PGE 383
Feature Transformations

Lecture outline . . .

- **Feature Transformations**
- **Feature Transformations Examples**

Michael Pyrcz, The University of Texas at Austin

Motivation for Feature Transformations

- **There are many reasons that we may want to perform feature transformations.**
  - **the make the features consistent for visualization and comparison**
  - **to avoid bias or impose feature weighting for methods (e.g. k nearest neighbours regression) that rely on distances calculated in predictor feature space**
  - **the method requires the variables to have a specific range or distribution**
    - » artificial neural networks may require all features to range from [-1,1]
    - » partial correlation coefficients require a Gaussian distribution.
    - » statistical tests may require a specific distribution
    - » geostatistical sequential simulation requires an indicator or Gaussian transform

PGE 383
Feature Transformations

Lecture outline . . .

- **Feature Transformations**

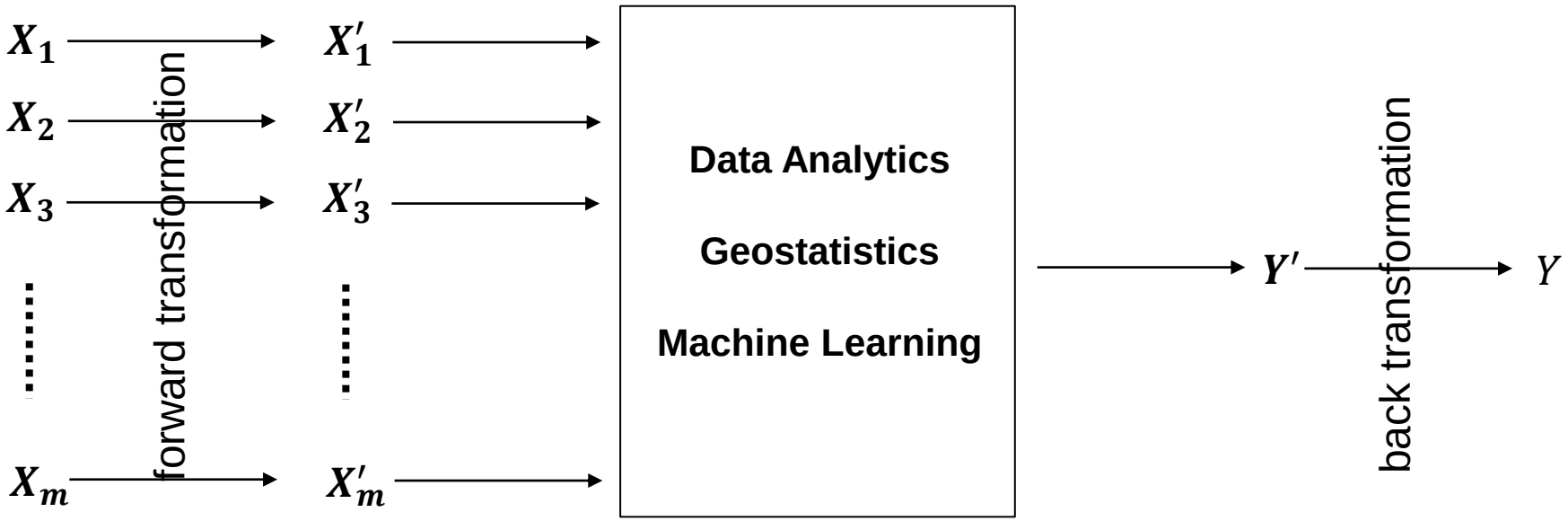
Michael Pyrcz, The University of Texas at Austin

What is a Feature Transformation?

- The application a transformation applied to the feature

$$x'_\alpha = f(x_\alpha)$$

- May be applied to a predictor feature prior to input into a predictive model
- May be applied to a response feature output from a predictive model
- May be applied to any feature to improve an inferential workflow
- Could be just applied to improve data visualization



An illustration of feature transformations to support data analytics, geostatistics and machine learning.

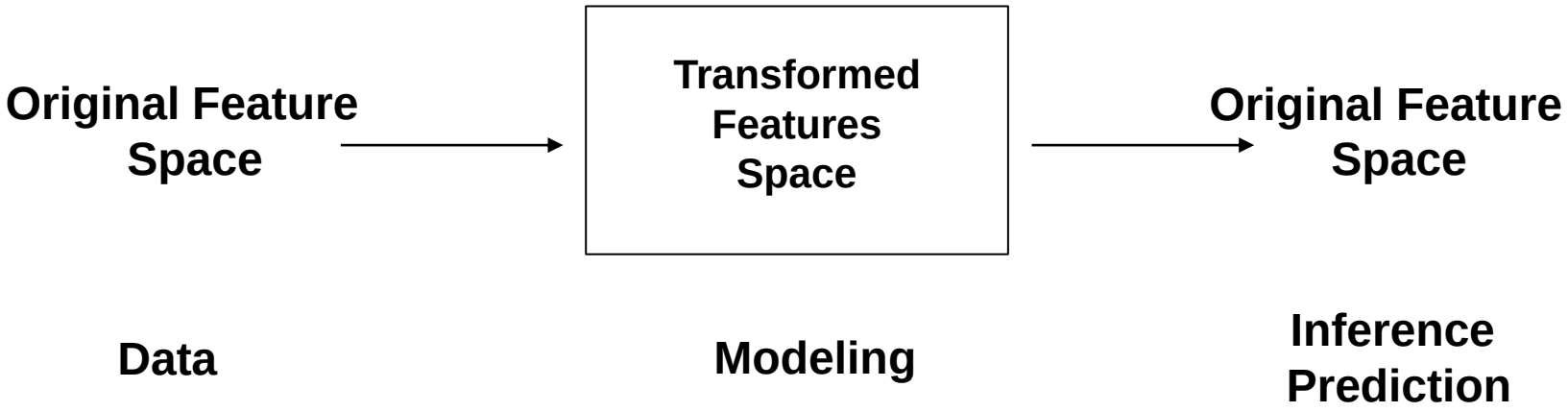
What is a Feature Transformation?

