Multivariate Modeling:
Feature Selection

Lecture outline . . .

- **Curse of Dimensionality**
- **Overfit / Model Complexity**
- **Feature Ranking**

Introduction

Prerequisites

Probability

Multivariate Analysis

Spatial Estimation

Feature Selection

Multivariate Modeling

Conclusions

Multivariate Modeling:
Feature Selection

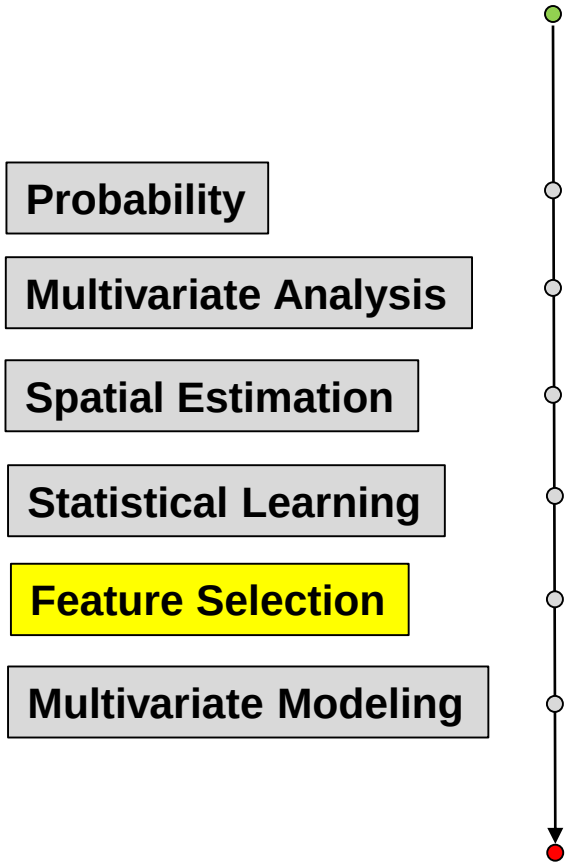
Lecture outline . . .

- **Curse of Dimensionality**

What Will You Learn?

Why Cover Feature Selection?

- Methods for feature selection.
- Careful feature selection results in better models.



Multivariate, Spatial Uncertainty

Motivation for Multivariate Methods

- **We typically need to build reservoir models of more than one property of interest.**
  - Expanded by whole earh modeling, closing loops with forward models
  - Expanded by unconventionals

- **Subsurface properties may include:**
  - Rock Classification: lithology, architectural elements, facies, depofacies
  - Petrophyscial: porosity, directional permeability, saturuations
  - Geophysical: density, p-wave and s-wave velocity
  - Gemechanical: compressibility / Poisson's ratio, Yong's modulus, brittleness, stress field
  - Paleo- / Time Control: fossil adundances, stratigraphic surfaces, ichnofacies, paleo-flow indicators

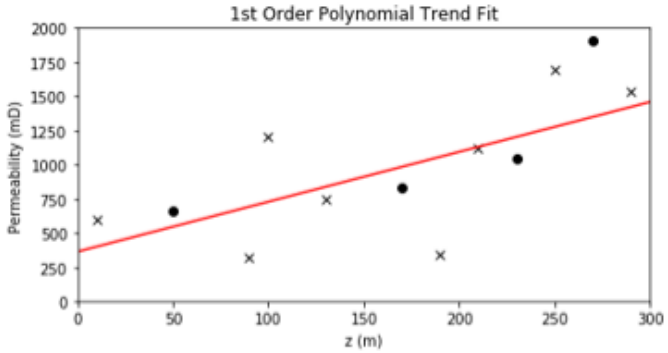
Curse of Dimensionality

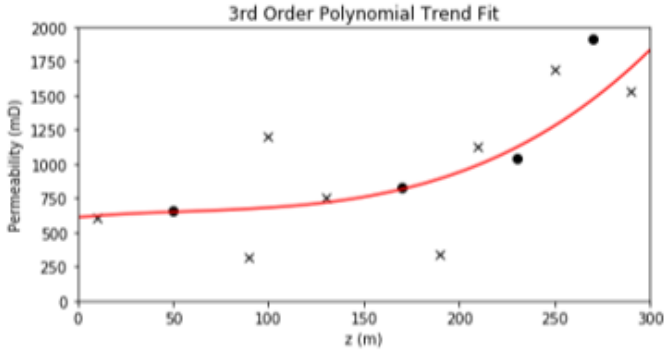
- **Working with more features / variables is harder!**

