Machine Learning and Pattern Recognition - Week 2 - Model Evaluation & Gaussians

Antonio León Villares

September 2022

Contents

Based on the online notes here.

1 Model Evaluation

1.1 Baseline Models

- **What is a baseline model?**
  - a model used to compare a new model in development
  - it can be indicative that the problem is **too hard** (if a fancy model can't generalise well) or that there are bugs in the code
- **What types of baselines are typically used?**
  - **"dummy" baselines**:
    - allow us to verify that our model works as expected
    - for instance:

      $$f(\underline{x}) = b$$

    - if our new model performs worse than this simple model, there is probably an issue with our code
  - **state-of-the-art**
    - typically used when proposing a new method
    - can look at papers, see which baselines they use and use them yourself

1.2 Test Sets & Generalisation Error

- **What is a test set?**
  - data not seen by the model
  - used to report the error that the model should attain when applied to **new data**
- **Why is the test set important?**
  - if we train a bunch of models, typically the most complicated one will get greater training accuracy, since it is more powerful at learning
  - for example:

    $$f(\underline{x}; \underline{w}, b) = \underline{w}^T \underline{x} + b$$

    will always outperform:

    $$f(\underline{x}; b) = b$$

    since the models are **nested**: anything produced by $f(\underline{x}; b)$ can be produced by $f(\underline{x}; \underline{w}, b)$
  - hence, training performance is not a good indicative of how well our model will perform with new data, so can't be used to select the "best" model
- **What is generalisation error?**
  - the **average error** that a model would achieve on **future** test cases produced by a distribution $p(\underline{x}, y)$:

    $$E_{gen} = \mathbb{E}_{p(\underline{x},y)}[L(y, f(\underline{x})]$$

  - the loss L will depend on the task of interest

  - depending on the data:

$$\mathbb{E}_{p(\underline{x},y)}[L(y, f(\underline{x})] = \int L(y, f(\underline{x}))p(\underline{x}, y)d\underline{x}dy$$

$$\mathbb{E}_{p(\underline{x},y)}[L(y, f(\underline{x})] = \sum_{\underline{x},y} L(y, f(\underline{x})p(\underline{x}, y)d\underline{x}dy$$

- **Why can't we compute the generalisation error?**
  - we would need to know $p(\underline{x}, y)$
  - but if we knew the distribution of future events, we wouldn't need a model in the first place
- **What is the Monte Carlo method?**
  - repeatedly sampling from a **distribution**, hoping that the sample is **representative** of the distribution, and so, allows us to predict is parameters
- **What do we use to approximate the generalisation error?**
  - if we **assume** that the **test set** contains M samples taken from $p(\underline{x}, y)$, then we can use a **Monte Carlo** estimate:

$$E_{test} = \frac{1}{M}\sum_{m=1}^{M} L(y^{(m)}, f(\underline{x}^{(m)})), \qquad \underline{x}^{(m)}, y^{(m)} \sim p(\underline{x}, y)$$

  - E_{test} provides an **unbiased** estimate of E_{gen}:

$$E_{gen} \approx E_{test}$$

    since **on average** the estimate is correct:

$$\mathbb{E}_{p(\underline{x},y)}[E_{test}] = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{p(\underline{x},y)}[L(y, f(\underline{x})] = \frac{ME_{gen}}{M} = E_{gen}$$

  - again, this is contingent on the data being **representative** of **future data**

1.3 Validation Data

- **What is a validation set?**
  - a set used to **validate** model hyperparameters
  - when developing a model, there are many parameters that need **tweaking** (i.e regularisation, types of basis function, etc ...)
  - we can train all these different models on a **training set**, and then use the **validation set** to select the most performant
- **Why can't we use the test set for selecting parameters?**
  - because then the parameters would be selected to **fit** to the **test data**
  - however, test data should **only** be used for **evaluating** a model, and should **never** be seen by the model
  - we use **validation** to select the "best" parameter, and **test** for seeing performance on future data

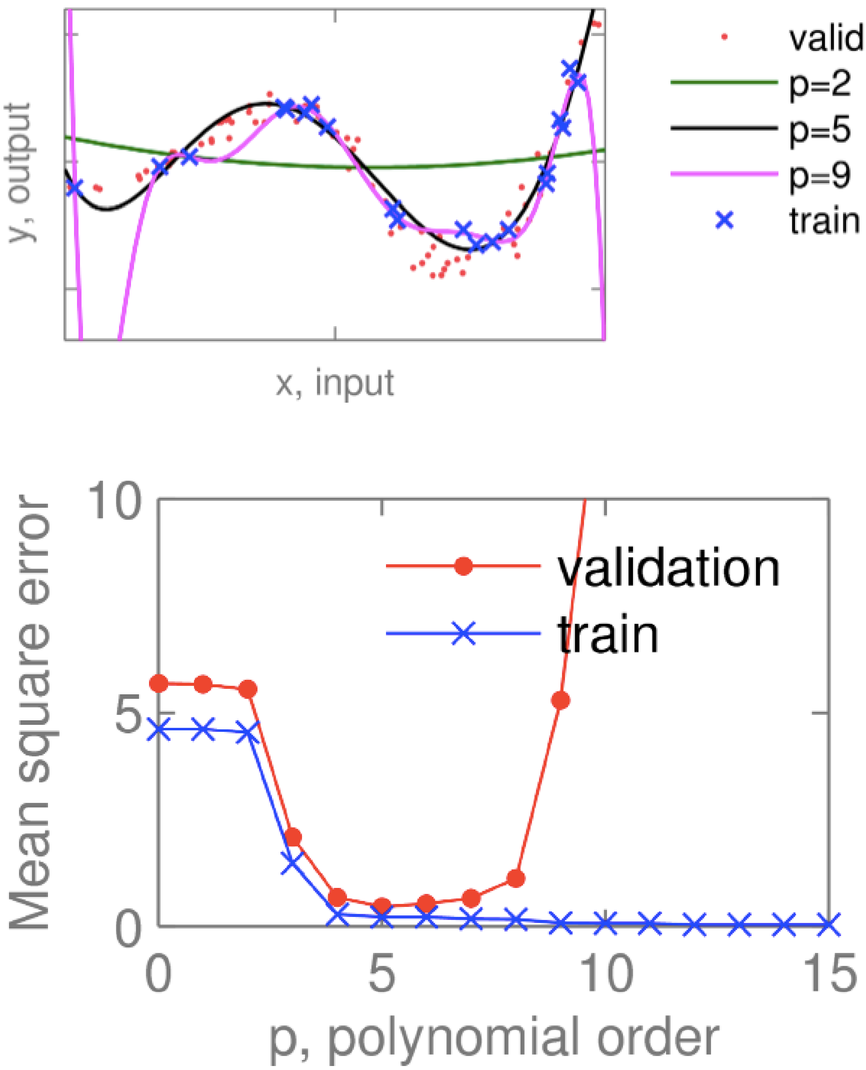


Figure 1: As expected, increasing the **order** of the polynomial monotonically decreases the **training error** (since the models are nested, so higher order polynomials can always attain **at least** the same performance as lower order ones). However, this is due to overfitting: on completely new data (**validation**), the validation error quicly increases. For this data, we can see that polynomials with orders between 4 and 7 will perform decently.

