

Data Analytics and Geostatistics: Multivariate Analysis

Lecture outline . . .

- **Multivariate Analysis**
- **Joints and Conditionals**
- **Feature Selection**
- **Multivariate Estimation**

Introduction

Modeling Prerequisites

Spatial Estimation

Spatial Uncertainty

Multivariate, Spatial

Multivariate Analysis

Machine Learning

Novel Workflows

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin



Data Analytics and Geostatistics: Multivariate Analysis

Lecture outline . . .

- Multivariate Analysis

Introduction

Modeling Prerequisites

Spatial Estimation

Spatial Uncertainty

Multivariate, Spatial

Multivariate Analysis

Machine Learning

Novel Workflows

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin

Motivation for Multivariate Methods



- **We typically need to build reservoir models of more than one property of interest.**
 - Expanded by whole earth modeling, closing loops with forward models
 - Expanded by unconventional
- **Subsurface properties may include:**
 - Rock Classification: lithology, architectural elements, facies, depofacies
 - Petrophyscial: porosity, directional permeability, saturuations
 - Geophysical: density, p-wave and s-wave velocity
 - Gemeanical: compressibility / Poisson's ratio, Yong's modulus, brittleness, stress field
 - Paleo- / Time Control: fossil adundances, stratigraphic surfaces, ichnofacies, paleo-flow indicators

Curse of Dimensionality

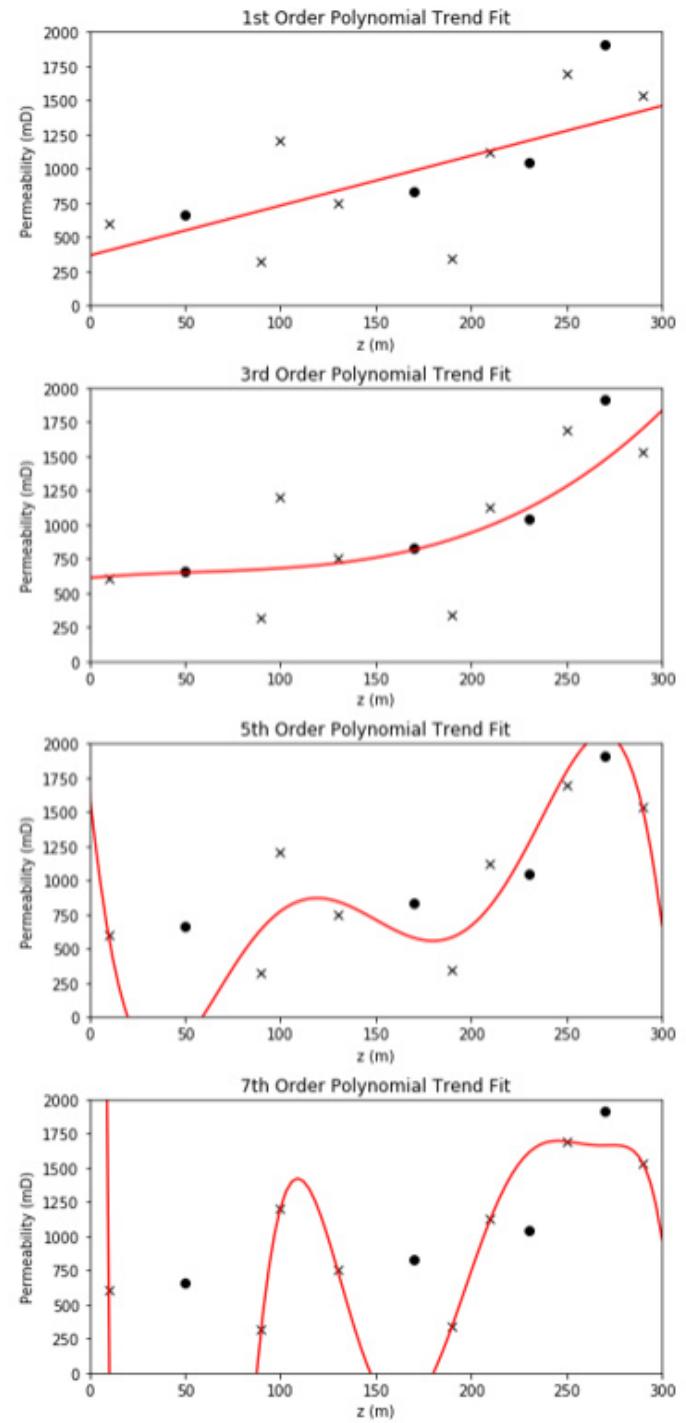


Working with more features / variables is harder!

1. More difficult to visualize
2. More data are required to infer the joint probabilities
3. Less coverage
4. More difficult to interrogate / check the model
5. More likely redundant
6. More complicated, more likely overfit

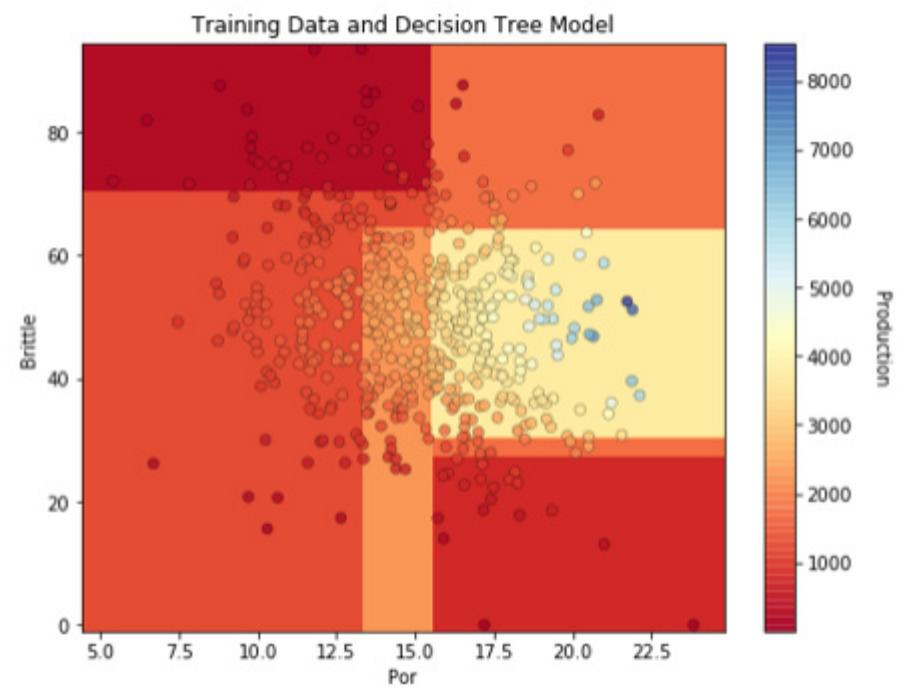
Visualization

- Consider this simple model:
 - 1 predictor feature
 - 1 response feature
- How's our model performing?
 - Accuracy in training and testing
- Range of Applicability?
 - Are we extrapolating?
- Overfit
 - Is the model defendable given the data?



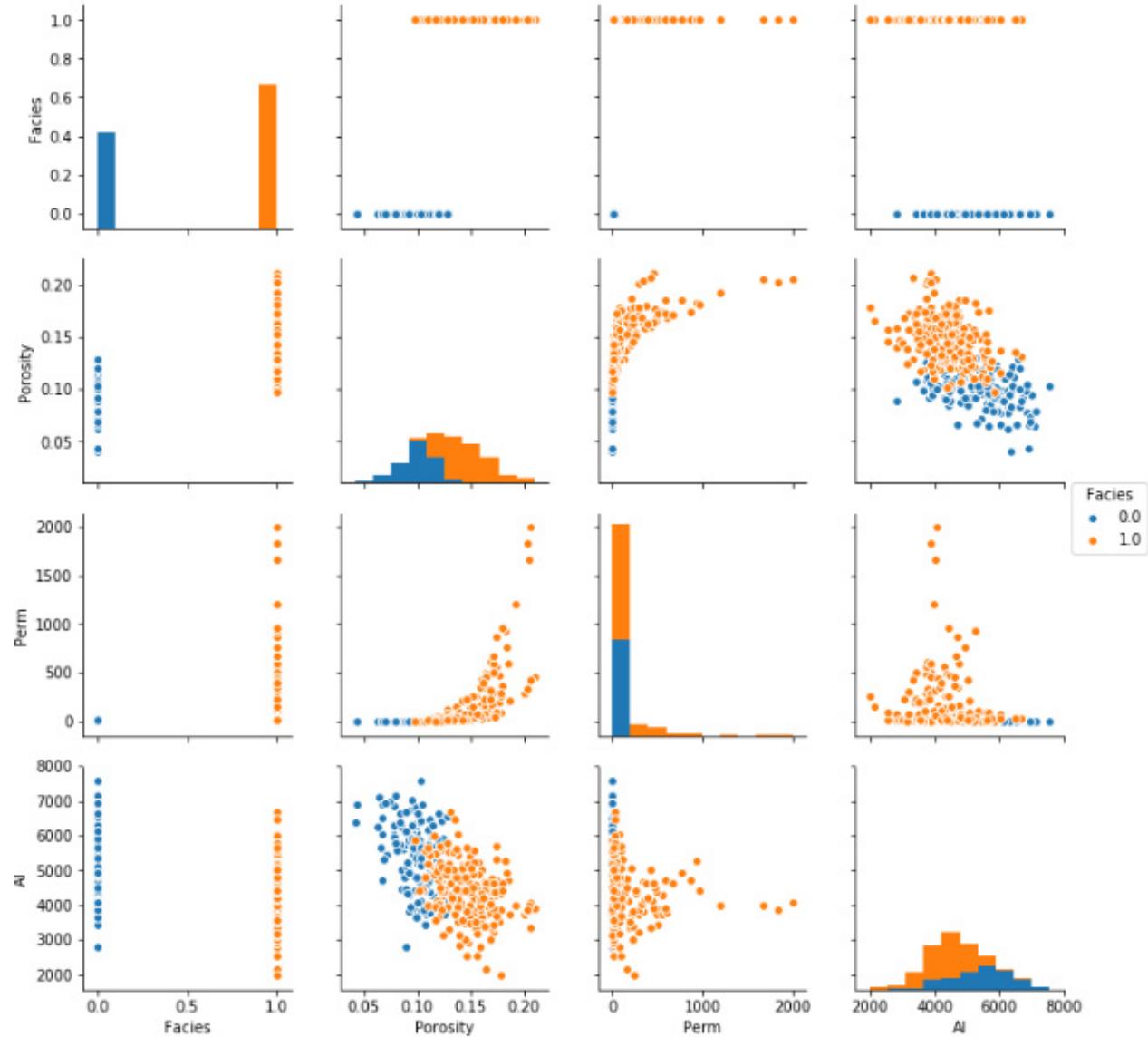
Visualization

- Consider this simple model:
 - 2 predictor features
 - 1 response feature
- How's our model performing?
 - Accuracy in training and testing
- Range of Applicability?
 - Are we extrapolating?
- Overfit
 - Is the model defendable given the data?

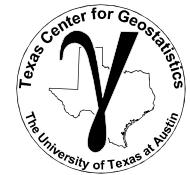


Visualization

- Consider this:
 - 4 predictor features
 - 1 response feature (not shown)
- What are the relationships between features?
- Are there constraints?



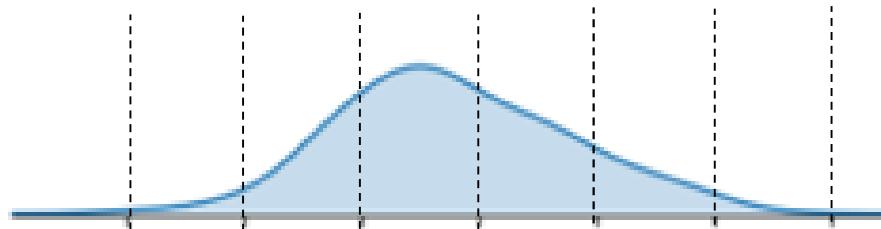
Inferring Joint Probabilities



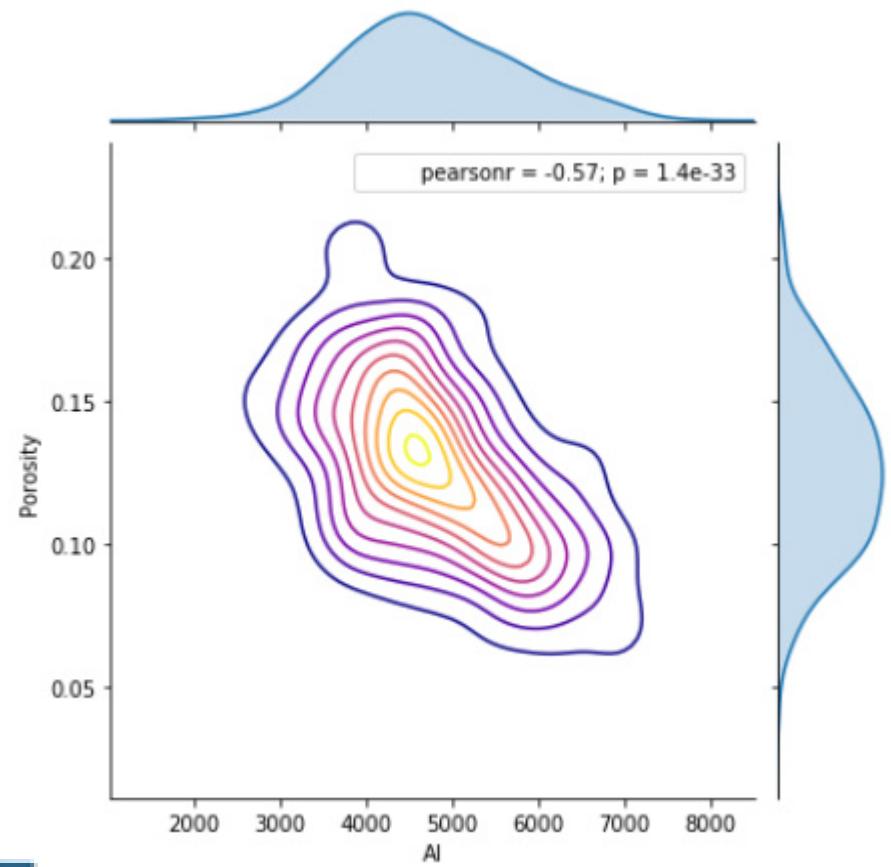
- Consider any joint probability:

$P(X_1 \cap \dots \cap X_m)$ the joint probability of X_1, \dots, X_m

- Let's start with 1 feature ($m=1$)

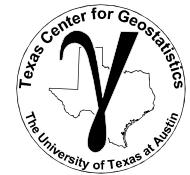


$$P(X_1^i \leq X \leq X_1^{i+1}) = \frac{n(X_1^i \leq X \leq X_1^{i+1})}{n}$$



In each bin we are estimating a probability!
10 data in each bin = 80 data?

Inferring Joint Probabilities



- Consider any joint probability:

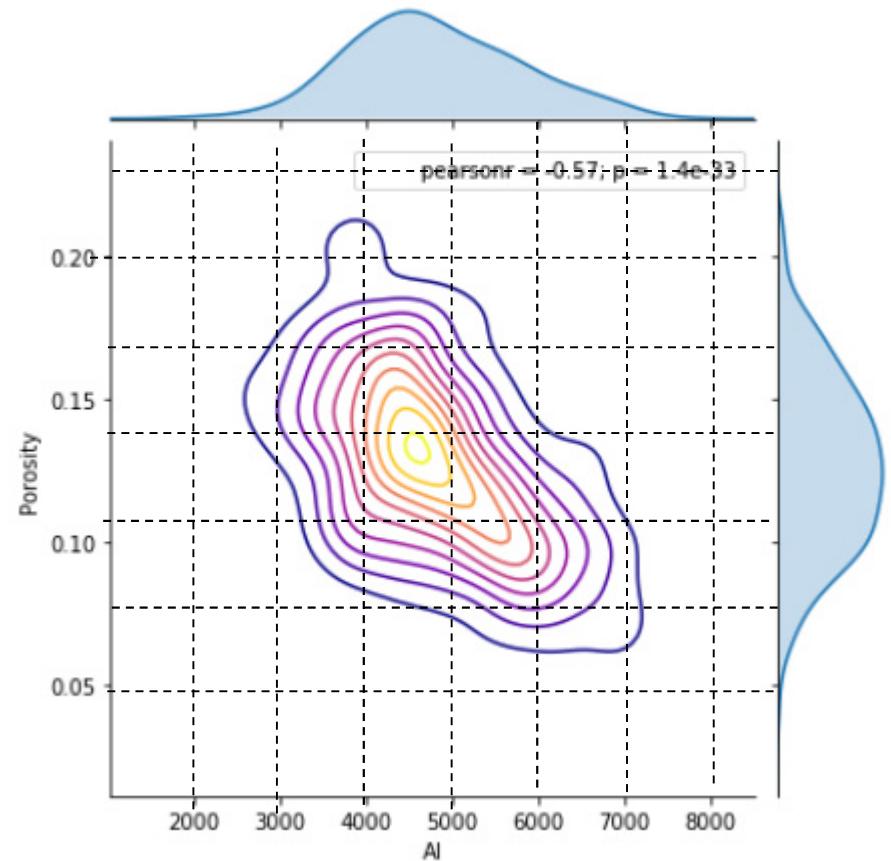
$P(X_1 \cap, \dots, \cap X_m)$ the joint probability of X_1, \dots, X_m

- Now move to 2 features ($m=2$)

$$P(X_1^i \leq X \leq X_1^{i+1}, X_2^j \leq X \leq X_2^{j+1}) = \frac{n(X_1^i \leq X \leq X_1^{i+1}, X_2^j \leq X \leq X_2^{j+1})}{n}$$

$$n = \text{Data/Bin} \cdot \text{Bins}^m$$

- This is optimistic, as it assumes uniform sampling

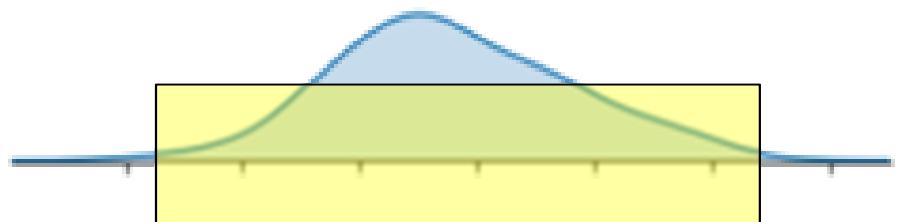


In each bin we are estimating a probability!
10 data in each bin = 640 data?

Coverage

Consider coverage:

- The range of the sample values
- The fraction of the possible solution space that is sampled.
- Let's return to 1 feature, and assume 80% coverage!
- That's pretty good right?