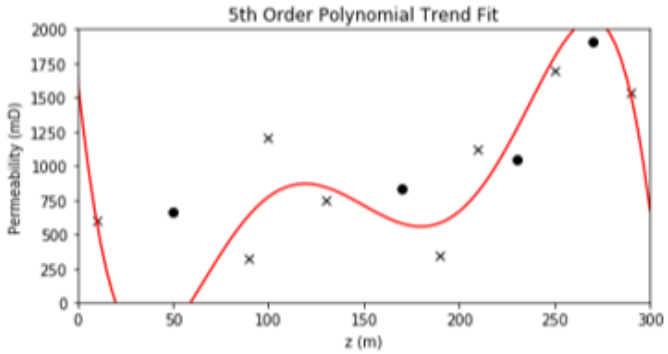
1. More difficult to visualize
2. More data are required to infer the joint probabilities
3. Less coverage
4. More difficult to interrogate / check the model
5. More likely redundant
6. More complicated, more likely overfit

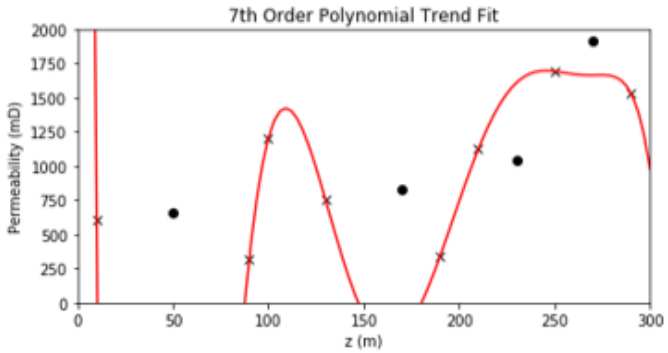
Visualization

- **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?



