$$

$$\Lambda^{-T} = \Sigma = \frac{1}{N}S \tag{54}$$

so

$$\hat{Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{x})(\mathbf{x}_i - \overline{x})^T \tag{55}$$

Note that this is only of rank N, so if $N < p$, $\hat{\Sigma}$ will be uninvertible.

In the case $p = 1$, this reduces to the standard result

$$\sigma^2_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{56}$$

In matlab, just type `Sigma = cov(X,1)`. If you use `Sigma = cov(X)`, you will get the unbiased estimate

$$\hat{\Sigma}_{unb} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{x})(\mathbf{x}_i - \overline{x})^T \tag{57}$$

N, $\sum_i \mathbf{x}_i$ and $\sum_i \mathbf{x}_i \mathbf{x}_i^T$ are called **sufficient statistics**, because if we know these, we do not need the original raw data X in order to estimate the parameters.

8 Bayesian parameter estimation

The multivariate analog of the normal inverse chi-squared (NIX) distribution is the normal inverse Wishart (NIW) (see also [GCSR04, p85]). Below, we state the results without proof. The inverse Wishart and multivariate T distributions are defined in the appendix.

8.1 Likelihood

The likelihood is

$$p(D|\mu,\Sigma) \propto |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^T \Sigma^{-1}(y_i - \mu)\right) \tag{58}$$

$$= |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2} tr(\Lambda S)\right) \tag{59}$$

$$\tag{60}$$

where S is the matrix of sum of squares (scatter matrix)

$$S = \sum_{i=1}^{N}(y_i - \overline{y})(y_i - \overline{y})^T \tag{61}$$

8.2 Prior

The natural conjugate prior is normal-inverse-wishart

$$\Sigma \sim IW(\Lambda_0^{-1}, \nu_0) \tag{62}$$

$$\mu|\Sigma \sim N(\mu_0, \Sigma/\kappa_0) \tag{63}$$

$$p(\mu,\Sigma) \stackrel{\text{def}}{=} NIW(\mu_0, \kappa_0, \Lambda_0, \nu_0) \tag{64}$$

$$\propto |\Sigma|^{-((\nu_0+d)/2+1)} \exp\left(-\frac{1}{2} tr(\Lambda_0 \Sigma^{-1}) - \frac{\kappa_0}{2}(\mu - \mu_0)^T \Sigma^{-1}(\mu - \mu_0)\right) \tag{65}$$

8.3 Posterior

The posterior is

$$p(\mu, \Sigma|D, \mu_0, \kappa_0, \Lambda_0, \nu_0) = NIW(\mu, \Sigma|\mu_n, \kappa_n, \Lambda_n, \nu_n) \tag{66}$$

$$\mu_n = \frac{\kappa_0 \mu + 0 + n\overline{y}}{\kappa_n} \tag{67}$$

$$\kappa_n = \kappa_0 + n \tag{68}$$

$$\nu_n = \nu_0 + n \tag{69}$$

$$\Lambda_n = \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0 + n}(\overline{y} - \mu_0)(\overline{y} - \mu_0)^T \tag{70}$$

The marginals are

$$\Sigma|D \sim IW(\Lambda_n^{-1}, \nu_n) \tag{71}$$

$$\mu|D = t_{\nu_n - d+1}(\mu_n, \frac{\Lambda_n}{\kappa_n(\nu_n - d + 1)}) \tag{72}$$

To see the connection with the scalar case, not that Λ_n plays the role of $\nu_n \sigma_n^2$ (posterior sum of squares), so

$$\frac{\Lambda_n}{\kappa_n(\nu_n - d + 1)} = \frac{\Lambda_n}{\kappa_n \nu_n} = \frac{\sigma^2}{\kappa_n} \tag{73}$$

8.4 Posterior predictive

$$p(x|D) = t_{\nu_n - d+1}(\mu_n, \frac{\Lambda_n(\kappa_n + 1)}{\kappa_n(\nu_n - d + 1)}) \tag{74}$$