Coverage

Consider coverage:

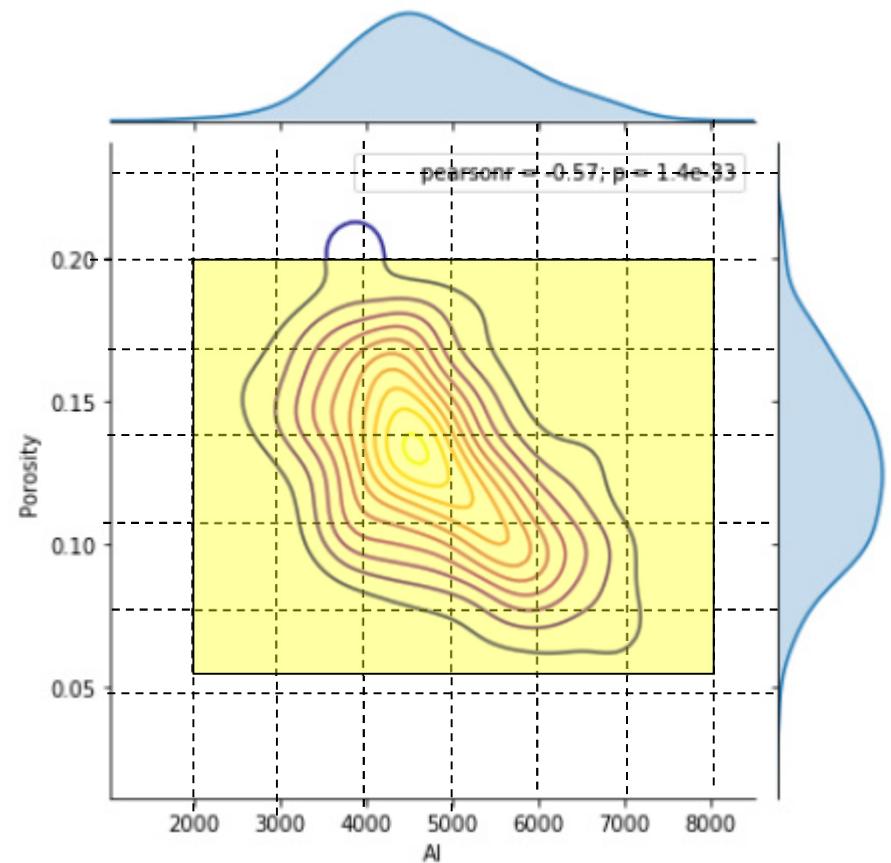
- Now let's move to 2 features, each with 80% coverage
- How much of the solution space is covered?

$$0.8^D, \quad e.g. 0.8^2 = 0.64$$

- Even with exponential increase in number of data:

$$n = Data/Bin \cdot Bins^m$$

coverage is decreasing as we increase the number of features!



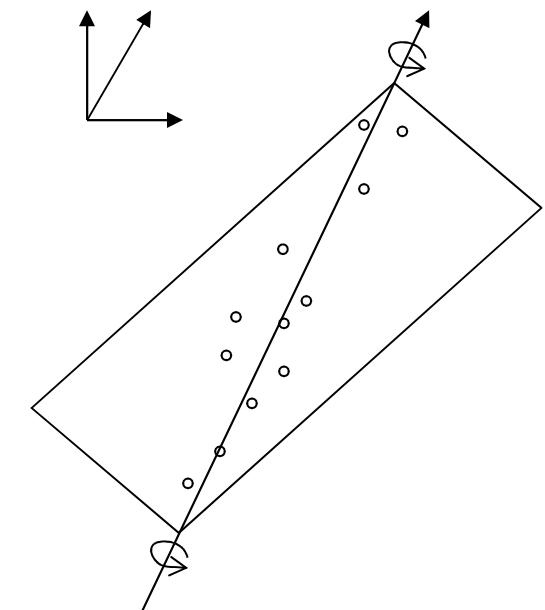
Multicollinearity Feature Redundancy



"the existence of such a **high degree of correlation between supposedly independent variables** being used to estimate a dependent variable that the contribution of each independent variable to variation in the dependent variable cannot be determined"

- Merriam-Webster Online Dictionary

"In statistics, **multicollinearity** (also collinearity) is a phenomenon in which one predictor variable in a **multiple regression** model can be linearly predicted from the others with a substantial degree of accuracy."



It is like fitting a plane to a line!

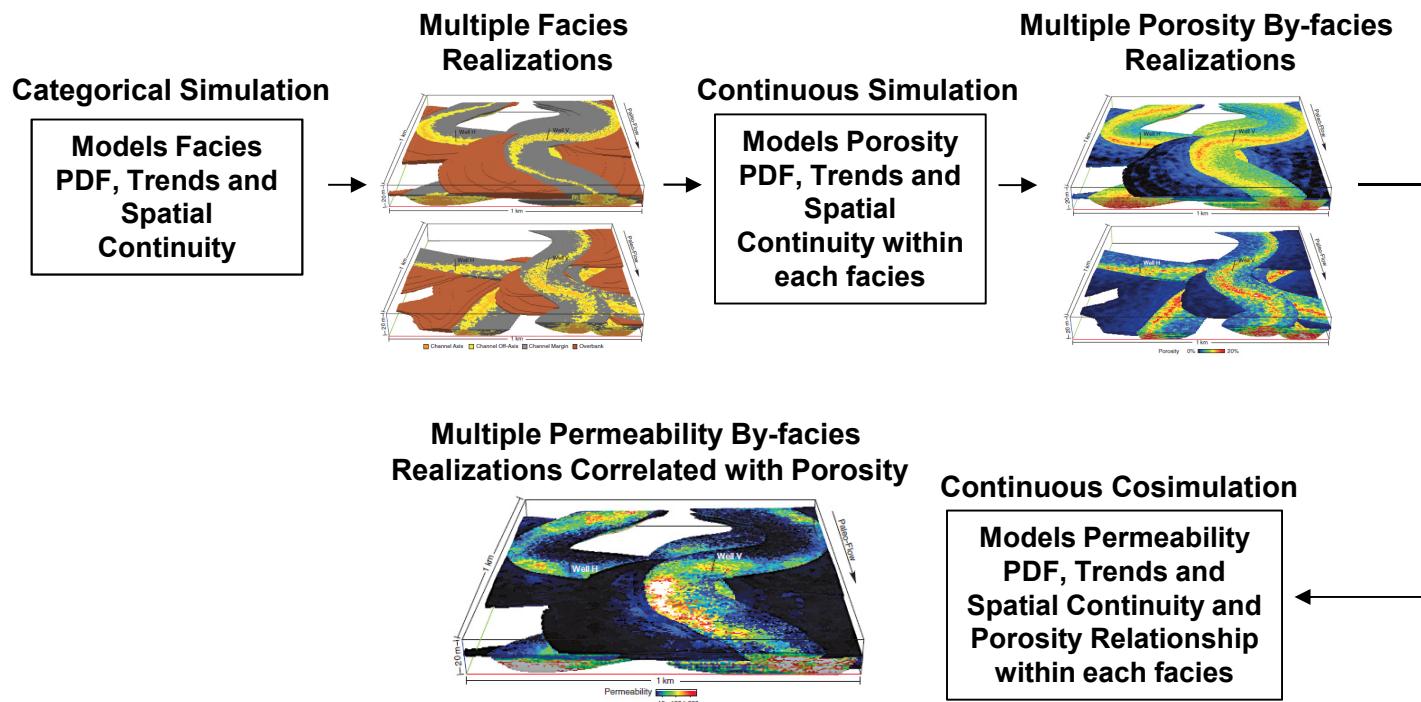
- Wikipedia

Motivation for Multivariate Methods



- **A Confession:**

- Standard geostatistical workflows are bivariate at most
 - » e.g. simulate permeability conditional to porosity



Motivation for Multivariate Methods



- **Emerging Multivariate Methods Include:**
 - Transforms – remove correlations and then model with independent variables and then back-transform to restore correlation (e.g. step-wise conditional transform).

This is beyond the scope of this course.

Bivariate Statistics

What is Bivariate Analysis?



- **Bivariate Analysis: Understand and Quantify the relationship between two variables**
 - Example: Relationship between porosity and permeability
 - How can we use this relationship?

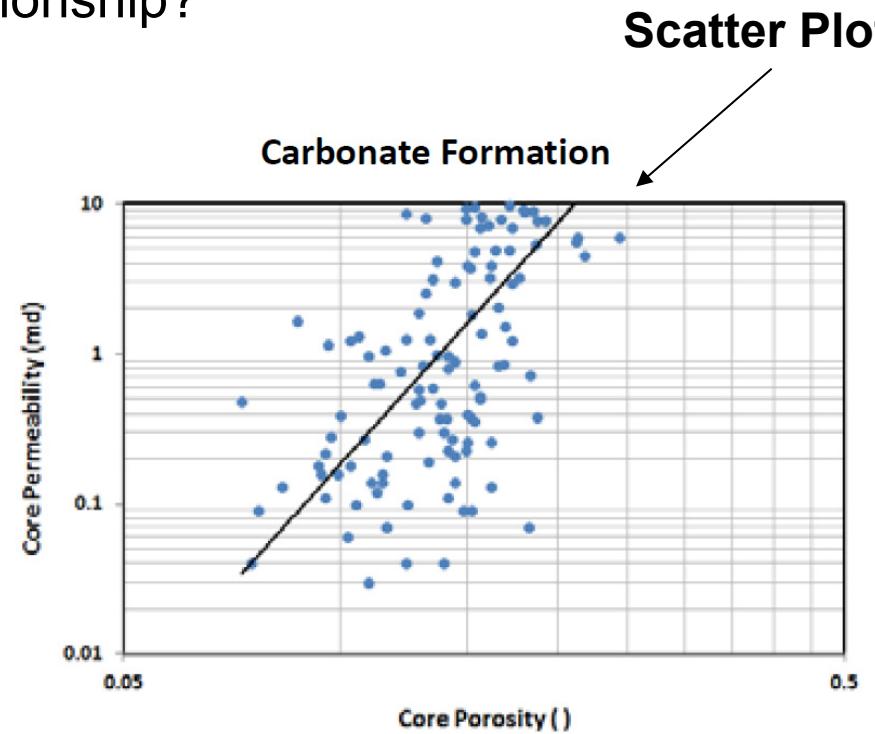
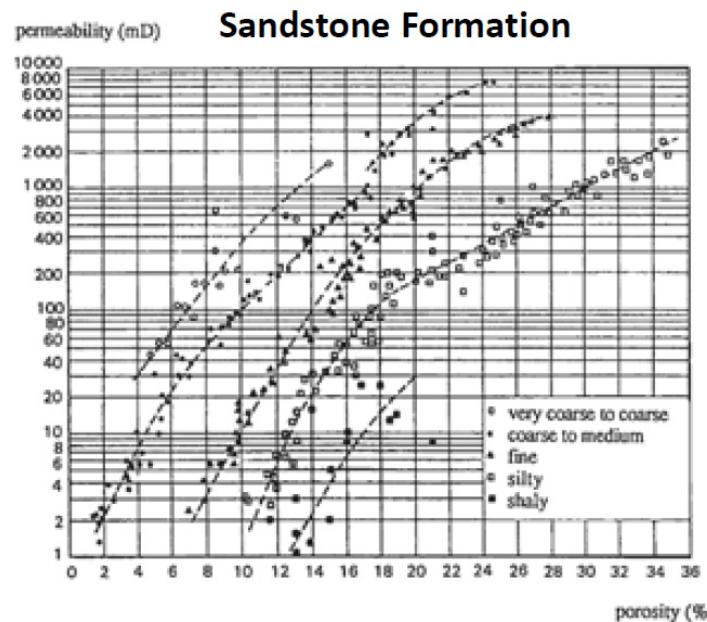


Figure from Peters, E. J., 2012, Advanced Petrophysics.

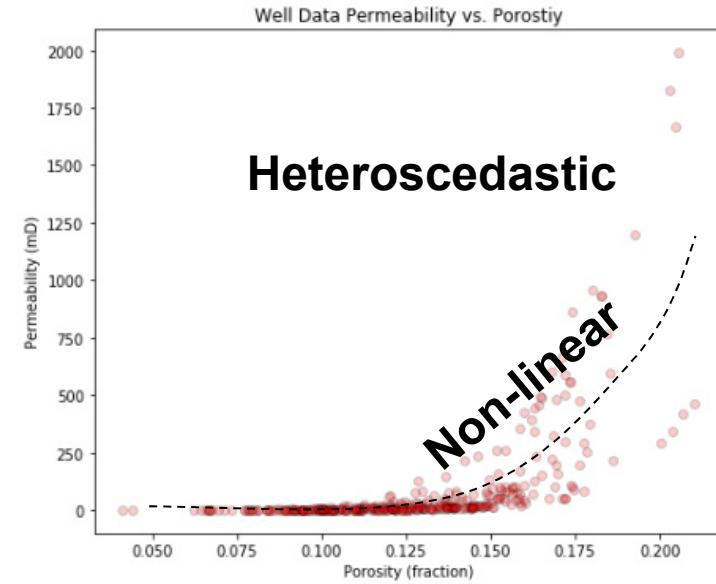
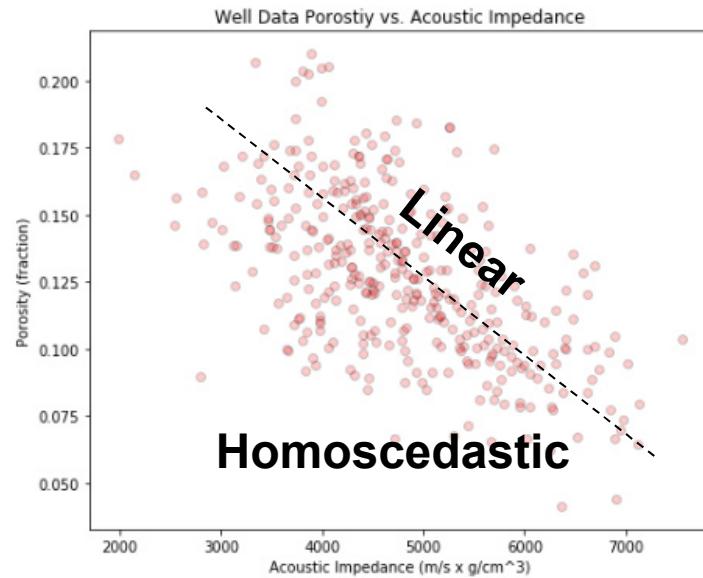
Slide from Dr. Zoya Heidari's PGE 337 Course

Bivariate Statistics

What is Bivariate Analysis?



- Examples of bivariate structures



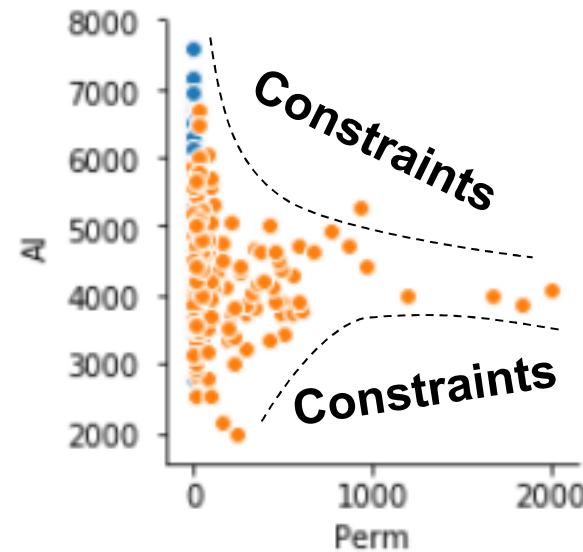
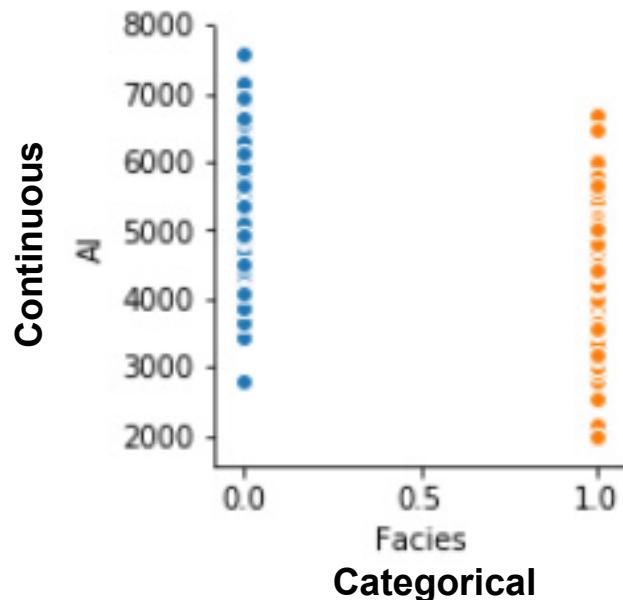
- Linear / Nonlinear – shape of the conditional expectation $Y | X$
- Homoscedastic / Heteroscedastic – conditional variance of $Y | X$

Bivariate Statistics

What is Bivariate Analysis?



- Examples of bivariate structures



- Categorical variables only have a specified number of possible outcomes, continuous takes on a range of possible outcomes.
- Constraints – specific combinations of variables are not possible.

Bivariate Statistics

Pearson's Correlation Coefficient



- **Definition: Pearson's Product-Moment Correlation Coefficient**
 - Provides a measure of the degree of linear relationship.

$$\rho_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)\sigma_x\sigma_y}, -1.0 \leq \rho_{xy} \leq 1.0$$

means of variables
x and y

Correlation coefficient of
variables x and y

number of
data pairs

standard deviation of
variables x and y

A diagram illustrating the components of the Pearson correlation coefficient formula. The formula is $\rho_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)\sigma_x\sigma_y}$. Red arrows point from labels to specific parts of the formula: an arrow points from 'Correlation coefficient of variables x and y' to ρ_{xy} ; another arrow points from 'means of variables x and y' to \bar{x} and \bar{y} ; a third arrow points from 'number of data pairs' to the summation symbol; and a fourth arrow points from 'standard deviation of variables x and y' to σ_x and σ_y .

- Correlation coefficient is a standardized covariance.

$$C_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}$$

Covariance

$$\rho_{xy} = \frac{C_{xy}}{\sigma_x\sigma_y}$$

Bivariate Statistics

Variance and Covariance



- We can see that covariance and variance are related.
 - Replace the second term in the square with another variable.
 - Covariance:

$$C_{xy} = \frac{1}{n-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

A measure of how 2 variables vary together.

- Variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})$$

A measure of how 1 variable varies with itself.