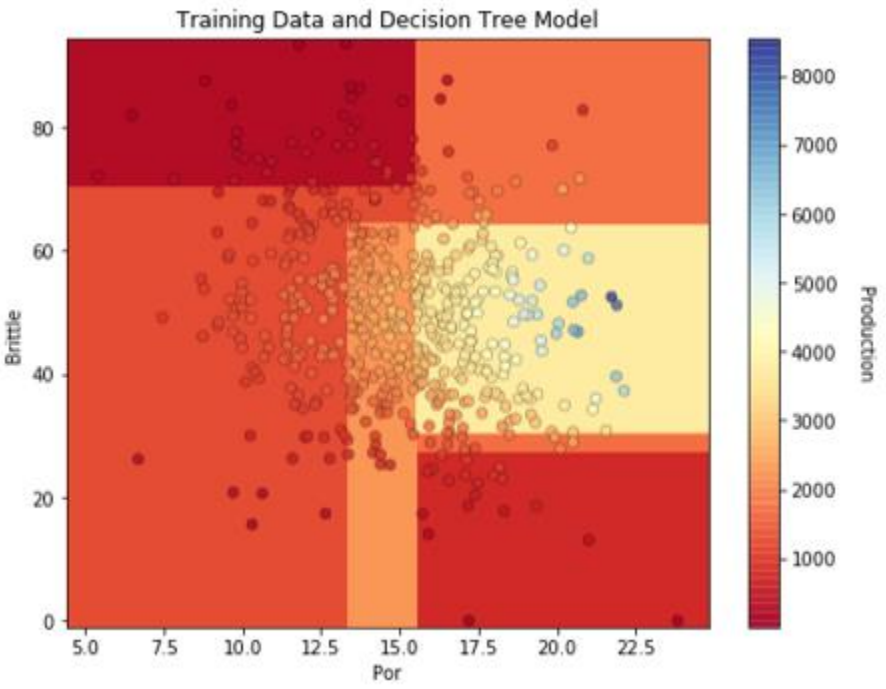






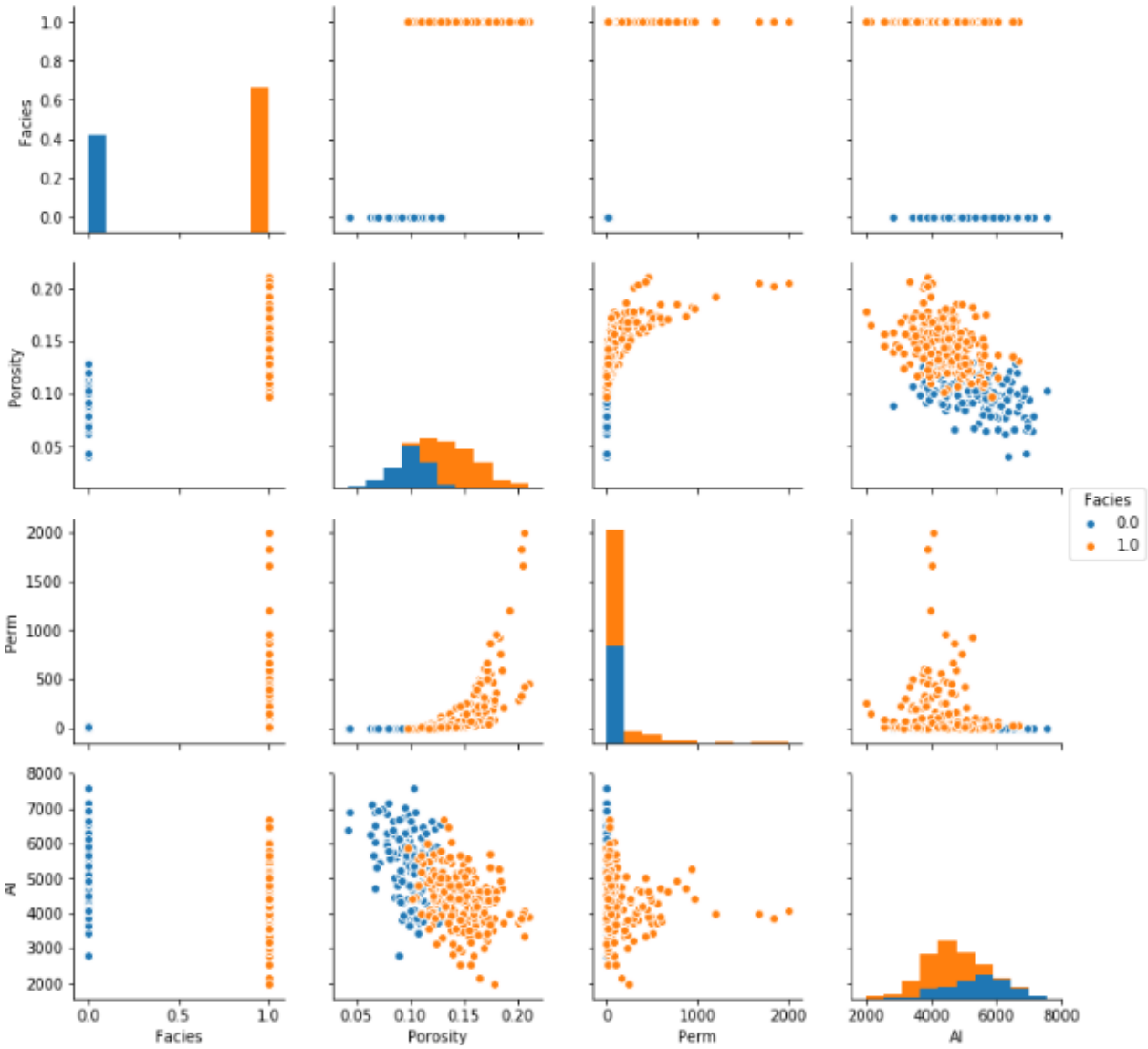
Visualization

- **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?