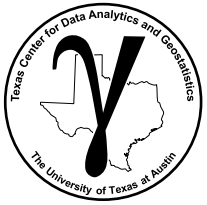
- Working in transformed space:

transform

$$x'_\alpha = f(x_\alpha)$$

transformed original





Feature Transformations

- We will start with very simple transformation and move to more complicated ones
- In general, this topic is not complicated and may not be very interesting!
- But, feature transformations are common in many data analytics and machine learning workflows
- You'll learn what they are and how to do them in Python with standard packages

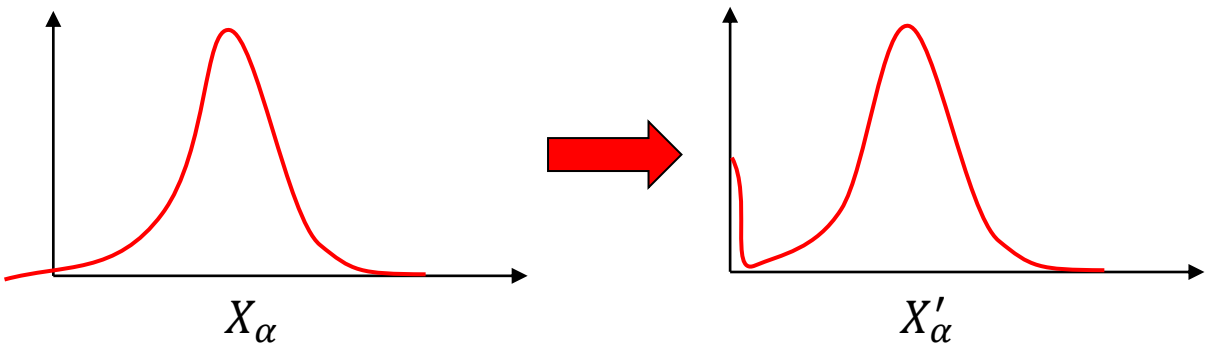
Feature Truncation

- Due to measurement error or imprecision of methods and workflows, it is possible to have feature values that are implausible
  - e.g. negative porosity, percentages outside of [0%, 100%] etc.
- We may also have outliers that exceed the range of the majority of the data set

- Truncation is the following operation:

$$x'_\alpha = \min(x_\alpha, x_t) \qquad \text{or} \qquad x'_\alpha = max(x_\alpha, x_t)$$

set the sample value to a threshold if less than or greater than.



Example of truncation of a feature distribution.

Feature Truncation

Methods that require truncation:

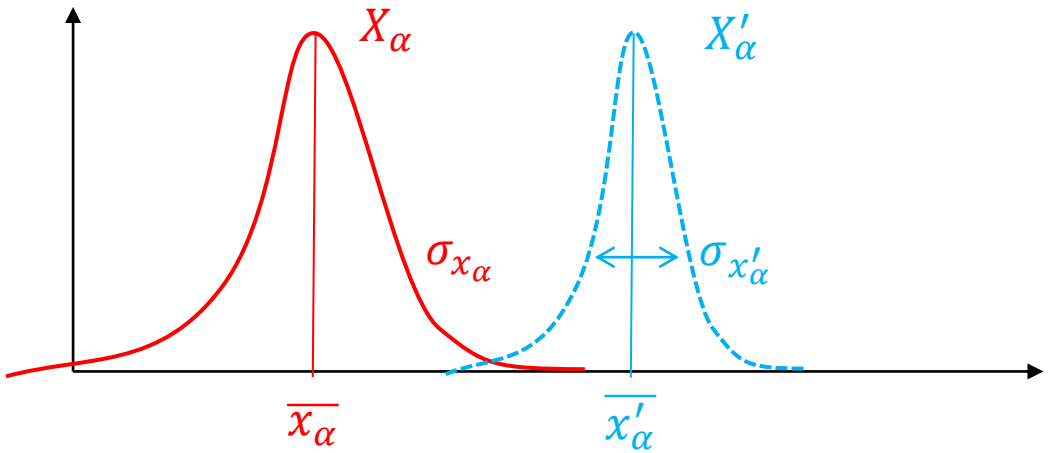
- feature values outside physical constraints (negative values, porosity exceeding geomechincal constraints)

- compositional data like mineral grades that are positive and sum to 100%

Affine Correction

- **The affine correction is the transform of the feature distribution to a new mean and variance.**
  - this is a shift and stretch / squeeze of the original property distribution
  - assumes no shape change

$$x'_\alpha = \frac{\sigma_{x'_\alpha}}{\sigma_{x_\alpha}} \cdot (x_\alpha - \overline{x_\alpha}) + \overline{x'_\alpha}$$



Example of affine correction of a feature distribution.