Bivariate Statistics

Spearman's Rank Correlation Coefficient



- **Definition: Spearman's Rank Correlation Coefficient**
 - Provides a measure of the degree of monotonic relationship.

$$\rho_{R_x, R_y} = \frac{\sum_{i=1}^n (R_{x_i} - \bar{R}_x)(R_{y_i} - \bar{R}_y)}{(n - 1)\sigma_{R_x}\sigma_{R_y}}, -1.0 \leq \rho_{xy} \leq 1.0$$

means of rank transform of variables
x and y

Rank correlation coefficient of
variables x and y

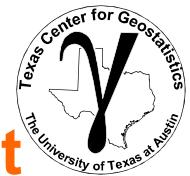
number of
data pairs

standard deviation of
Rank transform of variables x and y

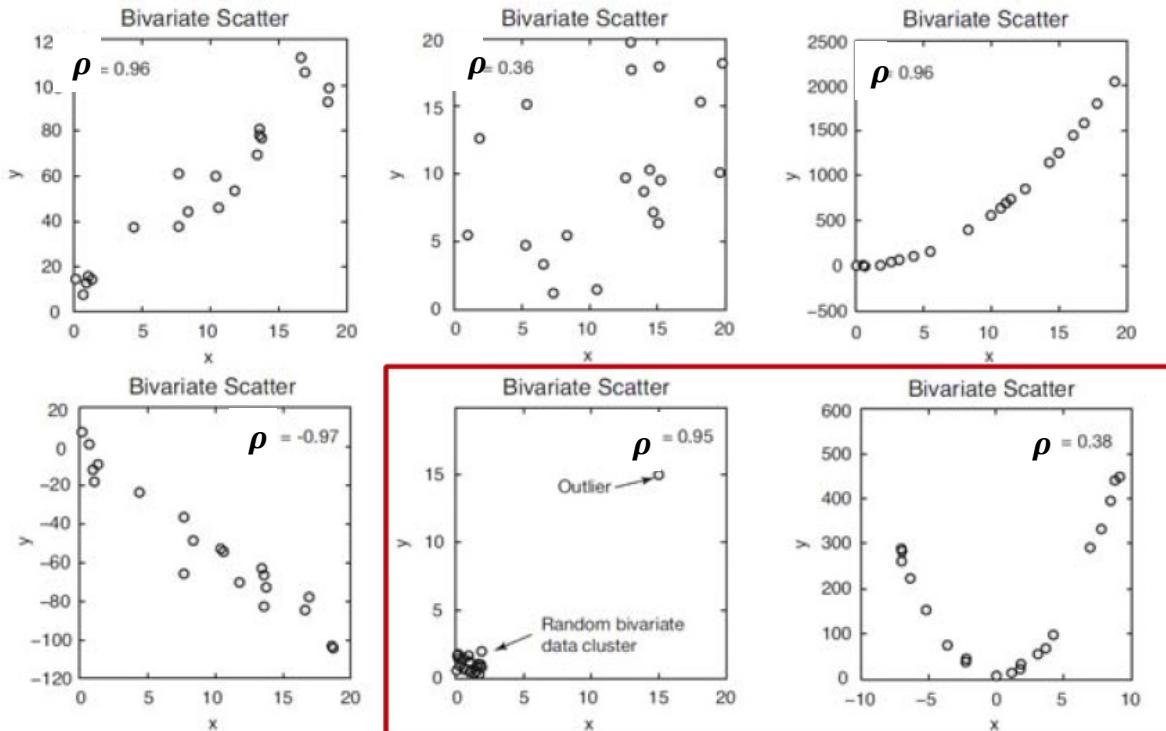
- Rank transform, e.g. R_{x_i} , sort the data in ascending order and replace the data with the index, $i = 1, \dots, n$.
- Spearman's rank correlation coefficient is more robust in the presence of outliers and some nonlinear features than the Pearson's correlation coefficient

Bivariate Statistics

Pearson's Correlation Coefficient



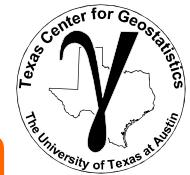
- Interpreting the correlation coefficient



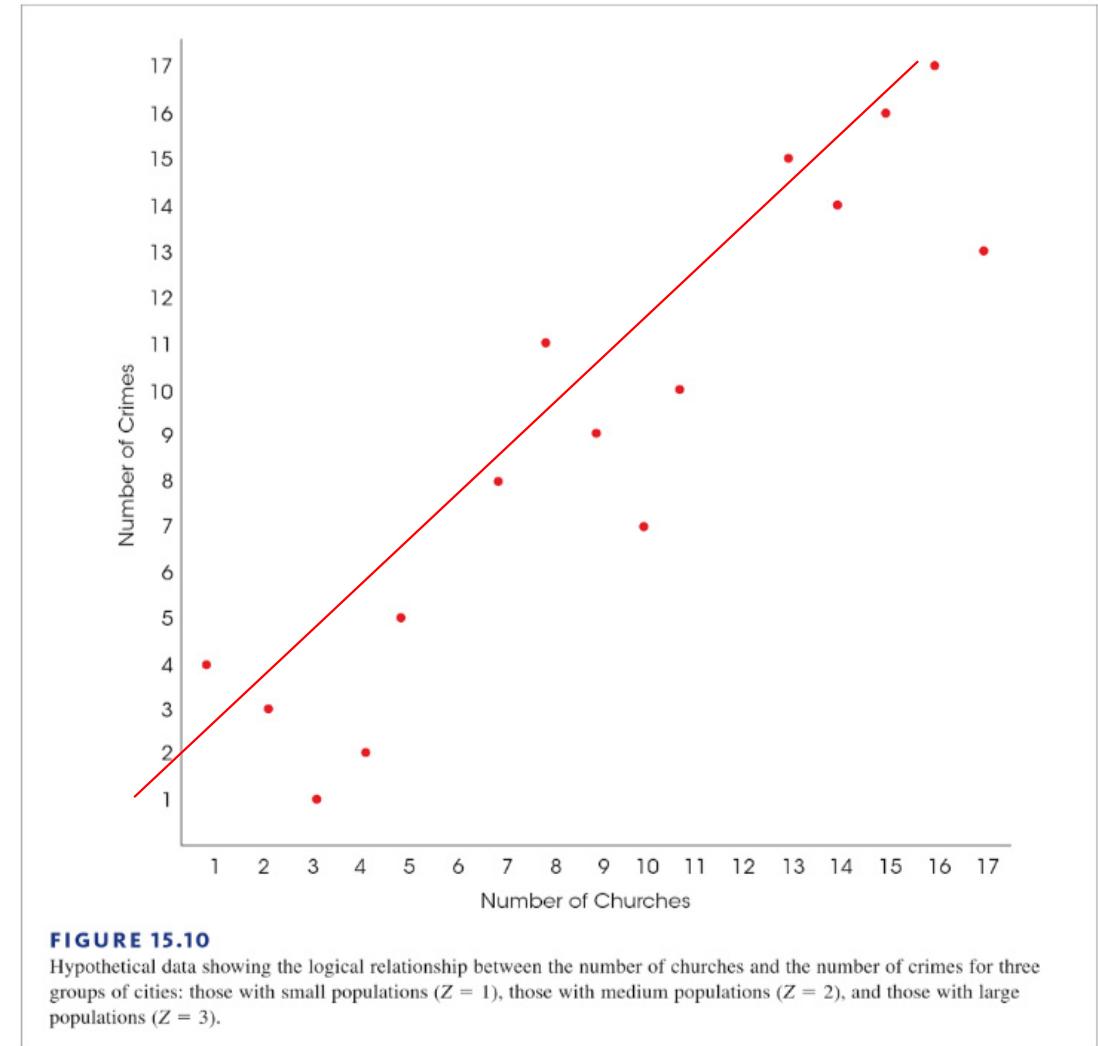
Is Pearson's correlation coefficient a reliable measure of correlation in these cases?

Bivariate Statistics

Correlation and Causation

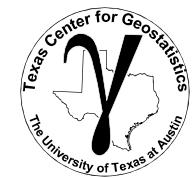


- Correlation does not imply causation!
 - We require a “true experiment” where one variable is manipulated and others are rigorously controlled!

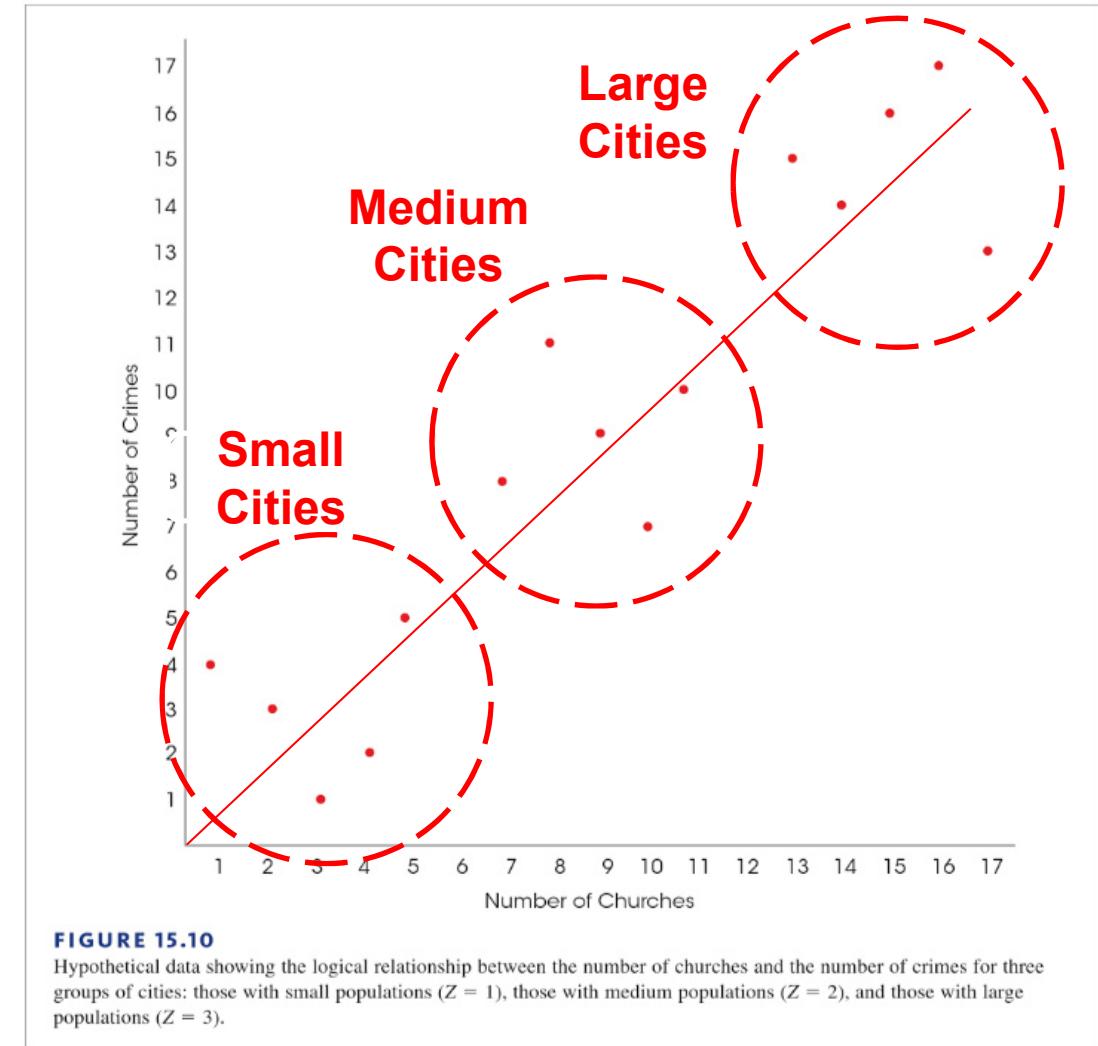


Bivariate Statistics

Correlation and Causation

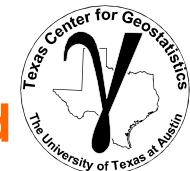


- Correlation does not imply causation!
 - Population was not controlled!
 - For each size of city the correlation is nearly zero.

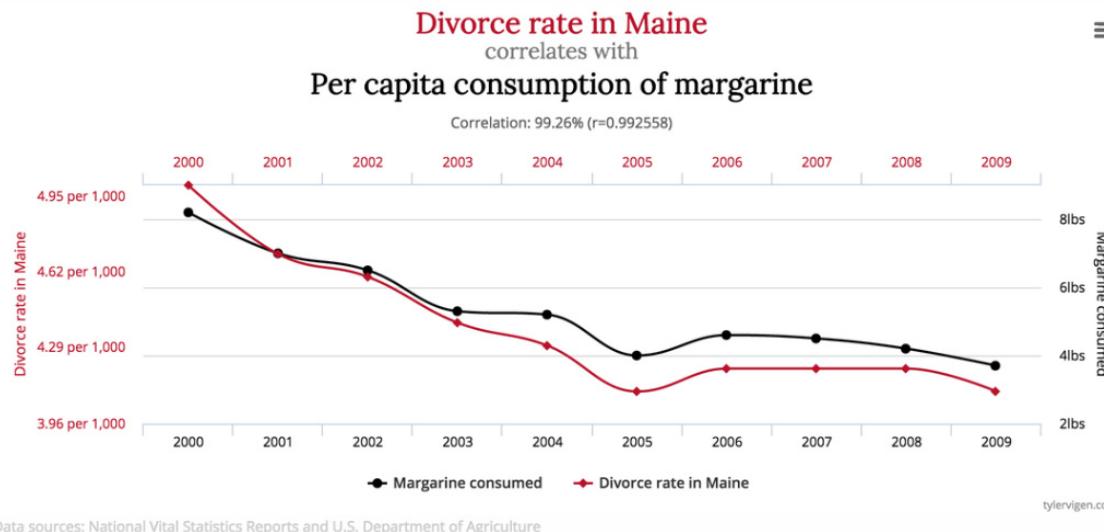


Bivariate Statistics

Comical Examples of Correlation and Causation



Margarine causes divorce?
or **divorce causes margarine?**



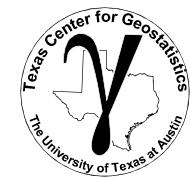
Spiders killing people causes longer words in spelling bees?

or longer words in spelling bees causes venomous spiders to kill people?

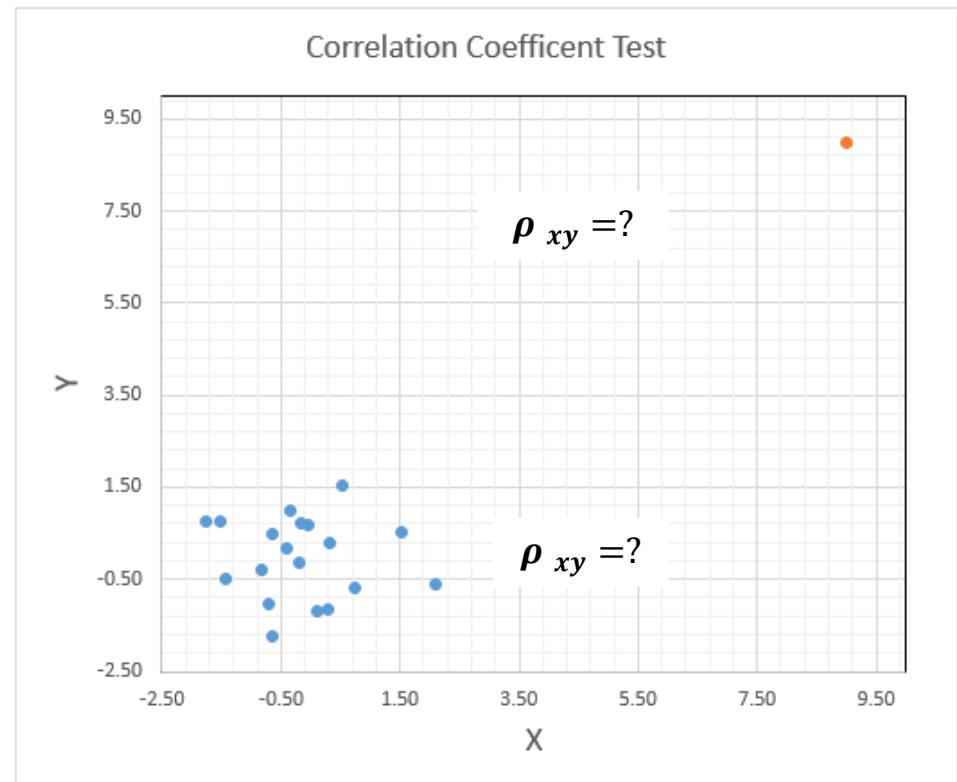


Bivariate Statistics

Exercise with Pearson's Correlation Coefficient



- Task 1: Generate a random data set of x and y variables and estimate their correlation coefficient (Hint: Rand() in Excel with $N[0,1]$).
- Task 2: Now add any desired outlier to the data and estimate the correlation coefficient (see example).
- How does this outlier affect the correlation coefficient?



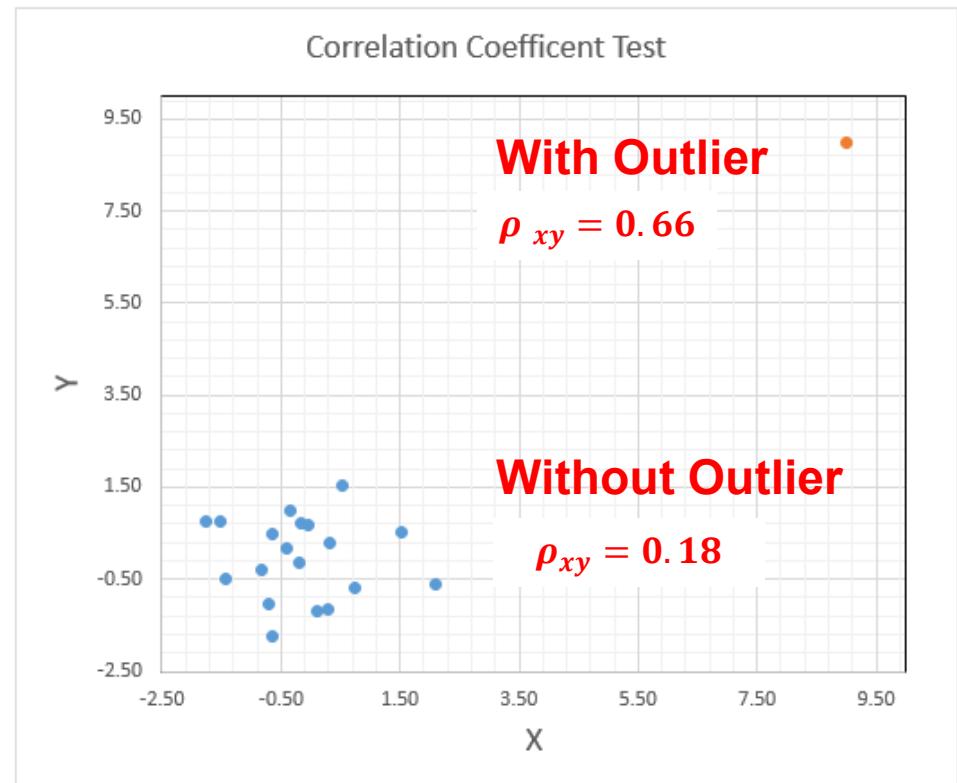
Excel Function NORM.INV(RAND(),0,1)

Bivariate Statistics

Exercise with Pearson's Correlation Coefficient

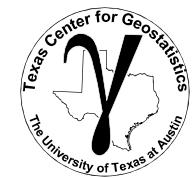


- Task 1: Generate a random data set of x and y variables and estimate their correlation coefficient (Hint: Rand() in Excel with N[0,1]).
- Task 2: Now add any desired outlier to the data and estimate the correlation coefficient (see example).
- How does this outlier affect the correlation coefficient?

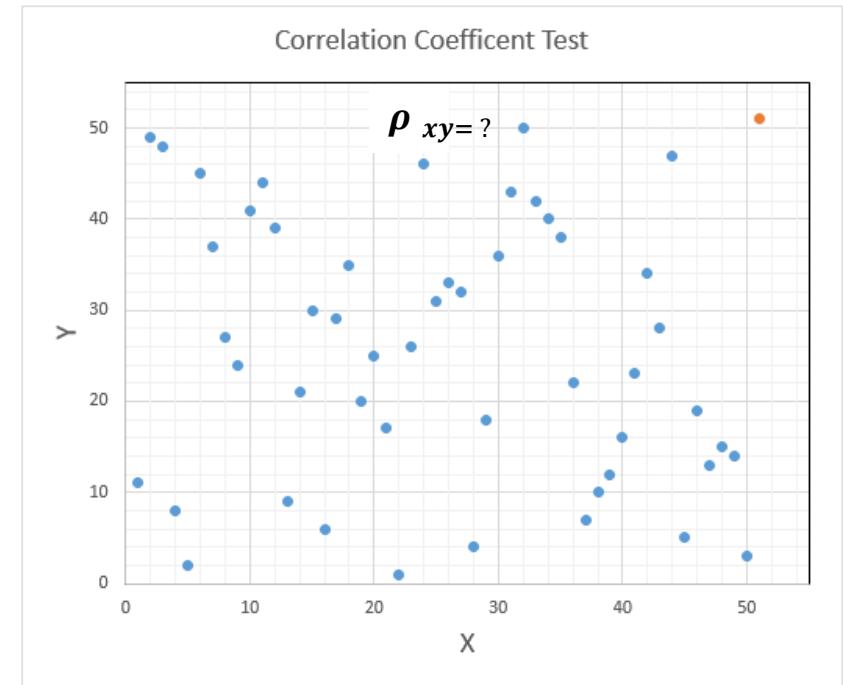


Bivariate Statistics

Exercise with Pearson's Correlation Coefficient

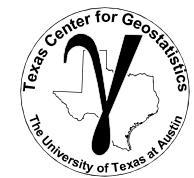


- Task 3: Apply the rank transform to the dataset (Hint: 21-Rank.Avg() in Excel).
- How does this outlier now affect the correlation coefficient?
- This is a more robust form of the correlation coefficient called the rank correlation coefficient.

