Reference URLS

Wednesday, April 19, 2023 10:45 AM

Environmental data resources

https://www.epa.gov/outdoor-air-quality-data/download-daily-data

https://eeaonline.eea.state.ma.us/portal/#!/home

Differential equation

https://activecalculus.org/single/sec-7-1-diff-eq-intro.html#:~:text=A%20differential%20equation%20is%20an%20equation%20that%20describes%20the%20derivative,that%20is%20unknown%20to%20us.

Kriging model:

https://www.youtube.com/watch?v=J-IB4_QL7Oc

Introduction

Wednesday, April 19, 2023 10:44 AM

Environmental data examples:

Air quality data at different time and location

Pollutant level in the air

Pollutant level in the water.

Weather and climate data

Forecast weather

Forecast extreme conditions such as storm, wind, snow etc.

Ocean dynamics

Currents and streams

Measuring algae level

Land

Forestation and deforestation

Monitoring crops

Monitoring wildlife

Why do we need to analyze environment data?

Understanding the underlying process and making inferences of what interacts with what ways helps in learning the global dynamics. It will also help in understanding the impact of change in environmental variables in the health, economics and society .

For example,

climate change is the statistics of weather over time. Weather is short-term while the climate is long-term. We will be able to simulate policies, and analyze their predictions and forecast. We could understand what outcome each policies will have before implementing them.

In terms of monitoring and forecasting storms, earthquakes are very important in early warnings.

In water and energy management, these analysis helps in managing the precious resources.

Understanding environmental data:

From the point of view of environmental data, there are many other related variables that one might be interested in, for example:

- Predicting animal populations: Usually, animal populations serve as indicator of healthy ecosystems.
- Predicting disease: Environmental data can provide contextual information about disease spreading. For example, how temperature changes and rainy seasons impact the spread of malaria in the tropics.
- Renewable energy predictions: Clear understanding of wind speeds, water levels in damps, or underground thermal activity provide information for renewable energy decisions.
- Predicting extreme events: extreme events like storms and floods are usually modeled and predicted to incorporate action plans.
- Policy making: for example pollution concentrations measurements in the air can help the design of guidelines for industry contaminants.

What type of common characteristics would you find in environmental data?

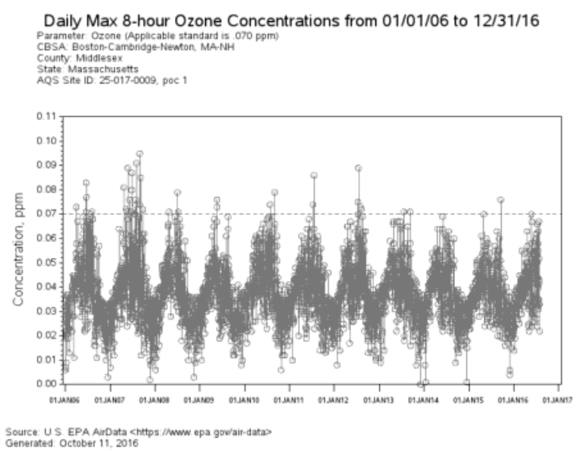
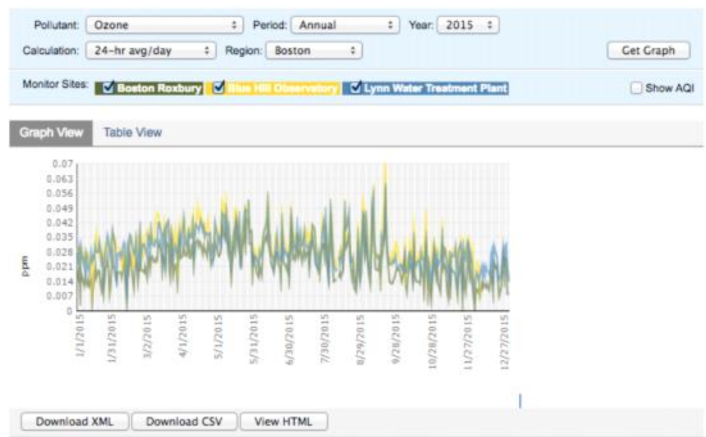
- Seasonality (correlation in time): Some variables like temperature, or weather systems in general have some temporal patterns. For example, seasons, or rain/drought times.