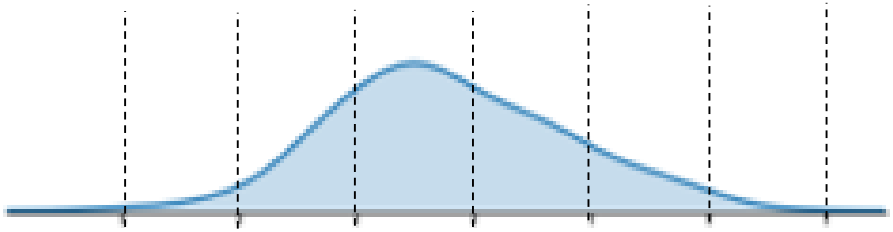


Inferring Joint Probabilities

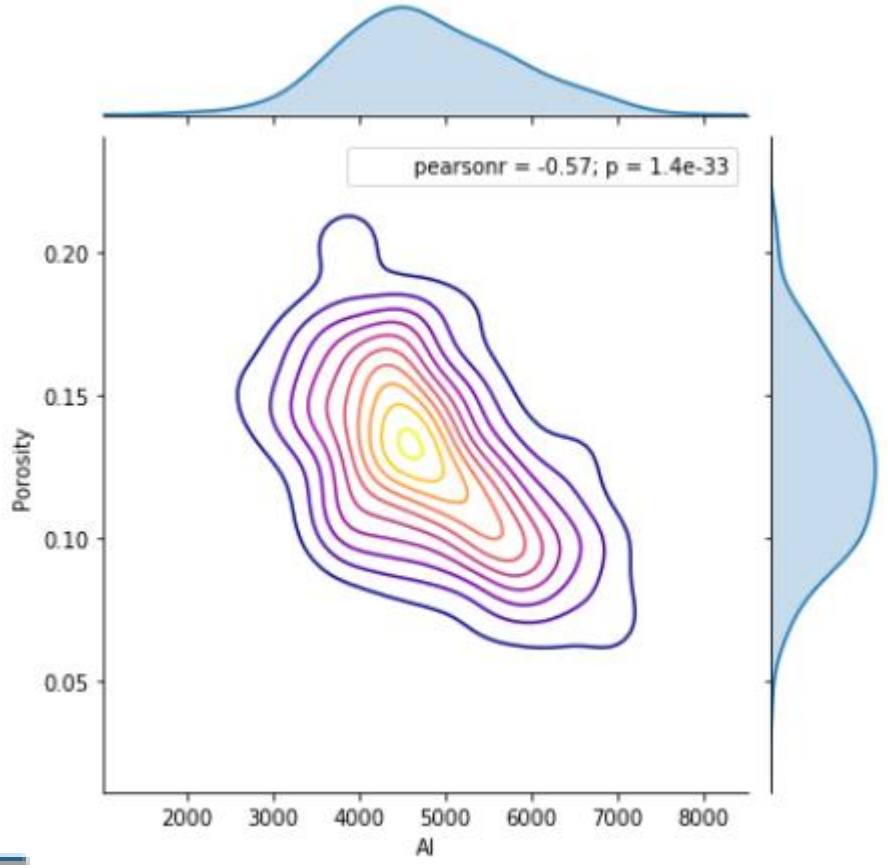
- **Consider any joint probability:**

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

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



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



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

Inferring Joint Probabilities

- **Consider any joint probability:**

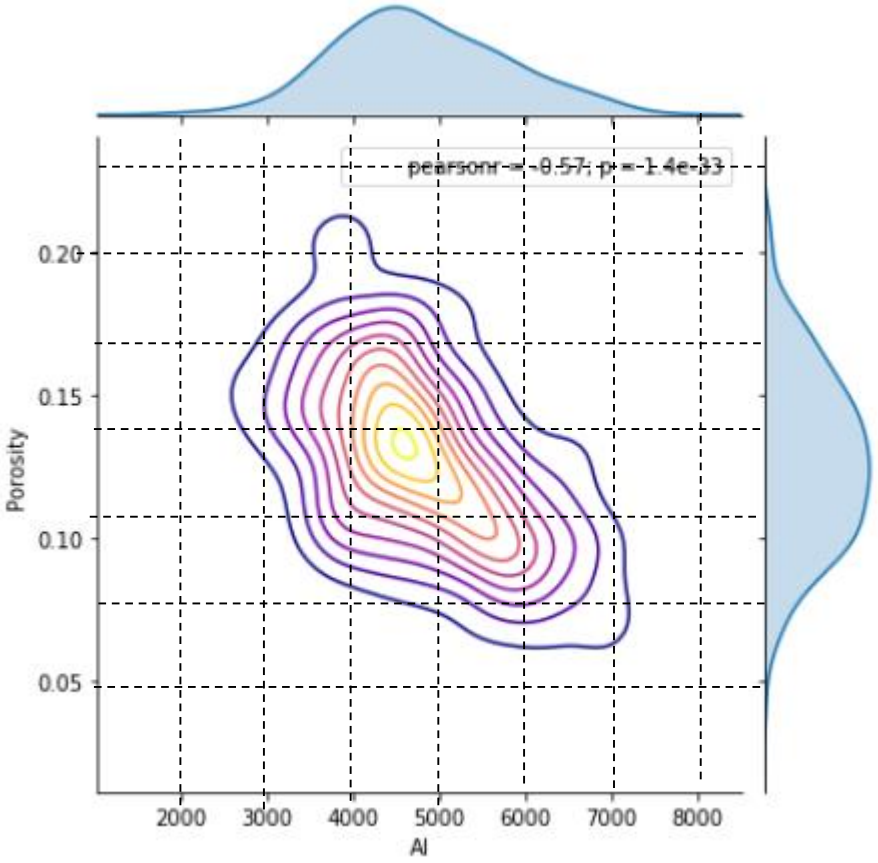
  $P(X_1 \cap, \ldots, \cap X_m)$ the joint probability of $X_1, \ldots, X_m$

- Now move to 2 features (m=2)