Affine Correction

Methods that require affine correction:

- debiased feature distributions, e.g. calculate a feature declustered mean and shift the distribution to the new mean

- bootstrap for the uncertainty in the feature mean and then shift the distribution mean to the P10 and P90 mean for low and high cases

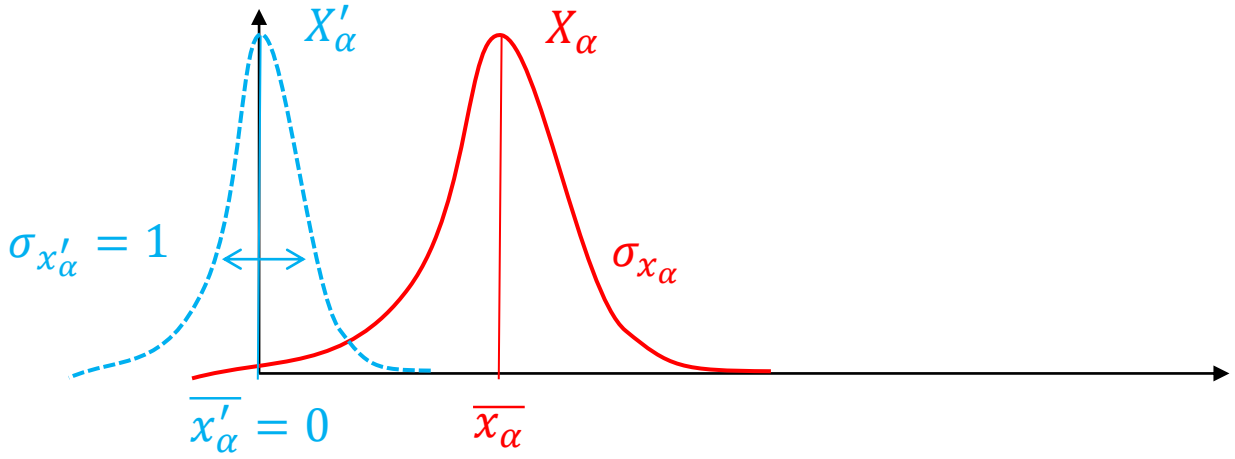
Standardization

- **Standardization is the transform of the feature distribution to a mean of 0 and variance of 1.**
  - this is a shift and stretch / squeeze of the original property distribution
  - assumes no shape change
  - specific form of the affine correction

$$x'_\alpha = \frac{\sigma_{x'_\alpha}}{\sigma_{x_\alpha}} \cdot (x_\alpha - \overline{x_\alpha}) + \overline{x'_\alpha}$$

($\sigma_{x'_\alpha} = 1$, $\overline{x'_\alpha} = 0$)

$$x'_\alpha = \frac{1}{\sigma_{x_\alpha}} \cdot (x_\alpha - \overline{x_\alpha})$$

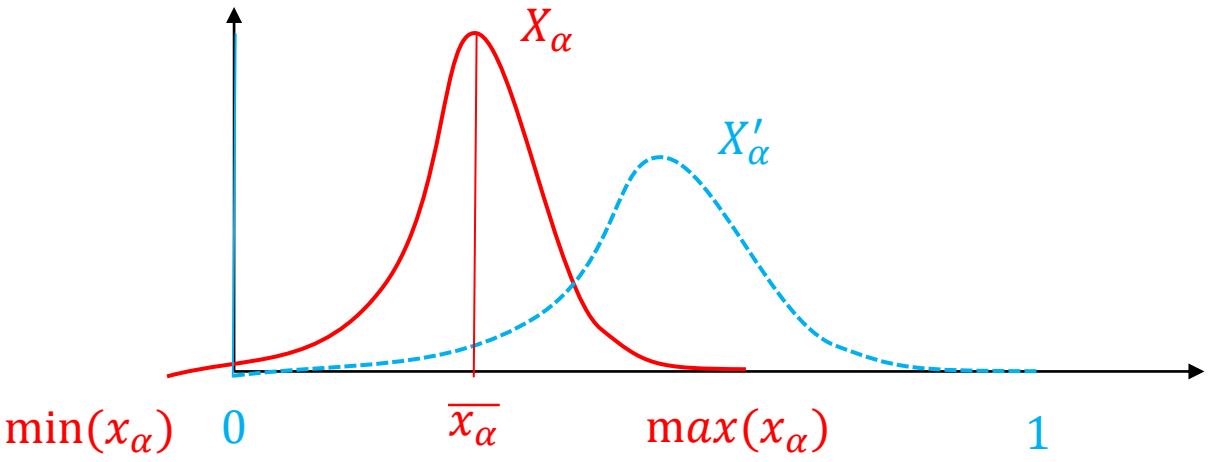


Example of standardization of a feature distribution.