- **How is validation data selected for time series?**
  - validation is selected **after** the data used for training
  - testing is selected **after** the data used for validation

1.4 Cross Validation

- **What is K-Fold Cross Validation?**
  - a method for **validation**, particularly useful when there is little data available
  - splits the data into K "folds"
  - then, uses $K - 1$ folds for training, and the Kth fold for validation

  - repeat, until all the folds have been used for validation once
  - the **validation error** will be the average of the validation errors for each fold

- **What are the issues with K-Fold Cross Validation?**
  - **expensive**: ultimately we have to retrain a new model K times, which is very expensive (especially if we use validation to pick a parameter)
  - **statistics**: hard to make any statistically rigorous statements about performance

1.5 Warning: Test Data

- **How should data be split?**
  - no "percentages" for train, validation and test
  - ideally, want as much training as possible to fit a good model

- **Why is it so important that a model never sees testing data?**
  - if we use the test data in **any** way to change the model, and thus improve testing performance, we are **fitting** our model to the test
  - thus our reference for "goodness of fit" gets lost
  - as a practical example:

> *Someone may have followed good practice for all of their analysis, but then the final test score is disappointing. They then realize that there was something they should probably have done differently, so they change that and try again. Then they have another realization, but after that change the test score gets worse, so they revert that change... and so on.*
>
> *Each minor re-run of a method, or peek at the test set, doesn't seem like it could cause any problems individually. But the effects build up. These problems with accidental overfitting are frequently seen on Kaggle. Their competitions display a public leaderboard, based on a test set, but the final rankings are based on a second test set. It is common for some competitors to fall many places when the leaderboard is re-ranked. One such competitor (Greg Park) wrote a reflective blog post on how they had fooled themselves. Despite knowing about cross-validation and the dangers of overfitting, they slowly but surely slipped into fitting the test set. They reached second place on the public leaderboard, but fell dramatically when it was re-ranked... embarrassingly beneath one of the available baselines.*

- **What are the limitations of test errors?**
  - whilst test errors are typically good at comparing model performance, they aren't fully reliable
  - for example, data from the future can **change** (i.e measured with different equipment, people, location, etc...)
  - ultimately, test errors are only reflective of how good a model is **under the given distribution**: if used for some other data, it might generalise poorly

- **How can online prediction reduce these limitations?**
  - typically training is performed using a bunch of data all at once

- with **online prediction**, we constantly feed the model new data, and it uses this “stream” to update its parameters

2 further notes on model development.

1. ***Interpreting Weights****: weights aren’t always meaningful, and we shouldn’t always seek to assign a meaning (for example, this paper discusses how, according to weights, Asthma should reduce health-related risks)*

2. ***Type of Testing****: how we test data depends on the task at hand. For example, this paper argues that AI still hasn’t caught up to human radiologists when screening for breast cancer. They argue that the testing data used may have inflated the results, since the data is taken from a subset of hospitals, in specific locations. They argue that this doesn’t make the AI “general”, since screenings from other hospitals might not get the same performance. Moreover, they argue that even if data has kept up with time (a model trained x years ago is still good for current screenings) this might not be indicative of generalisation, but rather a phenomenon caused by the fact that it is likely that the same people get screenings across times, so again, the AI isn’t generalising. They argue that we would need new data, from new locations and new patients. This just illustrates how we should* ***always*** *be sceptical of our models before sending them out to production, especially for important cases like these.*

2 Gaussians

2.1 Univariate Gaussians

2.1.1 The Standard, Univariate Gaussian

- **What is the standard, univariate normal distribution?**
  - a **probability distribution**:
  $$\mathcal{N}(\mu = 0, \sigma^2 = 1)$$
  - μ is the **mean** of the distribution
  - σ^2 is the **variance** of the distribution
  - if we **sample** from the distribution and plot a histogram we obtain a **bell-shaped** distribution, centered at 0, and with points of inflection at 1:

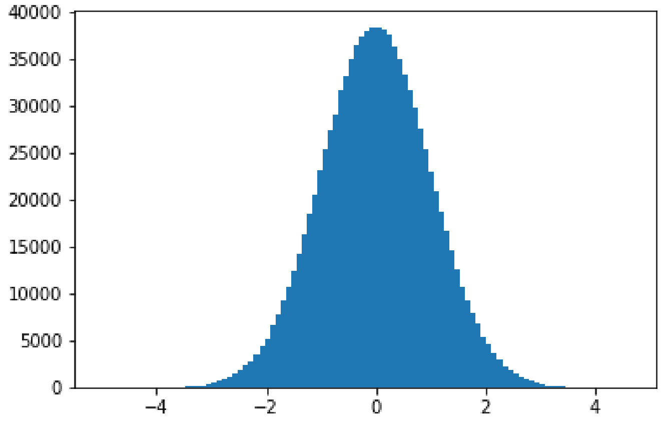


Figure 2: The above distribution has mean -0.00026 and variance 0.99802 (using 10^6 samples). If $Z \sim \mathcal{N}(0,1)$, we expect that as the number of samples goes to ∞, the mean and variance of the samples tends to $\mu = 0, \sigma^2 = 1$. Formally:

$$\mathbb{E}[Z] = \int zp(z)dz = 0, \qquad Z \sim \mathcal{N}(0,1)$$

$$Var[Z] = \mathbb{E}[(Z-\mu)^2] = \int z^2 p(z)dz = 1, \qquad X \sim \mathcal{N}(0,1)$$

where $p(x)$ is the **probability density function** of the normal distribution.