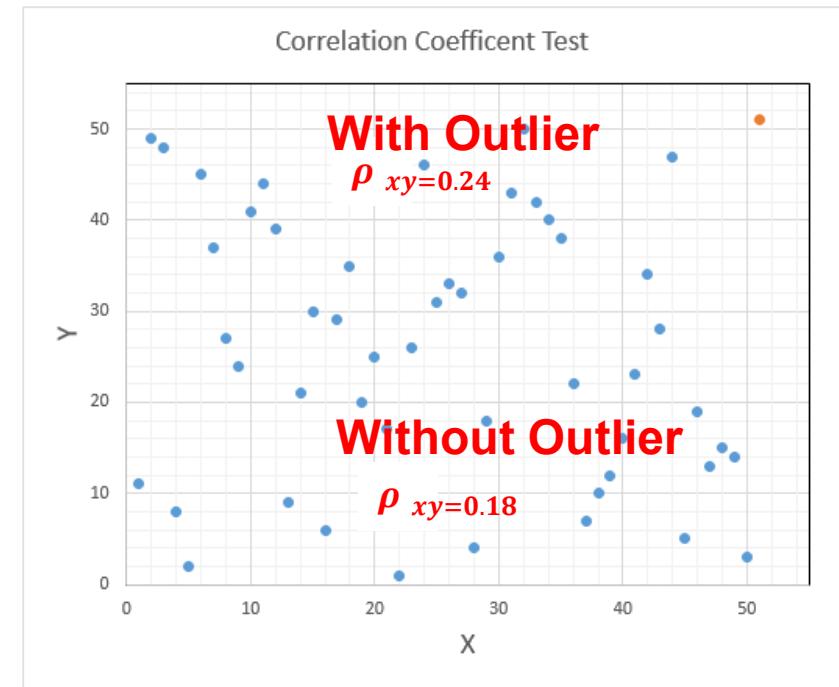


Bivariate Statistics

Exercise with Pearson's Correlation Coefficient



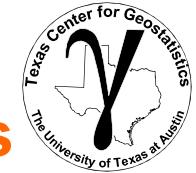
- Task 3: Applied the rank transform to the dataset
(Hint: **52-Rank.Avg()** in Excel).
- How does this outlier now affect the correlation coefficient?
- This is a more robust form of the correlation coefficient called the rank correlation coefficient.



Excel Function =21-RANK.AVG(value,array)

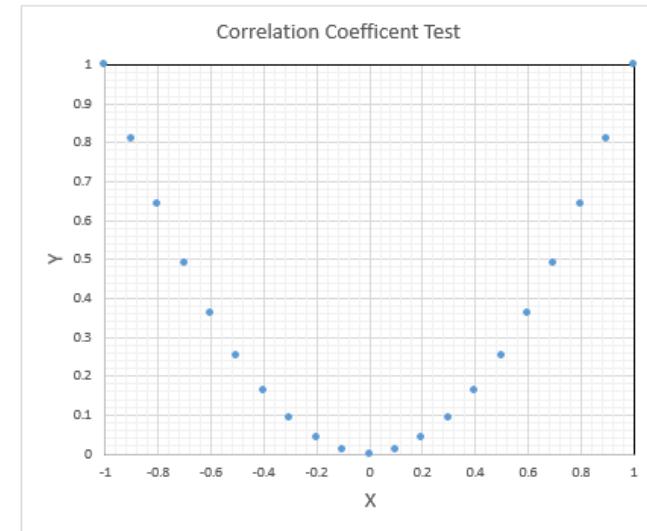
Bivariate Statistics

Measuring Linear Relationships with the Correlation Coefficient



Correlation / Covariance is a measure of linear relationship

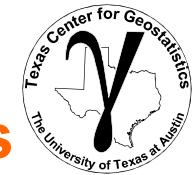
- What is the Correlation / Covariance of $y = x^2$ over range of $[-1, 1]$?



Excel Function Correl(array1,array2)

Bivariate Statistics

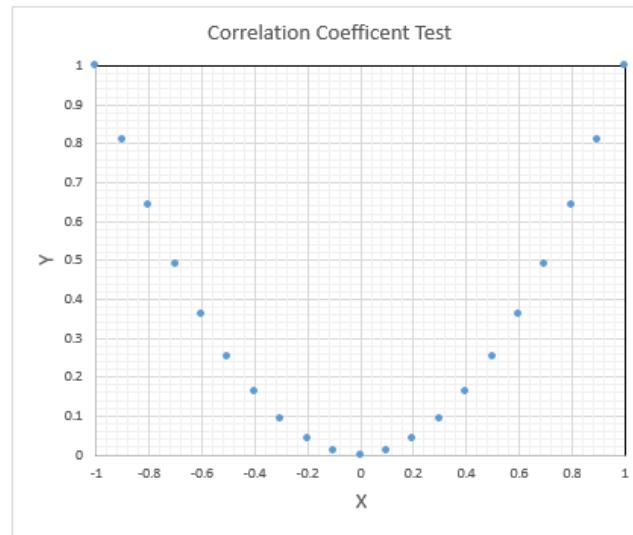
Measuring Linear Relationships with the Correlation Coefficient



Correlation / Covariance is a measure of linear relationship

- What is the Correlation / Covariance of $y = x^2$ over range of $[-1, 1]$?

Correlation Coefficient, $\rho_{xy} = 0.0!$



- Over range $[0,1]$?

Correlation Coefficient, $\rho_{xy} = 0.96$,
Rank Correlation Coefficient, $\rho_{RxRy} = 1.0$

Excel Function Correl(array1,array2)

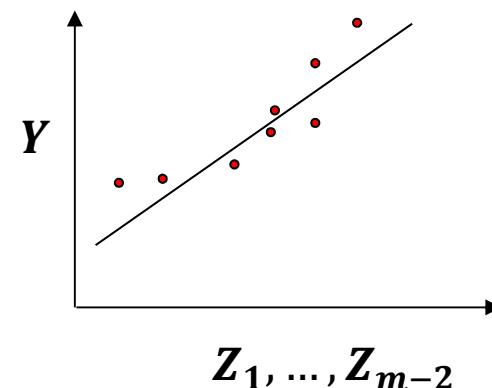
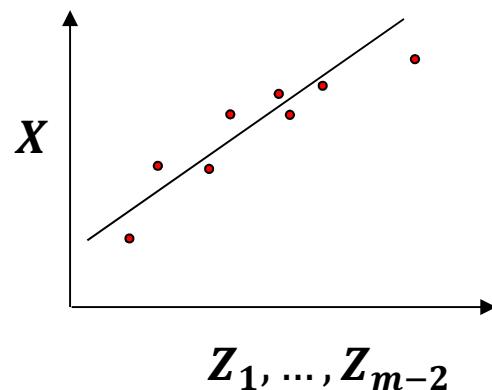
Bivariate Statistics

Partial Correlation



A method to calculate the correlation between X and Y after controlling for the influence of Z_1, \dots, Z_{m-2} other features on both X and Y .

1. perform linear, least-squares regression to predict X from Z_1, \dots, Z_{m-2} .
 X is regressed on the predictors to calculate the estimate, X^*
2. perform linear, least-squares regression to predict Y from Z_1, \dots, Z_{m-2} .
 Y is regressed on the predictors to calculate the estimate, Y^*

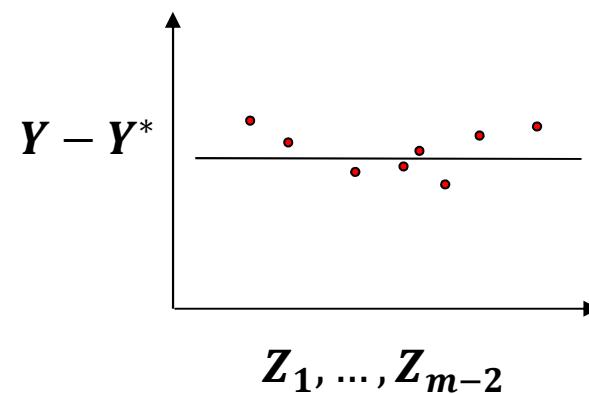
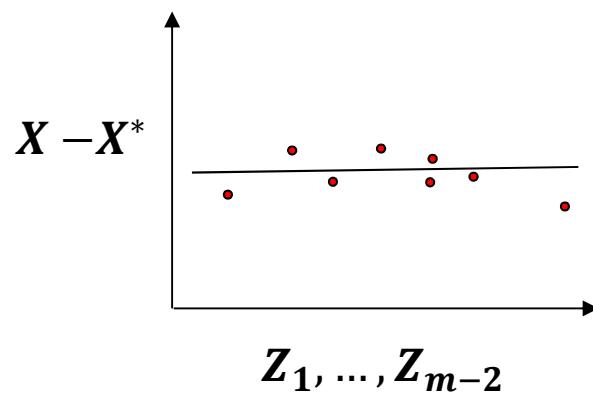


Bivariate Statistics

Partial Correlation

A method to calculate the correlation between X and Y after controlling for the influence of Z_1, \dots, Z_{m-2} other features on both X and Y .

3. calculate the residuals in Step #1, $X - X^*$, where $X^* = f(Z_1, \dots, Z_{m-2})$, linear regression model
4. calculate the residuals in Step #1, $Y - Y^*$, where $Y^* = f(Z_1, \dots, Z_{m-2})$, linear regression model



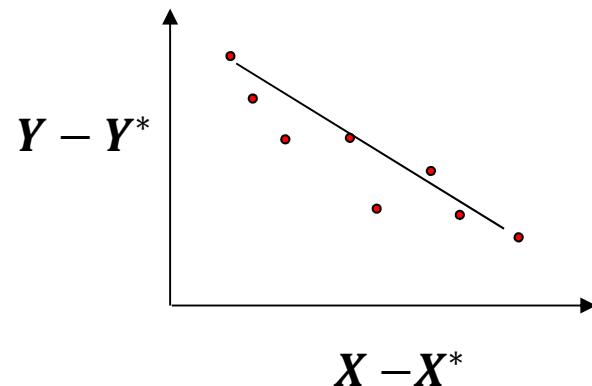


Bivariate Statistics

Partial Correlation

A method to calculate the correlation between X and Y after controlling for the influence of Z_1, \dots, Z_{m-2} other features on both X and Y .

5. calculate the correlation coefficient between the residuals from Steps #3 and #4, $\rho_{X-X^*, Y-Y^*}$



The partial correlation, provides a measure of the linear relationship between X and Y while controlling for the effect of Z_1, \dots, Z_{m-2} other features on both, X and Y .

Partial Correlation Hands-on in Excel

Experiment with Partial Correlation:

Partial Correlation Demonstration, Michael Pyrcz, University of Texas at Austin, @GeostatsGuy on Twitter

This is a demonstration of calculating the partial correlation. Partial correlation provides the linear correlation between two features after removing the influence of other variables.

In this case we calculate the correlation between TOC and production, while controlling for the influence of acoustic impedance.

Steps for partial correlation between X and Y controlling for all other variables Z

1. regress all other features on X, and calculate the residual of $X - \hat{X}$
2. regress all other features on Y, and calculate the residual of $Y - \hat{Y}$
3. calculate the correlation coefficient between $\hat{X} - \hat{\hat{X}}$ and $\hat{Y} - \hat{\hat{Y}}$ (partial correlation) and compare to regular correlation coefficient

Available Features

Prod	AI	TOC
1695.4	2.8	116
3007.1	3.22	0.99
2531.9	4.01	0.69
5026.5	2.82	1.00
2652.8	3.18	1.51
4017.4	2.68	0.94
2652.9	2.93	0.8
2670.9	3.25	0.69
12474	2.43	0.95
2722.9	3.71	114
3828.2	2.22	1.09
5095.8	2.28	153
4031.6	2.8	117
5631.7	2.55	187
2873.6	2.51	0.18
2423.8	3.43	0.57
5836.4	3.02	1.34
3225.1	2.53	135
3555	3.03	0.9
3074.2	3.54	0.9
6167.8	2.38	126
5026.3	2.41	121
1770.6	2.38	1
1781.7	3.29	0.38
6720.7	1.39	177
1087.1	3.41	-0.04
4778.1	2.55	137
3497.1	3.74	11
2250.4	4.03	0.72
3244.5	2.9	0.81
6528	2.85	145
3496.7	2.45	0.92
2962.4	3.43	0.79
3583.3	2.95	113
2427.7	4.18	0.43
1928.3	3.17	0.54

Derived Features

Prod [*] AI	TOC [*] AI	R _{prod}	R _{toc}
4034.31	1.07	-2338.95	0.09
3611.69	0.88	-604.59	0.01
2816.76	0.52	-284.83	0.37
4205.37	1.14	1063.14	-0.06
3651.94	0.90	-75.82	0.51
4145.00	1.12	127.62	-0.19
3803.50	1.01	-950.69	-0.21
3581.50	0.86	-910.57	-0.17
4406.62	1.23	-1932.57	-0.28
3118.64	0.66	395.74	0.48
4617.93	1.33	-789.68	-0.25
4547.49	1.30	-548.32	0.23
4034.31	1.07	57.33	0.10
4285.87	1.18	1345.87	0.69
4326.12	1.20	-1452.51	-1.02
3400.38	0.78	-976.61	-0.21
3752.56	0.94	2083.86	0.40
4305.99	1.19	-1060.91	0.16
3802.88	0.96	-647.85	-0.06
3289.70	0.73	-215.54	0.17
4456.93	1.26	1710.91	0.00
4426.74	1.24	599.60	-0.03
4456.93	1.26	-2666.34	-0.26
3514.25	0.85	-1759.56	-0.47
5433.10	1.70	1267.55	0.07
3420.51	0.79	-2333.45	-0.83
4285.87	1.18	492.21	0.19
3088.45	0.64	408.61	0.46
3933.69	1.02	-689.18	-0.21
3884.00	1.04	2543.99	0.41
4386.49	1.22	-889.77	-0.30
3400.38	0.78	-437.99	0.01
3883.37	1.00	-300.08	0.13
2645.70	0.44	-218.05	-0.01
3662.00	0.90	-1733.72	-0.36

Step 1: Regress AI on Production

Predict production from acoustic impedance, then calculate the residual ($\text{Prod}(\text{AI}) - \text{Prod}$), this is production controlled for acoustic impedance.

Linear Regression

Production vs. TOC

Step 2: Regress AI on TOC

Predict TOC from acoustic impedance, then calculate the residual ($\text{TOC}(\text{AI}) - \text{TOC}$), this is TOC controlled for acoustic impedance.

Linear Regression

LogPerm vs. Acoustic Impedance

Step 3: Calculate the Partial Correlation and Compare to Pearson Correlation

Calculate the regular Pearson correlation coefficient between TOC and production and compare to the partial correlation coefficient between TOC and production controlling for AI.

Production vs. TOC Controlled for AI

Things to try:

1. Increase the frequency over a region in the joint frequency distribution.
2. Add a TOC, Production outlier. TOC = 10, Production = 9999. What happened? Does Vsh inform porosity?

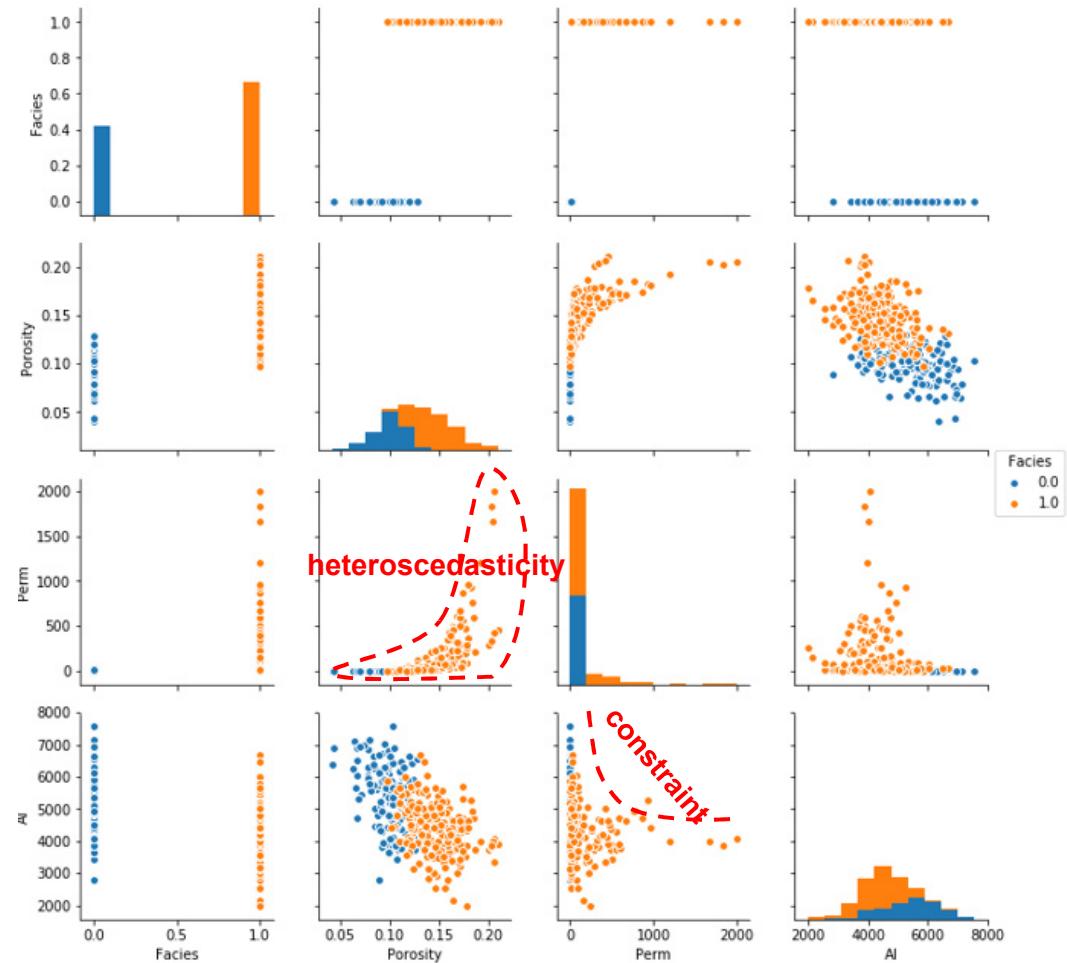
The file is Partial_Correlation_Demo.xlsx at location <https://git.io/fhA95>.

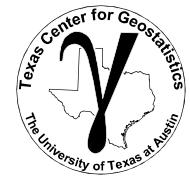
Bivariate Statistics

Matrix Scatter Plots



- **For more than two variables make matrix scatterplots**
 - By hand in Excel or packages in R and Python.
 - Look for linear / nonlinear features
 - Look for homoscedasticity (constant conditional variance) and heteroscedasticity (conditional variance changes with value)
 - Look for constraints





Data Analytics and Geostatistics: Multivariate Analysis

Lecture outline . . .

