Multivariate Normal Distribution and Maximum Likelihood Estimation

Eric Jacquier

Boston University Questrom School of Business
MF 840: Financial Econometrics

Spring 2021

Univariate Normal Distribution

Single random variable (RV) $y_i \sim N(\mu_i, \sigma_i)$ has pdf:

$$p(y_i \mid \mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-0.5(\frac{y_i - \mu_i}{\sigma_i})^2}$$

- The RV in the kernel is a χ^2
- This χ^2 is the squared distance of the random variable y_i to its mean in terms of standard deviations.
- We can use the known distribution of the χ^2 to compute the probability of getting a random variable that far (or farther) from its mean.

  Not really needed here, we can just use the properties of the univariate normal, but remember this for the multivariate case.
- Can be written: $(y_i - \mu_i)\,(\sigma_i^{-1})^2\,(y_i - \mu_i)$, a **quadratic form**

N Independent Normal Random Variables

- N-vector of independent RVs $\boldsymbol{Y} = (y_1, \ldots, y_N)$ has pdf:

$$p(\boldsymbol{Y}) = \prod_i p(y_i) \propto \frac{1}{\prod \sigma_i} e^{-0.5 \sum_i (y_i - \mu_i)\,(\sigma_i^{-1})^2\,(y_i - \mu_i)} \tag{1}$$

- Notes:
  - $\mathbf{Y}$'s covariance matrix is $D = diag\{\sigma_i^2\}$. The determinant of D is $|D| = \prod_i \sigma_i^2$
  - Let $\mu = (\mu_1, \ldots, \mu_N)$, the sum of quadratic forms in eq.(1) is:

$$(\boldsymbol{Y} - \mu)' D^{-1} (\boldsymbol{Y} - \mu)$$

- Therefore, the joint density of N independent normal RVs, with mean vector μ and diagonal variance covariance matrix D is:

$$p(\boldsymbol{Y} \mid \mu, D) \propto \frac{1}{|D|^{1/2}} e^{-0.5 (\boldsymbol{Y} - \mu)' D^{-1} (\boldsymbol{Y} - \mu)} \tag{2}$$

- Now the general case, correlated RVs with covariance matrix Σ.
  Can we just replace D in (2) by Σ?
  Yes we can.

From i.i.d. to Correlated Normal Random Variables

Random iid vector $\boldsymbol{Y} \sim N(\mu, I_N)$ has $p(\boldsymbol{Y}) \propto e^{-0.5(\boldsymbol{Y}-\mu)'\,(\boldsymbol{Y}-\mu)}$.

Want the density of N-vector $\boldsymbol{R} \sim N(\mu, \Sigma)$

Covariance matrix results needed

- Eigen value decomposition of a positive definite symmetric matrix Σ:

$$\Sigma = P\Lambda P' = P\Lambda^{0.5}\Lambda^{0.5}P' \equiv P^* P^{*\prime}$$

  P: matrix of Eigen vectors, $PP' = I$, Λ: diagonal matrix of Eigen values.

- Determinant of Σ:
  $|P^*| = |P^{*\prime}| \implies \quad |\Sigma| = |P^*|\,|P^{*\prime}| = |P^*|^2 \quad$ and: $\quad |P^*| = |\Sigma|^{0.5}$

- We also have: $\Sigma^{-1} = (P^{*\prime})^{-1}(P^*)^{-1}$ and $|(P^*)^{-1}| = |\Sigma|^{-0.5}$

Consider the linear transformation: $\boldsymbol{R} - \mu = P^*(\boldsymbol{Y} - \mu)$,

Covariance matrix of $\boldsymbol{R}$:

$$E[(R-\mu)(R-\mu)'] = E[P^*(Y-\mu)(Y-\mu)'P^{*\prime}] = P^* I P^{*\prime} = \Sigma$$

Joint density of a multivariate normal random variable

Recall $p(\boldsymbol{Y} \mid \mu, I_N) \propto e^{-0.5(\boldsymbol{Y}-\mu)'(\boldsymbol{Y}-\mu)}$.