- **What is the probability density function of a standard normal distribution?**
  - a function describing the **bell-shaped** curve of the normal distribution:

$$p(z) = \mathcal{N}(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

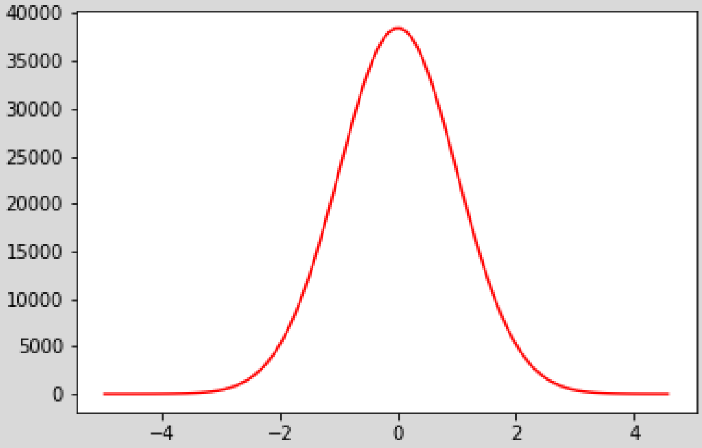
  - $p(z)$ does **not** give the probability of z under a **normal distribution**
  - PDFs are used to compute probabilities of the form:

$$P(Z \leq z) = \int_{-\infty}^{z} p(u)du$$

Whilst $p(z)$ doesn't give a probability, we can assume that as $\delta \to 0$, the probability of a sample $z \in \mathcal{N}(0,1)$ being in the range $\left[x - \frac{\delta}{2}, x + \frac{\delta}{2}\right]$ is close to:

$$p(x)\delta$$

(this is an approximation for the area under the curve in this range)
If we have generated N samples, we expect $p(x)\delta N$ of those samples to land within the bin $\left[x - \frac{\delta}{2}, x + \frac{\delta}{2}\right]$.
We can use this procedure to approximate the PDF:



We expect $\approx 68\%$ of the data to be in the range $[-\sigma, \sigma]$, and $\approx 95\%$ of the data to be in the range $[-2\sigma, 2\sigma]$.

2.1.2 General Univariate Gaussians

- **How can we compute a normal distribution with arbitrary parameters?**
  - say we want to define:
  $$x \sim \mathcal{N}(\mu, \sigma^2)$$
  - if $z \sim \mathcal{N}(0,1)$, then:
  $$x = z\sigma + \mu$$
  is such that:
  $$x \sim \mathcal{N}(\mu, \sigma^2)$$
  - in the other direction, if $x \sim \mathcal{N}(\mu, \sigma^2)$, then:
  $$z = \frac{x - \mu}{\sigma}$$
  is such that:
  $$z \sim \mathcal{N}(0,1)$$
- **What is the PDF of the general univariate Gaussian?**

- this is a matter of **shifting** and **scaling** the standard univariate Gaussian
- to go from **general** to **standard**, we apply the transformation:

$$x \mapsto \frac{x-\mu}{\sigma}$$

  so:

$$p(x) \propto e^{-\frac{\left(\frac{x-\mu}{\sigma}\right)^2}{2}} = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- however, this **increases** the area under the curve, since the curve has widened by a factor of σ; thus, we need to scale:

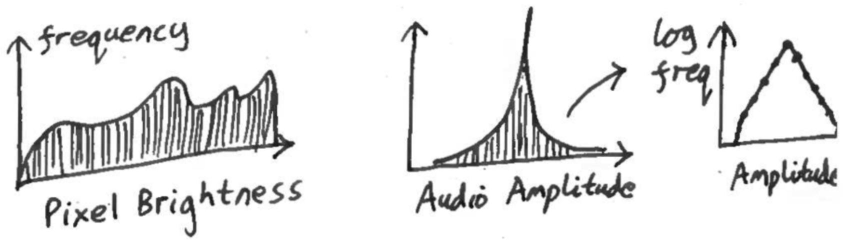
$$p(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- $p(x)$ now describe a curve centered at μ, with inflection points ("width") at $\pm\sigma$

> *The factor $\frac{1}{\sigma}$ is nothing but the **Jacobian** of the transformation. In this case, we are doing a change of variables via:*
>
> $$f(x) = \frac{x-\mu}{\sigma}$$
>
> *so the Jacobian is:*
>
> $$|f'(x)| = \frac{1}{\sigma}$$

2.2 The Central Limit Theorem

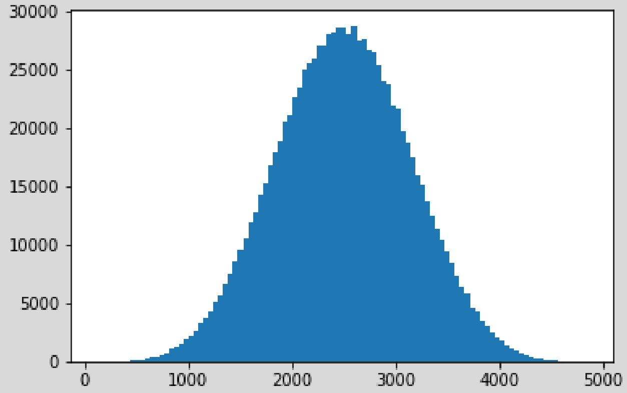
- **Are all probability distributions normal?**
  - no - real world data is not typically normally distributed:



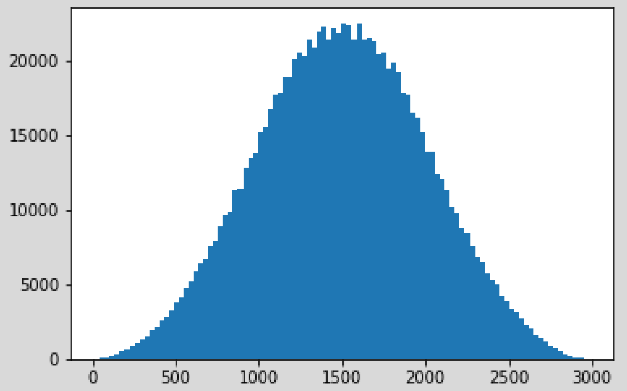
- **Why are normal distributions so important?**
  - they are simple to use, and still ubiquitous enough
  - even when things are **not** normally distributed, assuming that they are often leads to good results
- **What is the central limit theorem?**
  - under certain conditions, adding together many outcomes leads to data which is **approximately** Gaussian distributed
  - this means that understanding normal distributions will allow us to understand **a lot** of real data
- **What conditions are required for the CLT?**