Normalization
Min / Max Transform

- **Normalization is the transform of the feature distribution to a min of 0 and max of 1 (sometimes -1 to +1)**
  - this is a shift and stretch / squeeze of the original property distribution
  - assumes no shape change

$$x'_\alpha = \frac{x_\alpha - \min(x_\alpha)}{max(x_\alpha) - \min(x_\alpha)}$$



Example of normalization of a feature distribution.

Standardization and Normalization

Methods that require standardization and normalization:

- k-means clustering, k-nearest neighbour regression
- β coefficient's for feature ranking
- standardized variograms
- artificial neural networks forward transform of predictor features and back transform of response features

L1/L2 Normalizer

- **L1 / L2 Normalizer is performed across features over individual samples to constrain the sum**
  - The L1 Norm has the following constraint across samples

$$\sum_{\alpha=1}^{m} x'_{i,\alpha} = 1.0, \qquad i = 1, \ldots, n$$

  - The L1 normalizer transform:

$$x'_{i,\alpha} = \frac{x_{i,\alpha}}{\sum_{\alpha=1}^{m} x_{i,\alpha}}$$

  - The L2 Norm has the following constraint across samples

$$\sum_{\alpha=1}^{m} \left(x'_{i,\alpha}\right)^2 = 1.0, \qquad i = 1, \ldots, n$$

  - The L2 normalizer transform:

$$x'_{i,\alpha} = \sqrt{\frac{\left(x_{i,\alpha}\right)^2}{\sum_{\alpha=1}^{m} \left(x_{i,\alpha}\right)^2}}$$

  - applied in text classification and clustering, and L1 for compositional data

Binary / Indicator Transform

Indicator coding is transforming a feature to a probability relative to a category or a threshold.

- If $I\{\mathbf{u}: z_k\}$ is an indicator for a categorical variable,
  - What is the probability of a realization equal to a category?

$$I(\mathbf{u};\boldsymbol{z_k}) = \begin{cases} 1, & \text{if } Z(\mathbf{u}) = \boldsymbol{z_k} \\ 0, & \text{otherwise} \end{cases}$$

  - e.g. given threshold, $z_2 = 2$, and data at $\mathbf{u}_1, z(\mathbf{u}_1) = 2$, then $I\{\mathbf{u}_1; z_2\} = 1$
  - e.g. given threshold, $z_1 = 1$, and a RV away from data, $\mathrm{Z}(\mathbf{u}_2)$ then $I\{\mathbf{u}_2; z_1\} = 0.25$

- If $I\{\mathbf{u}: z_k\}$ is an indicator for a continuous variable,
  - What is the probability of a realization less than or equal to a threshold?

$$I(\mathbf{u};\boldsymbol{z_k}) = \begin{cases} 1, & \text{if } Z(\mathbf{u}) \leq \boldsymbol{z_k} \\ 0, & \text{otherwise} \end{cases}$$

  - e.g. given threshold, $z_1 = 6\%$, and data at $\mathbf{u}_1, z(\mathbf{u}_1) = 8\%$, then $I\{\mathbf{u}_1; z_1\} = 0$
  - e.g. given threshold, $z_4 = 18\%$, and a RV, $\mathrm{Z}(\mathbf{u}_2) = N[16\%, 3\%]$ then $I\{\mathbf{u}_1; z_k\} = 0.75$

Binary / Indicator Transform

Example of indicator transforms for a categorical variable.

Original Data	$I\{\mathbf{u}_\alpha; z_1 = 1\}$	$I\{\mathbf{u}_\alpha; z_2 = 2\}$	$I\{\mathbf{u}_\alpha; z_3 = 3\}$
$z(\mathbf{u}_1) = 3$	0	0	1
$z(\mathbf{u}_2) = 1$	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$z(\mathbf{u}_n) = 2$	0	1	0

Example of indicator transform of a categorical feature.

Our $z(\mathbf{u}_\alpha)$, $\alpha = 1, \ldots, n$, data become $k = 1, \ldots, K$ sets of n data, a new variable that indicates the probability of being each category.

Binary / Indicator Transform

Example of indicator transforms for a continuous variable.

Original Data	$I\{\mathbf{u}_\alpha; z_1 = 5\%\}$	$I\{\mathbf{u}_\alpha; z_2 = 10\%\}$	$I\{\mathbf{u}_\alpha; z_3 = 15\%\}$
$z(\mathbf{u}_1) = 12\%$	0	0	1
$z(\mathbf{u}_2) = 4\%$	1	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$z(\mathbf{u}_n) = 17\%$	0	0	0

Example of indicator transform of a continuous feature.