1. Can write the joint density of $\boldsymbol{R}$ by the change of variable formula:

$$p(\boldsymbol{R}|\mu, \Sigma) = |\frac{\partial Y}{\partial R}| \; p(Y(R))$$

2. The inverse transform is $\boldsymbol{Y} - \mu = (P^*)^{-1}(\boldsymbol{R} - \mu)$.

$$\begin{aligned} (\boldsymbol{Y} - \mu)'(\boldsymbol{Y} - \mu) &= (\boldsymbol{R} - \mu)'(P^{*\prime})^{-1}(P^*)^{-1}(\boldsymbol{R} - \mu) \\ &= (\boldsymbol{R} - \mu)'\Sigma^{-1}(\boldsymbol{R} - \mu) \end{aligned} \quad (3)$$

3. The Jacobian of the inverse transform is $\partial Y / \partial R = (P^*)^{-1}$. Its determinant is:

$$|\partial Y / \partial R| = |(P^*)^{-1}| = |\Sigma|^{-0.5} \quad (4)$$

Multivariate normal density

Substitute (3) and (4), the joint density of a N-vector $\boldsymbol{R} \sim N(\mu, \Sigma)$ is

$$p(\boldsymbol{R}|\mu, \sigma) \propto \frac{1}{|\Sigma|^{0.5}} e^{-0.5(\boldsymbol{R}-\mu)'\Sigma^{-1}(\boldsymbol{R}-\mu)} \quad (5)$$

Distance of vector R to the mean vector μ

How far is a RV from its mean ... in a probabilistic sense?
Was today's set of stock returns *far* from the average day?

- Distance of a univariate RV y to its mean μ: $d_0 = \frac{y_0 - \mu}{\sigma}$

$$Pr[(\frac{y-\mu}{\sigma})^2 > d_0^2] = Pr(\chi^2(1) > d_0^2)$$

- Distance d of the vector $\boldsymbol{R}$ to its mean μ:

$$(\boldsymbol{R} - \mu)'\Sigma^{-1}(\boldsymbol{R} - \mu) = d^2$$

  The equation of an ellipse for all the $\boldsymbol{R}$s at distance d from the mean μ.
  Recall $\Sigma = P\Lambda^{1/2}\Lambda^{1/2}P'$.
  Eigen vectors define the axes of the ellipse (rotation).
  Axes lengths are proportional to one over the square roots of the Eigen values.

- The quadratic form is a $\chi^2(N)$:
  Proof: It can be written as a sum of N squared iid unit normals, see eq.(3).

- Therefore:

$$Prob[(\boldsymbol{R} - \mu)'\Sigma^{-1}(\boldsymbol{R} - \mu) > d^2] = Prob.(\chi^2(N) > d^2). \quad (6)$$

Multivariate Normal: marginal and conditional distributions 1

Consider $\boldsymbol{R} = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$

Result on the Distribution of $R_1|R_2$ and R_2

$$R_1|R_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(R_2 - \mu_2)\ ,\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})) \quad (7)$$

$$R_2 \sim N(\mu_2, \Sigma_{22}) \quad (8)$$

- If (R_1, R_2) are jointly normal, the marginal density of R_2 and the conditional density of $R_1|R_2$ are both normal.
- The conditional mean of R_1 given R_2 is a linear function of R_2.
- That is, the linear regression of R_1 on R_2 is a correctly specified functional form.
- This is a multivariate regression as R_1 is a multivariate RV.