1. **Bounded Mean and Variance**: each summand should be sampled from a distribution with **bounded** mean and variance (if we allowed a lot of extreme values, the sum would be distorted and the distribution wouldn't be bell-shape)
2. **Constrained Value**: the resulting distribution will be **proportional** to a Gaussian PDF, but constrained to the possible values of the original distribution (i.e if we sample from the integers, the sum can only be an integer)
3. **Convergence in Distribution**: the CLT converges in the sense of "convergence in distribution" (this is a weak form of convergence). In other words, the convergence to the Gaussian will be quick a few standard deviations from the mean, whilst the extreme tails of the distribution won't be so quick.

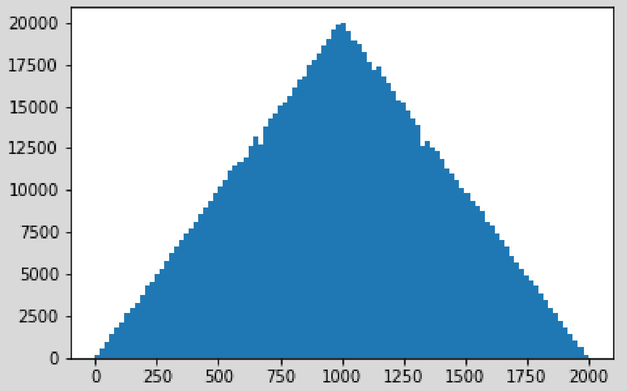
We generate 5 integers (in the range $[0, 1000]$*), and do this 1000 times. We then plot the histogram of the sums:*



If instead we add 3 integers we get:



Perhaps more interestingly, if we add 2 integers:



2.3 Standard Error on the Mean

2.3.1 The Standard Error on the Mean

- **How can we estimate the parameters of a distribution?**
  - given a sample produced from a distribution, the **maximum likelihood estimation** of the parameters are obtained by using the **sample mean** and **sample variance**:

$$\mu \approx \hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\sigma^2 \approx \hat{\sigma}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

  - the factor of $\frac{1}{N-1}$ for the variance makes it so that $\hat{\sigma}^2$ is an **unbiased** estimator

- **How good is the sample mean as an estimate of the population mean?**
  - notice, $\hat{\mu}$ is a **random variable**: its value depends on the samples
  - we have:

$$\mathbb{E}[\hat{\mu}] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n] = \frac{1}{N}\sum_{n=1}^{N}\mu = \mu$$

    so $\hat{\mu}$ is an **unbiased estimator**, and **on average**, $\hat{\mu}$ should be around μ

- **What is the standard error on the mean?**
  - gives us a measure of how good the estimate $\hat{\mu}$ is
  - if we compute its variance (assuming the samples are **independent**):

$$Var[\hat{\mu}] = \frac{1}{N^2}\sum_{n=1}^{N} Var[x_n] = \frac{1}{N^2}\sum_{n=1}^{N}\sigma^2 = \frac{\sigma^2}{N}$$

  - the value:

$$\frac{\sigma}{\sqrt{N}}$$

    is the **standard error on the mean**, which we typically approximate using the **sample variance**:

$$\frac{\hat{\sigma}}{\sqrt{N}}$$

  - the **standard error on the mean** gives a range on which μ is likely to lie:

$$\mu \in \left[x - \frac{\hat{\sigma}}{\sqrt{N}}, x + \frac{\hat{\sigma}}{\sqrt{N}}\right]$$

    68% of the time, μ should be within this range; 95% of the time it will lie 2 standard errors away

- **How can we reduce the standard error on the mean?**
  - to decreases the standard error by a factor of 10, we would need to increase the amount of samples by a factor of 100
  - hence, increasing the sampels increases our confidence in the sample mean as being similar to the true population mean

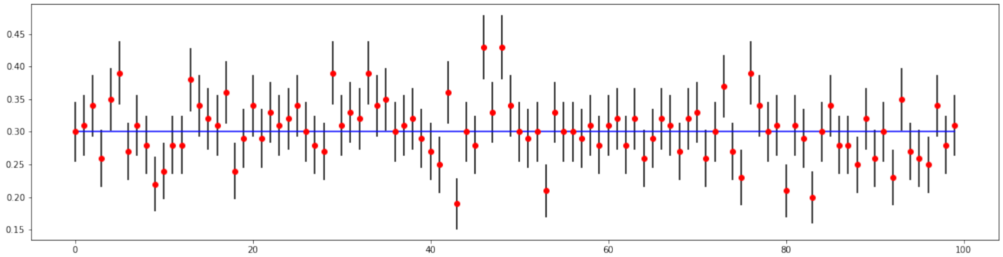


Figure 3: Here we perform 100 trials. At each trial, we take 100 samples from a Bernoulli distribution with mean $\mu = 0.3$. We then average these results, and compute the **standard error of the mean** (plotted as error bars). Notice how for **most** of the trials, the true mean 0.3 was within the error bars.

- **How can standard error on the mean be used to compare test error to generalisation error?**
  - the **average test error** is:

$$L_{test} = \frac{1}{M}\sum_{m=1}^{M} L(y^{(m)}, f(\underline{x}^{(m)})) = \frac{1}{M}\sum_{m=1}^{M} L_m$$

  - if we had infinitely many **different** test sets, then L_{test} could be treated as a random variable (albeit can't assume it to be normally distributed)
  - however, we can compute the **standard error on the mean**, to gauge how the performance of our model might deviate
  - again, this all depends on:
    - test cases being independent
    - loss being bounded (or at least finite variance)
    - future inputs coming from same distribution
    - relationship between input and output remaining the same

2.3.2 Model Reliability and Comparison

- **When is a model not robust?**
  - when its performance changes significantly across different fits
- **What can cause fit changes in a model?**
  - using new data
  - models might require **randomness** (i.e initialising a neural network)
  - randomness introduced by **training** (i.e GPUs perform a lot of parallel computations, which can lead to different results - even as significant as using randomness within the algorithm itself)
- **How can we report the variability of our model?**
  - give the **standard deviation** of **performance** across the different fits
  - a **robust** model will have a low σ, and so, will not vary much given future data
- **Can standard error be used to compare model performance?**