Our $z(\mathbf{u}_\alpha)$, $\alpha = 1, \ldots, n$, data become $k = 1, \ldots, K$ sets of n data, a new variable that indicates the probability of being less than or equal to each threshold.

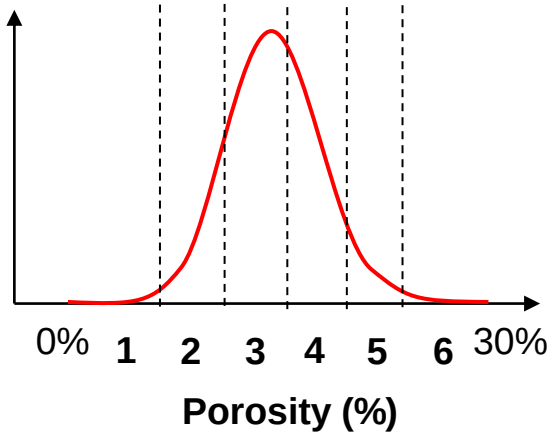
Binary / Indicator Transform

Methods that require binary / indicator transform:

- indicator variograms, indicator kriging and indicator simulation
- indicator maps
- environmental and economics thresholds and modeling probabilities of occurance

K Bins Discretization

- Bin the range of the feature into K bins
- Then for each sample assignment of a value of 1 if the sample is within a bin and 0 if outsize the bin
  - strategies include uniform width bins (uniform) and uniform number of data in each bin (quantile)

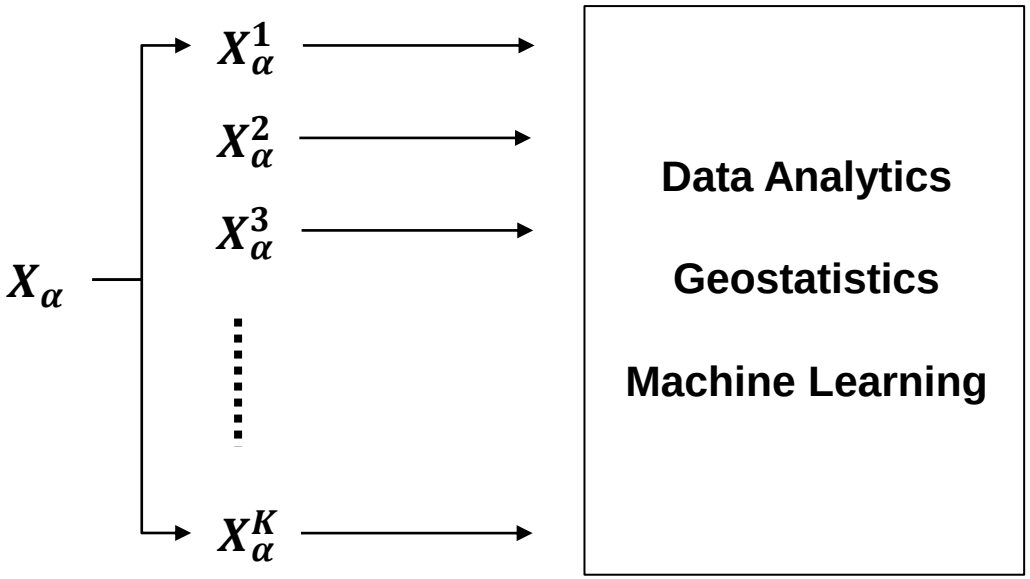


x_α	x_α^1	x_α^2	x_α^3	x_α^4	x_α^5	x_α^6
2%	1	0	0	0	0	0
16%	0	0	0	1	0	0
26%	0	0	0	0	0	1
8%	0	1	0	0	0	0

Simple example of K bins discretization

K Bins Discretization

- What is K bins discretation?
  - A probability coding, probability of the sample existing in each bin, could integrate sample uncertainty
  - A form of basis expansion (more during support vector machines)



Probability coding / basis expansion

K Bins Discretization

Methods that require K bins discretization:

- basis expansion to work in a higher dimensional space
- continuous to categorical for categorical methods such as naïve Bayes classifier
- histogram construction and Chi-square test for difference in distributions
- mutual information

Gaussian Anamorphosis

- Quantile transformation to a Gaussian distribution.
- Mapping feature values through their cumulative probabilities.

$$y = G_y^{-1}\left(F_x(x)\right)$$

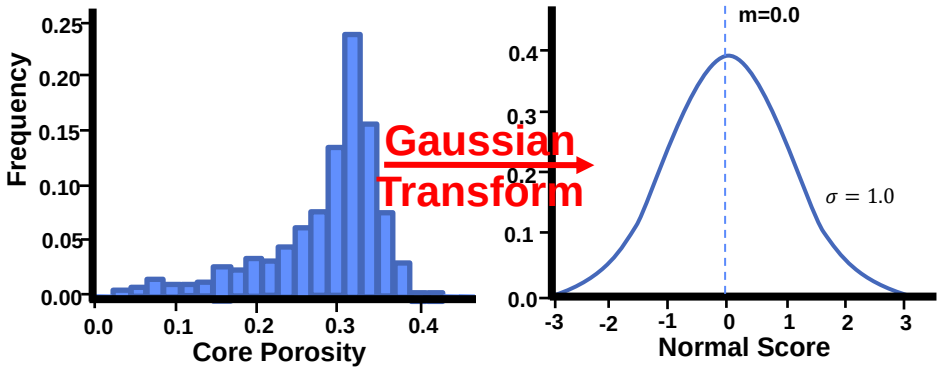
- where F_x is the original feature cumulative distribution function (CDF) and G_y is the Gaussian CDF
- Gaussian probability density function