To see the connection with the scalar case, note that

$$\frac{\Lambda_n(\kappa_n + 1)}{\kappa_n(\nu_n - d + 1)} = \frac{\Lambda_n(\kappa_n + 1)}{\kappa_n \nu_n} = \frac{\sigma^2(\kappa_n + 1)}{\kappa_n} \tag{75}$$

8.5 Marginal likelihood

$$p(D) = \frac{1}{\pi^{nd/2}} \frac{\Gamma_d(\nu_n/2)}{\Gamma_d(\nu_0/2)} \frac{|\Lambda_0|^{\nu_0/2}}{|\Lambda_n|^{\nu_n/2}} \left(\frac{\kappa_0}{\kappa_n}\right)^{d/2} \tag{76}$$

where where $\Gamma_p(a)$ is the generalized gamma function

$$\Gamma_p(\alpha) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left(\frac{2\alpha + 1 - i}{2}\right) \tag{77}$$

(So $\Gamma_1(\alpha) = \Gamma(\alpha)$.)

8.6 Reference analysis

A noninformative (Jeffrey's) prior is $p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}$ which is the limit of $\kappa_0 \rightarrow 0$, $\nu_0 \rightarrow -1$, $|\Lambda_0| \rightarrow 0$ [GCSR04, p88]. Then the posterior becomes

$$
\begin{aligned}
\mu_n &= \overline{x} && (78)\\
\kappa_n &= n && (79)\\
\nu_n &= n-1 && (80)\\
\Lambda_n &= S = \sum_i (x_i - \overline{x})(x_i - \overline{x})^T && (81)\\
p(\Sigma|D) &= IW_{n-1}(\Sigma|S) && (82)\\
p(\mu|D) &= t_{n-d}(\mu|\overline{x}, \frac{S}{n(n-d)}) && (83)\\
p(x|D) &= t_{n-d}(x|\overline{x}, \frac{S(n+1)}{n(n-d)}) && (84)
\end{aligned}
$$

Note that [Min00] argues that Jeffrey's principle says the uninformative prior should be of the form

$$\lim_{k \longrightarrow 0} \mathcal{N}(\mu|\mu_0, \Sigma/k) IW_k(\Sigma|k\Sigma) \propto |2\pi\Sigma|^{-\frac{1}{2}} |\Sigma|^{-(d+1)/2} \propto |\Sigma|^{-(\frac{d}{2}+1)} \quad (85)$$

This can be achieved by setting $\nu_0 = 0$ instead of $\nu_0 = -1$.

9 Appendix

9.1 Partitioned matrices

To derive the equations for conditioning a Gaussian, we need to know how to invert block structured matrices.

(In this section, we follow [Jor06, ch13].) Consider a general partioned matrix

$$M = \begin{pmatrix} E & F \\ G & H \end{pmatrix} \quad (86)$$

where we assume E and H are invertible. The goal is to derive an expression for M^{-1}. If we could block diagonalize M, it would be easier, since then the inverse would be a diagonal matrix of the inverse blocks. To zero out the top right we can pre-multiply as follows

$$\begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix} = \begin{pmatrix} E - FH^{-1}G & 0 \\ G & H \end{pmatrix} \quad (87)$$

Similarly, to zero out the bottom right we can post-multiply as follows

$$\begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix} \begin{pmatrix} I & 0 \\ -H^{-1}G & I \end{pmatrix} = \begin{pmatrix} E - FH^{-1}G & 0 \\ 0 & H \end{pmatrix} \quad (88)$$

The top left corner is called the **Schur complement** of M wrt H, and is denoted M/H:

$$M/H = E - FH^{-1}G \quad (89)$$

If we rewrite the above as

$$XYZ = W \quad (90)$$

where $Y = M$, we get the following expression for the determinant of a partitioned matrix:

$$
\begin{aligned}
|X||Y||Z| &= |W| && (91)\\
|M| &= |M/H||H| && (92)
\end{aligned}
$$

Also, we can derive the inverse as follows

$$Z^{-1}Y^{-1}X^{-1} = W^{-1} \tag{93}$$

$$Y^{-1} = ZW^{-1}X \tag{94}$$

hence

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -H^{-1}G & I \end{pmatrix} \begin{pmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{pmatrix} \begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix} \tag{95}$$

$$= \begin{pmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + G(M/H)^{-1}FH^{-1} \end{pmatrix} \tag{96}$$

