# Multivariate Distribution and Covariance

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST

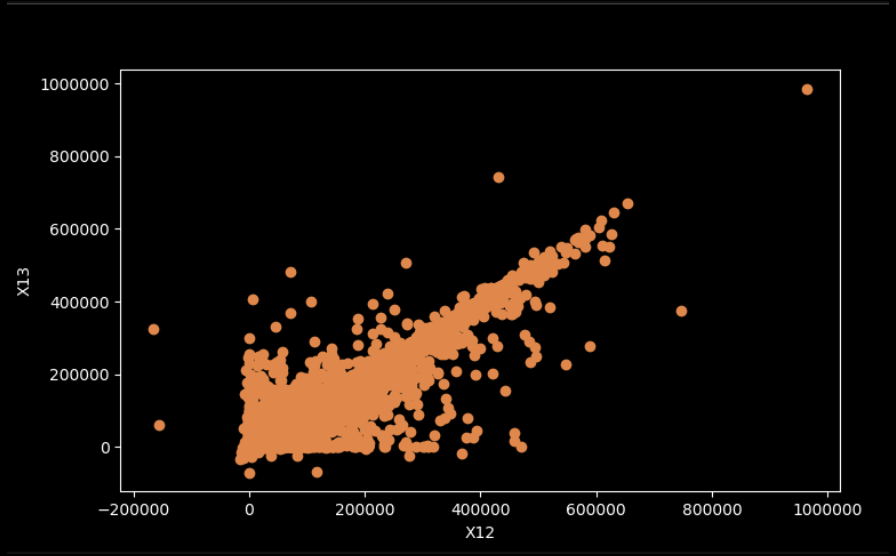
icmoon@kaist.ac.kr

# *Detour*) Dataset Introduction: Credit Score

- This research aimed at the case of customers' default payments in Taiwan
- Dependent variable :
  - a binary variable, default payment (Yes = 1, No = 0), as the response variable.
- Independent variable:
  - X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
  - X2: Gender (1 = male; 2 = female).
  - X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
  - X4: Marital status (1 = married; 2 = single; 3 = others).
  - X5: Age (year).
  - X6 - X11: History of past payment.
    - We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005.
    - The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
  - X12-X17: Amount of bill statement (NT dollar).
    - X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
  - X18-X23: Amount of previous payment (NT dollar)
    - X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.
- Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.
- Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

# Relation between Two Random Variables

- Scatter plot would be an intuitive visualization of two random variables.
- How to measure the relation between the two variables?
  - We need to model the distribution of multiple random variables
    - i.e. Multivariate Normal Distribution



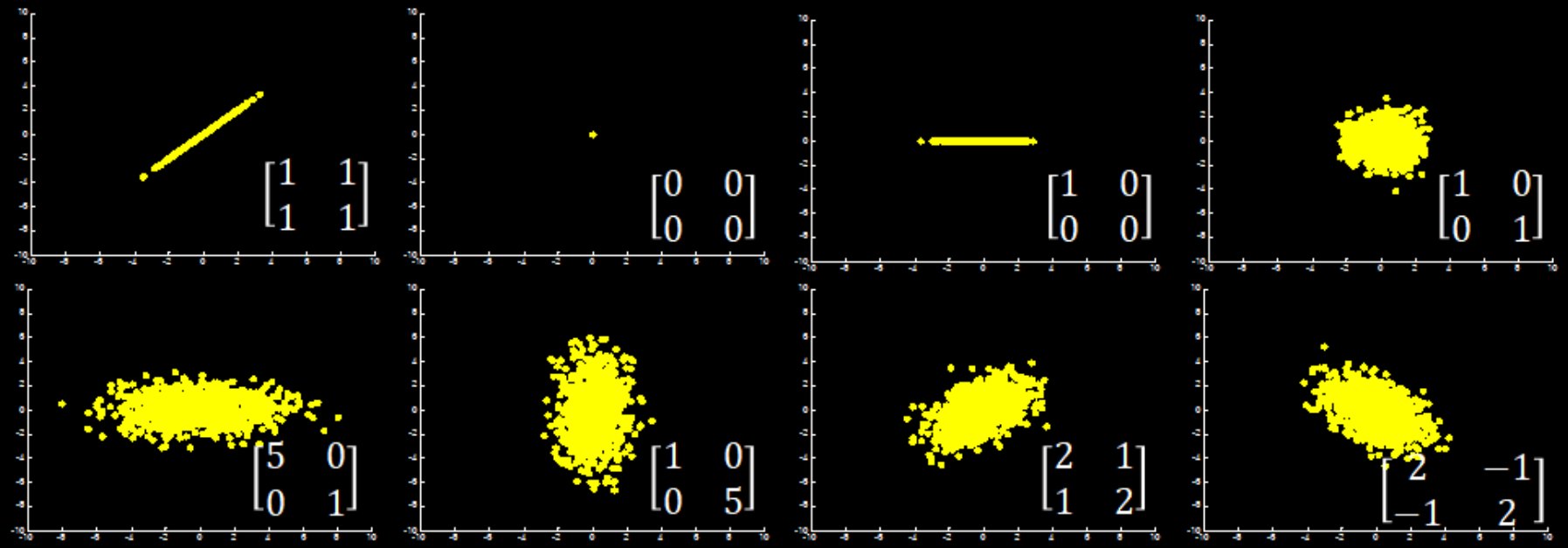

```
if __name__ == '__main__':
    objData = pd.read_excel('../data/ucidataset-credit score/default of credit card clients.xls',
                            skiprows=[1,2])
    plt.scatter(x=objData['X12'],y=objData['X13'])
    plt.xlabel('X12')
    plt.ylabel('X13')
    plt.show()
```

