1 Multivariate Normal Theory

See the notes under Supporting Materials on the course web site for much of the theory (and some that may not be so relevant to this course, but still part of the theory and relevant elsewhere). Some elaboration and additions to the theory outlined there, and ties into exploration of normal models and also simulation, are described here.

1.1 Density, Ellipses, Contours

Random quantity x is $p-$vector with a Gaussian/normal distribution, $x \sim N(m, V)$, with pdf

$$p(x) = ((2\pi)^p|V|)^{-1/2} \exp(-Q(x)/2)$$

with

$$Q(x) = (x-m)'V^{-1}(x-m).$$

The family of multivariate normals $x \sim N(m, V)$ is *elliptically symmetric*, being a function only of the quadratic form $Q(x)$ around the centroid m. Points x of constant density lie on elliptical contours (hyper-ellipses in p dimensions - simple ellipses when $p = 2$.) The shape of the density scales as the marginal variances change, but maintain elliptical shapes. The ellipses are oriented along the primary axes when V is diagonal.

Draw some density contours in the bivariate normal for various choices of (m, V).

1.2 Definition and m.g.f.s

The best definition of the multivariate normal is that based on arbitrary linear combinations being univariate normal, and it is best seen, and proven, using either moment generating functions Laplace transforms of p.d.f.s) or characteristic generating functions (Fourier transforms of p.d.f.s). For example, in terms of the moment generating function (m.g.f.):

- Univariate normal: $x \sim N(m, v)$ if and only if $E(\exp(tx)) = \exp(mt + vt^2/2)$ as a function of t, characterizing the normal distribution. This is trivially derived.
- Multivariate normal: in d dimensions, $x \sim N(m, V)$ has m.g.f. $E(\exp(t'x)) = \exp(t'm + t'Vt/2)$ as a function of the $p-$vector-valued argument t, characterizing the (multivariate)normal distribution.
- Any linear combination: scalar random quantity $y = a'x$ has m.g.f. given by

  $$E(\exp(ty)) = E(\exp((ta)'x)) = \exp(t(a'm) + (a'Va)t^2/2)$$

  so that $y \sim N(a'm, a'Va)$.
- Similar direct calculations deliver the results about the normal distributions any affine function $y = Ax + b$ and general linear forms. These are central results in much of statistics.
- The use of Laplace and Fourier transformations in statistical work is limited, but they are very important in theoretical development of probability models when interest lies in weighted averages of random quantities, and characterization of distributions, as this key example shows.

1.3 Partitioned Normal Distributions

The $p-$dimensional normal random quantity $x \sim N(m, V)$ is partitioned as

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

with each subvector x_i of dimension p_i. The mean and variance partition conformably:

$$m = \left(\begin{array}{c} m_1 \\ m_2 \end{array} \right), \qquad V = \left(\begin{array}{cc} V_1 & R \\ R' & V_2 \end{array} \right),$$

and of course $x_i \sim N(m_i, V_i)$ (by an application of the linear transformation of normals).

Assume that V is non-singular (invertible) so that each of the partitioned variance matrices is too. The *precision matrix* of x is

$$K = V^{-1}.$$

Standard linear algebraic results (or easy derivation) give the inverse of any partitioned matrix, here simplified due to the symmetry of V. The precision matrix K is partitioned conformably with V and given by

$$K = V^{-1} = \left(\begin{array}{cc} K_1 & H \\ H' & K_2 \end{array} \right),$$

with entries

- $K_1^{-1} = V_1 - RV_2^{-1}R'$,
- $H = -K_1RV_2^{-1}$,
- $K_2 = V_2^{-1} + H'K_1^{-1}H$.

1.4 Partitioned Normal Distributions: Precision Matrix

In statistical modeling and inference, a central role is played by the regression structure of conditional distributions, and these are trivially derived using standard linear algebraic expressions for the inverses of partitioned matrices. The conditional distributions (regressions) relate intimately to the structure of the precision matrix. Take the zero-mean case: $x \sim N(0, V) = N(0, K^{-1})$ with $p(x) \propto \exp(-Q(x)/2)$ where

$$Q(x) = x'Kx = x_1'K_1x_1 + 2x_1'Hx_2 + x_2'K_2x_2.$$

By inspection,

$$p(x_1|x_2) \propto \exp(-Q(x_1|x_2)/2)$$

where

$$Q(x_1|x_2) = x_1'K_1x_1 + 2x_1'Hx_2.$$

This is quadratic in x_1 (for any conditioning value of x_2) which implies the conditional normality. Center and complete the quadratic to obtain

$$\begin{aligned} Q(x_1|x_2) &= (x_1 + K_1^{-1}Hx_2)'K_1(x_1 + K_1^{-1}Hx_2) \\ &= (x_1 - RV_2^{-1}x_2)'K_1(x_1 - RV_2^{-1}x_2) \end{aligned}$$

based on the formula for H, whereupon $E(x_1|x_2) = A_1x_2$ and $V(x_1|x_2) = K_1^{-1}$ where A_1 has each of the forms $A_1 = RV_2^{-1}$ and $A_1 = -K_1^{-1}H$.