Alternatively, we could have decomposed the matrix M in terms of E and M/E, yielding

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix}^{-1} = \begin{pmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{pmatrix} \tag{97}$$

Equating these two expression yields the following two formulae, the first of which is known as the **matrix inversion lemma** (aka **Sherman-Morrison-Woodbury formula**)

$$(E - FH^{-1}G)^{-1} = E^{-1} + E^{-1}F(H - GE^{-1}F)^{-1}GE^{-1} \tag{98}$$

$$(E - FH^{-1}G)^{-1}FH^{-1} = E^{-1}F(H - GE^{-1}F)^{-1} \tag{99}$$

In the special case that $H = -1$, $F = u$ a column vector, $G = v'$ a row vector, we get the following formula for a **rank one update of an inverse**

$$(E + uv')^{-1} = E^{-1} + E^{-1}u(-I - v'E^{-1}u)^{-1}v'E^{-1} \tag{100}$$

$$= E^{-1} - \frac{E^{-1}uv'E^{-1}}{1 + v'E^{-1}u} \tag{101}$$

9.2 Marginals and conditionals of MVNs: derivation

We can derive the results in Section 5 using the techniques for inverting partitioned matrices (see Section 9.1). Let us factor the joint $p(x_1, x_2)$ as $p(x_2)p(x_1|x_2)$ by applying Equation 95 to the matrix inverse in the exponent term.

$$\exp\left\{ -\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right\} \tag{102}$$

$$= \exp\left\{ -\tfrac{1}{2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{pmatrix} \begin{pmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right\} \tag{103}$$

$$= \exp\left\{ -\tfrac{1}{2}(x_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2))^T(\Sigma/\Sigma_{22})^{-1}(x_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)) \right\} \tag{104}$$

$$\times \exp\left\{ -\tfrac{1}{2}(x_2 - \mu_2)^T\Sigma_{22}^{-1}(x_2 - \mu_2) \right\} \tag{105}$$

This is of the form

$$\exp(\text{quadratic form in } x_1, x_2) \times \exp(\text{quadratic form in } x_2) \tag{106}$$

Using Equation 92 we can also split up the normalization constants

$$(2\pi)^{(p+q)/2}|\Sigma|^{\frac{1}{2}} = (2\pi)^{(p+q)/2}(|\Sigma/\Sigma_{22}||\Sigma_{22}|)^{\frac{1}{2}} \tag{107}$$

$$= (2\pi)^{p/2}|\Sigma/\Sigma_{22}|^{\frac{1}{2}} \ (2\pi)^{q/2}|\Sigma_{22}|^{\frac{1}{2}} \tag{108}$$

Hence we have succesfully factorized the joint as

$$p(x_1, x_2) = p(x_2)p(x_1|x_2) \tag{109}$$

$$= \mathcal{N}(x_2|\mu_2, \Sigma_{22})\mathcal{N}(x_1|\mu_{1|2}, \Sigma_{1|2}) \tag{110}$$

where the parameters of the marginal and conditional distribution can be read off from the above equations, using

$$(\Sigma/\Sigma_{22})^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{111}$$

9.3 Bayes rule for linear Gaussian systems: derivation

The following section is based on [Bis06, p93]. Consider the following joint distribution.

$$\begin{aligned} p(x) &= \mathcal{N}(x|\mu, \Lambda^{-1}) && (112) \\ p(y|x) &= \mathcal{N}(y|Ax + b, L^{-1}) && (113) \end{aligned}$$

Let $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ and consider the log of the joint:

$$\log p(\mathbf{z}) = -\frac{1}{2}(\mathbf{x} - \mu)^T \Lambda(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - A\mathbf{x} - \mathbf{b})^T L(\mathbf{y} - A\mathbf{x} - \mathbf{b}) + \text{const} \quad (114)$$