Proof of Marginal and Conditional Distributions

- Must write $p(R)$ as $p(R_1|R_2)p(R_2)$
- Write $R-\mu = Z$ to simplify notation. Partition $\Sigma^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix}$
- The quadratic form in p(R) is:

$$Z'\Sigma^{-1}Z = Z_1'A^{11}Z_1 + 2Z_1'A^{12}Z_2 + Z_2'A^{22}Z_2 \tag{9}$$

We must now *"complete the square"*:
First two terms on RHS of (9) start a perfect square for $Z_1|Z_2$. We must find the Z_2-squared term to *complete* that square,

- Find the missing term **"?"**:

$$(Z_1 + \ ?\ Z_2)'A^{11}(Z_1 + \ ?\ Z_2)$$

  To match the cross-term in (9), we need: $? = (A^{11})^{-1}A^{12}$:

$$(Z_1 + (A^{11})^{-1}A^{12}Z_2)'A^{11}(Z_1 + (A^{11})^{-1}A^{12}Z_2)$$

- Don't forget to subtract it back.
  We added: $Z_2'A^{21}(A^{11})^{-1}A^{12}Z_2$, we subtract it back, (9) becomes:

$$(Z_1 + (A^{11})^{-1}A^{12}Z_2)'A^{11}(Z_1 + (A^{11})^{-1}A^{12}Z_2) + Z_2'(A^{22} - A^{21}(A^{11})^{-1}A^{12})Z_2$$

Proof of Marginal and Conditional Distributions

We now have $p(Z) \propto \exp -0.5\{..\}$, with in the exponential:

$$(Z_1 + (A^{11})^{-1}A^{12}Z_2)'A^{11}(Z_1 + (A^{11})^{-1}A^{12}Z_2) + Z_2'(A^{22} - A^{21}(A^{11})^{-1}A^{12})Z_2$$

That is:

$$\underset{p(Z_1|Z_2)}{e^{-0.5\{(Z_1+(A^{11})^{-1}A^{12}Z_2)'A^{11}(Z_1+(A^{11})^{-1}A^{12}Z_2)\}}} \quad \underset{p(Z_2)}{e^{-0.5\{Z_2'(A^{22}-A^{21}(A^{11})^{-1}A^{12})Z_2\}}} \tag{10}$$

- What is $E(R_1|R_2)$, $V(R_1|R_2)$, $E(R_2)$, $V(R_2)$?
- Still need to write the sub-matrices of Σ^{-1} in terms of sub-matrices of Σ
- Still need to rewrite $\frac{1}{|\Sigma|^{0.5}}$ as a function of the partitions of Σ as well

See the addendum at the end of the note on the inverse of a partitioned matrix to finish the proof.

Combinations of Multivariate Normal RVs

Consider $\boldsymbol{R} \sim N(\mu, \Sigma_N)$, a Nx1 vector ℓ, and a QxN matrix A.

1. $\ell' R$ is **one** linear combination of R and $\ell' R \sim N(\ell' \mu, \ell' \Sigma \ell)$.
2. AR is a set of Q linear combinations of R and $AR \sim N(A\mu, A\Sigma A')$.
3. As seen before, the quadratic form $(\boldsymbol{R} - \mu)' \Sigma^{-1} (\boldsymbol{R} - \mu)$ is $\chi^2(N)$.

Multivariate Normal Likelihood

- A $m \times 1$ vector of returns $\boldsymbol{R_t} \sim N(\mu, \Sigma)$ has density:

$$p(\boldsymbol{R}_t \mid \mu, \Sigma) \propto \frac{1}{|\Sigma|^{1/2}} e^{-0.5(R_t-\mu)'\Sigma^{-1}(R_t-\mu)}$$

- T i.i.d. observations of the vector $(\boldsymbol{R}_1, \ldots \boldsymbol{R_T})$. The likelihood of (μ, Σ) is:

$$\begin{aligned} \ell(\mu, \Sigma | \boldsymbol{R_1}, \ldots \boldsymbol{R_T}) &\propto \frac{1}{|\Sigma|^{T/2}} \prod_{t=1}^{T} e^{-0.5(\boldsymbol{R}_t-\mu)'\Sigma^{-1}(\boldsymbol{R}_t-\mu)} \\ &\propto \frac{1}{|\Sigma|^{\frac{T}{2}}} e^{-0.5 \sum_{t=1}^{T}(\boldsymbol{R}_t-\mu)'\Sigma^{-1}(\boldsymbol{R}_t-\mu)} \qquad (11) \end{aligned}$$