- **Joints and Conditionals**

Introduction

Modeling Prerequisites

Spatial Estimation

Spatial Uncertainty

Multivariate, Spatial

Multivariate Analysis

Machine Learning

Novel Workflows

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin

Probability Definitions

Conditional, Marginal and Joint Probability



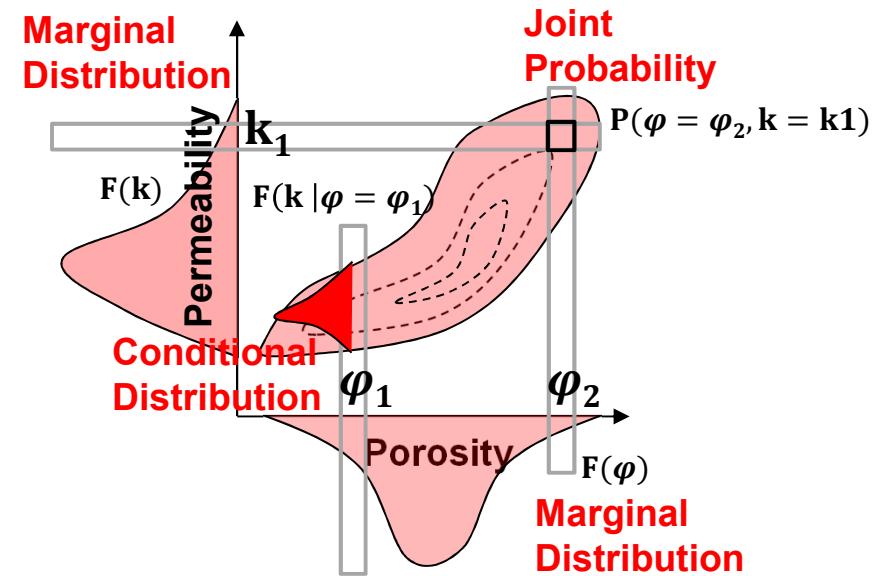
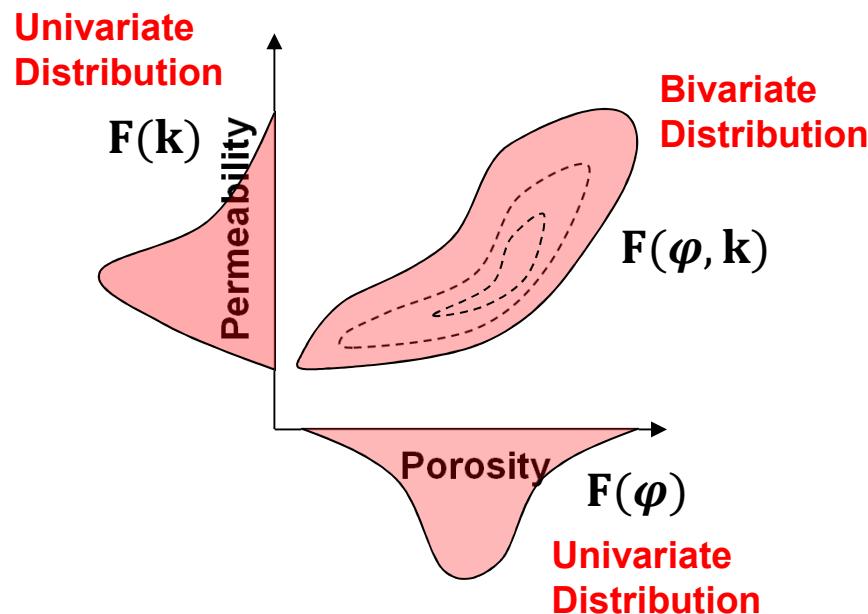
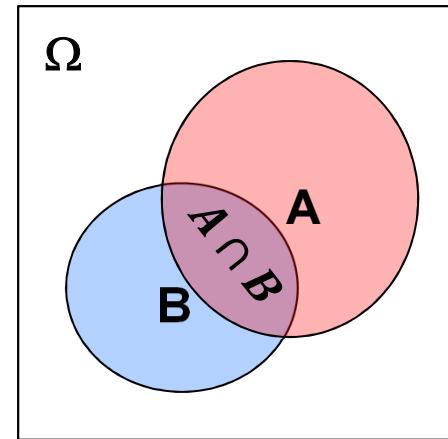
Probability of B given A occurred? $P(B | A)$

Conditional Probability

$$P(B | A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A \text{ and } B)}{P(A)}$$

Joint Probability

Marginal Probability



Probability Definitions

Conditional, Marginal and Joint Probability



Marginal Probability: Probability of an event, irrespective of any other event
 $P(X), P(Y)$

Conditional Probability: Probability of an event, given another event is already true.

$$P(X \text{ given } Y), P(Y \text{ given } X)$$

$$P(X | Y), P(Y|X)$$

Joint Probability: Probability of multiple events occurring together.

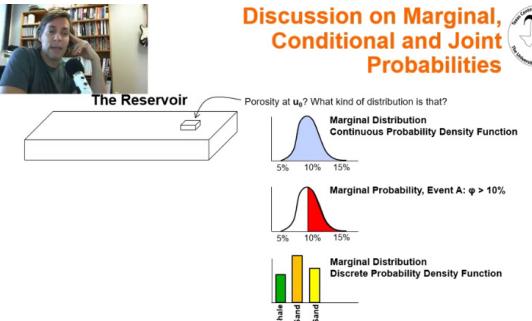
$$P(X \text{ and } Y), P(Y \text{ and } X)$$

$$P(X \cap Y), P(Y \cap X)$$

$$P(X, Y), P(Y, X)$$



Discussion on Marginal,
Conditional and Joint
Probabilities

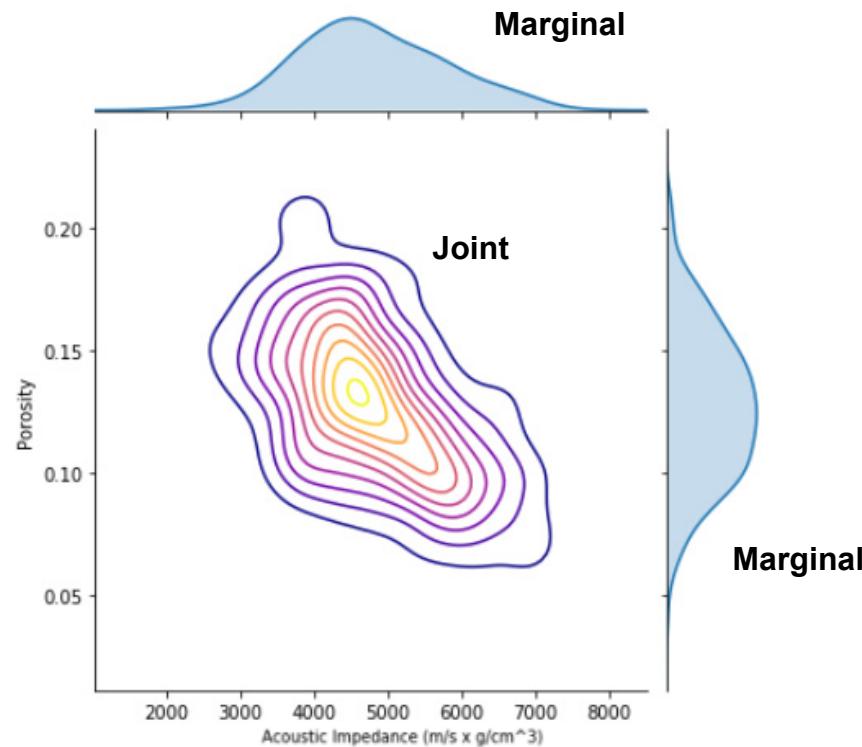


See YouTube Video on Marginals, Conditionals and Joints! <https://www.youtube.com/watch?v=bL2gPwMfYpc&index=5&t=0s&list=PLG19vXLQHvSB-D4XKYieEku9GQM0yAzjI>

Marginal, Conditional and Joint Probability



- Working directly with marginal, conditional and joint probability
 - If you have enough data, you can directly calculate all the required probabilities
 - Go beyond statistics like correlation coefficient



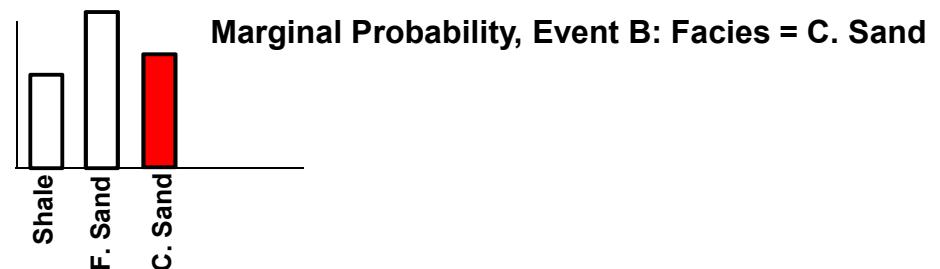
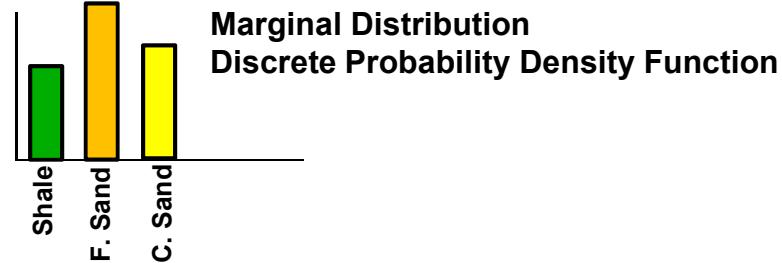
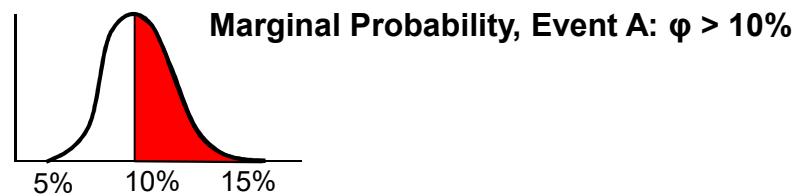
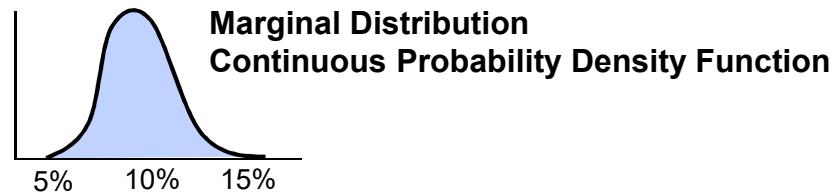
Marginal, Conditional and Joint Probability



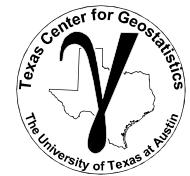
The Reservoir



Porosity at u_0 ? What kind of distribution is that?



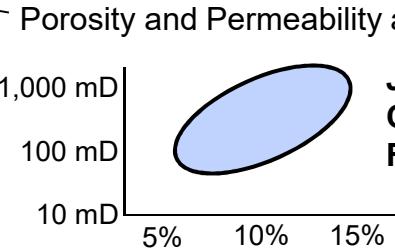
Marginal, Conditional and Joint Probability



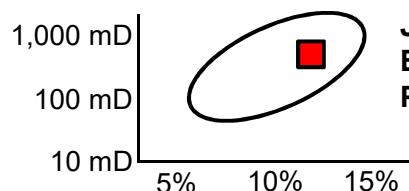
The Reservoir



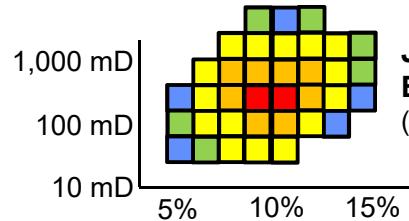
Porosity and Permeability at u_0 ? What kind of distribution is that?



Joint Distribution
Continuous Joint Probability Density Function

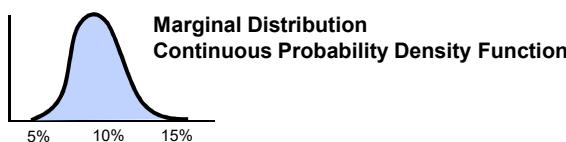


Joint Probability
Event A: $12\% < \varphi < 14\%$ and $600\text{mD} < k < 900\text{mD}$
 $P(12\% < \varphi < 14\% \cap 600\text{mD} < k < 900\text{mD})$

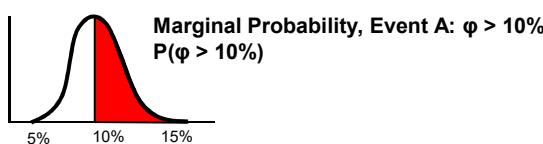


Joint Probability Density Function
Binned
(0% bins removed)

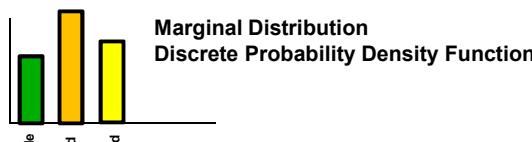
Univariate, Marginal Examples



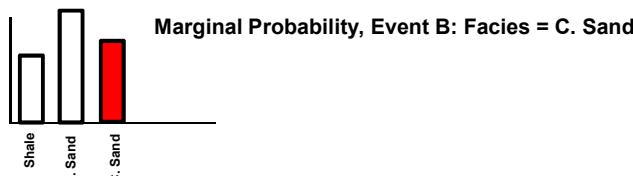
Marginal Distribution
Continuous Probability Density Function



Marginal Probability, Event A: $\varphi > 10\%$
 $P(\varphi > 10\%)$

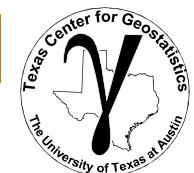


Marginal Distribution
Discrete Probability Density Function



Marginal Probability, Event B: Facies = C. Sand

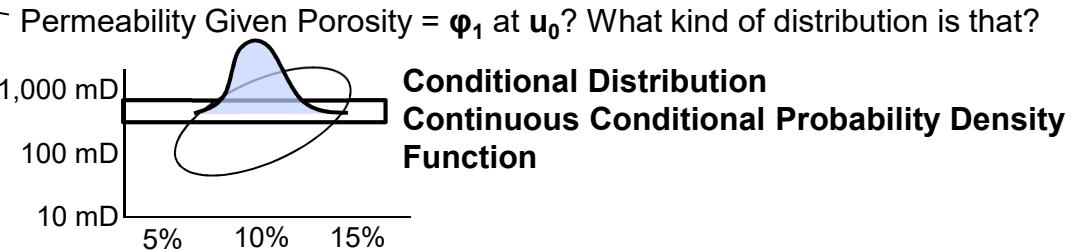
Marginal, Conditional and Joint Probability



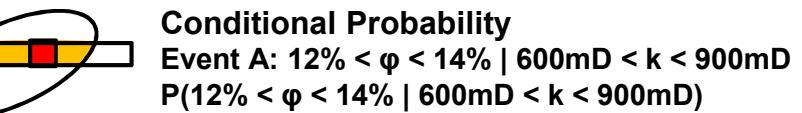
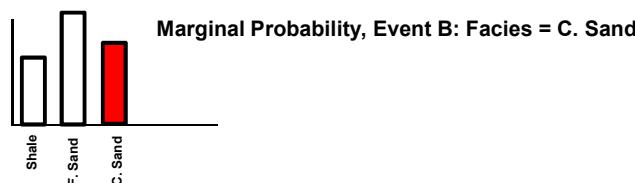
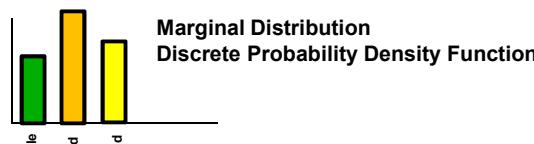
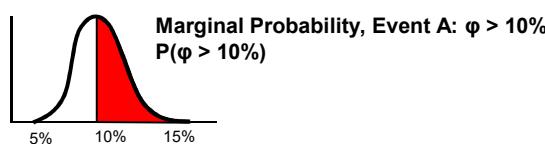
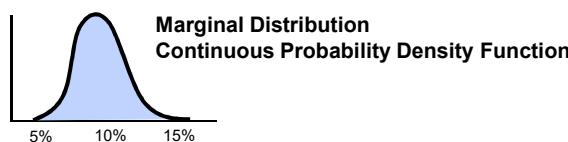
The Reservoir



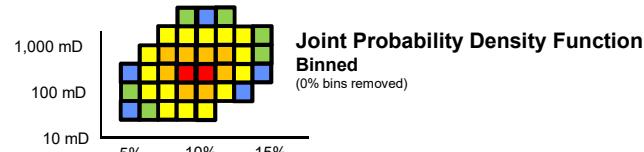
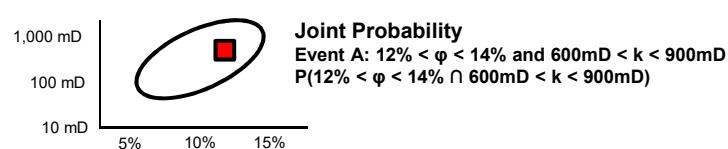
Permeability Given Porosity = φ_1 at u_0 ? What kind of distribution is that?



Univariate, Marginal Examples



Bivariate, Joint Examples



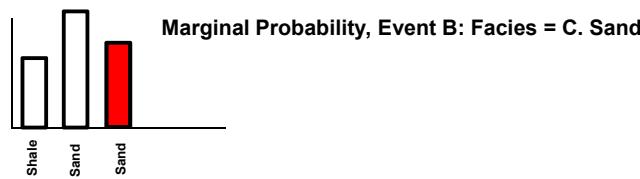
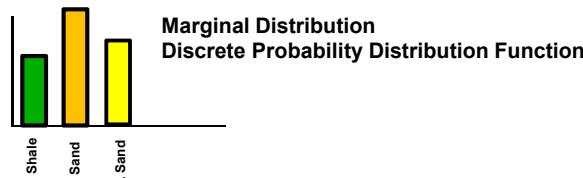
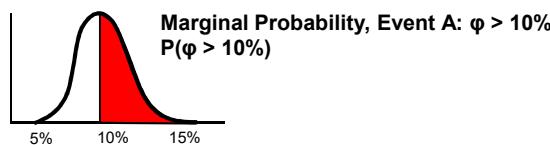
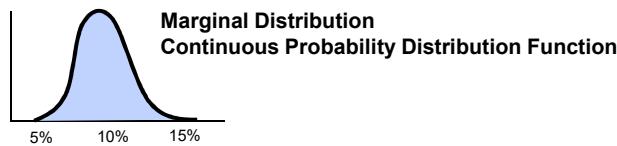
Marginal, Conditional and Joint Probability



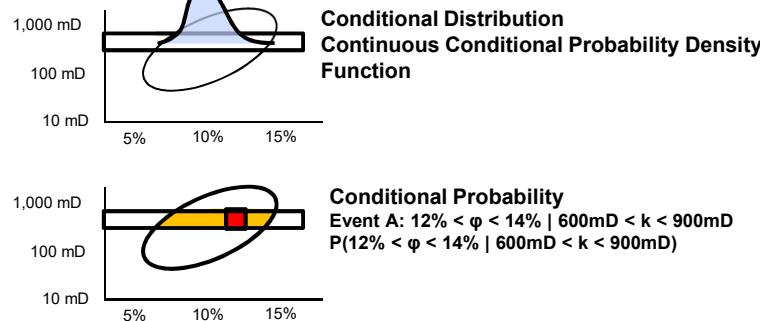
The Reservoir



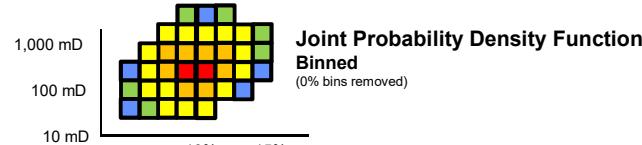
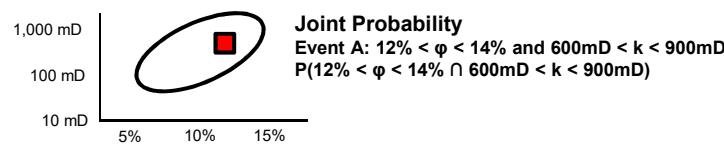
Univariate, Marginal Examples



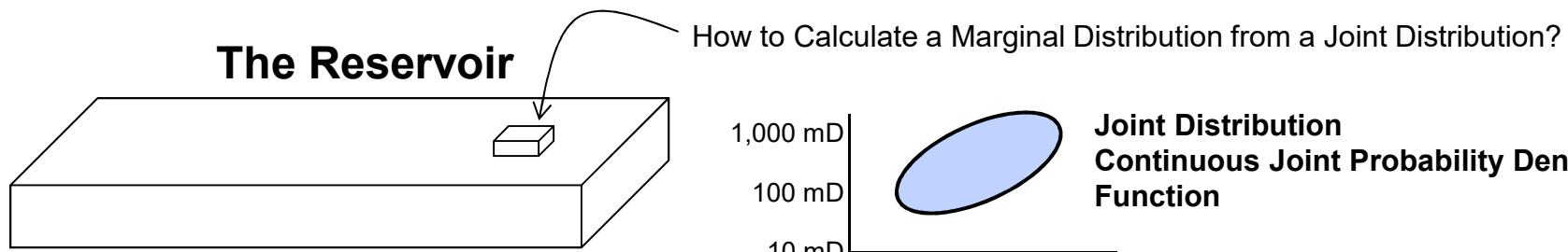
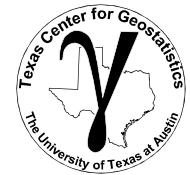
Bivariate, Conditional Examples



Bivariate, Joint Examples

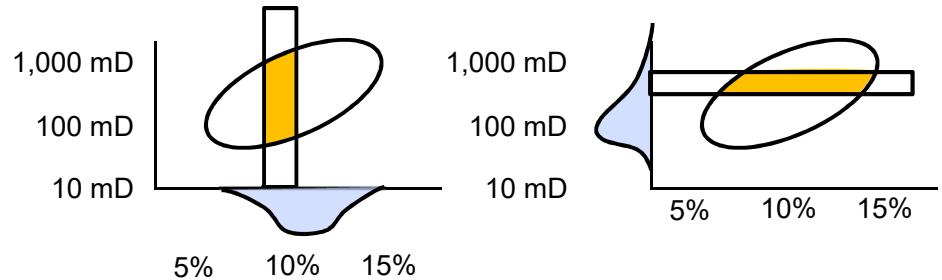


Marginal, Conditional and Joint Probability

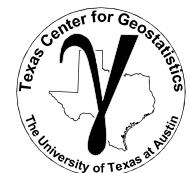


Definition of a Marginal Distribution

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy \quad \text{or} \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dx$$



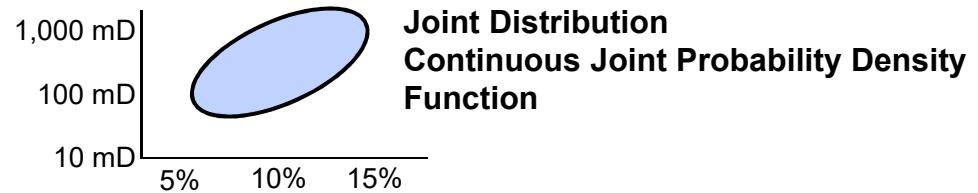
Marginal, Conditional and Joint Probability



The Reservoir

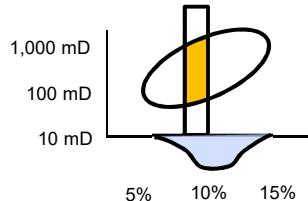


Calculate a Conditional Distribution from a Joint Distribution?