  - consider 2 models A, B, with standard error intervals overlapping (that is, the true generalisation error should be within 1 standard error of their test error) - could we assert that we can't tell if A is better than B (or viceversa)
  - this is false - for example, if A attains 0.1 better performance on **each** test case, we can confidently say that A is better than B
  - moreover, it could be the case that the standard error of B is **larger** than that of A

- **What is a paired comparison?**
  - a method for putting model performance head to head
  - we consider the **difference of losses**:

$$\delta_m = L_{test}^A - L_{test}^B$$

  - if we do this for each test set, we can compute the mean difference, and its corresponding standard error
  - if the mean is several **standard errors** greater than 0, we can be confident that A is a better model (the difference would be **statistically significant**)

2.4 Multivariate Gaussians

2.4.1 The Standard, Multivariate, Independent Gaussian

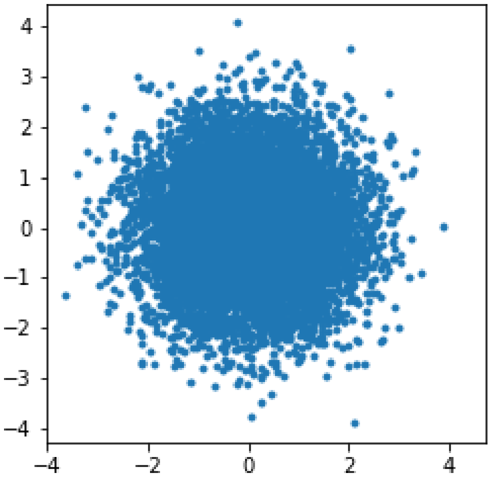
- **What is a standard multivariate gaussian?**
  - a distribution $\mathcal{N}(0, 1)$
  - however, we sample **vectors** $\underline{z}$, where each component z_i follows a standard normal distribution:

$$z_i \sim \mathcal{N}(0, 1)$$

- **What is the PDF of the standard multivariate gaussian?**
  - assuming each component is an **independent variable**, then the product of their PDFs will give the PDF of the mutlivariate gaussian:

$$\begin{aligned} p(\underline{z}) &= \prod_{d=1}^{D} p(z_d) \\ &= \prod_{d=1}^{D} \mathcal{N}(z_d; 0, 1) \\ &= \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi}} e^{-\frac{z_d^2}{2}} \\ &= \frac{1}{(2\pi)^{D/2}} e^{-\frac{1}{2}\sum_{d=1}^{D} z_d^2} \\ &= \frac{1}{(2\pi)^{D/2}} e^{-\frac{1}{2}\underline{z}^T \underline{z}} \end{aligned}$$

  - notice, this PDF is **proportional** to an RBF, albeit **scaled** to ensure it integrates to 1



2.4.2 The Covariance Matrix

- **What is the covariance matrix?**
  - a matrix Σ, the generalisation of **variance** for higher dimensions
  - in general, given a set of random variables:

$$\Sigma_{ij} = Cov(\underline{x}^{(i)}, \underline{x}^{(j)}) = \frac{1}{N-1}(\underline{x}^{(i)} - \underline{x}^{\bar{(i)}}) \cdot (\underline{x}^{(j)} - \underline{x}^{\bar{(j)}}) = \frac{1}{N-1}\sum_{n=1}^{N}(x_n^{(i)} - x^{\bar{(i)}})(x_n^{(j)} - x^{\bar{(j)}})$$

  - can also write, for a vector $\underline{x}$:

$$\Sigma = \mathbb{E}[\underline{xx}^T] - \mathbb{E}[\underline{x}]\mathbb{E}[\underline{x}]^T$$

  - the covariance matrix is **symmetric**, with the **variance** of a variable along the diagonal, and the **covariances** in the remaining entries

- **What is the covariance of 2 independent variables?**
  - the covariance is 0

- **What is the inverse of the covariance matrix?**
  - a matrix known as the **precision matrix**

- **What is a positive definite matrix?**
  - a **real, symmetric** matrix is positive definite **if and only if**:

$$\underline{z}^T\Sigma\underline{z} > 0, \qquad \forall\underline{z} \in \mathbb{R}^n, \underline{z} \neq \underline{0}$$

  - a **positive definite** matrix is **always** invertible, and the inverse is also **positive definite**:

$$\underline{z}^T\Sigma^{-1}\underline{z} > 0, \qquad \forall\underline{z} \in \mathbb{R}^n, \underline{z} \neq \underline{0}$$

- **What is a positive semi-definite matrix?**

  - a **real, symmetric** matrix is positive semi-definite **if and only if**:

$$\underline{z}^T \Sigma \underline{z} \geq 0, \qquad \forall \underline{z} \in \mathbb{R}^n, \underline{z} \neq \underline{0}$$

  - the **covariance** matrix is **positive semi-definite**
  - if $\underline{z}^T \Sigma \underline{z} = \underline{0}$, then Σ won't be **invertible**, since:

$$det(\Sigma) = |\Sigma| = 0$$

- **How can positive semidefinite matrices be generated?**
  - take any real-valued matrix A
  - then:

$$\Sigma = AA^T$$

    will be **positive semi-definite**
  - in fact, any symmetric, positive semi-definite matrix Σ can be written as a matrix product of A with its transpose (for some A)

2.4.3 The General, Multivariate Gaussian

- **What is the covariance matrix of vectors transformed linearly?**
  - consider $\underline{z} \sim \mathcal{N}(\underline{0}, \mathbb{I})$
  - we can apply a **linear transformation** to $\underline{x}$ via a matrix A:

$$\underline{y} = A\underline{z}$$

  - the **covariance** of this new random variable will be:

$$\begin{aligned} Cov[\underline{y}] &= \mathbb{E}[\underline{y}\underline{y}^T] - \mathbb{E}[\underline{y}]\mathbb{E}[\underline{y}]^T \\ &= \mathbb{E}[A\underline{x}\underline{x}^T A^T] - \mathbb{E}[A\underline{x}]\mathbb{E}[A\underline{x}]^T \\ &= A\mathbb{E}[\underline{x}\underline{x}^T]A^T - A\mathbb{E}[\underline{x}](A\mathbb{E}[\underline{x}])^T \\ &= A\mathbb{E}[\underline{x}\underline{x}^T]A^T, \qquad \textit{(since } \mathbb{E}[\underline{x}] = \underline{0}\textit{)} \\ &= ACov[\underline{x}]A^T \\ &= AA^T, \qquad \textit{(since } \mathbb{E}[\underline{x}\underline{x}^T] = Cov[\underline{x}]\textit{)} \end{aligned}$$