In the general case with a non-zero mean, $x \sim N(m, V)$ and then, similarly,

$$(x_1|x_2) \sim N(m_1 + A_1(x_2 - m_2), K_1^{-1}).$$

The expression $A_1 = RV_2^{-1}$ is the traditional form and shows how the regression of x_1 on x_2 is based on the covariance elements R being rotated and scaled by the variance V_2 of the conditioning variables. The second expression for A_1 relates to the elements H of the precision matrix K, and has critical uses, as follows.

1.5 Precision Matrix and Univariate Complete Conditional Distributions

An extremely important special case is when $p_1 = 1$ so that x_1 is scalar and $x_2 = x_{2:p} = x_{1:p\backslash 1}$. By extension the same theory holds for the *univariate complete conditional distribution* of any element x_i, that is $p(x_i|x_{1:p\backslash i})$, $(i = 1, \ldots, p)$.

Work now in terms of all the *univariate* elements $x = (x_1, \ldots, x_p)'$ and $m = (m_1, \ldots, m_p)'$, as well as the elements $K_{i,j}$ of the full precision matrix K $(i, j = 1, \ldots, p)$.

- The complete conditional normal distribution of x_1 has $V(x_1|x_{1:p\backslash 1}) = 1/K_{1,1}$ and

$$E(x_1|x_{1:p\backslash 1}) = m_1 - \sum_{j \in (1:p\backslash 1)} K_{1,j}(x_j - m_j)/K_{1,1}.$$

  The conditional regression of x_1 on the other variables is such that the regression coefficient on each x_j is $-K_{1,j}/K_{1,1}$.
- By extension, for any $i = 1, \ldots, p$, the complete conditional normal distribution of $(x_i|x_{1:p\backslash i})$ is normal with moments

$$E(x_i|x_{1:p\backslash i}) = m_i + \sum_{j \in (1:p\backslash i)} \gamma_{i,j}(x_j - m_j) \qquad \text{and} \qquad V(x_i|x_{1:p\backslash i}) = 1/K_{i,i}$$

  where

$$\gamma_{i,j} = -K_{i,j}/K_{i,i}.$$

This shows explicitly how the elements of the precision matrix K of the full joint distribution determine all the relevant conditional structure.

- Zeros in the precision matrix K define, and are defined by, *conditional independencies* in $p(x)$. The precision $K_{i,j} = 0$ if any only if the complete conditional distribution of x_i does not depend on x_j, equivalent to $x_i \perp\!\!\!\perp x_j$ conditional on all $x_k, k \in 1 : p \setminus (i, j)$.
- This is a key aspect of analysis of multivariate models and, in particular, underlies basic ideas in *Gaussian graphical models*.

1.6 Example: AR(1) in Noise - A Hidden Markov Model

$$\begin{aligned} y_t &= x_t + \nu_t \\ x_t &\leftarrow AR(1|(\phi, v)) \end{aligned}$$

with $\nu_t \sim N(0, w)$ and with $\nu_t \perp\!\!\!\perp \nu_s$ and $\nu_t \perp\!\!\!\perp \epsilon_s$ for all t, s.

Consider the following: suppose the parameters (ϕ, v, w) are specified and we want to explore what the data y tells us about the hidden (latent, unobservable, unknown) process x. The objective is then to compute and understand, and perhaps simulate, from distributions for sets of x values conditional on some observations from the y series. We will do this for a consecutive n time points, so we want to find and interpret

$$p(x_{1:n}|y_{1:n}).$$

This is the $n-$dimensional conditional distribution from the $2n-$dimensional joint distribution of $x_{1:n}$ and $y_{1:n}$ jointly, so we find that first and then use the above theory to condition.