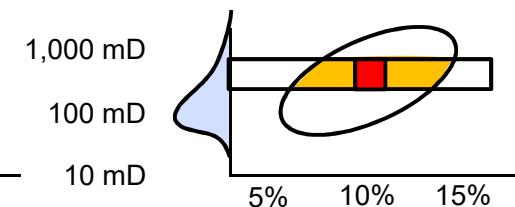
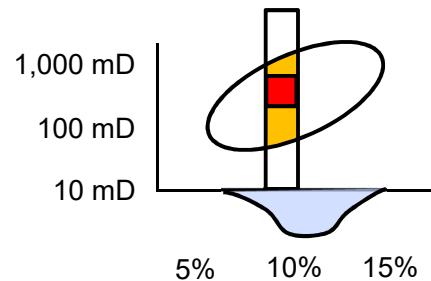
Definition of a Marginal Distribution

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy \quad \text{or} \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dx$$



Definition of a Conditional Distribution

$$f_{Y|X}(y | x) = \frac{f_{XY}(x, y)}{f_X(x)} \quad \text{or} \quad f_{X|Y}(x | y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$



Marginal, Conditional and Joint Probability



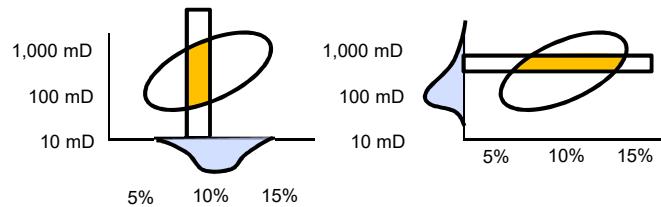
The Reservoir

How to Calculate a Joint Distribution?



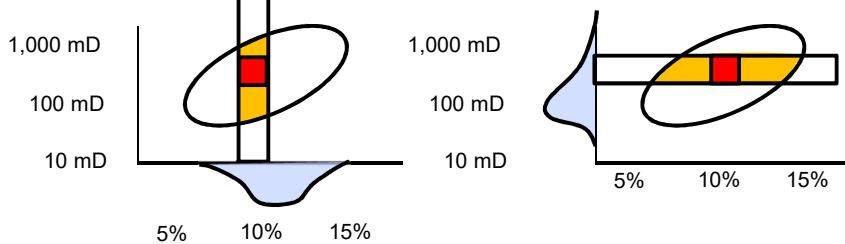
Definition of a Marginal Distribution

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy \quad \text{or} \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dx$$

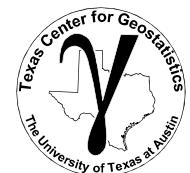


Definition of a Conditional Distribution

$$f_{Y|X}(y | x) = \frac{f_{XY}(x, y)}{f_X(x)} \quad \text{or} \quad f_{X|Y}(x | y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

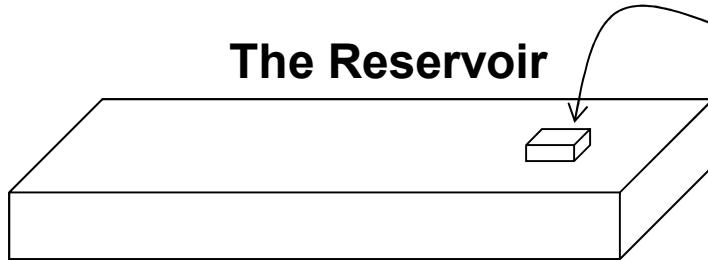


Marginal, Conditional and Joint Probability



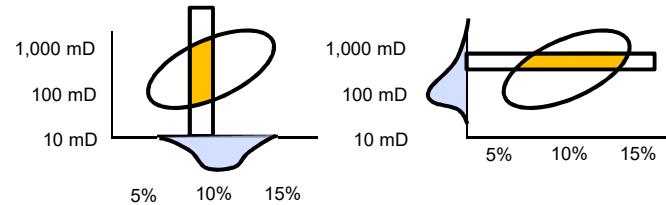
The Reservoir

How to Calculate a Joint Distribution?



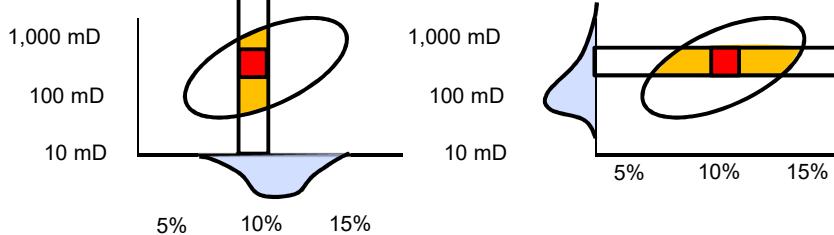
Definition of a Marginal Distribution

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y) dy \quad \text{or} \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y) dx$$



Definition of a Conditional Distribution

$$f_{Y|X}(y | x) = \frac{f_{XY}(x,y)}{f_X(x)} \quad \text{or} \quad f_{X|Y}(x | y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$



Non-parametric - Counting Samples in Bins

	1,000 mD	100 mD	10 mD	
1,000 mD	0	0	• 1	1
100 mD	0	1 • 2	3 • 1	1
10 mD	• 1	• 2 • 3	• 1 1	0
	1	1 1	1 0	0

5% 10% 15%

Fitting a Parametric Model

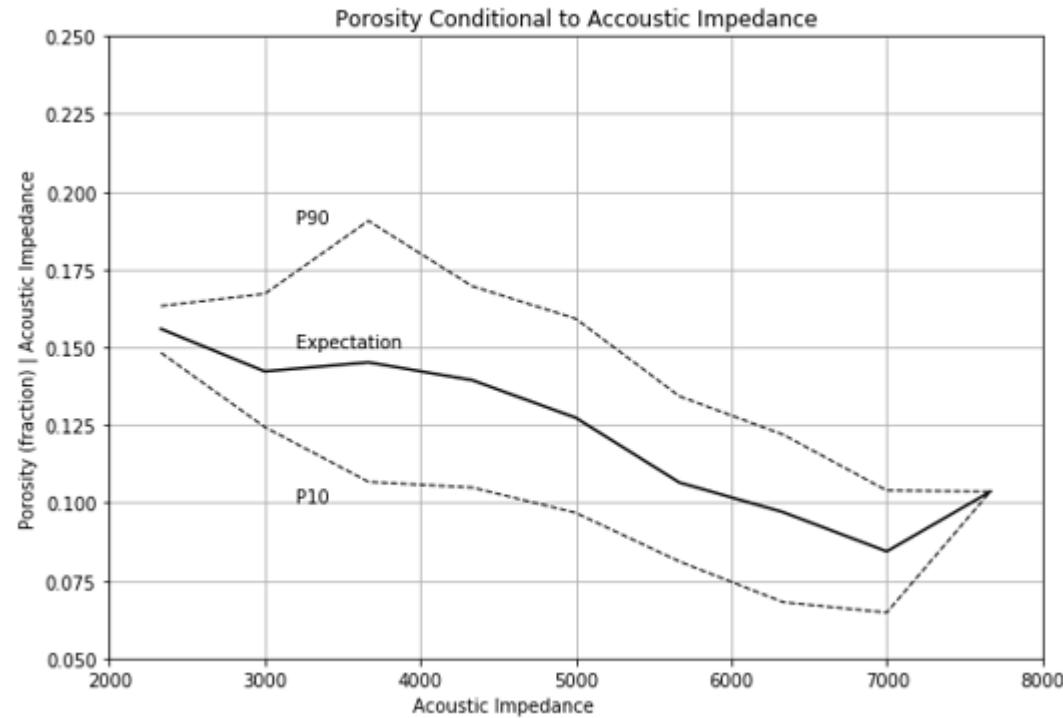
	1,000 mD	100 mD	10 mD	
1,000 mD	0	0	• 4% 4%	4%
100 mD	0	4% 8% 12%	• 12% 10%	10%
10 mD	• 4%	12% 8% 4%	4% 0	0
	4%	4% 4%	4% 0	0

5% 10% 15%

Marginal, Conditional and Joint Probability



- Consider working with conditional statistics.
 - Powerful, flexible assessment of multivariate relationships, without linear assumption



Conditional, Marginal and Joint Hands-on



Joint Distribution:

$$f_{XY}(x, y)$$

Marginal Distribution:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy$$

Conditional Distribution:

$$f_{X|Y}(x | y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

Table of Frequencies

		Porosity (%)				
		25%	20%	15%	10%	5%
		1	1	0	0	0
		2	3	2	0	0
		1	2	2	1	0
		0	0	2	3	2
		0	0	1	1	1
Fraction Shale (%)						
10%		30%	50%	70%	90%	

Conditional, Marginal and Joint Hands-on



Joint Distribution:

$$f_{XY}(x, y)$$

Marginal Distribution:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy$$

Conditional Distribution:

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

Table of Joint Probabilities

Porosity (%)	10%	30%	50%	70%	90%
5%	0	0	4%	4%	4%
10%	0	0	8%	12%	8%
15%	4%	8%	8%	4%	0
20%	8%	12%	8%	0	0
25%	4%	4%	0	0	0

Conditional, Marginal and Joint Hands-on



Given these joint probabilities calculate the: **Table of Joint Probabilities**

Marginal Distributions:

Vsh	10%	30%	50%	70%	90%
$f_{Vsh}(v_{sh})$					

Porosity	5%	10%	15%	20%	25%
$f_{\varphi}(\varphi)$					

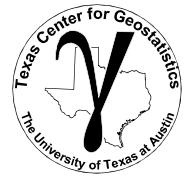
Conditional Distribution:

Vsh	10%	30%	50%	70%	90%
$f_{Vsh \varphi}(v_{sh} \varphi = 15\%)$					

$$f_{Vsh|\varphi}(v_{sh} | \varphi = 15\%) =$$

		25%	20%	15%	10%	5%
		4%	8%	12%	16%	20%
		0	0	0	0	0
Vsh	10%	4%	8%	12%	16%	20%
Porosity (%)	30%	0	0	0	0	0
Fraction Shale (%)	50%	0	0	0	0	0
	70%	0	0	0	0	0
	90%	0	0	0	0	0

Conditional, Marginal and Joint Hands-on



Given these joint probabilities calculate the: **Table of Joint Probabilities**

Marginal Distributions:

Vsh	10%	30%	50%	70%	90%
$f_{Vsh}(v_{sh})$	16%	24%	28%	20%	12%
Porosity	5%	10%	15%	20%	25%
$f_{\varphi}(\varphi)$	12%	28%	24%	28%	8%

Conditional Distribution:

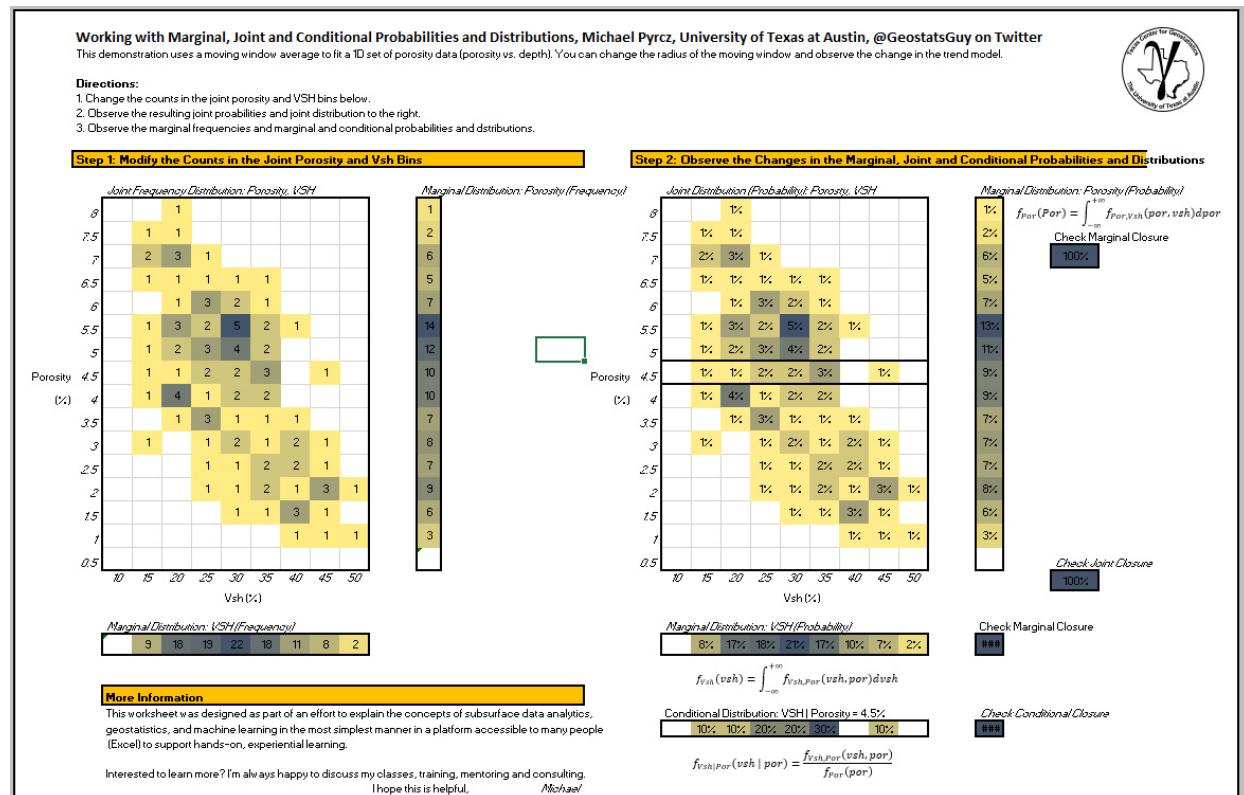
Vsh	10%	30%	50%	70%	90%
	1/6	1/3	1/3	1/6	0

		Porosity (%)				
		4%	8%	12%	15%	20%
		4%	8%	8%	4%	0
		0	0	8%	12%	8%
		0	0	4%	4%	4%
		10%	30%	50%	70%	90%
		25%	4%	0	0	0
		20%	8%	0	0	0
		15%	8%	0	0	0
		10%	0	0	12%	8%
		5%	0	0	4%	4%
		0	0	0	0	0
		Fraction Shale (%)	10%	30%	50%	70%

$$f_{Vsh|\varphi}(v_{sh} | \varphi = 15\%) = f_{Vsh,\varphi}(v_{sh}, \varphi = 15\%) / f_{\varphi}(\varphi = 15\%)$$

Spatial Calculation in Hands-on in Excel

Experiment with Marginal, Joint and Conditionals:

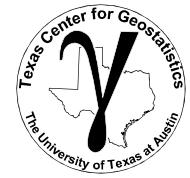


Things to try:

1. Increase the frequency over a region in the joint frequency distribution.
2. Does Vsh inform porosity?

The file is Marginal_Joint_Conditional.xlsx at location <https://git.io/fhA9X>.

Multivariate Analysis Demo



GeostatsPy: Multivariate Analysis for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

[Twitter](#) | [GitHub](#) | [Website](#) | [Google Scholar](#) | [Book](#) | [YouTube](#) | [LinkedIn](#)

PGE 383 Exercise: Multivariate Analysis for Subsurface Data Analytics in Python

Here's a simple workflow, demonstration of multivariate analysis for subsurface modeling workflows. This should help you get started with building subsurface models that integrate uncertainty in the sample statistics.

Bivariate Analysis

Understand and quantify the relationship between two variables

- example: relationship between porosity and permeability
- how can we use this relationship?

What would be the impact if we ignore this relationship and simply modeled porosity and permeability independently?

- no relationship beyond constraints at data locations
- independent away from data
- nonphysical results, unrealistic uncertainty models

Bivariate Statistics

Pearson's Product-Moment Correlation Coefficient

- Provides a measure of the degree of linear relationship.
- We refer to it as the 'correlation coefficient'

Let's review the sample variance of variable x . Of course, I'm truncating our notation as x is a set of samples at locations in our modeling space, $x(u_\alpha)$, $\forall \alpha = 0, 1, \dots, n - 1$.

$$\sigma_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n - 1)}$$

We can expand the squared term and replace one of them with y , another variable in addition to x .

$$C_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}$$

We now have a measure that represents the manner in which variables x and y co-vary or vary together. We can standardize the covariance by the product of the standard deviations of x and y to calculate the correlation coefficient.

$$\rho_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)\sigma_x\sigma_y}, -1.0 \leq \rho_{xy} \leq 1.0$$

In summary we can state that the correlation coefficient is related to the covariance as:

$$\rho_{xy} = \frac{C_{xy}}{\sigma_x\sigma_y}$$

The Pearson's correlation coefficient is quite sensitive to outliers and departure from linear behavior (in the bivariate sense). We have an alternative known as the Spearman's rank correlation coefficient.

Demo workflow for
Multivariate Analysis
<https://git.io/fh2DR>



Multivariate Topics

- Other Topics that Could be Covered
 - Methods to remove correlation and model variables independently
 - Methods for dimensional reduction
 - Methods for clustering analysis

Multivariate New Tools



Topic	Application to Subsurface Modeling
Multivariate Analysis	<p>In the presence of multivariate relationships, must jointly model variables.</p> <p><i>Summarize with bivariate statistics, and visualize and use conditional statistics to go beyond linear measures.</i></p>
Limitations of Correlation	<p>Correlation indicates degree of linear correlation and does not imply causation.</p> <p><i>Visualize and use rank correlation coefficient when needed and apply careful experiments (controlled) to establish causation.</i></p>
Use Conditional Statistics	<p><i>Use conditional distributions to communicate the influence of variables on each other. Provides the value of knowing X to predict Y.</i></p> <p><i>Assess the influence of acoustic impedance on predicting porosity away from wells with conditional distributions.</i></p>



Data Analytics and Geostatistics: Multivariate Analysis

Lecture outline . . .