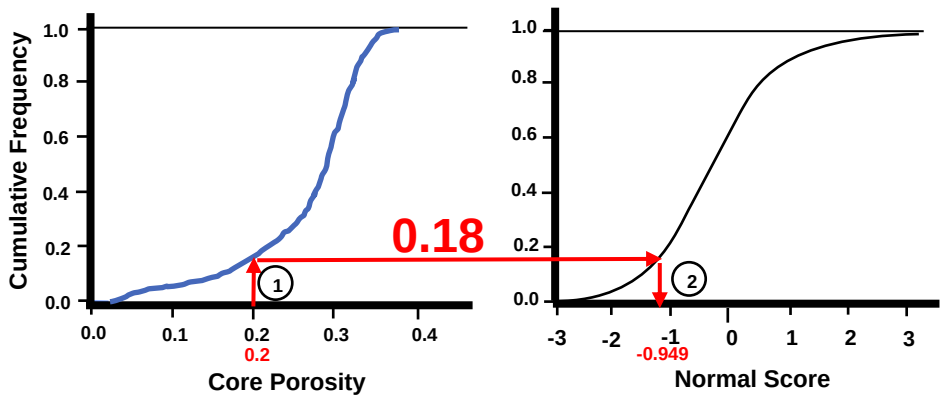
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

- Gaussian CDF

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] dy$$

$$-\infty < x < +\infty$$





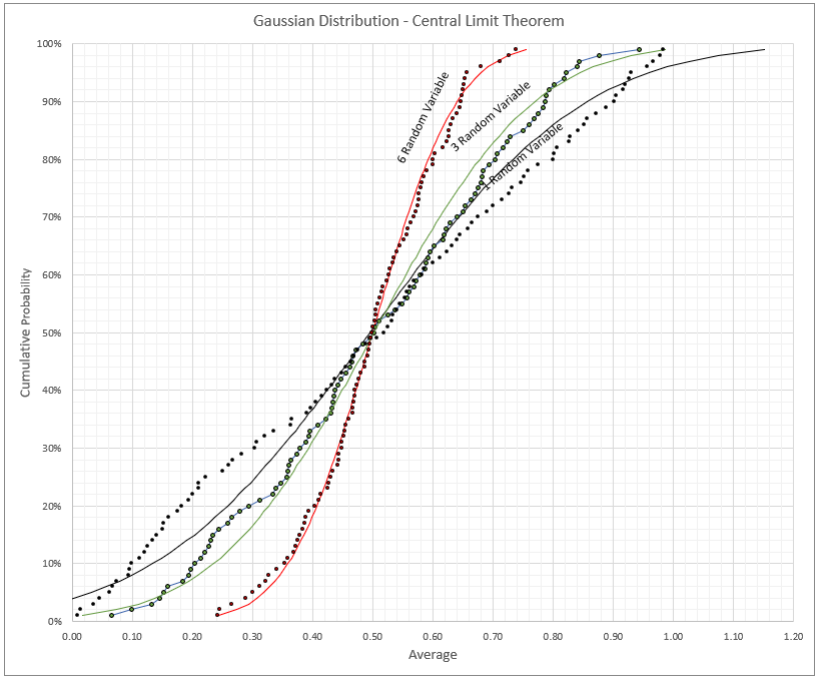
Gaussian Anamorphosis

More on the Gaussian distribution

- Shorthand for a Normal Distribution is N[mean, st.dev], N(μ,σ^2).
- Much of “natural variation” / measurement error is Gaussian
- Parameterized fully by mean, variance and correlation coefficient (if multivariate)
- distribution is unbounded, no min nor max
  - extremes are very unlikely, some type of truncation is often applied

Gaussian Anamorphosis

- Central Limit Theorem
  - the summation / average of multiple random variables tends towards a Gaussian distributed
  - this occurs quickly with 3-4 independent variables
  - some reservoir properties may be Gaussian distributed (e.g. porosity is the average of pore space vs. grains over smaller volumes.



Experimental demonstration of the Central Limit Theorem

Gaussian Anamorphosis

The Multivariate Gaussian distribution:

$$f_X(x_1, \ldots, x_m) = \frac{exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^m |\Sigma|}}$$

where μ is the m vector of means and Σ is the $m \times m$ matrix of all pairwise covariances.

$$\mu = [\mu_1, \ldots, \mu_1] \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & C_{1,m} \\ \vdots & \ddots & \vdots \\ C_{m,1} & \cdots & \sigma_m^2 \end{bmatrix}$$

Very compact parameterization!