Expanding out the second order and cross terms we have

$$\begin{aligned} & -\frac{1}{2}\mathbf{x}^T(\Lambda + A^T LA)\mathbf{x} - \frac{1}{2}\mathbf{y}^T L\mathbf{y} + \frac{1}{2}\mathbf{y}^T LA\mathbf{x} + \frac{1}{2}\mathbf{x}^T A^T L\mathbf{y} && (115) \\ &= -\frac{1}{2}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^T \begin{pmatrix}\Lambda + A^T LA & -A^T L \\ -LA & L\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = -\frac{1}{2}\mathbf{z}^T R\mathbf{z} && (116) \end{aligned}$$

where the precision matrix is defined as

$$R = \begin{pmatrix}\Lambda + A^T LA & -A^T L \\ -LA & L\end{pmatrix} \quad (117)$$

The covariance of the joint is found using the matrix inversion lemma:

$$\Sigma_z = R^{-1} = \begin{pmatrix}\Lambda^{-1} & -\Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T\end{pmatrix} \quad (118)$$

The mean of the joint is given by

$$E[\mathbf{z}] = (E[\mathbf{x}], E[A\mathbf{x} + b]) = (\boldsymbol{\mu}, A\boldsymbol{\mu} + \mathbf{b}) \quad (119)$$

To compute the marginal $p(\mathbf{y})$, we use the moment form results:

$$\begin{aligned} E[\mathbf{y}] &= A\boldsymbol{\mu} + \mathbf{b} && (120) \\ \text{Cov}[\mathbf{y}] &= \Sigma_{22} = L^{-1} + A\Lambda^{-1}A^T && (121) \end{aligned}$$

To compute the conditional $p(\mathbf{x}|\mathbf{y})$ we use the canonical form results:

$$\begin{aligned} E[\mathbf{x}|\mathbf{y}] &= \Sigma_{1|2}\eta_{1|2} = \Sigma_{1|2}(\eta_1 - \Lambda_{12}(\mathbf{x}_2 - \mu_2)) && (122) \\ &= \Sigma_{1|2}(\Lambda_{11}\mu_1 + A^T L(\mathbf{y} - \mathbf{b})) && (123) \\ &= (\Lambda + A^T LA)^{-1}(A^T L(\mathbf{y} - \mathbf{b}) + \Lambda\boldsymbol{\mu}) && (124) \\ Cov[\mathbf{x}|\mathbf{y}] &= \Sigma_{1|2} = \Lambda_{1|2}^{-1} = \Lambda_{11}^{-1} = (\Lambda + A^T LA)^{-1} && (125) \end{aligned}$$

9.4 Inverse Wishart

This is the multidimensional generalization of the inverse Gamma. Consider a $d \times d$ positive definite (covariance) matrix $\mathbf{X}$ and a dof parameter $\nu > d - 1$ and psd matrix $\mathbf{S}$. Some authors (eg [GCSR04, p574]) use this parameterization:

$$IW_\nu(\mathbf{X}|\mathbf{S}^{-1}) = \left(2^{\nu d/2}\Gamma_d(\nu/2)\right)^{-1} |\mathbf{S}|^{\nu/2}|\mathbf{X}|^{-(\nu+d+1)/2} \exp\left(-\frac{1}{2}Tr(\mathbf{S}\mathbf{X}^{-1})\right) \quad (126)$$

which has mean

$$E\, X = \frac{\mathbf{S}}{\nu - d - 1} \quad (127)$$

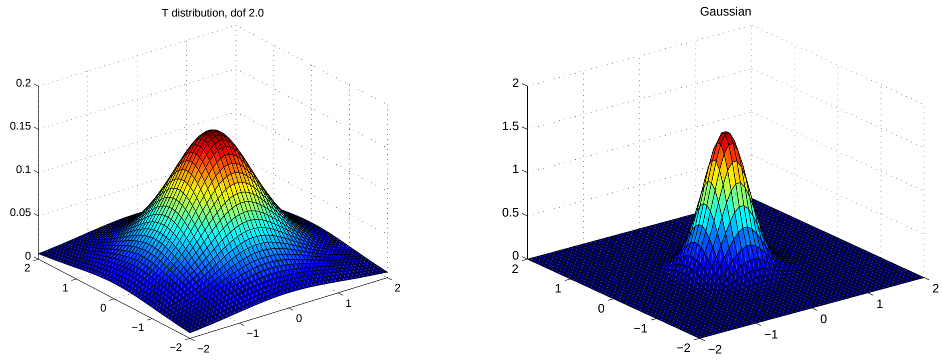

Figure 4: Left: T distribution in 2d with dof=2 and $\Sigma = 0.1 I_2$. Right: Gaussian density with $\Sigma = 0.1 I_2$ and $\mu = (0, 0)$; we see it goes to zero faster. Produced by `multivarTplot`.