- Feature Selection

Introduction

Modeling Prerequisites

Spatial Estimation

Spatial Uncertainty

Multivariate, Spatial

Multivariate Analysis

Machine Learning

Novel Workflows

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin

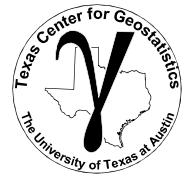
Feature Ranking Motivation



Variable Ranking

- There are often many predictor features, input variables, available for us to work with for subsurface prediction.
- There are good reasons to be selective, throwing in every possible feature is not a good idea!
- In general, for the best prediction model, careful selection of the fewest features that provide the most amount of information is the best practice.

Feature Ranking Motivation



More Motivation to Work with Fewer Variables:

- more variables result in more complicated workflows that require more professional time and have increased opportunity for blunders
- higher dimensional feature sets are more difficult to visualize
- more complicated models may be more difficult to interrogate, interpret and QC
- inclusion of highly redundant and colinear variables increases model instability and decreases prediction accuracy in testing
- more variables generally increase the computational time required to train the model and the model may be less compact and portable
- the risk of overfit increases with the more variables, more complexity

What is Feature Ranking?



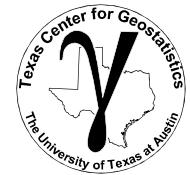
More Motivation to Work with Fewer Variables:

- Feature ranking is a set of metrics that assign relative importance or value to each feature with respect to information contained for inference and importance in predicting a response feature.
- There are a wide variety of possible methods to accomplish this.
- My recommendation is a **wide-array** approach with multiple metric, while understanding the assumptions and limitations of each metric.

Here's the general types of metrics that we will consider for feature ranking:

1. Visual Inspection of Data Distributions and Scatter Plots
2. Statistical Summaries
3. Model-based
4. Recursive Feature Elimination

What is Feature Ranking?



Expert Knowledge:

- Also, we should not neglect expert knowledge.
- If additional information is known about physical processes, causation, reliability and availability of features this should be integrated into assigning feature ranks.
- We should be learning as we perform our analysis, testing new hypotheses.

Feature Ranking Metrics



Metric #1: Visual Inspection

- In any multivariate work we should start with the univariate analysis, summary statistics of one variable at a time. The summary statistic ranking method is qualitative, we are asking:
 - are there data issues?
 - do we trust the features? do we trust the features all equally?
 - are there issues that need to be taken care of before we develop any multivariate workflows?

Feature Ranking Metrics



Summary statistics are a critical first step in data checking.

	count	mean	std	min	25%	50%	75%	max
Well	200.0	100.500000	57.879185	1.000000	50.750000	100.500000	150.250000	200.000000
Por	200.0	14.991150	2.971176	6.550000	12.912500	15.070000	17.402500	23.550000
Perm	200.0	4.330750	1.731014	1.130000	3.122500	4.035000	5.287500	9.870000
AI	200.0	2.968850	0.566885	1.280000	2.547500	2.955000	3.345000	4.630000
Brittle	200.0	48.161950	14.129455	10.940000	37.755000	49.510000	58.262500	84.330000
TOC	200.0	0.991950	0.478264	0.000000	0.617500	1.030000	1.350000	2.180000
VR	200.0	1.964300	0.300827	0.930000	1.770000	1.960000	2.142500	2.870000
Prod	200.0	3864.407081	1553.277558	839.822063	2686.227611	3604.303507	4752.637556	8590.384044
const	200.0	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000

- the number of valid (non-null) values for each feature
- general behaviors such as central tendency, mean, and dispersion, variance.
- issues with negative values, extreme values, and values that are outside the range of plausible values for each property.

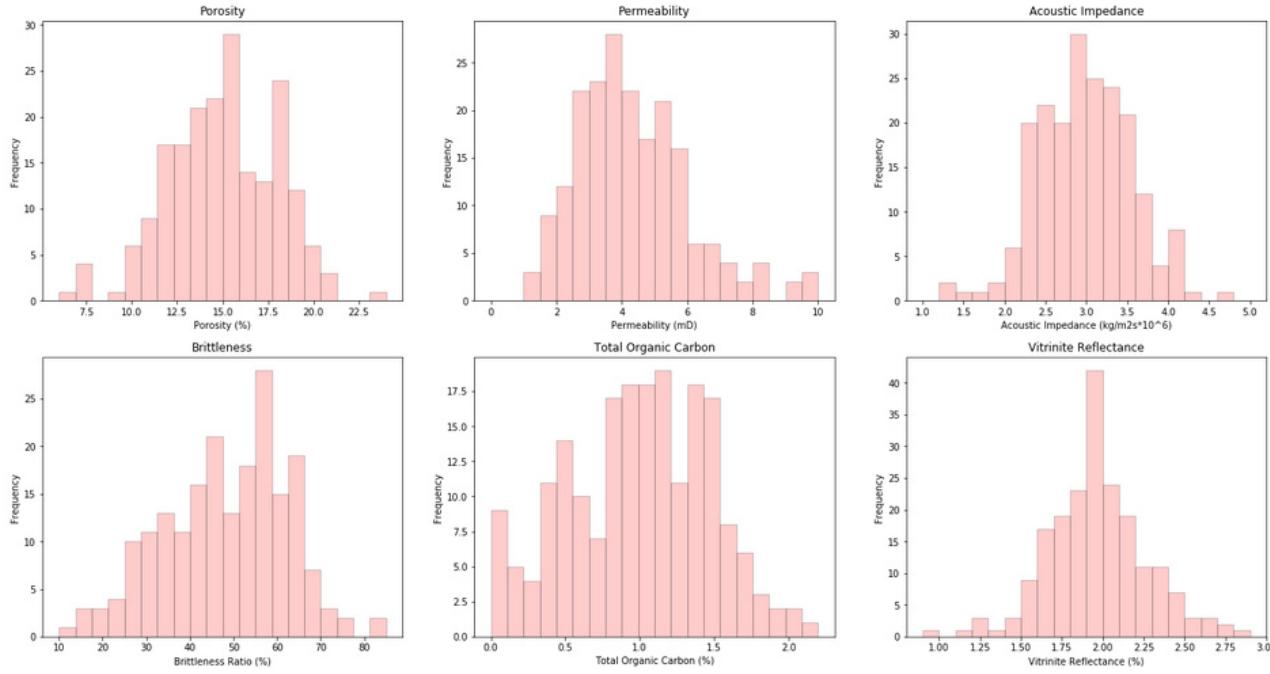
Feature Ranking Metrics



Metric #2: Univariate Distributions

- As with summary statistics, this ranking method is a qualitative check for issues with the data and to assess our confidence with each feature.
- It is better to not include a feature with low confidence of quality as it may be misleading (while adding to model complexity as discussed previously).
- Assess our ability to use methods that have distribution assumptions

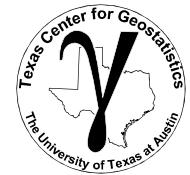
Feature Ranking Metrics



The univariate distributions look good:

- there are no obvious outliers
- the permeability is positively skewed as often observed
- the corrected TOC has a small zero truncation spike, but it's reasonable
- some departure from Gaussian form, could transform

Feature Ranking Metrics



Metric #3: Bivariate

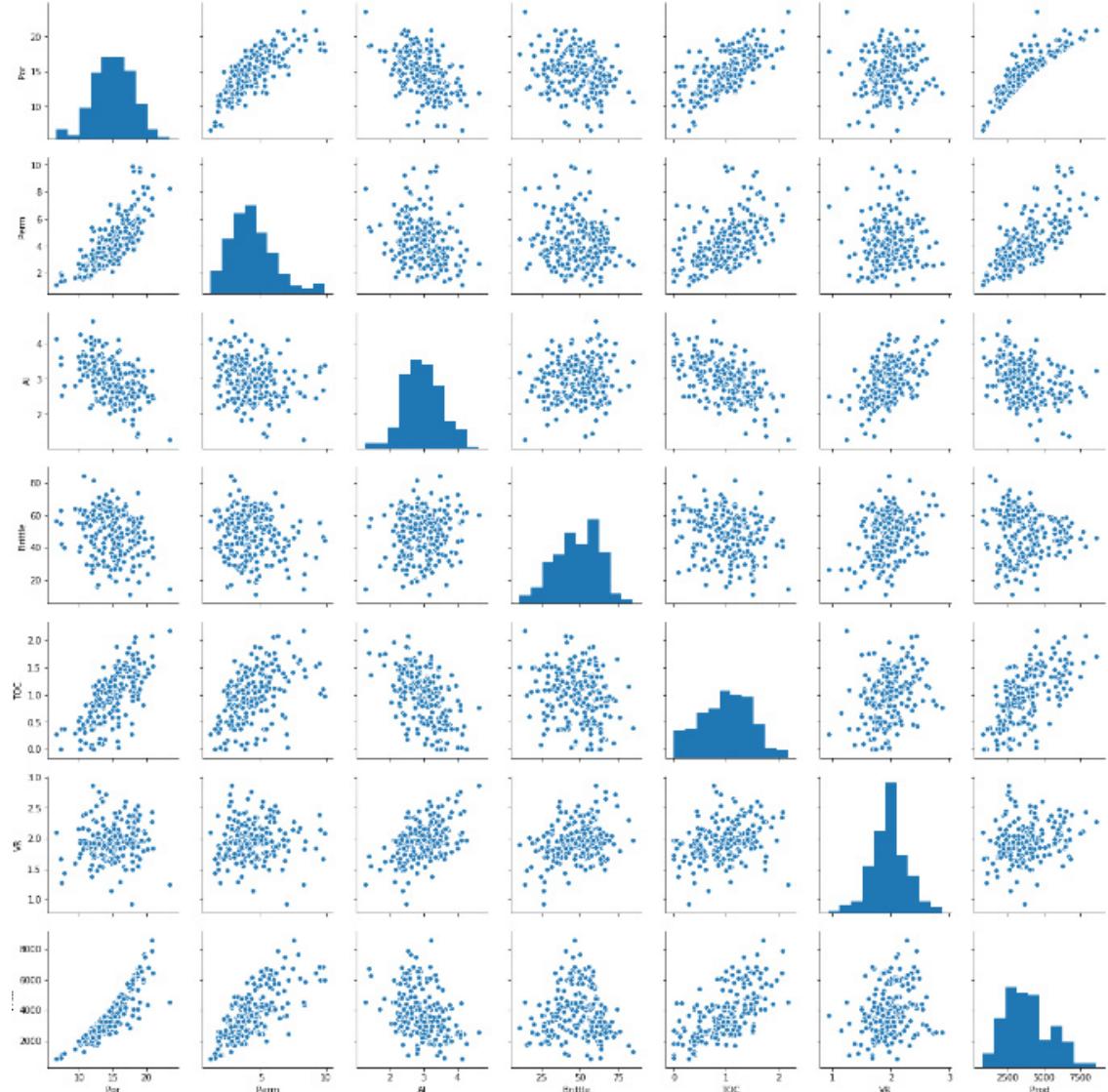
- matrix scatter plots are a very efficient method to observe the bivariate relationships between the variables.
- this is another opportunity through data visualization to identify data issues, outliers
- we can assess if we have collinearity, specifically the simpler form between two features at a time
- Bivariate Gaussian is assumed for methods such as correlation and partial correlation

Feature Ranking Metrics

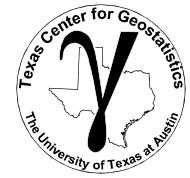


How could we use this plot for variable ranking?

- variables that are closely related to each other.
- linear vs. non-linear relationships
- constraint relationships and heteroscedasticity between variables.



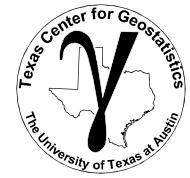
Feature Ranking Metrics



Metric #3: Bivariate

- bivariate visualization and analysis is not sufficient to understand all the multivariate relationships in the data
- multicollinearity includes strong linear relationships between 2 or more features.
- higher order nonlinear features, outliers and coverage?
- these may be hard to see with only bivariate plots.

Feature Ranking Metrics



Ranking Method #4 - Pairwise Covariance

- Pairwise covariance provides a measure of the strength of the linear relationship between each predictor feature and the response feature.
- We now specify our goal of this study is to predict production, our response variable, from the other available predictor features.
- We are thinking predictively now, not inferentially, we want to estimate the function, \hat{f} to accomplish this

Covariance:

- measures the strength of the linear relationship between features
- sensitive to the dispersion / variance of both the predictor and response

Feature Ranking Metrics



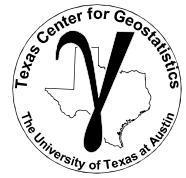
Ranking Method #4 - Pairwise Covariance

- Sensitive to feature variance
- Feature variance is somewhat arbitrary.
 - For example, what is the variance of porosity in fraction vs. percentage or permeability in Darcy vs. milliDarcy. We can show that if we apply a constant multiplier, c , to a variable, X , that the variance will change according to this relationship (the proof is based on expectation formulation of variance):

$$\sigma_{cX}^2 = c^2 \sigma_X^2$$

- By moving from percentage to fraction we decrease the variance of porosity by a factor of 10,000!
- The variance of each variable is potentially arbitrary, with the exception when all the features are in the same units.

Feature Ranking Metrics



Ranking Method #5 - Pairwise Correlation Coefficient

- Pairwise correlation coefficient provides a measure of the strength of the linear relationship between each predictor feature and the response feature.
- The correlation coefficient:
 - measures the linear relationship
 - removes the sensitivity to the dispersion / variance of both the predictor and response features, by normalizing by the product of the standard deviation of each feature

Feature Ranking Metrics



Ranking Method #6 – Rank Correlation Coefficient

- The rank correlation coefficient applies the rank transform to the data prior to calculating the correlation coefficient. To calculate the rank transform simply replace the data values with the ranks, where n is the maximum value and 1 is the minimum value.
- The rank correlation:
 - measures the monotonic relationship, relaxes the linear assumption
 - removes the sensitivity to the dispersion / variance of both the predictor and response, by normalizing by the product of the standard deviation of each.

Feature Ranking Metrics

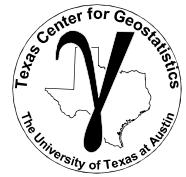


Ranking Method #7 – Partial Correlation Coefficient

This is a linear correlation coefficient that controls for the effects all the remaining variables

- $\rho_{XY.Z}$ and is the partial correlation between X and Y after controlling for Z .
1. perform linear, least-squares regression to predict X from $Z_{1,...,m-2}$.
 2. calculate the residuals in Step #1, $X - X^*$
 3. perform linear, least-squares regression to predict Y from $Z_{1,...,m-2}$.
 4. calculate the residuals in Step #1, $Y - Y^*$
 5. calculate the correlation coefficient, $\rho_{XY.Z} = \rho_{X - X^*, Y - Y^*}$

Feature Ranking Metrics



Ranking Method #7 – Partial Correlation Coefficient

The partial correlation, provides a measure of the linear relationship between X and Y while controlling for the effect of Z other features on both, X and Y

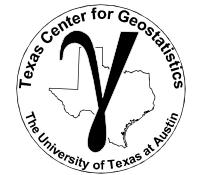
To use this method we must assume:

- two variables to compare, X and Y
- other variables to control, $Z_{1,\dots,m-2}$.
- linear relationships between all variables
- no significant outliers
- approximately bivariate normality between the variables

We are in pretty good shape, but we have some departures from bivariate normality.

- We apply a Gaussian transform in the demonstration

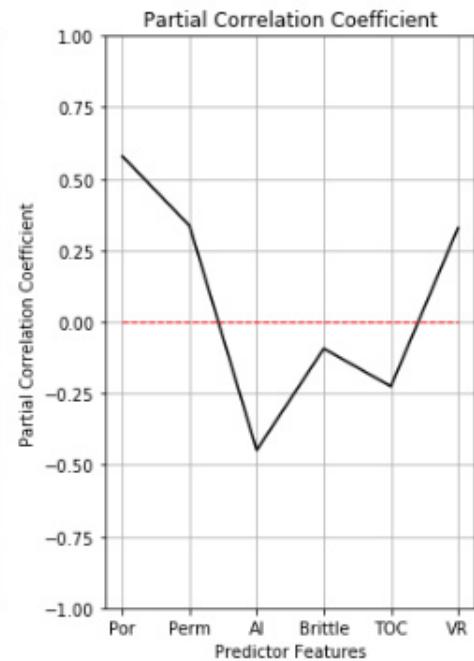
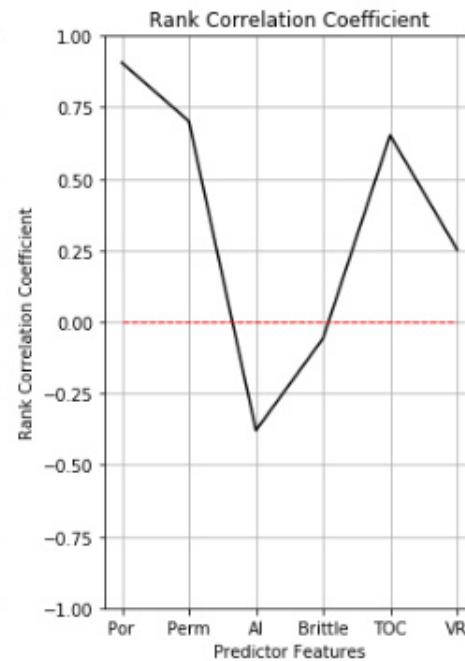
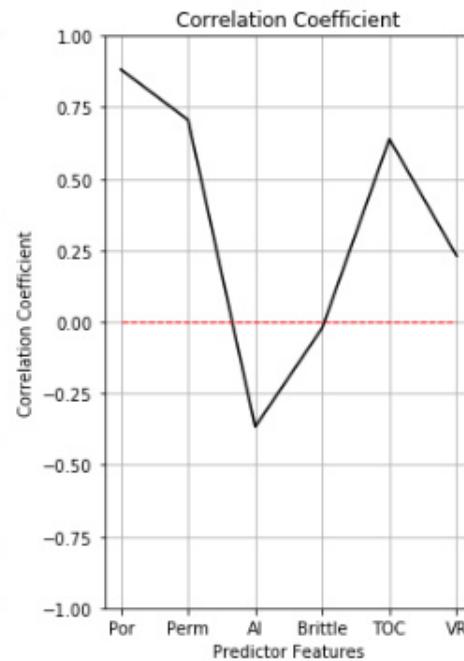
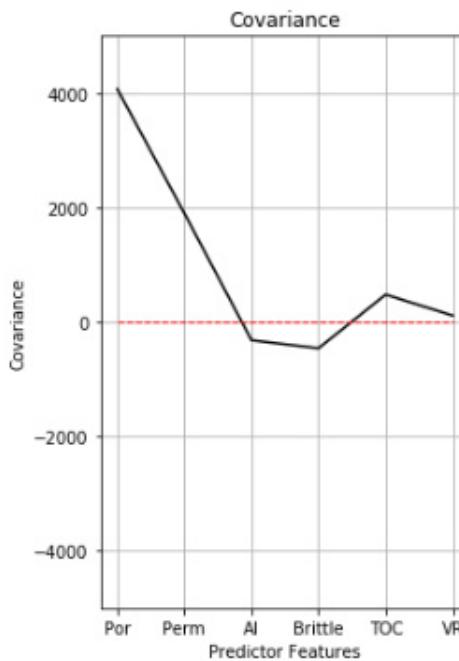
Feature Ranking Metrics



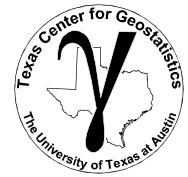
Ranking Methods #4 - #7 – Results

Are we converging on porosity, permeability and vitrinite reflectance as the most important variables with respect to linear relationships with the production?

- What about brittleness?