Gaussian Anamorphosis

The Marginal Distributions:

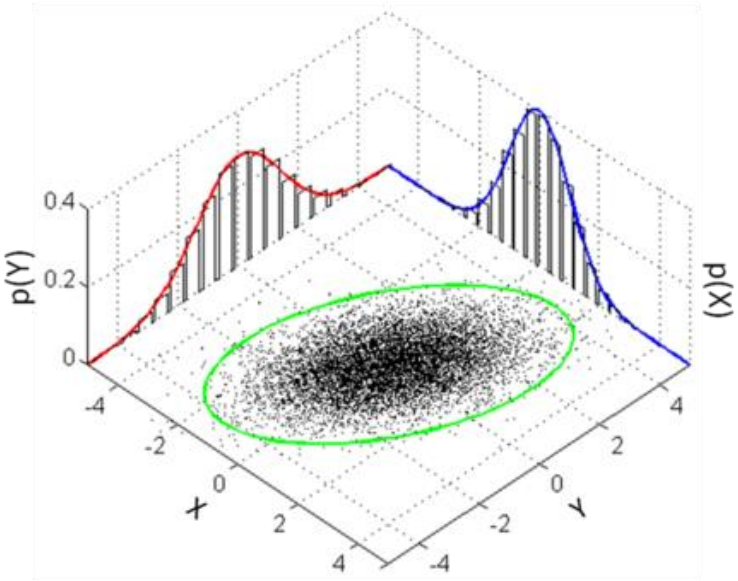
All marginal distributions are Gaussian:

$$f_{X_1}(x_1) \sim N(\mu_1, \sigma_1^2)$$

$$f_{X_2}(x_2) \sim N(\mu_2, \sigma_2^2)$$

$$\vdots$$

$$f_{X_m}(x_m) \sim N(\mu_m, \sigma_m^2)$$



Gaussian joint and marginal distributions.

Image from https://en.wikipedia.org/wiki/Multivariate_normal_distribution#/media/File:MultivariateNormal.png

Gaussian Anamorphosis

The Conditional Distributions:

All conditional distributions are Gaussian, we just show the bivariate case:

$$f_{X_1|X_2}(x_1 \,|X_2 = x_2) \sim N\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2), (1-\rho^2)\sigma_1^2\right)$$

- the conditional variance is homoscedastic, does not depend on the mean!

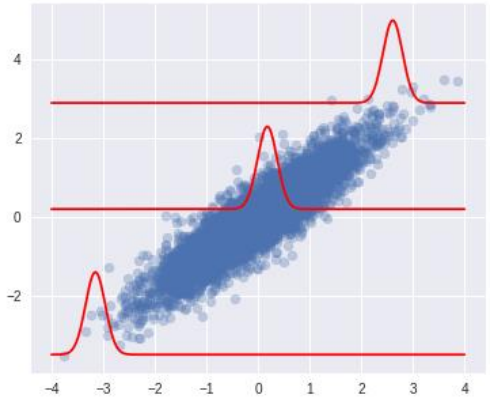


Image from https://stats.stackexchange.com/questions/385823/variance-of-conditional-multivariate-gaussian

Gaussian Anamorphosis

Methods that require a Gaussian distribution:

- Pearson product momment correlations completely characterize multivariate relationships when Gaussian
- partial correlations require bivariate Gaussian
- sequential simulation needs Gaussian to reproduce the global distribution
- student's t test for difference in means
- Chi-square distributions is derived from sum of squares of Gaussian distributed random variables

Uniform General Distribution Transform

- Quantile transformation to a uniform distribution (e.g cumulative probabilies).
- Mapping feature values through their cumulative probabilities.

$$y = F_y^{-1}\left(F_x(x)\right)$$

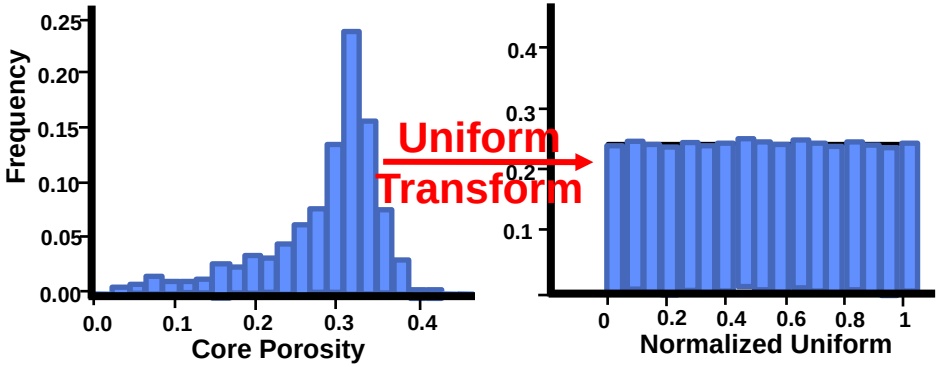
- where F_x is the original feature cumulative distribution function (CDF) and F_y is the uniform CDF
- uniform probability density function