- Space relations, or spatial correlations: Under the assumption that some of the variables that we are interested in have some form of smooth behavior, one might expect that the physical distance between two points induces or is an indication that the date generated at those points have some relation between them.
- Continuous changes: Continuity or smoothness is a special assumption for environmental data, where we suppose that the variables will not have sudden changes that generate abrupt data discontinuities.

Correlations

Saturday, April 22, 2023 8:52 PM

Weather is a classic example of a dynamical system from which one can measure a number of variables to generate environmental data, see the above figure. One can obtain physical and chemical models for such processes to understand parts of such systems. This allows us to use data generated from first-principles simulations. Moreover, one can observe the seasonality, spatial relations, and continuity of the generated data.
The below figure is a clear example of seasonality or temporal correlations in data. In this case, we observed ozone levels in Boston in 2016. A pattern is recognized: periodically, the levels go up and down, which corresponds to yearly variations.

	
Ozone concentration in Boston	Spatial correlation example

The above figure presents an example of correlations in space. In this case, we see ozone levels of three different locations in the Boston area. The measuring stations are close together; thus, some correlation is expected. One can see that there is also fluctuations in time, but near-by stations produce correlated data. There is some temporal correlation evident in the above spatial correlation example figure. This correlation has a **timescale** : this is the period of time over which we can expect the measurements at each location to be strongly correlated to each other.
The timescale of the correlation observed is **"Day".**

Advantage and disadvantage of correlation

Saturday, April 22, 2023 9:13 PM

Advantage and disadvantage of correlation:
Correlated data helps with predictions of unknown variables based on observations of other variables that are highly correlated with the unknown variables. If you have measurements at one point in space, and correlations are expected with some other point from which data is not available, one can quantify or predict unobserved data. However, the estimation of means with highly correlated data might not be useful, as the correlation prevents a reduction in the variance of the mean.
Negative correlation:

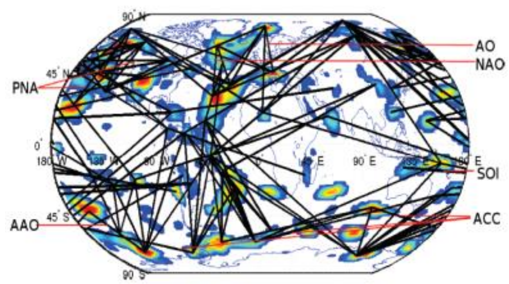


Figure 1.8. *Dipoles in NCEP sea-level data for the period 1948–1967. The color background shows the regions of high activity. The edges represent dipole connections between regions.*

Dipole connections between regions, an example of negative correlation

The figure on the left shows global correlations in space: in this case, the temperate between two points that are anti-correlated. Two points are connected if their temperatures are correlated negatively, i.e., high temperature at one location suggests a low temperature in the other.

Example: We can estimate the temperature off the coast of Madagascar (the Mozambique Channel), based on measurements on Caracas, Venezuela

The figure below shows the temperature measurements for two towns in Québec. Assuming some proximity in these two towns, some correlation is expected and observed in the data. However, around 1962 there is a discontinuity in the measurements. By data observations only, this will imply a sudden decrease in the temperature of Sherbrooke, is this possible under some continuity assumptions? A more mundane reason explains such change. The station that was taking such measurements was moved from the inner to the outer city. As expected, the temperature outside the city is lower. This leads us also to consider context as well as new phenomena that can be present in data. Taking into account such possibly unknown dynamics is a part of environmental data analysis, as different sensors will have different biases and each measurement needs to be individually calibrated.

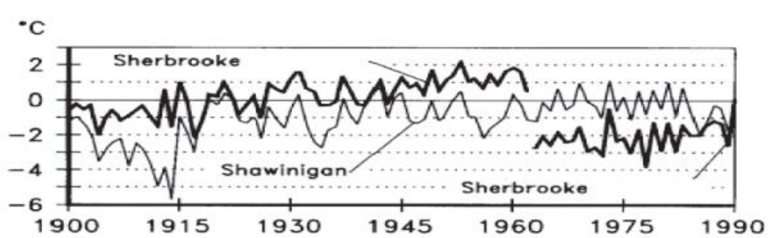


Figure 8.1. *Temperature records for two towns in Québec* [117].

From the examples above, we can conclude a number of common characteristics about environmental data

- Correlations in time: temporal patterns such a seasons.
- Correlations in space: linear dependencies or relations between variables measured at close-by locations.

Modeling flow data

Saturday, April 22, 2023 9:19 PM

Now we will see how to model the dynamic and flaw in the ocean.
Flow data means that we have a vector at each particular location that indicates the direction and magnitude of the flow, i.e., displacement in time

Flow Field:

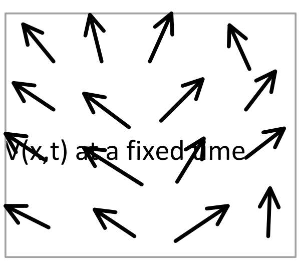
(Ref:

https://aia.springeropen.com/articles/10.1186/s42774-022-00113-1

https://en.wikipedia.org/wiki/Vector_field

)

Flow visualization is the visualization of the vector data generated in fluid dynamics studies, for example, the simulations of combustion models, aerodynamic models, and climate models. The flow field data consists of one or more fields, some of which have time-varying properties, representing the magnitude and direction of the velocity at each location. As the size of flow data becomes larger, and the internal structure becomes more complex, the analysis of flow data becomes complicated. Visualization of flow data is an important method for understanding flow data.
A vector field in the plane can be visualized as a collection of arrows with a given magnitude and direction, each attached to a point in the plane. A Vector field is often used to model the speed and direction of a moving fluid throughout 3D space as it changes from one point to another. A Vector field vector can be 2D (fixed time, integrated coordinates) or 3D . It is represented as V(x,t) : Flow at location x at time t.



Each element in the vector shows the direction of the flow at the location x at time t. The length of the vector tells us the velocity of the flow (how quickly an object is transported from point x to another if dropped at point x).
We will have a velocity vector for every space and time.

We can use the velocity vector for the following tasks:

1) Forward prediction:

   If we drop an object at point x that can float. Where would we find it after a certain amount of time. This will help in finding the trajectory in space and time of that item.

Flow dynamics:

The flow refers to the displacement in time , that is if x(t) denotes the location of a particle at a time t, then we define the flow vector as

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = V(x(t), t)$$

Or similarly, their explicit Euler discretization as , for a small e>0,

x(t+e) = x(t) + eV(x(t),t)

We set the initial time as 0, then the location of a particle at a time t+e is defined as

x(e) = x(0) + e* V(X(0),0)

Spatial correlation

Sunday, April 23, 2023 7:06 PM

Many Environmental data has a strong spatial and temporal correlation. That means, in short distances, the measurements are very strongly positively correlated.

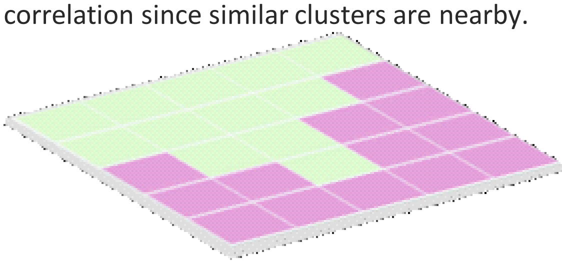
Temporal correlation measures the change of a variable over time while spatial correlation measures the change of two variables - observations (values like income, rainfall etc.) and location.

Moran's I

Spatial auto-correlation is measured by Moran's I. Moran's I is a correlation coefficient used to measure the overall spatial correlation in your data set. Moran I's can be classified as positive, negative and no spatial autocorrelation:

1. Positive correlation:

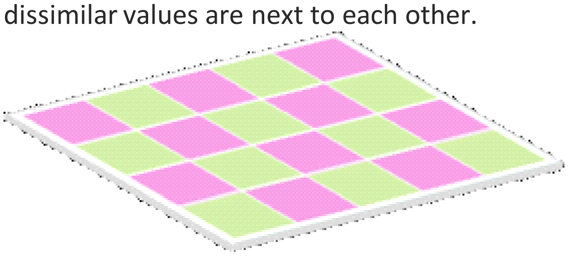
Spatial correlation is positive when similar values cluster together on a map. Positive autocorrelation occurs when Moren I is close to +1. The image below shows the land cover in an area and it is an example of a positive correlation since similar clusters are nearby.



Land cover image depicting the positive correlation

2. Negative correlation:

Spatial correlation is negative when dissimilar values cluster together on a map. A negative spatial autocorrelation occurs when Moran's I value is -1. A checkerboard is a good example of negative auto-correlation because dissimilar values are next to each other.



Dissimilar objects cluster together hence the negative correlation

3. Zero correlation:

A Moran's I value of 0 denotes no spatial autocorrelation.

https://medium.com/locale-ai/spatial-autocorrelation-how-spatial-objects-affect-other-nearby-spatial-objects-e05fa7d43de8

Correlation Definitions

Sunday, April 23, 2023 9:37 PM

Normal Distribution:

A normal distribution is a continuous probability distribution for a real-valued random variable. This is also called as Gaussian distribution.

Univariant probability density function:

The unidimensional probability density function is defined as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Where, μ = mean or expectation of the distribution

σ = Standard deviation

σ^2 = variance

In the normal distribution with mean μ and standard deviation σ, the median and mean are $\boldsymbol{same}$.

The standard Normal distribution is defined as the Normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The maximum value attained by the probability density function of the standard normal distribution is $\frac{1}{\sqrt{2\pi}}$

Multivariate distribution:

The probability density function of a Multivariate Gaussian random variable is defined as

$$P(x|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2} \cdot |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Where, $\mu = E[x]$

$|\Sigma| = determinant\ of\ the\ covariance\ matrix\ \Sigma$

Covariance:

Covariance quantifies the joint variation between two random variables. Formally, the covariance between two real-values random variables X and Y, both with finite second moments, is defined as

$\text{cov}(x, y) = E\left[(x - \mu_X)(y - \mu_y)\right] = E[xy] - E[x]E[Y]$

$\text{cov}(X, Y) = \sigma_{XY} = \sigma(X, Y)$

Correlation:

Correlation is defined as any statistical relation between two random variables. There are many ways we can measure correlation. The Pearson correlation coefficient between two random variables X and Y, with expected values $E[x] = \mu_x$ and $E[Y] = \mu_y$, and standard deviations σ_x and σ_y, is defined as

$$\rho \times y = Corr(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \cdot \sigma_y} = \frac{E\left[(x-\mu_x)(y-\mu_y)\right]}{\sigma_x \sigma_y}$$

Below are the two properties of the covariance operator:

- $\text{cov}(x, x) = \sigma_x^2 = variance$
- $\text{cov}(aX + bY, cW + \mathrm{d}V) = ac \cdot \text{cov}(X, W) + ad \cdot \text{cov}(X, V) + bc \cdot \text{cov}(y_1 W) + bd \cdot \text{cov}(y_1 V)$
- Covariance operator is bilinear.

Exercise:

Given 2 random variables X,Y with covariance COV(X,Y) = 0.5, COV(X,X) = 1 and COV(Y,Y) = 2. Define a new random variable Z = 2X+Y. What is the value of COV(Z,Z)?

COV(Z,Z) = COV(2X+Y, 2X+Y) = 4COV(X,X)+2COV(X,Y)+2COV(Y,X)+COV(Y,Y)

=4*1+2*0.5+2*0.5+2

= 8

Random Variables

Two random variables X and Y are independent if and only if, for every x and y, the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent. Recall that two events A and B are independent if and only if their joint probability equals the product of their probabilities, i.e., $P(A \cap B) = P(A)P(B)$. For Random variables, this is equivalent to
$F_{X,Y}(x, y) = F_X(x)F_Y(y)$
Where $F_X(x)$ and $F_Y(y)$ are cumulative functions of the random variables X and Y.

The cumulative distribution function (CDF) of random variable X is defined as
F_X (x)=P(X≤x), for all x∈R.

Exercise-1:
If X and Y are independent random variables, and both are defined on a finite sample space omega=z1,z2,…zn, then E[XY]=E[X]E[Y] as

E[XY] = $\sum_{i,j=1}^{n} z_i z_j P\left(X = X_i, Y = y_j\right)$

= $\left(\sum_{i=1}^{n} P(X = z_i)\right)\left(\sum_{j=1}^{n} P\left(Y = z_j\right)\right)$

What is E[XY] when X and Y are 2 unbiased independent six sided dice?

Here omega = 1,2,3,4,5,6 since X and Y are dice.
X = {1,2,3,4,5,6}, Y={1,2,3,4,5,6}
P(X=x) = 1/6
P(Y=y) = 1/6
E(XY) = (1/6+2/6+3/6+4/6+5/6+6/6)(1/6+2/6+3/6+4/6+5/6+6/6)
= 12.25

Exercise-2:
Assume you are given two unbiased six sided dice modeled as random variables X_1 and X_2 respectively. However, die 2 depends on the output of die 1. If the output of the die 1 is even, then die 2 can only output odd numbers. Note there is no condition if die 1 id odd. What is $E[X_1X_2]$?

$$E[x_1x_2] = \frac{E[x_1x_2|x_1\text{even}] + E[x_1x_2|x,\text{odd}]}{2}$$

$$E[x_1x_2] = \frac{E[x_1|x_1\text{even}]E[x_2|x_1\text{even}] + E[x_1|x_1\text{odd}]\,E[x_2]}{2}$$

```
e_x1_x1_even = (2+4+6)/3
e_x2_x1_even = (1+3+5)/3
e_x1_x1_odd = (1+3+5)/3
e_x2 = (1+2+3+4+5+6)/6
e_x1_x2 = ((e_x1_x1_even*e_x2_x1_even)+(e_x1_x1_odd*e_x2))/2

Result = 11.25
```

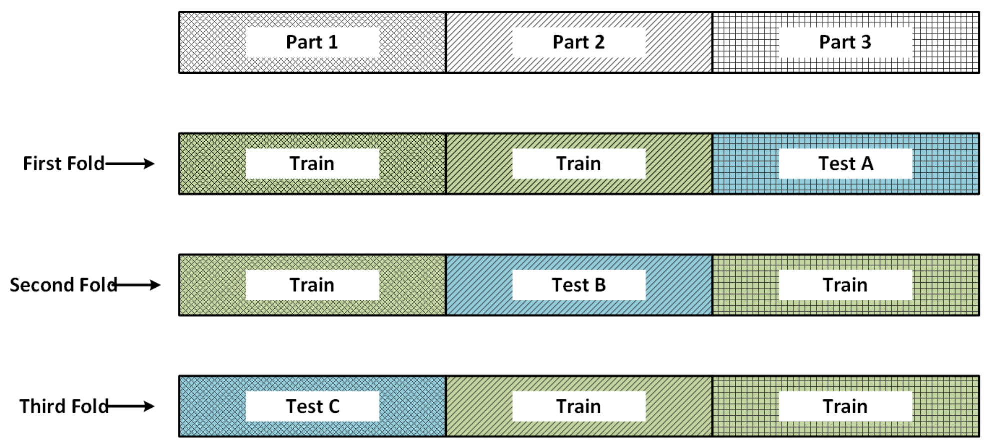
Note:
Define two random variables as $x \sim \mathrm{U}[-1,1]$ and $\mathrm{Y} = X^2. \textit{Then they are not independent. They are not correlated.}$

Cross-validation

Tuesday, April 25, 2023 8:40 PM

Cross-validation is a statistical technique for validating the generalization of a model when applied to independent data. This kind of validation consists of splitting the available data in various disjoint train and test subsets. Then, the model is used to form a prediction of the test data using the train data, and the performance evaluated. The goal is to validate the model's ability to predict new data that was not part of the training set.

K-fold Cross-validation assumes we partition our data set into K disjoint subsets. Then, we train K models, each one with one of the partitions selected as the test set and the remaining subsets are used to form the training set. The below figure shows an example for 3-fold validation, where the data was partitioned into three disjoint subsets. At each fold, two of the partitions shown in green are used for training and the remaining partition shown in blue are used for testing.



An example of 3-fold cross validation.

The testing step on each of the folds produces a score, which will be defined according to the particular requirements of the problem. Recall that following the conditional distribution approach we have studied in the previous lecture, testing will produce the marginal distribution on the set of random variables marked for testing, from which we need to make a prediction and generate a score.
One immediate prediction could be the mean of the conditional distribution, and the score could be the mean square error to the true data point. The average score over all of the folds is computed and used to evaluate the performance of that particular choice of model. This process is repeated for each of the models being tested. Finally, one can select the model with the best performance according to the chosen score.

Leave-one-out Cross-validation refers to the case where the testing dataset is composed of a single data point. Going back to the 3-fold example the above figure. Assume we have observed $X_2 = x_2$, and we have at our disposal μ_2 . Moreover, let us assume we are working with the kernel function

$$K(z_i, z_j) = \exp\left(-\frac{\|z_i - z_j\|^2}{2l^2}\right)$$

With parameter $\theta = \{l\}$

Working With Gaussians

Tuesday, April 25, 2023 8:57 PM

Gaussian Function:
Gaussian functions are sometimes called normal function.

1-D Gaussian function:

The generic form of one dimensional data in Gaussian function is

$$f(x) = a \cdot \exp\left(-\frac{[x-b]^2}{2c^2}\right)$$

When we work with probability or random variables, the same function is also written as

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right)$$

Even through, they look different, they are same.

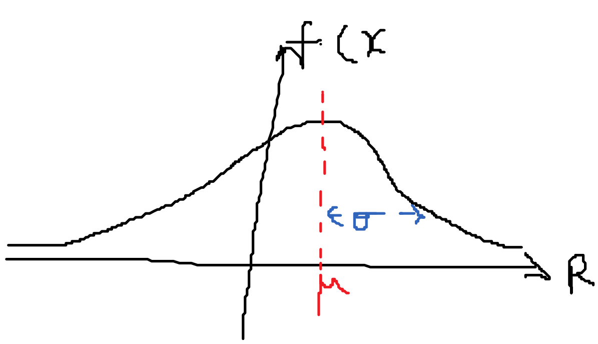
Gaussian random variable:
we are going to say a random variable X is Gaussian which is represented as Normal distribution of mean and variance , if the probability distribution looks like the formula below.

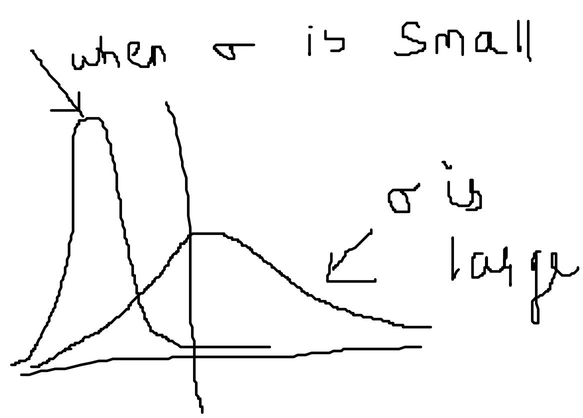
$$\mathrm{X} \sim N(\mu, \sigma^2)\ if\ f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Since we are working with the 1D random variable, , we will say that

$\mu \in \mathbb{R} \Rightarrow$ Mean
$\sigma > 0 \Rightarrow$ Standard deviation

If we plot the Gaussian random variables it will look something like this.





Multivariate Gaussian

Friday, April 28, 2023 6:34 PM

It is the Gaussian for multi-dimensional random variable.

$X = \{x_1, x_2 \ldots x_k\}^T$

Where each variable $x_1, x_2, \ldots$ are individually Gaussian. In this case ,
$X \sim N(\mu, \Sigma)$

X is a Gaussian distribution of parameters μ (mean of vector dimension k), and sigma becomes covariance (**Σ**).
Covariance is a vector of dimension k x k.

The multivariate Gaussian function can be written as $f_X(x_1, x_2 \cdots x_k) = \dfrac{\exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)\right)}{\sqrt{2\pi^k \det(\Sigma)}}$

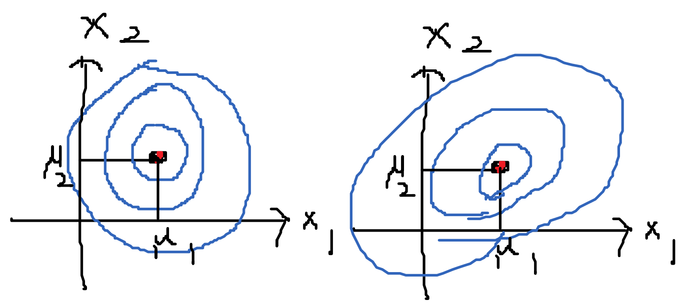
When k=2,

X= $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $X \sim N(\mu, \Sigma)$

$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$

Hence, $X \sim N(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix})$

σ_{12}, σ_{21} = cross terms thats are usually called cross − correlation terms.
σ_{11} = standard deviation of x_1
σ_{22} = standard deviation of x_2



The Circles in the graphs are perfect circles if their diagonal term is Zero. Here $\sigma_{12} = \sigma_{21} = 0$

If these values are large, the circle can become ellipse.

Example:
Let's consider the temperatures of two cities as random variables such as
x1 = Temperature of city1 with $x1 \sim N(\mu_1, \sigma_1^2)$
x2 = Temperature of city2 with $x2 \sim N(\mu_2, \sigma_2^2)$
Now, depending on the distance between the two cities, we can assume that the temperatures are related. The covariance matrix plays a role in this decision.

The two cross-terms in the covariance matrix represent how the two variables relates to each other.
If the two cities are farther away from each other ,their temperatures do not relate to each other and the cross-terms will be small.
If the two cities are closer to each other, their temperature will be related and the cross-term will be large.

If we have the value of X in one city and the covariance measurement, will we be able to find temperature in the other city? To do this, we will use conditional distribution of the Gaussian variables.

Conditional distribution

Friday, April 28, 2023 7:28 PM

Conditional distribution of the Gaussian random variable is represented as

$$\mu_{x_1|x=x_2} = \mu_1 + \frac{\sigma_{12}}{\sigma_2}(x_2 - \mu_2)^2$$

$$\sigma_{x_1|x=x_2} = \sigma_1 - \frac{\sigma_{12}^2}{\sigma_2}$$

Now, the conditional distribution of the multivariate Gaussian variable is given as

$$X = \begin{bmatrix} x_1 & \in & R^d \\ x_2 & \in & R^{N-d} \end{bmatrix}, \mathrm{X} \in R^N \qquad \mu_x = \begin{bmatrix} \mu_{x_1} & \epsilon & R^d \\ \mu_{x_2} & \epsilon & R^{N-d} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \Rightarrow \begin{matrix} \Sigma_{11} \in R^{dxd} \\ \Sigma_{21} \in R^{(N-d)xd} \end{matrix} \qquad \begin{matrix} \Sigma_{12} \in R^{dx(dN-)} \\ \Sigma_{22} \in R^{(N-d)x(N-d)} \end{matrix}$$

Hence the conditional distribution of the multivariate Gaussian distribution is

The conditional mean,
$\mu_{x_1|x_2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$

Where,
$\mu_{x_1|x_2}$ = New estimate of the mean of X1 conditioned on X2=x2
μ_1 =Old mean of X1
Σ_{12} = Covariance between X1 and X2
Σ_{22}^{-1} = Inverse of the variance of X2
X_2 = Observation of X2
μ_2 = Old mean of X2

The conditional variance,
$\sum_{x_1|x=x_2} = \Sigma_{11} - \Sigma_{12} \cdot \Sigma_{22}^{-1} . \Sigma_{21}$

Where,
$\sum_{x_1|x=x_2}$ = New estimate of the variance of X1 conditioned on X=x2
Σ_{11} = Old variance of X1
Σ_{12}, Σ_{21} = Covariance between X1 and X2
Σ_{22}^{-1} = Inverse of the variance of X2

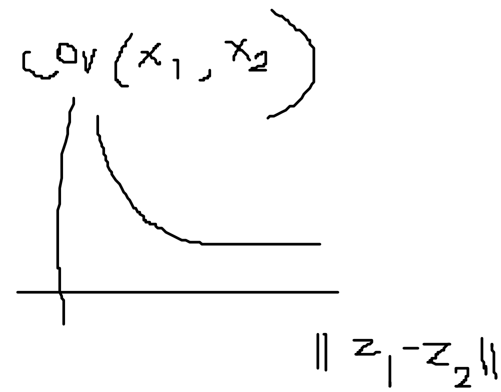
Kernel functions

Friday, April 28, 2023 9:00 PM

Estimation of the covariance can be done using the kernel function.

Let's consider the example of temperature between two cities (X1, X2).
We might not always get enough data to compute the covariance of two random variables.
WE could model the covariance between the two temperature variables using the kernel function.

$$\mathrm{cov}(x_1, x_2) = k(z_1, z_2) = \exp\left(-\frac{\|z_1 - z_2\|^7}{2l^2}\right)$$



$k(z_1, z_2) = kernel\ functione\ between\ two\ variable$
$z_1 = Location\ of\ city1$
$z_2 = Location\ of\ city2$
l = *parameter value*

This exponential function that decays between z1 and z2.
Z1 is close to Z2 => large Covariance
Z1 is away from Z2 => small covariance
The parameter value l is going to determine the result of covariance. If we pick the small l value, the result might be large.

We can also multiply the function with parameter 'a' to manage the cross -term

$$\mathrm{cov}(x_1, x_2) = k(z_1, z_2) = a.\exp\left(-\frac{\|z_1 - z_2\|^7}{2l^2}\right)$$

We can pick one of the methods to estimate the covariance.
Hence the parameter will be $\theta = \{a, l\}$ *or* $\theta = \{l\}$ based on which method is used for computing the covariance.

How to estimate the right parameter values , θ? Let's consider the two variables:
X1 = variable with no observation
X2 = variable with observation
We will use X2 to get a good estimation of θ. We will partition X2 into K partitions. Then we will run Cross-validation to estimate the good parameters of theta.

Cross- validation:
X_2^i = X2 datain partition 'I'
X_2^{-i} = all data in X2 that are not in partition 'i'

Now, let's compare the parameters θ_1 and θ_2 .

1) Estimate X_2^i from X_2^{-i} and $\theta_1 \Rightarrow \hat{x}_2^i(\theta_1)$
2) Use a performance function ($l\left(x_2^i, \hat{x}_2^i(\theta_1)\right)$) which returns a number value.
3) Performance of $\theta_1 = \sum_{i=1}^{k} l\left(x_2^i, \hat{x}_2^i(\theta_1)\right)$ Performance of $\theta_2 = \sum_{i=1}^{k} l\left(x_2^i, \hat{x}_2^i(\theta_2)\right)$
4) Pick the $\theta \in \{\theta_1, \theta_2\}$ with better performance