- **What is the PDF of a normal distribution whose vectors are transformed linearly?**
  - consider $\underline{z} \sim \mathcal{N}(\underline{0}, \mathbb{I})$ transformed by A
  - assuming the covariance matrix $\Sigma = AA^T$ of this transformation is **positive definite**, then:

$$\underline{y} = A\underline{z} \implies \underline{z} = A^{-1}\underline{y}$$

  - plugging this into the PDF of the standard multivariate Gaussian:

$$p(\underline{y}) \propto e^{-\frac{1}{2}(A^{-1}\underline{y})^T(A^{-1}\underline{y})} = e^{-\frac{1}{2}\underline{y}^T(A^{-1})^T A^{-1}\underline{y}} = e^{-\frac{1}{2}\underline{y}^T \Sigma^{-1}\underline{y}}$$

  - we need to rescale this so that the area remains the same; the volume by which a matrix changes a region after transforming it is its determinant, so:

$$p(\underline{y}) = \frac{1}{|A|(2\pi)^{D/2}} e^{-\frac{1}{2}\underline{y}^T \Sigma^{-1}\underline{y}}$$

- alternatively we can write:

  $$p(\underline{y}) = |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}\underline{y}^T\Sigma^{-1}\underline{y}}$$

  since:

  $$|2\pi\Sigma| = (2\pi)^D |A|$$

- **What is the PDF of a general, multivariate normal distribution?**
  - the last step is to apply a translation:

    $$\underline{x} = \underline{y} + \underline{\mu}$$

    so that the PDF gets centered at μ:

    $$p(\underline{x}) = \mathcal{N}(\underline{x}; \underline{\mu}, \Sigma) = \frac{1}{|A|(2\pi)^{D/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})}$$

- **How can we sample from a multivariate Gaussian?**
  - say we want to sample:

    $$\underline{x} \sim \mathcal{N}(\underline{0}, \Sigma)$$

  - the easiest way to do this is to use the fact that Σ will be positive definite, and so $\exists A$ such that:

    $$\Sigma = AA^T$$

  - A is what we can use to transform data:

    $$\underline{x} = A\underline{z}, \qquad \underline{z} \sim \mathcal{N}(\underline{0}, \mathbb{I})$$

  - however, A won't be unique:

    $$A = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \qquad A' = \begin{pmatrix} \sqrt{3} & 1 \\ -1 & \sqrt{3} \end{pmatrix} \implies AA^T = (A')(A')^T$$

  - in practice we can use the **Cholesky decomposition**, which gives us a triangular matrix decomposition

- **What is the shape of a multivariate Gaussian?**
  - data will look **elliptical**
  - the axes defining the ellipse are determined by the **eigenvectors** of Σ
  - below are some plots of this

2.4.4 Intuition: Non-Invertible Covariance and the Gaussian

Consider the transformation matrix:

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

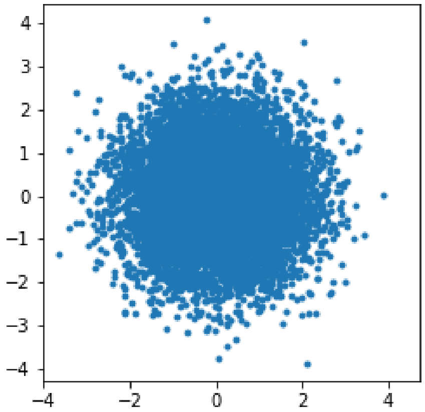
We can see A is not invertible, since:

$$|A| = 1 - 1 = 0$$

However, it is nicer to undertand this geometrically. The effect of A on any vector (x_1, x_2) is the mapping:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1 + x_2 \\ x_1 + x_2 \end{pmatrix}$$

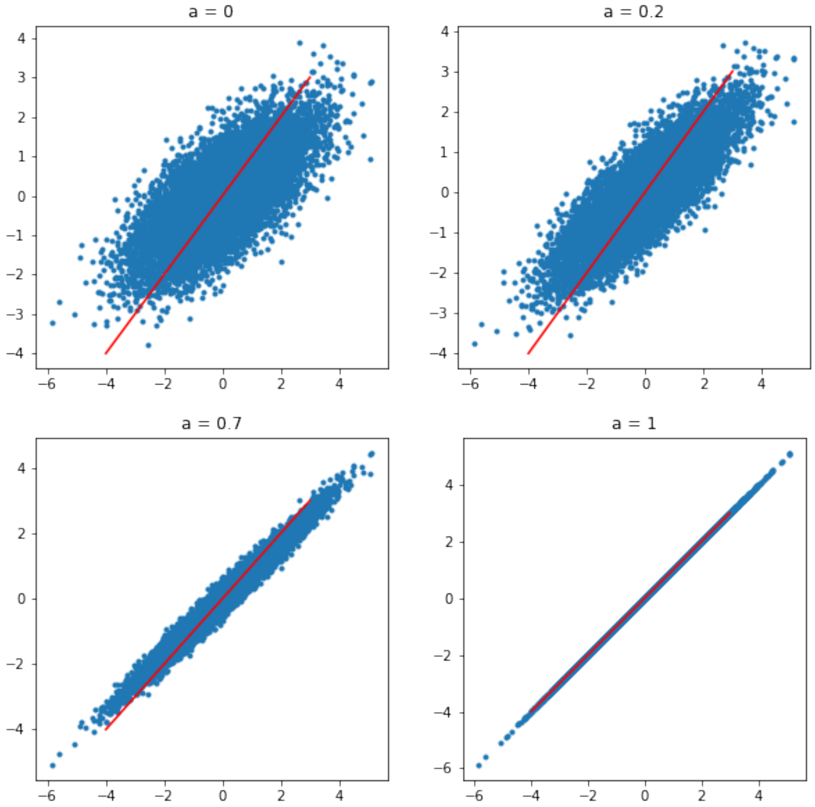
In other words, A has the effect of mapping **any** point in the plane to the single line $x_1 + x_2 = x_1 + x_2$. This means A can't be invertible, since given the point $(5, 5)$, there are infinitely many vectors which could've mapped to it $((1, 4), (4, 1), (2, 3), (4.9, 0.1)$ etc...).
Now, say we have a variety of points sampled from a standard normal distribution:



We can visualise the effect of:

$$B = \begin{pmatrix} 1 & 1 \\ 1 & a \end{pmatrix}$$

and see how it affects the above distribution as $a \to 1$:



This means that our transformed random variable “loses” its probability density: the probability of a Gaussian generating a point will be 0, unless the point lies **exactly** in the line through the origin..
How does this affect the shape of the gaussian? As $a \to 1$, $p(\underline{x}) = 0$ almost everywhere (since $\underline{x}$ will be constrained to some neighbourhood of the diagonal line). We still require that:

$$\int p(\underline{x})dx = 1$$

which means that $p(\underline{x})$ must approach infinite density along the aforementioned line. Intuitively, this means that the shape of the Gaussian distribution will approach the **Dirac Delta Function**
Since $\Sigma = AA^T$, Σ will also not have an inverse, so a non-invertible covariance matrix leads to a Dirac-like distribution.

3 Questions

3.1 Notes Question

1. **What is the MSE of the baseline which predicts the average of data each time?**

$$f(\underline{x}) = \frac{1}{N}\sum_{n=1}^{N} y^{(n)} = \bar{y}$$

We can compute:

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \bar{y})^2 = \sigma^2$$

Thus, if the MSE of a model is greater than the **variance** of the training labels, then there is an issue with our model.

2. **Consider the family of transformations:**

$$A = \begin{pmatrix} 1 & 0 \\ a & (1-a) \end{pmatrix}$$

   (a) **What does this transformation do?**
   If $\underline{x} = A\underline{z}$, then.

$$x_1 = z_1$$
$$x_2 = az_1 + (1-a)z_2$$

   (b) **For what values of a are the variables dependent?**
   The covariance matrix is:

$$\Sigma = \begin{pmatrix} 1 & a \\ a & a^2 + (1-a)^2 \end{pmatrix}$$

   Notice, if $a \neq 0$, the covariance of the variables will be non-zero, so they are dependent whenever $a \neq 0$.

   (c) **When are the variables maximally dependent?**
   When $a = 1$, the covariance will be 1, so maximally dependent.

   (d) **What happens to the PDF as $a \to 1$?**
   As $a \to 1$, the Gaussian approaches a Dirac distribution, with the probability density tending to infinity.

   (e) **Does the covariance matrix always have an inverse?**
   No. This is clear from computing the determinant:

$$|\Sigma| = a^2 + (1-a)^2 - a^2 = (1-a)^2$$

   so if $a = 1$, Σ^{-1} doesn't exist. Alternatively, notice that if $a = 1$, $x_2 = az_1 = ax_1$, so all points will lie on the same line, and we lose information about z_2, so this transformation can't be reversed.

4 Tutorial

1. **If $\underline{a}, \underline{b} \in \mathbb{R}^D$ and $M \in \mathbb{R}^{D\times D}$ is a symmetric matrix, show that:**

$$\underline{a}^T M \underline{b} = \underline{b}^T M \underline{a}$$

$$\underline{a}^T M \underline{b} = \underline{a}^T M^T \underline{b} = \underline{b}^T M \underline{a}$$

2. **Suppose that:**

$$P(\underline{x}) \propto \exp\left(-\underline{x}^T A \underline{x} - \underline{x}^T \underline{c}\right)$$

   where A is a symmetric invertible matrix. This distirbution is Gaussian, since it is proporrtional to the exponential of a quadratic in $\underline{x}$. Identify the Gaussian from which $\underline{x}$ comes from, by identifying the mean $\underline{mu}$ and covariance Σ in terms of A and $\underline{c}$.

We can just match terms. For a "normal" Gaussian:

$$\begin{aligned} P(\underline{x}) &\propto \exp\left(-\frac{1}{2}((\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})\right) \\ &= \exp\left(-\frac{1}{2}\left[\underline{x}^T\Sigma^{-1}\underline{x} - \underline{x}^T\Sigma^{-1}\underline{\mu} - \underline{\mu}\Sigma^{-1}\underline{x} + \underline{\mu}^T\Sigma^{-1}\underline{\mu}\right]\right) \\ &= \exp\left(-\frac{1}{2}\left[\underline{x}^T\Sigma^{-1}\underline{x} - 2\underline{x}^T\Sigma^{-1}\underline{\mu} + \underline{\mu}^T\Sigma^{-1}\underline{\mu}\right]\right) \end{aligned}$$

Thus, comparing with

$$P(\underline{x}) \propto \exp\left(-\underline{x}^T A\underline{x} - \underline{x}^T\underline{c}\right)$$

we see that:

$$A = \frac{1}{2}\Sigma^{-1} \implies \Sigma = \frac{1}{2}A^{-1}$$

$$-\underline{c} = \Sigma^{-1}\underline{\mu} \implies \underline{\mu} = -\Sigma\underline{c} = -\frac{1}{2}A^{-1}\underline{c}$$

3. **The first element of a vector has:**

$$x_1 \sim \mathcal{N}(m, \sigma^2)$$

the second element is generated via:

$$x_2 = \alpha x_1 + \nu, \qquad \nu \sim \mathcal{N}(0, n^2)$$

The joint distribution of the vector $\underline{x} = (x_1, x_2)$ is Gaussian. Identify the $\underline{\mu}, \Sigma$ such that:

$$\underline{x} \sim \mathcal{N}(\underline{\mu}, \Sigma)$$

1. Using expectations and probability theory:

$$\begin{aligned} \mathbb{E}[x_1] &= m \\ \mathbb{E}[x_1^2] &= Var(x_1) + (\mathbb{E}[x_1])^2 \\ &= \sigma^2 + m^2 \\ \mathbb{E}[x_2] &= \mathbb{E}[\alpha x_1 + \nu] \\ &= \alpha\mathbb{E}[x_1] + \mathbb{E}[\nu] \\ &= \alpha m \\ \mathbb{E}[x_1 x_2] &= \mathbb{E}[x_1(\alpha x_1 + \nu)] \\ &= \alpha\mathbb{E}[x_1^2] + \mathbb{E}[x_1]\mathbb{E}[\nu] \\ &= \alpha(\sigma^2 + m^2) \\ Var(x_2) &= Var[\alpha x_1 + \nu] \\ &= \alpha^2 Var[x_1] + Var[\nu] \\ &= \alpha^2\sigma^2 + n^2 \end{aligned}$$