For any $n > 0$,

$$\begin{aligned} y_{1:n} &= x_{1:n} + \nu_{1:n} \\ x_{1:n} &\sim N(0, \Sigma_n) \end{aligned}$$

with Σ_n as earlier derived, $\Sigma_n = s\Phi_n$ with $s = v/(1-\phi^2)$ and correlation matrix

$$\Phi_n = \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{n-1} \\ \phi & 1 & \phi & \cdots & \phi^{n-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \cdots & 1 \end{pmatrix}.$$

Also, $x_{1:n} \perp\!\!\!\perp \nu_{1:n}$ and

$$\nu_{1:n} \sim N(0, wI).$$

Notice that we are working in multivariate normals now, for vectors of length n. Let's now move to $p = 2n$ dimensions and look at the joint distribution of $x_{1:n}$ and $y_{1:n}$. Based on the linearity of y in x and the normal, independence structure, we know this is normal, so we just need the moments. Clearly $E(y_{1:n}) = 0$. Then

- $V(y_{1:n}) = V(x_{1:n} + \nu_{1:n}) = \Sigma_n + wI$, and
- $C(x_{1:n}, y_{1:n}) = E(x_{1:n} y'_{1:n})$ (since the means are zero), so that the covariance is

$$E(x_{1:n}x'_{1:n} + x_{1:n}\nu'_{1:n}) = V(x_{1:n}) + 0 = \Sigma_n.$$

Hence

$$\begin{pmatrix} x_{1:n} \\ y_{1:n} \end{pmatrix} \sim N(0, V_n) \qquad \text{with} \qquad V_n = \begin{pmatrix} \Sigma_n & \Sigma_n \\ \Sigma_n & \Sigma_n + wI \end{pmatrix}.$$

This is a very special example with a highly structured variance matrix. Apply the conditional normal theory to produce the required distribution: in the earlier notation we have $V_1 = \Sigma_n, R = R' = \Sigma_n$ and $V_2 = \Sigma_n + wI$. It follows that $A \equiv A_1 = \Sigma_n(\Sigma_n + wI)^{-1}$ and $K_1^{-1} = wA$, so that

$$(x_{1:n}|y_{1:n}) \sim N(Ay_{1:n}, wA).$$

Explore some examples, including evaluation of the conditional distribution here for some specified AR parameter values and signal-to-noise ratios s/q where $q = s + w$.

Key points to note:

- In the conditional mean $E(x_{1:n}|y_{1:n}) = Ay_{1:n}$ as a set of point estimates of the signal, each x_t is estimated by a weighted linear combination of y_t and other values, with highest weight on y_t and weights decaying away from t. This is *smoothing* - y_t is an unbiased estimate of x_t, but neighbouring y values also provide information since they are correlated with x_t.
- The conditional mean $E(x_{1:n}|y_{1:n}) = Ay_{1:n}$ as a set of point estimates of the signal is "shrunk" towards zero relative to the data point estimates (zero being the "initial" or prior point estimate under the marginal normal distribution for the signal. The *shrinkage effect* is greater for larger values of w (smaller signal-to-noise). That is, the more noise we have in the measurement error process, the smoother the estimated signal will be (the harder it is to "track" the real signal). Play with some simulations and varying parameter values to explore and appreciate this.

Notice that we have now discussed core aspects of theory, analysis and simulation for (a) inference on AR(1) model parameters when we actually observe the x process (the earlier Bayesian reference analysis and simulation of the posterior for $(\phi, v|x_{1:n}))$, and now (b) inference on the x process itself when it is latent, through the conditional $p(x_{1:n}|y_{1:n}, (\phi, v, w))$. These two components are almost all - but not quite all - that we need to move to the final stages of full analysis, inference and prediction in the HMM AR(1) model. That will include posterior simulations to generate inference on all the three parameters, (ϕ, v, w), together with the latent $x_{1:n}$. More on that later, when we have the relevant concepts and initial tools of Gibbs sampling, the first - and absolutely central - example of MCMC simulation methods.

1.7 Linear Transforms and Cholesky for Simulation

If $x \sim N(m, V)$ then $x = m + L\epsilon$ where the $p-$vector $\epsilon \sim N(0, I)$ and $LL' = V$. This holds for any *matrix square-root* L of V. The Cholesky decomposition of any (strictly) positive definite symmetric matrix V has L as a lower triangular matrix, and is compute efficiently using only the diagonal and lower (or upper) off-diagonal elements of V. The diagonal elements of L are positive. Samples of x are generated by this transformation, sampling ϵ as a vector of p independent, standard univariate normals. The correlation structure and scaling of $p(x)$ is accounted for through the Cholesky component L.

There are other matrix decompositions, including eigen-decompositions, of use in multivariate normal models and including alternative simulation methods. The Cholesky is, however, standard and numerically most efficient in cases of non-singular normal distributions.

1.8 Singular/Degenerate Normal Distributions

What if V is singular - non-negative definite? Consider the bivariate normal with unit variances and correlation 0.999' as a key example. In general cases, V is non-singular and so rank-deficient.