In Matlab, use `iwishrnd`. In the 1d case, we have

$$\chi^{-2}(\Sigma|\nu_0, \sigma_0^2) = IW_{\nu_0}(\Sigma|(\nu_0\sigma_0^2)^{-1}) \tag{128}$$

Other authors (e.g., [Pre05, p117]) use a slightly different formulation (with $2d < \nu$)

$$IW_\nu^2(\mathbf{X}|\mathbf{Q}) = \left(2^{(\nu-d-1)d/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma((\nu-d-j)/2)\right)^{-1} \tag{129}$$

$$\times|\mathbf{Q}|^{(\nu-d-1)/2}|\mathbf{X}|^{-\nu/2}\exp\left(-\frac{1}{2}Tr(\mathbf{X}^{-1}\mathbf{Q})\right) \tag{130}$$

which has mean

$$E\,\mathbf{X} = \frac{\mathbf{Q}}{\nu - 2d - 2} \tag{131}$$

9.5 Multivariate t distributions

The multivariate T distribution in d dimensions is given by

$$t_\nu(x|\mu, \Sigma) = \frac{\Gamma(\nu/2 + d/2)}{\Gamma(\nu/2)}\frac{|\Sigma|^{-1/2}}{v^{d/2}\pi^{d/2}} \times \left[1 + \frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]^{-(\frac{\nu+d}{2})} \tag{132}$$

where Σ is called the scale matrix (since it is not exactly the covariance matrix). This has fatter tails than a Gaussian: see Figure 4. In Matlab, use `mvtpdf`.

The distribution has the following properties

$$E\,x = \mu \text{ if } \nu > 1 \tag{133}$$

$$\text{mode } x = \mu \tag{134}$$

$$\text{Cov } x = \frac{\nu}{\nu - 2}\Sigma \text{ for } \nu > 2 \tag{135}$$

(The following results are from [Koo03, p328].) Suppose $Y \sim T(\mu, \Sigma, \nu)$ and we partition the variables into 2 blocks. Then the marginals are

$$Y_i \sim T(\mu_i, \Sigma_{ii}, \nu) \tag{136}$$

and the conditionals are

$$\begin{aligned} Y_1|y_2 &\sim T(\mu_{1|2}, \Sigma_{1|2}, \nu + d_1) && (137) \\ \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2) && (138) \\ \Sigma_{1|2} &= h_{1|2}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T) && (139) \\ h_{1|2} &= \frac{1}{\nu + d_2}\left[\nu + (y_2 - \mu_2)^T\Sigma_{22}^{-1}(y_2 - \mu_2)\right] && (140) \end{aligned}$$

We can also show linear combinations of Ts are Ts:

$$Y \sim T(\mu, \Sigma, \nu) \;\; \Rightarrow \;\; AY \sim T(A\mu, A\Sigma A', \nu) \tag{141}$$

We can sample from a $y \sim T(\mu, \Sigma, \nu)$ by sampling $x \sim T(0, 1, \nu)$ and then transforming $y = \mu + R^T x$, where $R = \text{chol}(\Sigma)$, so $R^T R = \Sigma$.

References

[Bis06] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[GCSR04] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman and Hall, 2004. 2nd edition.

[Jor06] M. I. Jordan. *An Introduction to Probabilistic Graphical Models*. 2006. In preparation.

[Koo03] Gary Koop. *Bayesian econometrics*. Wiley, 2003.

[Min00] T. Minka. Inferring a Gaussian distribution. Technical report, MIT, 2000.

[Pre05] S. J. Press. *Applied multivariate analysis, using Bayesian and frequentist methods of inference*. Dover, 2005. Second edition.