- Whether MLE or Bayesian, **sufficient statistics** are:
  Sample mean vector $\overline{\boldsymbol{R}}$
  Sample covariance matrix $\boldsymbol{S}$

$$\boldsymbol{S} = \frac{1}{T} \sum_{t=1}^{T} (\boldsymbol{R}_t - \overline{\boldsymbol{R}})(\boldsymbol{R}_t - \overline{\boldsymbol{R}})'$$

Convenient rewrite of the likelihood 1

- Use known results for the trace of a matrix:
  For a vector $\boldsymbol{x}$ and a symmetric matrix A with Eigen values λ_is:
  1. $x'Ax = tr(\boldsymbol{x}'A\boldsymbol{x}) = tr(A\boldsymbol{x}\boldsymbol{x}')$
  2. $tr(A) = \sum \lambda_i$.

  Equality (1): obvious since the quadratic form is a scalar.
  Equality (2): recall that $A = P'\Lambda P$, where $PP' = I$, and use (1).

- Now rewrite the likelihood

$$\ell(\mu, \Sigma | \boldsymbol{R_1}, \ldots \boldsymbol{R_T}) \propto \frac{1}{|\Sigma|^{T/2}} e^{-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{R}_t - \mu)'\Sigma^{-1}(\boldsymbol{R}_t - \mu)}$$

as

$$\ell(\mu, \Sigma | \boldsymbol{R_1}, \ldots \boldsymbol{R_T}) \propto \frac{1}{|\Sigma|^{T/2}} e^{-\frac{1}{2} tr\{\Sigma^{-1}\sum_{t=1}^{T}(\boldsymbol{R}_t - \mu)(\boldsymbol{R}_t - \mu)'\}}$$

Convenient rewrite of the likelihood 2

- Same as univariate mean problem, introduce the sample mean $\overline{\boldsymbol{R}}$. The quadratic form becomes

$$
\begin{aligned}
& tr\{\Sigma^{-1}\sum_{t=1}^{T}(\boldsymbol{R}_t - \overline{\boldsymbol{R}} + \overline{\boldsymbol{R}} - \mu)(\boldsymbol{R}_t - \overline{\boldsymbol{R}} + \overline{\boldsymbol{R}} - \mu)'\} \\
= \quad & tr\{\Sigma^{-1}\sum_{t=1}^{T}\left((\overline{\boldsymbol{R}} - \mu)(\overline{\boldsymbol{R}} - \mu)' + (\boldsymbol{R}_t - \overline{\boldsymbol{R}})(\boldsymbol{R}_t - \overline{\boldsymbol{R}})'\right)\} \\
= \quad & tr\{T\Sigma^{-1}(\overline{\boldsymbol{R}} - \mu)(\overline{\boldsymbol{R}} - \mu)' + \Sigma^{-1}\sum_{t=1}^{T}(\boldsymbol{R}_t - \overline{\boldsymbol{R}})(\boldsymbol{R}_t - \overline{\boldsymbol{R}})'\} \qquad (12)
\end{aligned}
$$

First equality follows because the cross-product terms are zero (as in regression and univariate mean problems).

- Now recognize the sample covariance matrix **S** in Eq. (12). The quadratic form is:

$$
tr\{T\Sigma^{-1}(\overline{\boldsymbol{R}} - \mu)(\overline{\boldsymbol{R}} - \mu)' + T\Sigma^{-1}\boldsymbol{S}\} = T(\overline{\boldsymbol{R}} - \mu)'\Sigma^{-1}(\overline{\boldsymbol{R}} - \mu) + T\ tr\{\Sigma^{-1}\boldsymbol{S}\}
$$

MLE for μ

The Likelihood function to maximize is:

$$\ell(\mu, \Sigma | \boldsymbol{R_1}, \ldots \boldsymbol{R_T}) \propto \frac{1}{|\Sigma|^{T/2}} \exp\left\{ -\frac{T}{2} \left(tr\{\Sigma^{-1}\boldsymbol{S}\} + (\overline{\boldsymbol{R}} - \mu)'\Sigma^{-1}(\overline{\boldsymbol{R}} - \mu) \right) \right\}$$

- Result: $\widehat{\mu}_{MLE} = \overline{\boldsymbol{R}}$
  - μ is only in the second quadratic form.
  - .. which is never negative because Σ is a covariance matrix.
  - Therefore $\ell(\mu, \Sigma)$ is maximized when this quadratic form is zero.

