Recitation 5: More Multivariate Normal

1 Representations, roughly

- Distributions are typically defined by functions (such as density functions, characteristic functions, moments, etc.)
- However, we can also represent a density in terms of other random variables[1]
- If you're an algorithms buff, this is somewhat analogous to reducing a seemingly complicated problem to a problem in P and then realizing that there's a simple solution

2 Representation of Multivariate Normal

- y is Multivariate normal, or

  $$y \sim N(\mu, \Sigma)$$

  if we can write

  $$\underset{k\times 1}{y} = \underset{k\times m}{A}\ \underset{m\times 1}{z} + \underset{k\times 1}{\mu}$$

  $$z_i \sim N(0,1)\ , \quad \text{and} \quad \Sigma = AA^T$$

  Note that this means $z \sim N(0, I_{m\times m})$

  It's easy to show that this definition is equivalent to the function definitions of MVN.

- **Question:** Using its representation, find the mean and variance of y

$$\begin{aligned} E(y) &= E(Az + \mu) \\ &= AE(z) + \mu \\ &= \mu \end{aligned}$$

$$\begin{aligned} Var(y) &= Var(Az + \mu) \\ &= A\ Var(z)\ A^T + 0 \\ &= AA^T \end{aligned}$$

[1] Blitzstein and Morris (upcoming book)

Joy Yang, October 12, 2018

3 Linear Combinations of MVN

- **Question:** Are linear combinations of MVN random variables MVN themselves?

$$\begin{aligned} y^* &= By \\ &= B(Az + \mu) \\ &= BAz + B\mu \end{aligned}$$

Note that this is now in the form of a MVN random variable,

$$\begin{aligned} y^* &\sim N(B\mu, BAA^T B^T) \\ y^* &\sim N(B\mu, B\Sigma B^T) \end{aligned}$$

4 Subvectors of MVN

- Last week, Jennifer showed that subvectors of MVN are, themselves MVN

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

  implies that

$$y_1 \sim N(\mu_1, \Sigma_{11})$$

- Showing this required marginalizing out y_2

$$P(y_1) = \int P(y_1, y_2)\ dy_2$$

  This got real involved real fast. We needed to work with block matrix inverses, somewhere, we did a vector complete-the-square, and then somehow, after pages of work, the integral turned out to be 1, and the constants we pulled out of the integral turned out to specify an MVN.[2]

- **Question:** Derive the above property using the representation of y.

  Hint: Can we write y_1 as a linear combination of y?

$$y_1 = [I\ 0]\, y$$

  So, y_1 is MVN

$$\begin{aligned} y_1 &\sim N([I\ 0]\,\mu, [I\ 0]\,\Sigma\,[I\ 0]^T) \\ y_1 &\sim N(\mu_1, \Sigma_{11}) \end{aligned}$$

[2] A nice reference: http://cs229.stanford.edu/section/more_on_gaussians.pdf

5 The Uncorrelation Trick

- Last week, Jennifer also showed that conditional distributions of MVN are MVN

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

  implies that

$$y_2|y_1 \sim N(\mu^*, \Sigma^*)$$

$$\text{where} \quad \begin{aligned} \mu^* &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1) \\ \Sigma^* &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned}$$

- Showing this required using the law of conditional probabilities

$$P(y_2|y_1) = \frac{P(y_1, y_2)}{P(y_1)}$$

  Again, this got real involved real fast. We again needed to use block matrix inverses, complete-the-square, etc.

- Instead, today, we'll show this using the "uncorrelation trick"[3]:

  We can write y_2 in terms of y_1 and $y_{2\cdot 1}$, a centered MVN that's independent of y_1

$$y_2 = \underbrace{y_{2\cdot 1} + \mu_2}_{\text{independent of } y_1} + \underbrace{B(y_1 - \mu_1)}_{\text{function of } y_1}$$

  But what is B?

$$\begin{aligned} cov(y_2, y_1) &= cov(y_{2\cdot 1} + \mu_2 + B(y_1 - \mu_1), y_1) \\ cov(y_2, y_1) &= Bcov(y_1, y_1) \\ \Sigma_{21} &= B\Sigma_{11} \\ B &= \Sigma_{21}\Sigma_{11}^{-1} \end{aligned}$$

  In lecture, Stefanie mentioned a few times that getting this term is like doing linear regression. We're essentially regressing y_2 against y_1.

  Taking $\tilde{y}_1 = y_1 - \mu_1$ and $\tilde{y}_2 = y_2 - \mu_2$, we can write down a model

$$\tilde{y}_2 = B\tilde{y}_1 + \epsilon$$

  This leads us to the interpretation that $y_{2\cdot 1}$ is then the residual ϵ after linear regression

  And the least squares solution for B (see section 2 notes) is

$$\begin{aligned} B &= \tilde{y}_2\tilde{y}_1^T(\tilde{y}_1\tilde{y}_1^T)^{-1} \\ &= \Sigma_{21}\Sigma_{11}^{-1} \end{aligned}$$

[3] Blitzstein & Morris, upcoming book; other handy references: *Multivariate Analysis* (Mardia, Kent, Bibby), or Stanford Stat 306 Notes - http://statweb.stanford.edu/ lpekelis/306a/

6 Conditional Distributions

- Reminding ourselves of the uncorrelation trick:

$$y_2 = \underbrace{y_{2\cdot 1} + \mu_2}_{\text{independent of } y_1} + \underbrace{B(y_1 - \mu_1)}_{\text{function of } y_1}$$

In light of this decomposition, $y_2|y_1$ must be MVN.

- It's straightforward now to find the conditional expectation and variance

$$\begin{aligned} E(y_2|y_1) &= E(y_{2\cdot 1} + \mu_2 + B(y_1 - \mu_1) \mid y_1) \\ &= E(y_{2\cdot 1}|y_1) + \mu_2 + B(E(y_1|y_1) - \mu_1) \\ &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1) \end{aligned}$$

$$\begin{aligned} Var(y_2|y_1) &= Var(y_{2\cdot 1} + \mu_2 + B(y_1 - \mu_1) \mid y_1) \\ &= Var(y_{2\cdot 1}|y_1) + Var(\mu_2 + B(y_1 - \mu_1)|y_1) \\ &= Var(y_{2\cdot 1}) \end{aligned}$$

So now, in order to find $Var(y_2|y_1)$, we just need to find $Var(y_{2\cdot 1})$

$$\begin{aligned} Var(y_2) &= Var(y_{2\cdot 1} + \mu_2 + B(y_1 - \mu_1)) \\ Var(y_2) &= Var(y_{2\cdot 1}) + BVar(y_1 1)B^T \\ \Sigma_{22} &= Var(y_{2\cdot 1}) + \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{11}\Sigma_{11}^{-1}\Sigma_{12} \\ Var(y_{2\cdot 1}) &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned}$$

And so plugging this back in

$$Var(y_2|y_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

7 Sherman-Morrison-Woodbury Formula

- This might come in handy for your problem sets depending on what route you choose to go for the first question
- The formula:

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

- Applications: Recursive least squares, low rank decompositions, etc. Here's a very short application, just to show how it can be useful.
- **Example:** Matrix perturbations

  Suppose the inverse of a matrix A is known.

  We perturb each element to get a new matrix B

$$\begin{aligned} B_{ij} &= A_{ij} + \Delta x_i \Delta y_i \\ B &= A + xy^T \end{aligned}$$

  As a simple, concrete example, perhaps

$$x = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{then the perturbation matrix is} \quad y = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

  If we want to invert B, we don't need to re-compute the entire matrix inverse. Instead, just plug into the Woodbury formula:

$$B^{-1} = A^{-1} - A^{-1}x(1 + y^T A^{-1}x)^{-1}y^T A^{-1}$$

  Notice, $(1 + y^T A^{-1}x)$ is a scalar (in this case, it's simply 1), so that's easy to invert.