From this, we can just read off:

$$\underline{\mu} = \begin{pmatrix} m \\ \alpha m \end{pmatrix} \qquad \Sigma = \begin{pmatrix} Var[x_1] & \mathbb{E}[x_1x_2] - \mathbb{E}[x_1]\mathbb{E}[x_2] \\ \mathbb{E}[x_1x_2] - \mathbb{E}[x_1]\mathbb{E}[x_2] & Var[x_2] \end{pmatrix} = \begin{pmatrix} \sigma^2 & \alpha\sigma^2 \\ \alpha\sigma^2 & \alpha^2\sigma^2 + n^2 \end{pmatrix}$$

2. Using Gaussian distribution knowledge. Say that:

$$\underline{z} \sim \mathcal{N}(\underline{0}, \mathbb{I})$$

then:

$$\underline{x} = \underline{\mu} + L\underline{z}$$

where L is such that $\Sigma = LL^T$. If we have $z_1 \sim \mathcal{N}(0,1)$ and $z_2 \sim \mathcal{N}(0,1)$, then:

$$x_1 = m + \sigma z_1$$

$$x_2 = \alpha x_1 + \nu = \alpha(m + \sigma z_1) + n z_2$$

Then, we can read:

$$\underline{\mu} = \begin{pmatrix} m \\ \alpha m \end{pmatrix}$$

$$L = \begin{pmatrix} \sigma & 0 \\ \alpha\sigma & n \end{pmatrix}$$

which implies that:

$$\Sigma = LL^T = \begin{pmatrix} \sigma^2 & \alpha\sigma^2 \\ \alpha\sigma^2 & \alpha^2\sigma^2 + n^2 \end{pmatrix}$$

as above

4. **We can sample from:**

$$\underline{x} \sim \mathcal{N}(\underline{0}, \Sigma)$$

by drawing a vector of standard normals:

$$\underline{\nu} \sim \mathcal{N}(\underline{0}, \mathbb{I})$$

and setting:

$$\underline{x} = A\underline{\nu}$$

for any matrix A, where $AA^T = \Sigma$.
Real symmetric matrices, like covariance matrices, can always be written in the form:

$$\Sigma = Q\Lambda Q^T$$

where Λ is a diagonal matrix of eigenvalues, and the columns of Q are the eigenvectors of Σ.

(a) **Describe how to sample from $\mathcal{N}(\underline{0}, \Sigma)$ using this decomposition.**

- since Λ is a diagonal matrix, we can write:

$$\Lambda = \Lambda^{1/2}\Lambda^{1/2}$$

  where $\Lambda^{1/2}$ denotes the matrix obtained by taking an elementwise square root of Λ
- then we have that:

$$\Sigma = Q\Lambda^{1/2}\Lambda^{1/2}Q^T = Q\Lambda^{1/2}(Q\Lambda^{1/2})^T$$

- thus, using $A = Q\Lambda^{1/2}$ we can sample from the distribution with covariance Σ

(b) **Q is an orthogonal matrix, corresponding to a rigid rotation (and possible a reflection). Describe geometrically how you sampling process transforms a cloud of points drawn from a standard normal.**

- Q is orthogonal - its columns are orthonormal vectors, which thus give rise for a new basis of space; hence, multiplying by Q is equivalent to mapping data to this new basis, which is nothing but rotations (and potential reflections)
- to sample, we sample $\underline{x} \sim \mathcal{N}(\underline{0}, \mathbb{I}$ and:

$$A\underline{x} = Q(\Lambda^{1/2}\underline{x})$$

- $\Lambda^{1/2}$ is a diagonal matrix, which stretches each x_i by a factor of $\sqrt{\lambda_i}$; this converts the sphere of points into an ellipsoid of points
- Q then rotates the points, by mapping them to the orthogonal basis spanned by its columns; this rotates the ellipsoid, such that its principal axes align with the eigenvectors of Σ

5. **The notes introduced the lower-triangular Cholesky decomposition $\Sigma = LL^T$, which can be applied to symmetric positive- definite (but not semi-definite) matrices. As well as being useful for sampling, common computations involving triangular matrices (determinants, matrix inverses, solving equations) are quick, so many library routines involving Gaussians use the Cholesky decomposition.**

(a) **Sometimes instead of decomposing th covaraince matrix, we have the Cholesky decomposition of the precision matrix:**

$$\Sigma^{-1} = CC^T$$

where C is lower-triangular. How could we use C to sample from $\mathcal{N}(\underline{0}, \Sigma)$?

Since $\Sigma^{-1} = CC^T$, then:

$$\Sigma = (C^{-1})^T C^{-1}$$

We can then sample by using:

$$L = (C^{-1})^T$$

in our standard strategy.

(b) **Yet another possible decomposition is the principal square root:**

$$\Sigma = \Sigma^{1/2}\Sigma^{1/2}$$

where $\Sigma^{1/2}$ is symmetric. We now try to understand how these decompositions are related.

i. **Consider 2 different decompositions:**

$$\Sigma = AA^T = BB^T$$

We'll assume the matrices are full rank, so that we can write $B = AU$. Show that:

$$UU^T = \mathbb{I}$$

and so, U is orthogonal.

$$AA^T = BB^T = (AU)(AU)^T = A(UU^T)A^T \implies UU^T = \mathbb{I}$$

since A is invertible.

ii. **Explain geometrically why if computing $A\underline{\nu}$ from $\underline{\nu} \sim \mathcal{N}(\underline{0}, \mathbb{I})$ is a way to sample from $\mathcal{N}(\underline{0}, \Sigma)$, computing:**

$$B\underline{\nu} = AU\underline{\nu}$$

will be as well.

- U just rotates the points of $\underline{\nu}$, so they remain spherically distributed
- hence, $U\underline{\nu} \sim \mathcal{N}(\underline{0}, \mathbb{I}$
- hence, applying A to $U\underline{\nu}$ to sample is equivalent to sampling by applying A to $\underline{\nu}$ directly