- The *concentrated* likelihood function is

$$\frac{1}{|\Sigma|^{T/2}} \exp\{-\frac{T}{2} tr(\Sigma^{-1}\boldsymbol{S})\}$$

- Now find $\widehat{\Sigma}_{MLE}$.

MLE for Σ

Concentrated Likelihood function:

$$\frac{1}{|\Sigma|^{T/2}} \exp\{-\frac{T}{2} tr(\Sigma^{-1}\boldsymbol{S})\}$$

- Use the following convenient result (Johnson and Wichern, Multivariate Statistics p. 139):
  For known positive symmetric definite matrix B of size m, and scalar b, the function
  $$f(\Sigma) = \frac{1}{|\Sigma|^b} e^{-\frac{1}{2} tr(\Sigma^{-1} B)}$$
  is maximized for $\Sigma^* = \frac{B}{2b}$.
  You can then verify that its maximized value is $\frac{1}{|B|^b}(2b)^{mb} e^{-bm}$
- Here: $b = T/2$, $B = T\boldsymbol{S}$. The likelihood is maximized at:
  $$\widehat{\Sigma}_{MLE} = \boldsymbol{S}.$$

Maximized Likelihood and Likelihood Ratio Test

- The maximized Likelihood is

$$\ell^*(\widehat{\mu}_{MLE}, \widehat{\Sigma}_{MLE}) = \frac{1}{\sqrt{2\pi}^{Tm} |\boldsymbol{S}|^{T/2}} e^{-\frac{Tm}{2}}$$

- The value of the maximized likelihood is important.
  It is used to compute the **likelihood ratio** of two nested models H_0 and H_1.
- Under the null hypothesis, the likelihood ratio $\lambda_{0/1}$ has the asymptotic distribution (no proof):

$$-2Log(\frac{\ell_0^*}{\ell_1^*}) \sim \chi^2(dof)$$

  with *dof* degrees of freedom equal to the number of restrictions in the parameters from Model H_1 to Model H_0.

Properties of $\widehat{\mu}$ and $\widehat{\Sigma}$

- For the mean:

$$\widehat{\mu} \sim N(\mu, \frac{1}{T}\Sigma)$$

- For the covariance matrix:

$$T\widehat{\Sigma}_{MLE} \sim W(\Sigma, T-1, m)$$

  a **Wishart** with $\nu = T-1$ degrees of freedom.

- The Sum-of-Squares matrix is Wishart, not the sample covariance matrix.
- It is an exact result derived from the normality of the data.
- $\widehat{\mu}$ and $\widehat{\Sigma}_{MLE}$ are independent

The Wishart Distribution

A $m \times m$ PDS random matrix A follows a Wishart distribution with ν degrees of freedom if its density is:

$$p(A|\Sigma, \nu, m) \propto |A|^{(\nu-m-1)/2} e^{-\frac{1}{2} trA\Sigma^{-1}}, \quad \nu \geq m$$

- The Wishart distribution only exists if $\nu \geq m$.
- Mean of the Wishart: E(A) $= \nu\Sigma$.
- Think of it as a multivariate generalization of $T\widehat{\sigma}^2_{MLE} \sim \sigma^2\chi^2(\nu)$

For the MLE $\widehat{\Sigma}_{MLE}$:

- $\nu = T - 1$. Need $T \geq m + 1$ for invertibility.
- $T\ E(\widehat{\Sigma}) = (T-1)\Sigma$