2 Eigenstructure of Variance Matrices

2.1 Non-singular case

$p \times p$ positive definite and symmetric matrix V.

- V has p positive eigenvalues $d_1 > \ldots > d_p > 0$. Write $D = \text{diag}(d_1, \ldots, d_p)$.
- The p eigenvectors - $p-$vectors $e_1, \ldots e_p$ - defined by $Ve_j = d_j e_j$ for $j = 1, \ldots, p$, are orthogonal: $e_j' e_j = 1$ and $e_i' e_j = 0$ for $i \neq j$. The (eigenvector column) matrix $E = [e_1, \ldots e_p]$ is orthogonal: $E'E = EE' = I$.
- From definition $VE = ED$ we have the key eigen-decomposition (or spectral decomposition):

$$V = EDE'.$$

- Principal components decomposition, also related to singular value decomposition (SVD) as we shall see later.
- $V(x) = V = V(E\epsilon)$ for any $p-$vector random quantity ϵ such that $V(\epsilon) = D$. In the normal case, take elements of ϵ as p independent normals with variances d_j. Otherwise, in non-normal cases the variance matrix analysis is the same, though the elements will generally be dependent (though uncorrelated). The transform

$$x = E\epsilon$$

is key to principal components analysis (PCA) and other statistical computations. For example, simulation of x can be done in the normal case by simulation of p independent normals in ϵ, an alternative to the Cholesky method.

- Reciprocally, $\epsilon = E'x$ has variance matrix D.
- $\epsilon_i = e_i'x$ is the i^{th} principal component transformation of x. Note that if $x \sim N(0, V)$ then $\epsilon_i \sim N(0, d_i)$.
- In normal (and other elliptically symmetric) distributions, this principal component transformation represents a rotation of the axis of the ellipses: the density $p(\epsilon)$ has the same shape as $p(x)$ but the elliptical contours are aligned with the primary axes ϵ_1, ϵ_2, etc - decorrelation is ellipse realignment.
- $Tr(V) = \sum_{i=1}^{p} d_i = d'1$ is *the total variation* under $p(x)$. Note also that $|V| = |D| = \prod_{i=1}^{p} d_i$. The total variation under $p(x)$ is the same as that under $p(\epsilon)$. The i^{th} principal component ϵ_i contributes $100d_i/(d'1)\%$ of this total variation; the first eigenvalue is the largest, so the first component is the "dominant" component, and so on. If, for example, d_1 is really large compared to the rest of the eigenvalues, then $p(x)$ is very heavily concentrated around that one-dimensional subspace; a bivariate normal with very high correlation is a simple and useful example. If a number of eigenvalues are relatively very small, then $p(x)$ is coming close to concentrating in fewer than p dimensions, consistent with V approaching singularity.
- Precision matrix $K = V^{-1} = ED^{-1}E'$, so that K has same eigenvectors as V and the reciprocal eigenvalues.

2.2 Singular case

$p \times p$ symmetric matrix V of rank $k < p$, so V^{-1} is undefined. Now V has just k positive eigenvalues and the remaining $p - k$ are zero.

- Eigenvalues are $d_1 > \ldots > d_k > d_{k+1} = 0, \ldots, d_p = 0$.
- Eigen-decomposition of V is now
$$V = EDE'$$
where E is $p \times k$ of full rank k, and has columns that are the eigenvectors of the positive eigenvalues of the $k \times k$ matrix $D = \text{diag}(d_1, \ldots, d_k)$.
- Now $E'E = I$ $(k \times k)$ as before, but the $p \times p$ matrix EE' is not the identity (it is of rank k so non-singular.)
- Only $k < p$ principal components matter: $p - k$ constraints on the elements of x are implied.
- $x = E\epsilon$ where now the action is in the $k-$ dimensions of ϵ with $V(\epsilon) = D$.
- Generalized inverse: $V^- = ED^{-1}E'$ plays the role of the precision matrix, again of rank $k < p$.
- The p.d.f. of the singular normal $x \sim N(m, V)$ is defined in terms of this generalized inverse and the reduced dimension:
$$p(x) = ((2\pi)^k|D|)^{-1/2}\exp(-Q(x)/2)$$
with
$$Q(x) = (x - m)'V^-(x - m).$$
- Standard software (Matlab, R/Splus) generally delivers full eigen-decompositions, including the zero eigenvalues and (some) eigenvectors they correspond to in the null space of V. We generally need to reduce the output to the dimension of relevance, k. Numerical instabilities creep in (quickly) in higher dimensional problems. Note also that software packages differ in how they choose to order the eigenvalues and eigenvectors - Matlab, for example, orders them in decreasing rather than increasing order.