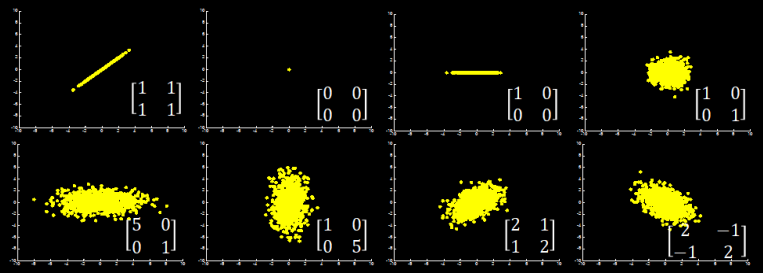
# Multivariate Gaussian Distribution

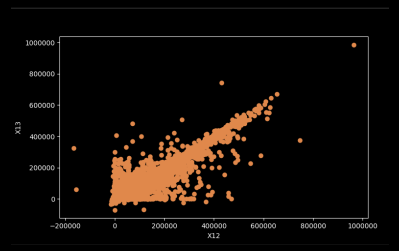
- Probability density function of the Gaussian distribution
  - $N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$
  - $\boldsymbol{N(x|\mu,\Sigma)} = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$
    - $\boldsymbol{\ln N(x|\mu,\Sigma)} = -\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \boldsymbol{C}$
    - $\boldsymbol{\ln N(X|\mu,\Sigma)} = -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}) + \boldsymbol{C}$
      - $\propto -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}Tr[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T]$
      - $= -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}Tr[\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}\big((\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T\big)]$
    - $\frac{d}{d\mu}\ln N(X|\mu,\Sigma) = 0 \rightarrow -\frac{1}{2}\times 2\times -1\times\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}(\boldsymbol{x}_n-\widehat{\boldsymbol{\mu}}) = 0 \rightarrow \widehat{\boldsymbol{\mu}} = \frac{\sum_{n=1}^{N}\boldsymbol{x}_n}{N}$
    - $\frac{d}{d\boldsymbol{\Sigma}^{-1}}\ln N(X|\mu,\Sigma) = 0 \rightarrow \widehat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\widehat{\boldsymbol{\mu}})(\boldsymbol{x}_n-\widehat{\boldsymbol{\mu}})^{\boldsymbol{T}}$
      - Beyond the scope of the course
      - Use "trace trick" and 1) $\frac{d}{dA}\log|A| = A^{-T}$, 2) $\frac{d}{dA}Tr[AB] = \frac{d}{dA}Tr[BA] = B^T$

# Samples of Multivariate Gaussian Distribution

- Samples of multivariate Gaussian distributions
  - With various covariance matrixes
  - Covariance matrix should a positive semi-definite matrix
    - $z^T \Sigma z \geq 0$ for every non-zero column vector $z$
    - $[a\ b]\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}a\\b\end{bmatrix} = \begin{bmatrix}a & 0\\0 & b\end{bmatrix}\begin{bmatrix}a\\b\end{bmatrix} = a^2 + b^2 \geq 0$ when a,b are non-zero






# Covariance Matrix

- Covariance Matrix
  - A matrix of variances between two random variables
    - A square matrix
    - Symmetric and positive semi-definite
      - $a^T \Sigma a \geq 0, for\ all\ a \in \mathrm{R}^n$
    - Diagonal values are variances
- Defination
  - $\Sigma_{X_i,X_j} = cov[X_i, X_j] = E[(X_i - E[X_i])(X_j - E[X_j])]$
- Relation to correlation
  - Pearson correlation coefficient, a.k.a. correlation coefficient
  - $\rho_{X_i,X_j} = \frac{cov(X_i,X_j)}{\sigma_{X_i}\sigma_{X_j}} = (diag(\Sigma))^{-\frac{1}{2}}\Sigma(diag(\Sigma))^{-\frac{1}{2}}$
  - Inverse of covariance matrix == precision matrix

# *Detour*) Some Properties of Expectations

- $E(k) = k$, if $k$ is a constant
- $E(kx) = kE(x)$, if $k$ is a constant and $x$ is a random variable
- $E(x + y) = E(x) + E(y)$, if $x$ and $y$ are both random variables
- $Var(x) = E\big(x - E(x)\big)^2$
- $$\begin{aligned} Var(x+y) &= E\big((x+y) - E(x+y)\big)^2 = E\big((x+y) - E(x) - E(y)\big)^2 \\ &= E\Big(\big(x - E(x)\big) + \big(y - E(y)\big)\Big)^2 \\ &= E\Big(\big(x - E(x)\big)^2 + \big(y - E(y)\big)^2 + 2\big(x - E(x)\big)\big(y - E(y)\big)\Big) \\ &= E\big(x - E(x)\big)^2 + E\big(y - E(y)\big)^2 + E\Big(2\big(x - E(x)\big)\big(y - E(y)\big)\Big) \\ &= Var(x) + Var(y) + 2Cov(x, y) \end{aligned}$$
  - a.k.a. Variance Sum Law
  - If $x$ and $y$ are independent, $Cov(x, y) = 0$
- We will visit $Cov(x, y)$ in the future by
  - Expanding the univariate Gaussian distribution to the multivariate Gaussian distribution

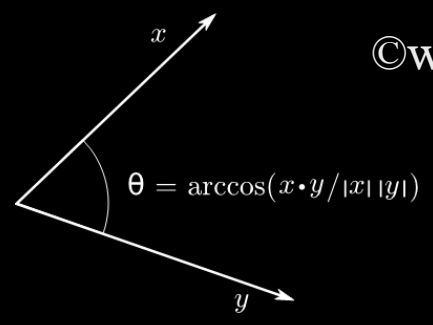
# Pearson Correlation

- $\rho_{X_i,X_j} = \frac{cov(X_i,X_j)}{\sigma_{X_i}\sigma_{X_j}} = \frac{E[(X_i-E[X_i])(X_j-E[X_j])]}{\sqrt{E[(X_i-E[X_i])(X_i-E[X_i])]}\sqrt{E[(X_j-E[X_j])(X_j-E[X_j])]}}$

$$= \frac{E[X_iX_j] - E[X_i]E[X_j]}{\sqrt{E[X_i^2] - (E[X_i])^2}\sqrt{E[X_j^2] - (E[X_j])^2}}$$

  - By utilizing $\mathrm{var}_n(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - E_n(X)^2$
- Range of $\rho_{X_i,X_j}$
  - Variance sum law
    - $Var(x+y) = Var(x) + Var(y) + 2Cov(x,y)$
  - $0 \le Var\left(\frac{X_i}{\sigma_{x_i}} \pm \frac{X_j}{\sigma_{x_j}}\right) = Var\left(\frac{X_i}{\sigma_{x_i}}\right) + Var\left(\frac{X_j}{\sigma_{x_j}}\right) \pm 2Cov\left(\frac{X_i}{\sigma_{x_i}}, \frac{X_j}{\sigma_{x_j}}\right)$

    $= \frac{1}{\sigma_{x_i}^2}Var(X_i) + \frac{1}{\sigma_{x_j}^2}Var(X_j) \pm 2Cov\left(\frac{X_i}{\sigma_{x_i}}, \frac{X_j}{\sigma_{x_j}}\right) = 1 + 1 \pm 2Cov\left(\frac{X_i}{\sigma_{x_i}}, \frac{X_j}{\sigma_{x_j}}\right) = 2 \pm 2Corr(X_i, X_j)$
  - $-1 \le \rho_{X_i,X_j} \le 1$




©wikipedia

# Pearson Correlation and its Geometric Interpretation

- $\rho_{X_i,X_j} = \frac{cov(X_i,X_j)}{\sigma_{X_i}\sigma_{X_j}} = \frac{E[(X_i-E[X_i])(X_j-E[X_j])]}{\sqrt{E[(X_i-E[X_i])(X_i-E[X_i])]}\sqrt{E[(X_j-E[X_j])(X_j-E[X_j])]}}$

$$= \frac{E[X_iX_j] - E[X_i]E[X_j]}{\sqrt{E[X_i^2] - (E[X_i])^2}\sqrt{E[X_j^2] - (E[X_j])^2}}$$

- Comparison to the cosine similarity
  - Cosine Similarity : $\cos(\theta) = \frac{x_i \cdot x_j}{\|x_i\|\|x_j\|} = \frac{\sum_{d=1}^{n} x_{i,d}x_{j,d}}{\sqrt{\sum_{d=1}^{n} {x_{i,d}}^2}\sqrt{\sum_{d=1}^{n} {x_{j,d}}^2}}$
  - Let's assume $E[X_i] = E[X_j] = 0$
    - $\rho_{X_i,X_j} = \frac{cov(X_i,X_j)}{\sigma_{X_i}\sigma_{X_j}} = \frac{E[X_iX_j]}{\sqrt{E[X_i^2]}\sqrt{E[X_j^2]}} = \cos(\theta)$
    - Another way to prove the range of $\rho_{X_i,X_j}$

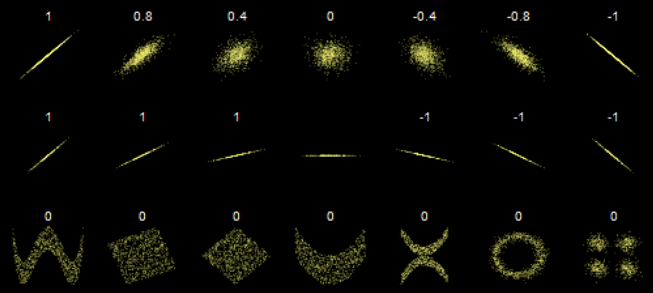
©wikipedia

# Some Properties of Pearson Correlation

- $\rho_{X_i,X_j} = Corr(X_i, X_j)$

$$= \frac{E[X_i X_j] - E[X_i]E[X_j]}{\sqrt{E[X_i^2] - (E[X_i])^2}\sqrt{E[X_j^2] - (E[X_j])^2}}$$

  - $\rho_{X_i,X_i} = 1$
  - $\rho_{X_i,X_j} = \rho_{X_j,X_i}$
  - $Corr(aX_i + b, X_j) = \text{sign}(a)Corr(X_i, X_j)$ for any $a, b \in R$
  - $-1 \leq \rho_{X_i,X_j} \leq 1$
  - $\left|\rho_{X_i,X_j}\right| = 1 \leftrightarrow X_j = aX_i + b, for\ some\ a, b \in R, and\ a \neq 0$
  - If $X_i$ and $X_j$ are independent then $\rho_{X_i,X_j} = 0$
  - If $\rho_{X_i,X_j} = 0$ then $X_i$ and $X_j$ need not be independent.
  - If $(X_i, X_j)$ is a bivariate Normal distribution with $\rho_{X_i,X_j} = 0$, then they are independent

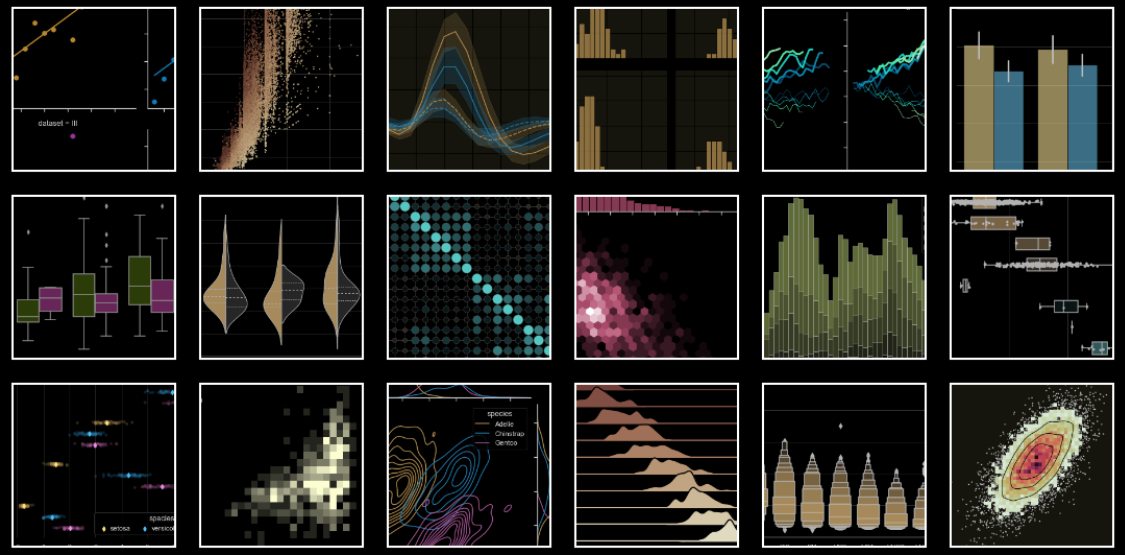


# *Detour*) seaborn

- *seaborn* is another data visualization library for Python
  - Higher level interface than the matplotlib
  - Good interoperation with pandas, matplotlib, numpy, etc
    - matplotlib becomes a basis for seaborn library
      - Usually, seaborn simplifies the graph drawing procedure in matplotlib
  - https://seaborn.pydata.org/

```
import seaborn as sns
```

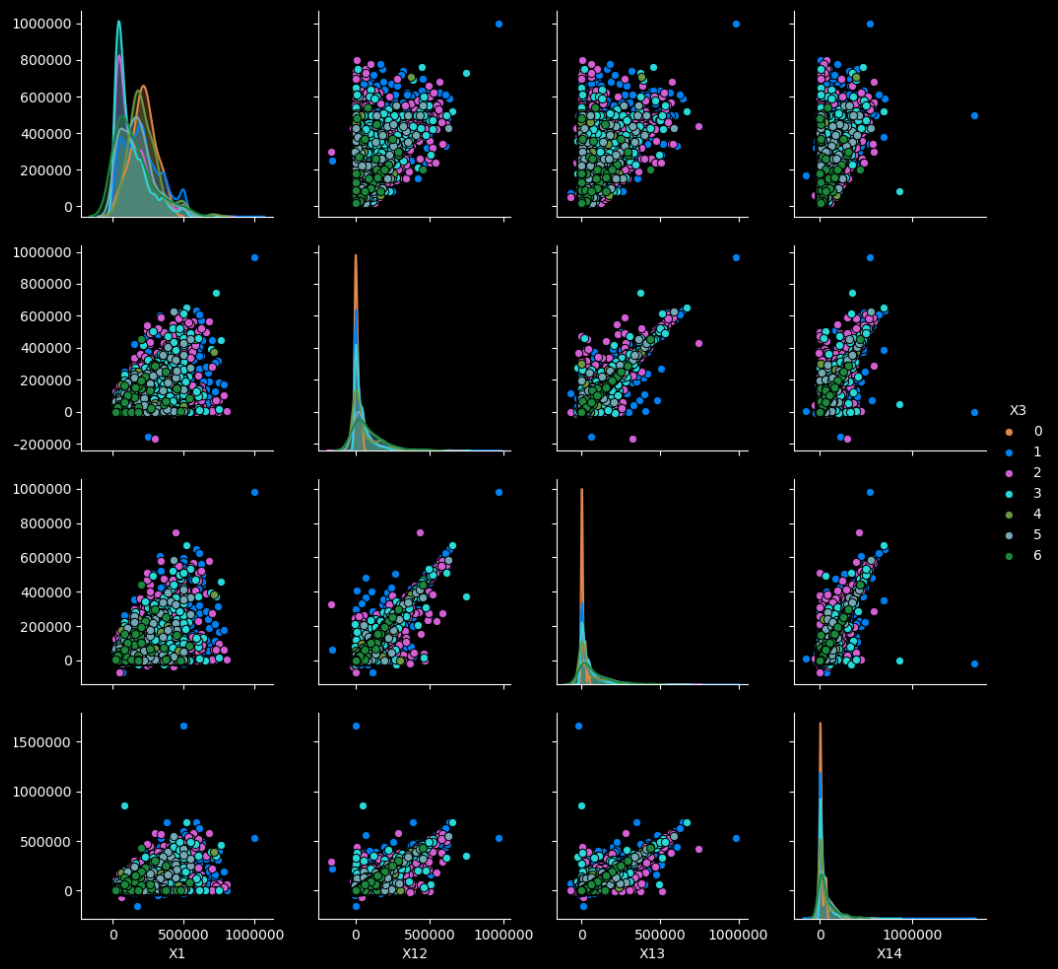
Example gallery #

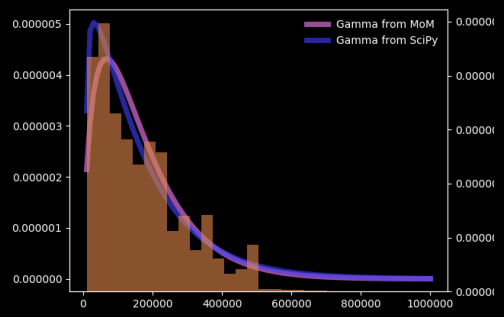


# Scatter Plot Matrix

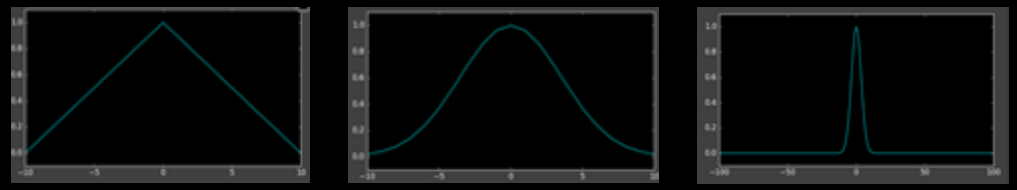
```
objFigure = sns.pairplot(objData,diag_kind='hist',hue='X3',vars=['X1','X12','X13','X14'])
plt.show()
```

- Symmetric aspect of the scatter plot
  - Diagonal can be chosen as a visualization on a single random variable
    - i.e. Histogram, Kernel Density Estimation, etc



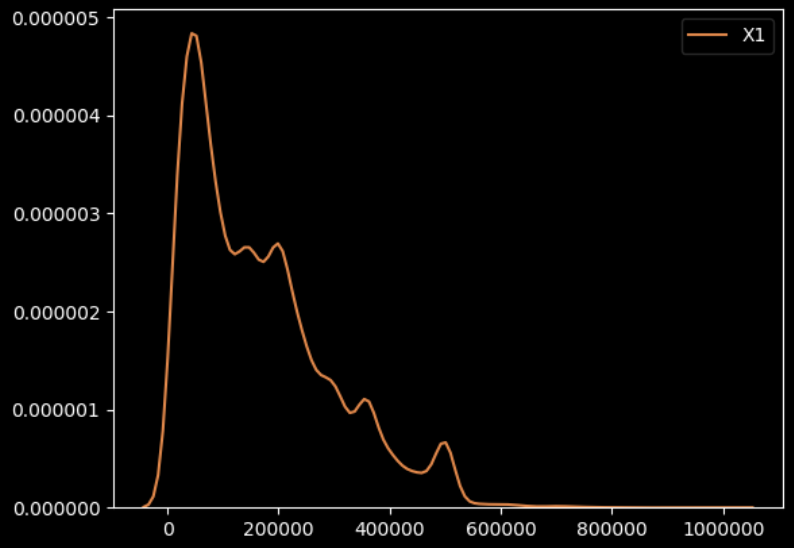



# *Detour*) Kernel Density Estimation

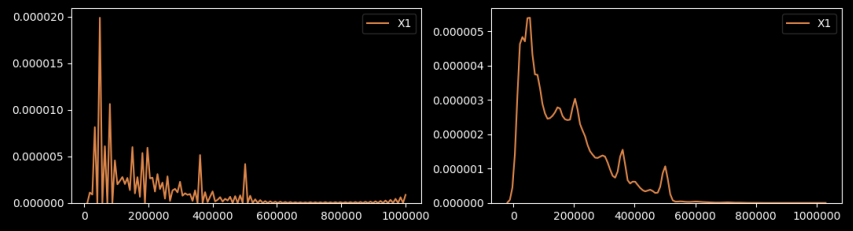


- $\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$
  - $K$ : Kernel function
    - $K_h(x - x_i) = \frac{1}{h}K(\frac{Z}{h})$
    - Coud be diverse selections
      - Uniform Kernel : $K\left(\frac{z}{h}\right) \propto \frac{1}{2}1_{\left\{-1<\frac{z}{h}<1\right\}}$
      - Gaussian Kernel : $K\left(\frac{Z}{h}\right) \propto \exp\left(-\frac{1}{2}\left(\frac{z}{h}\right)^2\right)$
- A form of non-parameteric density estimation
  - No selection on the prob. distribution
  - However, is there really no selection at all?

```
objFigure = sns.kdeplot(data=objData['X1'])
plt.show()
```




```
plt.figure(1)
plt.subplot(121)
objFigure = sns.kdeplot(data=objData['X1'], bw=1)
plt.subplot(122)
objFigure = sns.kdeplot(data=objData['X1'], bw=10000)
plt.show()
```



# Correlation Table

- Correlation table provides
  - The correlation between every possible combination of two random variables in the dataset
  - Correlation table is symmetric
    - Only lower or upper triangle is needed
    - Diagonal becomes the perfect correlation
- How to compute?
  - pandas provides the numpy-based correlation method, *corr*
    - Can call as a method of a data frame
    - There are more than "pearson" correlations
      - Kendall's tau, Spearman's rank correlation, etc

```
lstVariable = ['X1','X12','X13','X14']
objCorr = objData[lstVariable].corr(method='pearson')
print(objCorr)
```

| | X1 | X12 | X13 | X14 |
|---|---|---|---|---|
| X1 | 1.000000 | 0.285414 | 0.278298 | 0.283219 |
| X12 | 0.285414 | 1.000000 | 0.951483 | 0.892277 |
| X13 | 0.278298 | 0.951483 | 1.000000 | 0.928325 |
| X14 | 0.283219 | 0.892277 | 0.928325 | 1.000000 |

# Spearman's Rank Correlation

- Ordinary correlation : $\rho_{X_i,X_j} = \frac{cov(X_i,X_j)}{\sigma_{X_i}\sigma_{X_j}}$
- Spearman's rank correlation : $r_s = \rho_{R(X_i),R(X_j)} = \frac{cov\left(R(X_i),R(X_j)\right)}{\sigma_{R(X_i)}\sigma_{R(X_j)}}$
  - $R(X_i)$ : the rank of $X_i$ out of $X$
  - If we assume that $n$ ranks are all distinct integers, the Spearman's rank correlation becomes
    - $r_s = 1 - \frac{6\sum_{i=1}^{n} d_{i,j}^2}{n(n^2-1)}$, $d_i = R(X_i) - R(X_j)$
- When to utilize?
  - Pearson's correlation measures the linear relation between two random variables
  - Spearman's rank correlation measures the monotonic relation without linearity
    - Kendall's tau is another type of the rank correlation for the ordinal associations.