Feature Ranking Metrics



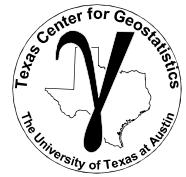
Ranking Method # 9 – Model-based Ranking – B coefficients

- We could also consider B coefficients from linear regression.

$$Y^* = \sum_{i=1}^m B_i X_i + c$$

- These are the linear regression coefficients without standardization of the variables.
- Sensitive to feature variance.
- We are capturing interactions between variables.

Feature Ranking Metrics



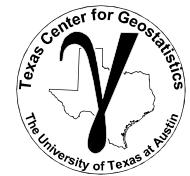
Ranking Method # 9 – Model-based Ranking – B (beta) coefficients

- We could also consider B coefficients from linear regression

$$Y^{s*} = \sum_{i=1}^m B_i X_i^s + c$$

- These are the linear regression coefficients with standardization of the variables, X_i^s and Y^{s*} (variance = 1)
- Not sensitive to variance of the features
- We are capturing interactions between variables.

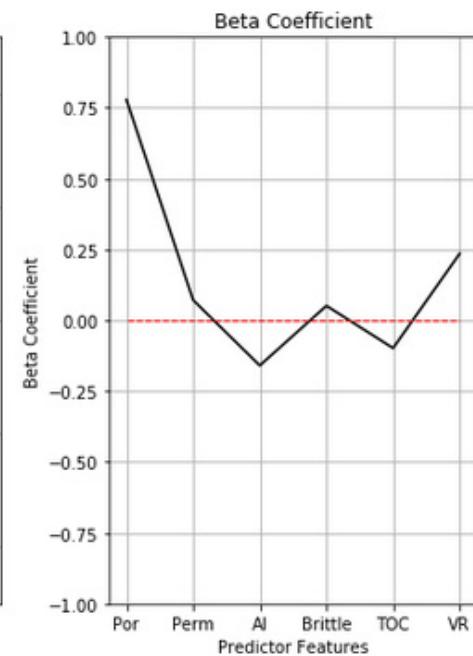
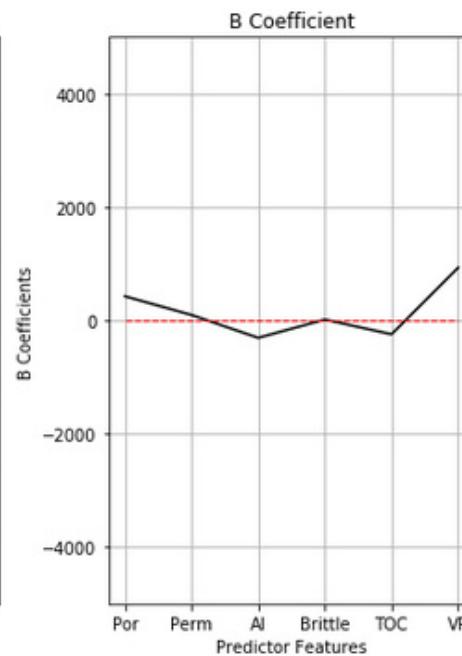
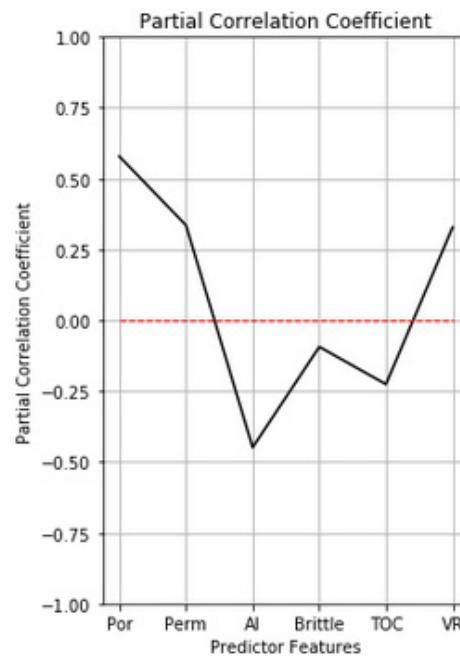
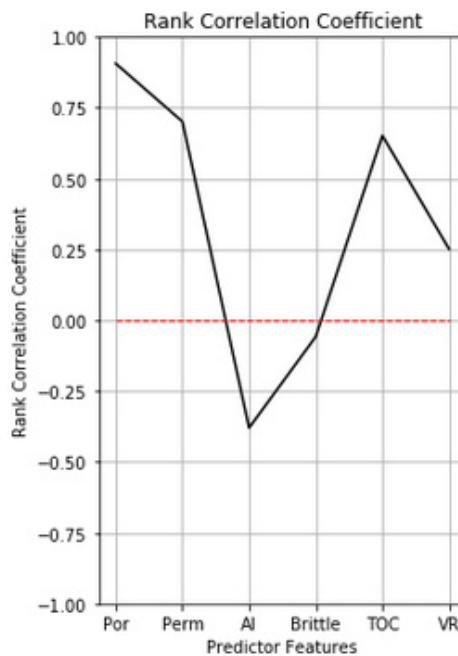
Feature Ranking Metrics



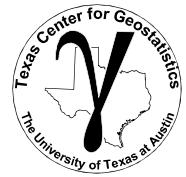
Ranking Methods #4 - #9 – Results

Now what do we see?

- Beta demotes permeability!
- Porosity, acoustic impedance and vitrinite reflectance retain high metrics



Feature Ranking Metrics

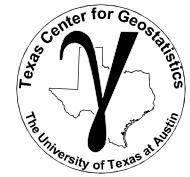


Ranking Methods #11– Recursive Feature Elimination

Recursive Feature Elimination (RFE) method works by recursively removing features and building a model with the remaining features.

- model accuracy is applied to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute
- any model could be used!
- in this example the prediction model based on multilinear regression and indicate that we want to find the best feature based on recursive feature elimination.
- the method assigns rank $1, \dots, m$ for all features.

Feature Ranking Metrics



Ranking Methods #11 – Recursive Feature Elimination

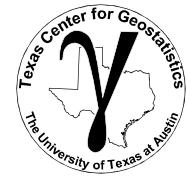
The recursive feature elimination method with a linear regression model provides these ranks:

1. Total Organic Carbon
2. Vitrinite Reflectance
3. Acoustic Impedance
4. Porosity
5. Permeability
6. Brittleness

A couple of the features moved from our previous assessment, but we are close. The advantages with the recursive elimination method:

- the actual model can be used in assessing feature ranks
- the ranking is based on accuracy of the estimate

Feature Ranking Metrics



Ranking Methods #11– Recursive Feature Elimination

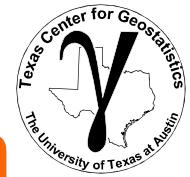
The recursive feature elimination method with a linear regression model provides these ranks, but this method is sensitive to:

- choice of model
- training dataset

This method may be applied with cross validation (k fold iteration of training and testing datasets)

- optimize variable selection for prediction with testing data after training with training data

Feature Ranking Demonstration in Python



Demonstration of the wide array approach with a documented workflow.

GeostatsPy: Multivariate Analysis for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

[Twitter](#) | [GitHub](#) | [Website](#) | [Google Scholar](#) | [Book](#) | [YouTube](#) | [LinkedIn](#)

Subsurface Machine Learning: Feature Ranking for Subsurface Data Analytics

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

Variable Ranking

There are often many predictor features, input variables, available for us to work with for subsurface prediction. There are good reasons to be selective, throwing in every possible feature is not a good idea! In general, for the best prediction model, careful selection of the fewest features that provide the most amount of information is the best practice.

Here's why:

- more variables result in more complicated workflows that require more professional time and have increased opportunity for blunders
- higher dimensional feature sets are more difficult to visualize
- more complicated models may be more difficult to interrogate, interpret and QC
- inclusion of highly redundant and colinear variables increases model instability and decreases prediction accuracy in testing
- more variables generally increase the computational time required to train the model and the model may be less compact and portable
- the risk of overfit increases with the more variables, more complexity

What is Feature Ranking?

Feature ranking is a set of metrics that assign relative importance or value to each feature with respect to information contained for inference and importance in predicting a response feature. There are a wide variety of possible methods to accomplish this. My recommendation is a 'wide-array' approach with multiple metric, while understanding the assumptions and limitations of each metric.

Here's the general types of metrics that we will consider for feature ranking.

1. Visual Inspection of Data Distributions and Scatter Plots
2. Statistical Summaries

Workflow at

https://github.com/GeostatsGuy/PythonNumericalDemos/blob/master/GeostatsPy_variable_ranking.ipynb



Data Analytics and Geostatistics: Multivariate Analysis

Lecture outline . . .

- Multivariate Estimation

Introduction

Modeling Prerequisites

Spatial Estimation

Spatial Uncertainty

Multivariate, Spatial

Multivariate Analysis

Machine Learning

Novel Workflows

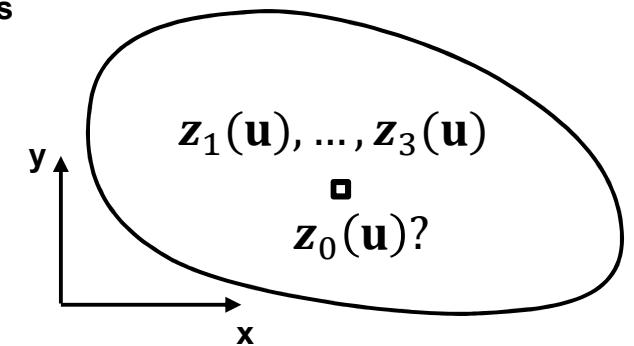
Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin

Multivariate Kriging

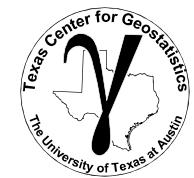
- Simple kriging may be applied to make estimates given a set of collocated secondary variables at the location to estimate the primary variable.
- This is not spatial estimation, but multivariate estimation!

$$\begin{array}{ll}
 \text{Covariance between secondary} & \text{Covariance between secondary} \\
 \text{variables} & \text{and primary variables} \\
 \left[\begin{array}{ccc} C(\mathbf{z}_1, \mathbf{z}_1) & C(\mathbf{z}_1, \mathbf{z}_2) & C(\mathbf{z}_1, \mathbf{z}_3) \\ C(\mathbf{z}_2, \mathbf{z}_1) & C(\mathbf{z}_2, \mathbf{z}_2) & C(\mathbf{z}_2, \mathbf{z}_3) \\ C(\mathbf{z}_3, \mathbf{z}_1) & C(\mathbf{z}_3, \mathbf{z}_2) & C(\mathbf{z}_3, \mathbf{z}_3) \end{array} \right] \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} C(\mathbf{z}_o, \mathbf{z}_1) \\ C(\mathbf{z}_o, \mathbf{z}_2) \\ C(\mathbf{z}_o, \mathbf{z}_3) \end{bmatrix} & \\
 \text{redundancy} & \text{closeness}
 \end{array}$$



- Given the assumption of Gaussian distributed variables we have a complete model of uncertainty for the primary variable at location \mathbf{u} !
- We can back transform for uncertainty in the original variable units.

Multivariate Kriging Hands-on



Here's an opportunity for experiential learning with Simple Kriging for multivariate estimation and uncertainty.

- Things to try:

Observe the multivariate weights, estimator and variance. Walk through the steps.

Kriging-based Multivariate Prediction This is an example, demonstration of kriging for multivariate prediction, Michael Pyrcz, the University of Texas at Austin

X	Y	Facies	Porosity	Perm	AI	St_Por	St_Per	St_AI	St_Por_est	Per_std	P10	P90
100	900	1	0.1002	1.36389	510.7	-1.0078	-0.5279	0.6876	-0.6195	0.025	0.088	0.151
100	800	0	0.0794	12.5768	4671.454	-0.8519	-0.4955	0.3546	-0.3568	0.1326	0.025	0.164
100	700	0	0.0854	5.98452	6127.548	-1.3057	-0.5145	1.4858	-1.2193	0.09	0.025	0.058
100	600	0	0.1085	2.44668	5021.636	-0.8416	-0.5247	0.7566	-0.6723	0.1163	0.025	0.095
100	500	0	0.1025	1.95226	385.27	-0.962	-0.5262	-0.2793	0.1273	0.1567	0.025	0.125
100	400	0	0.1106	3.69191	5235.267	-0.7931	-0.5211	0.8275	-0.7265	0.1142	0.025	0.082
100	300	0	0.0883	10.7356	6744.996	-1.2338	-0.5287	1.9266	-1.5717	0.0718	0.025	0.040
100	200	0	0.1021	2.39619	5347.338	-0.9695	-0.5249	1.3219	-1.1092	0.0951	0.025	0.063
100	100	1	0.1375	5.7276	5823.247	-0.2592	-0.5152	1.2278	-0.1043	0.0988	0.025	0.067
200	900	1	0.1371	14.7713	5621.147	-0.2671	-0.4891	1.0746	-0.9123	0.105	0.025	0.073
200	800	0	0.126	10.6754	4232.701	-0.4896	-0.5003	0.0675	-0.1358	0.1436	0.025	0.112
200	700	0	0.1218	3.08583	5397.4	-0.5746	-0.5229	0.905	-0.7867	0.1112	0.025	0.079
200	600	0	0.0951	0.96257	4619.786	-1.1031	-0.523	0.3155	-0.33216	0.1338	0.025	0.102
200	500	0	0.0875	18.2327	4949.881	-1.2627	-0.5265	0.5657	-0.52513	0.1242	0.025	0.092
200	400	0	0.0986	4.57102	5789.623	-1.04	-0.5188	1.2023	-0.1076	0.0998	0.025	0.068
200	300	0	0.1074	13.5819	7818.899	-0.8621	-0.4925	2.7885	-2.23717	0.039	0.025	0.007
200	200	0	0.0935	0.4088	6104.843	-1.1413	-0.5306	1.4413	-1.20245	0.0905	0.025	0.053
200	100	0	0.0793	0.73455	6485.732	-1.1444	-0.5237	1.73	-1.42543	0.0794	0.025	0.048
300	900	1	0.1115	27.9398	4183.467	-0.7802	-0.4508	-0.0153	-0.06347	0.1472	0.025	0.175
300	800	1	0.1195	61.0054	5224.544	-0.6205	-0.3552	0.7733	-0.65742	0.1176	0.025	0.086
300	700	1	0.1342	44.5353	5656.541	-0.3246	-0.4027	0.2675	-0.27401	0.1367	0.025	0.169
300	600	1	0.1512	190.74	3397.081	0.2179	0.0203	-0.1566	0.12447	0.1566	0.025	0.188
300	500	1	0.1072	13.3016	5684.318	-0.7771	-0.4933	1.1229	-0.9502	0.1031	0.025	0.071
300	400	1	0.1117	10.8511	5493.753	-0.7771	-0.5003	0.9826	-0.8429	0.1084	0.025	0.077
300	300	0	0.1052	6.3122	4715.684	-0.9064	-0.5135	0.3862	-0.38575	0.1312	0.025	0.093
300	200	0	0.1033	2.87135	6217.196	-0.9461	-0.5237	1.5265	-1.2671	0.0873	0.025	0.055
300	100	0	0.0991	3.18878	6456.338	-1.0305	-0.5226	1.7078	-1.40703	0.0603	0.025	0.048
400	900	0	0.0785	1.1038	5440.283	-1.4436	-0.5286	0.9375	-0.81278	0.1039	0.025	0.078
400	800	0	0.1104	4.11622	5956.235	-0.8017	-0.5139	1.3287	-1.11363	0.0949	0.025	0.063
400	700	1	0.1228	16.827	4937.337	-0.5538	-0.4773	0.5562	-0.50356	0.125	0.025	0.093
400	600	1	0.1208	5.57043	4040.573	-0.5933	-0.5143	-0.1236	0.009597	0.1508	0.025	0.183
400	500	1	0.114	132.124	4459.013	-0.7311	-0.4933	0.1936	-0.17455	0.1417	0.025	0.174
400	400	1	0.0838	5.55878	5933.782	-1.3362	-0.5157	1.3268	-1.1146	0.095	0.025	0.127
400	300	1	0.118	139.268	5557.636	-0.6503	-0.5261	1.0265	-0.8811	0.1065	0.025	0.075
400	200	0	0.169	6.63131	6735.723	-0.6721	-0.5126	1.3196	-1.5693	0.0722	0.025	0.104
400	100	0	0.0684	0.03361	6124.087	-1.6868	-0.5317	1.4553	-1.2139	0.0893	0.025	0.058
500	900	0	0.0934	13.0794	4873.351	-1.1448	-0.526	0.5081	-0.48089	0.1264	0.025	0.095
500	800	0	0.084	0.7714	5264.836	-1.3326	-0.5236	0.8196	-0.72187	0.1144	0.025	0.083
500	700	0	0.1098	52.5009	6581.836	-0.8143	-0.3798	1.8029	-1.4567	0.0778	0.025	0.046
500	600	1	0.122	8.22227	5296.785	-0.5702	-0.508	0.8287	-0.72527	0.1143	0.025	0.082

Workflow Steps:

0. Standardize the variables, the mean and variance have been corrected to 0 and 1 respectively with affine correction (we should use Gaussian transform for the complete workflow).
- 1-7. Multivariate kriging to calculate kriging estimate and kriging variance.
8. Estimate standardized porosity at each location.
9. Calculate the back transform of the kriging estimate and the kriging variance and 10. the local P10 and P90.

1. Calculate the correlation

Por	Perm	AI	
1.00	0.54	-0.85	
Perm	0.54	1.00	-0.49
All	-0.85	-0.49	1.00

2. Calculate the covariance of the standardized variables

Por	Perm	AI	
1.00	0.54	-0.85	
Perm	0.54	1.00	-0.49
All	-0.85	-0.49	1.00

3. Redundancy Matrix

Perm	AI	
1	-0.49	
Perm	-0.49	1

4. Closeness Matrix

Por	
0.54	
AI	-0.85

5. Invert Redundancy Matrix

Perm	AI	
1.31	0.64	
AI	0.64	1.31

6. Calculate the Weights

Weights	
Perm	0.17
AI	-0.77

7. Calculate the kriging variance

Kriging Variance: 0.25

Comparison to Multilinear Regression

b2: slope of fit	se2: standard error of slope	r2: proportion var. explained	ssreg: explained variance	dffit: d.f. used
-0.774	0.036	0.75	195.2	un
0.0356	0.0356	0.3010267	388.4	un
0.7507	0.5013	#N/A	258	#N/A
388.45	258	#N/A	195.2	#N/A

Test Significance of Coefficients

H0: bi = 0	tstat bi = b1/se1	H1: bi ≠ 0	tstat probability	critical
H0: b1 = 0	21.72	H1: b1 ≠ 0	4.68	2.25
H1: b1 ≠ 0	probability	#####	#####	critical

Result from Hypothesis tests for coefficients: Reject H0: Slope | Reject H0: Intercept = 0

Al Perm

File Name: Kriging_Multivariate_Estimation_Demo.xlsx File is at: <https://git.io/fhALF>

Multivariate New Tools



Topic	Application to Subsurface Modeling
Curse of Dimensionality	<p>Reduce problem to lowest dimension possible.</p> <p><i>Feature ranking determined that porosity may be predicted from acoustic impedance and rock type alone.</i></p>
Feature Selection	<p>Apply wide array methods to explore the importance of each predictor feature with respect to the response feature.</p> <p><i>Partial correlation reveals that rock type provides little additional information to acoustic impedance.</i></p>
Multivariate Kriging	<p><i>Multivariate kriging combines secondary information sources while accounting for closeness and redundancy.</i></p> <p><i>Given secondary data the likelihood distribution for local porosity is mean of 15% and standard deviation of 2.5% with a Gaussian distribution.</i></p>



Data Analytics and Geostatistics: Multivariate Analysis

Lecture outline . . .

- **Multivariate Analysis**
- **Joints and Conditionals**
- **Feature Selection**
- **Multivariate Estimation**

Introduction

Modeling Prerequisites

Spatial Estimation

Spatial Uncertainty

Multivariate, Spatial

Multivariate Analysis

Machine Learning

Novel Workflows

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin