INFSCI 2595

Fall 2019

Information Sciences Building: Room 403

Lecture 06

In lecture 05, we made use of the Multivariate Normal (MVN) distribution

- In this lecture, we will introduce fitting the MVN.

- Afterwards, we will continue working with multivariate distributions, but introduce the multivariate analog to the binomial distribution -> the Multinomial distribution.

- We will conclude by discussing non-parametric density estimation methods.

MVN density for D variables, $\mathbf{x} = \{x_1, \dots, x_D\}$

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

The likelihood function for a single observation of the D variables is proportional to:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-1/2} \cdot \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

How do we write the likelihood for N observations?

- When we were considering a single variable, x, we denoted the N observations as a vector $\mathbf{x} = \{x_1, \ldots, x_n, \ldots x_N\}$.

- Build off of this idea to organize the D elements of our multivariate vector $\mathbf{x} = \{x_1, \ldots, x_d, \ldots, x_D\}$.

Consider the D variables separately…

- We can write out a vector of N observations for each the D variables.

- The n-th observation of the d-th variable: $x_{n,d}$

- The N element vectors for the separate variables can be written as:

$$\mathbf{x}_{:,1} = \{x_{1,1}, x_{2,1}, \ldots, x_{n,1}, \ldots, x_{N,1}\}$$

Consider the D variables separately…

- We can write out a vector of N observations for each the D variables.

- The n-th observation of the d-th variable: $x_{n,d}$

- The N element vectors for the separate variables can be written as:

$$\mathbf{x}_{:,d} = \left\{ x_{1,d}, x_{2,d}, \ldots, x_{n,d}, \ldots, x_{N,d} \right\}$$

In general, for the d-th variable.

Consider the D variables separately...

- We can write out a vector of N observations for each the D variables.
- The n-th observation of the d-th variable: $x_{n,d}$
- The N element vectors for the separate variables can be written as:

$$\mathbf{x}_{\boxed{:},d} = \{x_{1,d}, x_{2,d}, \dots, x_{n,d}, \dots, x_{N,d}\}$$

MATLAB-like notation to represent **ALL** of the observations

In general, for the d-th variable.

Alternatively, consider the N observations separately…

- Write a vector of D variables for each of the N observations.

- Continue to use the notation: $x_{n,d}$

- The D element vector for each observation:

$$\mathbf{x}_{1,:} = \left\{x_{1,1}, x_{1,2}, \dots, x_{1,d}, \dots, x_{1,D}\right\}$$

Alternatively, consider the N observations separately…

- Write a vector of D variables for each of the N observations.

- Continue to use the notation: $x_{n,d}$

- The D element vector for each observation:

$$\mathbf{x}_{n,:} = \left\{x_{n,1}, x_{n,2}, \dots, x_{n,d}, \dots, x_{n,D}\right\}$$

In general, for the n-th observation.

Regardless of how we write out the variables, we can organize all observations together into a matrix

- Thus, the $N \times D$ matrix $\mathbf{X}$ can be viewed two different ways.

- "Stacking" the N **row-vectors** on top of each other.

- "Binding" the D **column-vectors** side-by-side.

The two styles are equivalent!

"Stacking" rows together

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{1,:} \\ \mathbf{x}_{2,:} \\ \vdots \\ \mathbf{x}_{N,:} \end{bmatrix}$$

"Binding" columns together

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{:,1} & \mathbf{x}_{:,2} & \cdots & \mathbf{x}_{:,D} \end{bmatrix}$$

Now, for N observations of the D variables

- Assume the observations are *conditionally independent* given the MVN parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

- We can factor the "complete" joint distribution into the product of N separate multivariate likelihoods.

$$p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{n=1}^{N} \left\{ p\left(\mathbf{x}_{n,:}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right) \right\}$$

The likelihood is proportional to:

$$p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-N/2} \cdot \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\left[\left(\mathbf{x}_{n,:}^{T} - \boldsymbol{\mu}\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_{n,:}^{T} - \boldsymbol{\mu}\right)\right]\right\}$$

The likelihood is proportional to:

$$p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-N/2} \cdot \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\left[\left(\boxed{\mathbf{x}_{n,:}^{T}} - \boldsymbol{\mu}\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\boxed{\mathbf{x}_{n,:}^{T}} - \boldsymbol{\mu}\right)\right]\right\}$$

$\boldsymbol{\mu}$ is structured as a column-vector.

Since we specified $\mathbf{x}_{n,:}$ as a row-vector, needed to transpose it to have the correct format above.

Be careful about the dataset structure across textbooks!!!!!

How can we fit the MVN given $\mathbf{X}$?

- We will focus on the case where the covariance matrix, $\mathbf{\Sigma}$, is known.

- This is analogous to the fitting the univariate normal model with unknown mean and known variance!

Proceed with a Bayesian formulation.

- In the univariate case, we saw that the conjugate prior to the normal likelihood is a normal distribution.

- The same holds for the multivariate case!

- The conjugate prior for $\boldsymbol{\mu}$ with a multivariate likelihood is a multivariate normal!

Conjugate prior

- The multivariate normal prior distribution on $\boldsymbol{\mu}$ will be specified as:

$$p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$$

- The hyperparameter $\boldsymbol{\mu}_0$ is the D-dimensional vector of prior means.

- The hyperparameter $\boldsymbol{\Lambda}_0$ is a $D \times D$ prior covariance matrix.

Joint posterior on all D unknown means

$$p(\boldsymbol{\mu}|\mathbf{X}, \boldsymbol{\Sigma}) \propto p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$$

Joint posterior on all D unknown means

$$p(\boldsymbol{\mu}|\mathbf{X}, \boldsymbol{\Sigma}) \propto p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$$

$$p(\boldsymbol{\mu}|\mathbf{X}, \boldsymbol{\Sigma}) \propto \prod_{n=1}^{N} \left\{ \mathcal{N}\left(\mathbf{x}_{n,:}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \right\} \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$$

Joint posterior on all D unknown means

Substitute in the quadratic terms within the exponential

$$\propto \exp\left\{-\frac{1}{2}\left(\sum_{n=1}^{N}\left[\left(\mathbf{x}_{n,:}^{T}-\boldsymbol{\mu}\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_{n,:}^{T}-\boldsymbol{\mu}\right)\right]+(\boldsymbol{\mu}-\boldsymbol{\mu}_{0})^{T}\boldsymbol{\Lambda}_{0}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_{0})\right)\right\}$$

This should look familiar…

- The posterior distribution on $\boldsymbol{\mu}$ is a MVN distribution.

$$p(\boldsymbol{\mu}|\mathbf{X}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_N, \boldsymbol{\Lambda}_N)$$

- As with the univariate case the posterior mean is a **precision weighted average** between the prior mean and the sample average.

The precision is now represented by the precision matrix

- The precision matrix is the inverse of the covariance matrix.

$$\mathbf{\Lambda}_N^{-1} = \mathbf{\Lambda}_0^{-1} + N\mathbf{\Sigma}^{-1}$$

- The posterior precision is still the sum of the prior and the data precision!

To calculate the posterior mean, we need multivariate sample average.

- Each variable's sample average can be computed without regard to the other variables.

$$\bar{x}_d = \frac{1}{N} \sum_{n=1}^{N} x_{n,d}$$

- The D-dimensional sample average vector is then: $\bar{\mathbf{x}}$

Posterior mean $\boldsymbol{\mu}_N$

- The multivariate precision weighted average:

$$\boldsymbol{\mu}_N = [\boldsymbol{\Lambda}_0^{-1} + N\boldsymbol{\Sigma}^{-1}]^{-1}[\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0 + N\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}]$$

The asymptotic trends we discussed for the univariate case are still valid!

- In the limit of infinite prior uncertainty, $\left|\Lambda_0^{-1}\right| \rightarrow 0$, the posterior distribution on the unknown means converges to:

$$\boxed{\boldsymbol{\mu} | \mathbf{X}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\bar{\mathbf{x}}, \frac{1}{N} \boldsymbol{\Sigma}\right)}$$

With one important caveat: $N \geq D$!!!

What about when Σ is also unknown?

- Even the conjugate analysis is rather involved.

- Introduces the Inverse-Wishart distribution as the conjugate prior on the covariance matrix.

- We will not go through this analysis presently.

- Please see PRML Section 2.3.4 for the MLE derivation and discussion.

Side note…what about the standard normal?

- We had seen how in the univariate case a general Gaussian:

$$x|\mu,\sigma \sim \mathrm{normal}(x|\mu,\sigma)$$

- Can be equivalently defined through the reparameterization:

$$z \sim \mathrm{normal}(z|0,1)$$
$$x = \sigma \cdot z + \mu$$

We can equivalently define a general MVN distribution through independent standard normals!

- Need to make use of the following reparameterization:

$$\mathbf{x}_{n,:}^{T} = \mathbf{L}\mathbf{z}_{n,:}^{T} + \boldsymbol{\mu}$$

$$z_{n,d} \sim \text{normal}\left(z_{n,d} | 0,1\right), d = 1, \ldots, D$$

$$\mathbf{L}\mathbf{L}^{T} = \boldsymbol{\Sigma}$$

We can equivalently define a general MVN distribution through independent standard normals!

- Need to make use of the following reparameterization:

$$\mathbf{x}_{n,:}^T = \mathbf{L}\mathbf{z}_{n,:}^T + \boldsymbol{\mu}$$

$$z_{n,d} \sim \text{normal}\left(z_{n,d} | 0,1\right), d = 1, \dots, D$$

The matrix $\mathbf{L}$ a lower triangular matrix and is known as the **Cholesky decomposition**…and represents a "matrix square root"

$$\boxed{\mathbf{L}\mathbf{L}^T} = \boldsymbol{\Sigma}$$

Multinomial distribution

The multivariate normal extends the Gaussian to higher dimensions

- Analogously, the <u>**Multinomial distribution**</u> extends the Binomial distribution to higher dimensions!

- But, how does the dimensionality increase for a discrete variable?

Number of states

- The Binomial distribution is associated with **BINARY** outcomes.
- The variable can take 2 possible states, $x \in \{0,1\}$
- With a multinomial distribution, we are dealing with a random variable that can take on **MORE** than 2 states!

Number of states

- With a multinomial distribution, we are dealing with a random variable that can take on **MORE** than 2 states!
- Examples:
  - Canonical example – rolling a 6 sided die
  - Voting with more than 2 political parties

1-of-K encoding

- Denote the total number of states as K.

- The random variable is represented as a K-dimensional vector.

$$\mathbf{x} = \{x_1, x_2, \dots, x_k, \dots, x_K\}$$

- The observed state is then assigned a value of 1: $x_k = 1$
- All other states are set to 0

For example, if we roll a 4 from a 6 sided die

- The 6 possible states (1 through 6) are encoded as:

$$\mathbf{x} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$

- If we observe a 4 the elements in the vector take on the values:

$$\mathbf{x} = \{x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 0, x_6 = 0\}$$

For example, if we roll a 4 from a 6 sided die

- The 6 possible states (1 through 6) are encoded as:

$$\mathbf{x} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$

- If we observe a 4 the elements in the vector take on the values:

$$\mathbf{x} = \{0, 0, 0, 1, 0, 0\}$$

Define the probability $x_k = 1$ as μ_k

- The distribution of $\mathbf{x}$ is therefore:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

- Where $\boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots, \mu_k, \ldots, \mu_K\}$ is the vector of probabilities for each state.

Now consider observing N <u>independent</u> observations of the random variable

- Similar to the Multivariate normal we can organize the observation of the K states in a matrix, $\mathbf{X}$.

- The n-th observation of the k-th state, $x_{n,k}$, will be 0 or 1.

The likelihood of $\mathbf{X}$ given $\boldsymbol{\mu}$ can be factored into the product of N separate likelihoods

$$p(\mathbf{X}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\left\{p\left(\mathbf{x}_{n,:}^{T}|\boldsymbol{\mu}\right)\right\} = \prod_{n=1}^{N}\left\{\prod_{k=1}^{K}\mu_{k}^{x_{n,k}}\right\}$$

The likelihood can be rearranged as

$$\prod_{n=1}^{N}\left\{\prod_{k=1}^{K}\mu_k^{x_{n,k}}\right\} = \prod_{k=1}^{K}\mu_k^{x_{1,k}} \times \mu_k^{x_{2,k}} \times \cdots \times \mu_k^{x_{n,k}} \times \cdots \times \mu_k^{x_{N,k}}$$

The likelihood can be rearranged as

$$\prod_{n=1}^{N}\left\{\prod_{k=1}^{K}\mu_k^{x_{n,k}}\right\} = \prod_{k=1}^{K}\underbrace{\mu_k^{x_{1,k}} \times \mu_k^{x_{2,k}} \times \cdots \times \mu_k^{x_{n,k}} \times \cdots \times \mu_k^{x_{N,k}}}$$

$$\prod_{n=1}^{N}\left\{\prod_{k=1}^{K}\mu_k^{x_{n,k}}\right\} = \prod_{k=1}^{K}\mu_k^{\left(\sum_{n=1}^{N} x_{n,k}\right)}$$

Sufficient statistics…are just counting!

- Define the number of times $x_k = 1$ as:

$$m_k = \sum_{n=1}^{N} x_{n,k}$$

The likelihood of the observations given the state probabilities is therefore:

$$p(\mathbf{X}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \left\{ p\left(\mathbf{x}_{n,:}^{T} | \boldsymbol{\mu}\right) \right\} = \prod_{k=1}^{K} \mu_k^{m_k}$$

What are we still missing…remember how we went from the Bernoulli to the Binomial for the binary outcome case?

- Just as we saw with the binary outcome situation, there are multiple potential sequences for observing exactly m_k counts out of N trials.

- Therefore, we need to account for the number of ways of partitioning N objects into K groups of size $m_1, m_2, \dots, m_K$.

The multinomial distribution

$$p(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \cdots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

Without deriving the MLE on $\boldsymbol{\mu}$ can you guess what it is?

HINT: The basic definition of probability…

The MLE on the vector probabilities per state

$$\hat{\boldsymbol{\mu}} = \{\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_K\} = \left\{\frac{m_1}{N}, \frac{m_2}{N}, \dots, \frac{m_K}{N}\right\}$$

Bayesian formulation – prior specification

- We saw in the Binary case, that the conjugate prior for the Binomial likelihood is the Beta distribution.

- Since the Multinomial is a multivariate generalization of the Binomial, we can expect that the corresponding conjugate prior is a multivariate generalization of the Beta…

Bayesian formulation – prior specification

- We saw in the Binary case, that the conjugate prior for the Binomial likelihood is the Beta distribution.

- Since the Multinomial is a multivariate generalization of the Binomial, we can expect that the corresponding conjugate prior is a multivariate generalization of the Beta…

Dirichlet distribution

The Dirichlet distribution

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

The Dirichlet distribution…is confined to a <u>simplex</u>

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

The simplex results from the summation constraint on the state probabilities: $\sum_k \mu_k = 1$

The posterior distribution on μ is…

The posterior distribution on μ is…a Dirichlet!

- Define the vector $\mathbf{m} = \{m_1, m_2, \ldots, m_K\}$

$$p(\boldsymbol{\mu}|\mathbf{m}, N, \boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

Interpretation of the $\boldsymbol{\alpha}$ hyperparameter

- Each α_k is added to the number of times we saw $x_k = 1, m_k$.

- Thus, each α_k is the a-priori effective number of times we saw each state!

Non-parametric density estimation

Histograms

- How many bins should we use?
- https://shiny.rstudio.com/gallery/faithful.html
- Try out different numbers of bins…does our interpretation change?

Kernel density estimation

- Can we smooth out bumps or discontinuities?

- Kernel smoothing!

- Try out the kernel density estimate on the faithful histogram app.

In class example