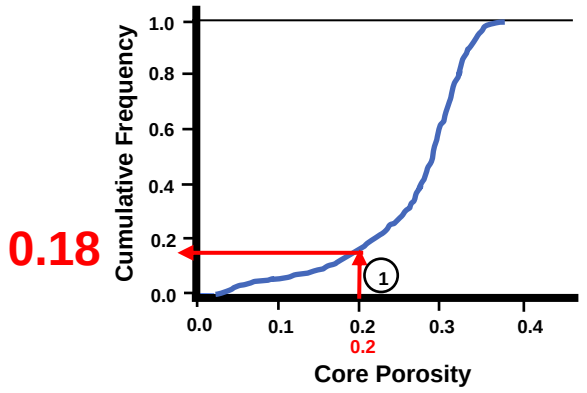
$$f_x(x) = \frac{1}{N} = constant$$

- Uniform CDF

$$F_x(x) = \frac{1}{N}x$$

$$x_{min} < x < x_{min}$$





PGE 383
Feature Transformations

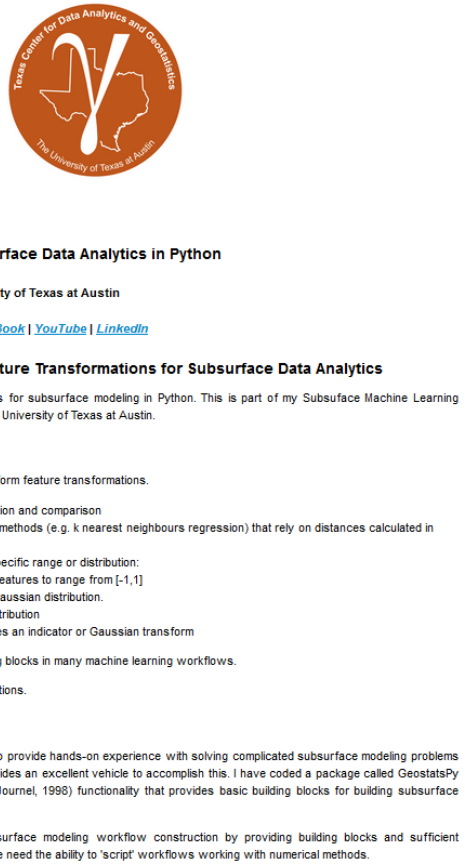
Lecture outline . . .

- **Feature Transformations Examples**

Michael Pyrcz, The University of Texas at Austin

Feature Transformations Demonstration in Python

Demonstration of feature transformations with a documented workflow.



Subsurface Data Analytics

Feature Transformations for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

Twitter | *GitHub* | *Website* | *GoogleScholar* | *Book* | *YouTube* | *LinkedIn*

Subsurface Machine Learning: Feature Transformations for Subsurface Data Analytics

Here's a demonstration of feature transformations for subsurface modeling in Python. This is part of my Subsuface Machine Learning Course at the Cockrell School of Engineering at the University of Texas at Austin.

Feature Transformations

There are many reasons that we may want to perform feature transformations.

- the make the features consistent for visualization and comparison
- to avoid bias or impose feature weighting for methods (e.g. k nearest neighbours regression) that rely on distances calculated in predictor feature space
- the method requires the variables to have a specific range or distribution:
  - artificial neural networks may require all features to range from [-1,1]
  - parital correlation coefficients require a Gaussian distribution.
  - statistical tests may require a specific distribution
  - geostatistical sequential simulation requires an indicator or Gaussian transform

Feature transformations is a common basic building blocks in many machine learning workflows.

- Let's learn how to perform feature transformations.

Objective

In the Stochastic Machine Learning class, I want to provide hands-on experience with solving complicated subsurface modeling problems with data analytics, machine learning. Python provides an excellent vehicle to accomplish this. I have coded a package called GeostatsPy with GSLIB: Geostatistical Library (Deutsch and Journel, 1998) functionality that provides basic building blocks for building subsurface modeling workflows.

The objective is to remove the hurdles of subsurface modeling workflow construction by providing building blocks and sufficient examples. This is not a coding class per se, but we need the ability to 'script' workflows working with numerical methods.

File SubsurfaceDataAnalytics_Feature_Transformation.ipynb at https://git.io/fj7ea.

Multivariate New Tools

Topic	Application to Subsurface Modeling
Curse of Dimensionality	Reduce problem to lowest dimension possible. *Feature ranking determined that porosity may be predicted from acoustic impedance and rock type alone.*
Feature Transformation	Apply feature transformations to improve the ability of your models to robustly infer patterns and predict away from training data. *Know what transformation are helpful and required for your modeling workflow.*

Michael Pyrcz, The University of Texas at Austin

PGE 383
Feature Selection

Lecture outline . . .

- **Feature Selection**
- **Feature Selection Hands-on**

Michael Pyrcz, The University of Texas at Austin