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

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

- This is optimistic, as it assumes uniform sampling

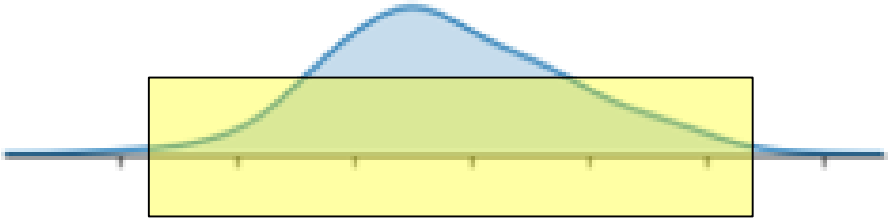


In each bin we are estimating a probability!
10 data in each bin = 640 data?

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:

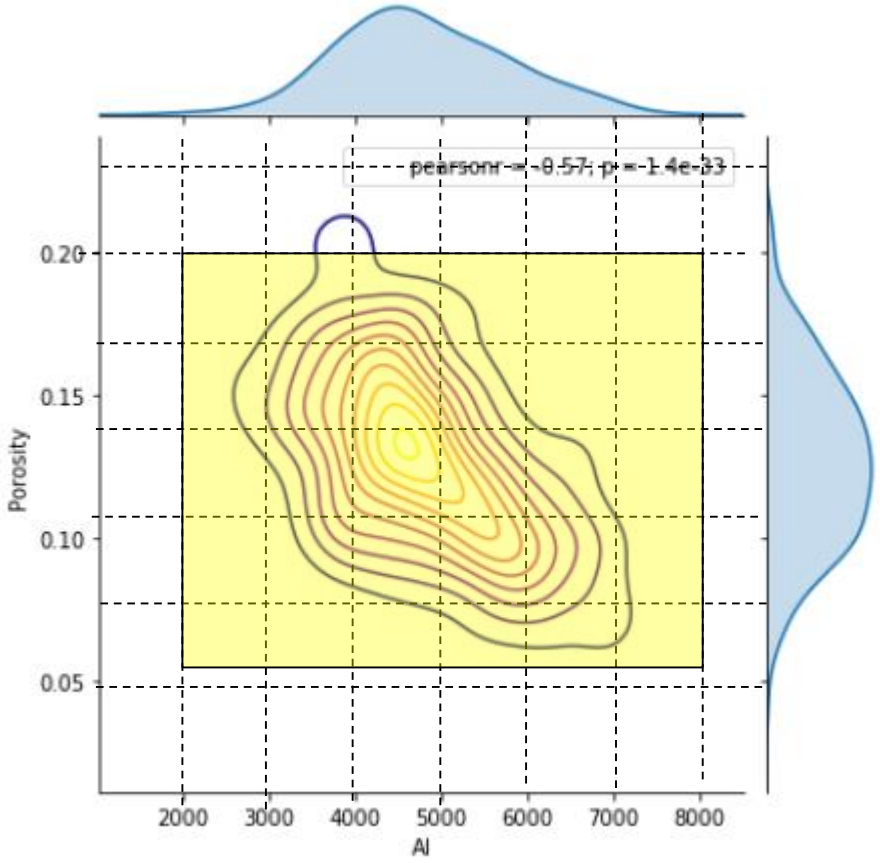
- Now let's move to 2 features, each with 80% coverage
- How much of the solution space is covered?

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

- Even with exponential increase in number of data:

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

coverage is decreasing as we increase the number of features!



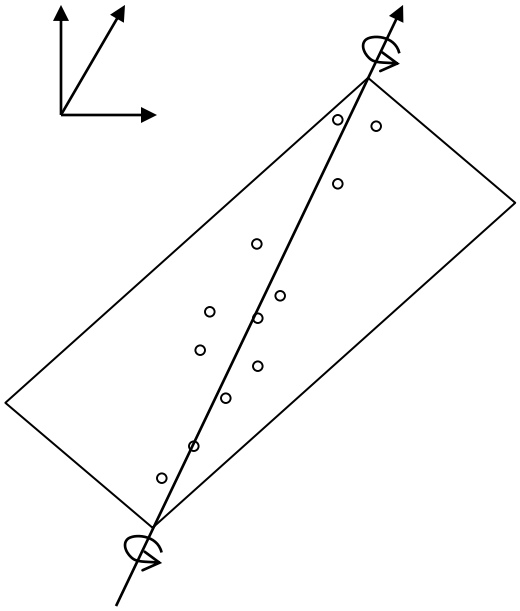
Multicollinearity
Feature Redundancy

"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."

- Wikipedia



It is like fitting a plane to a line!

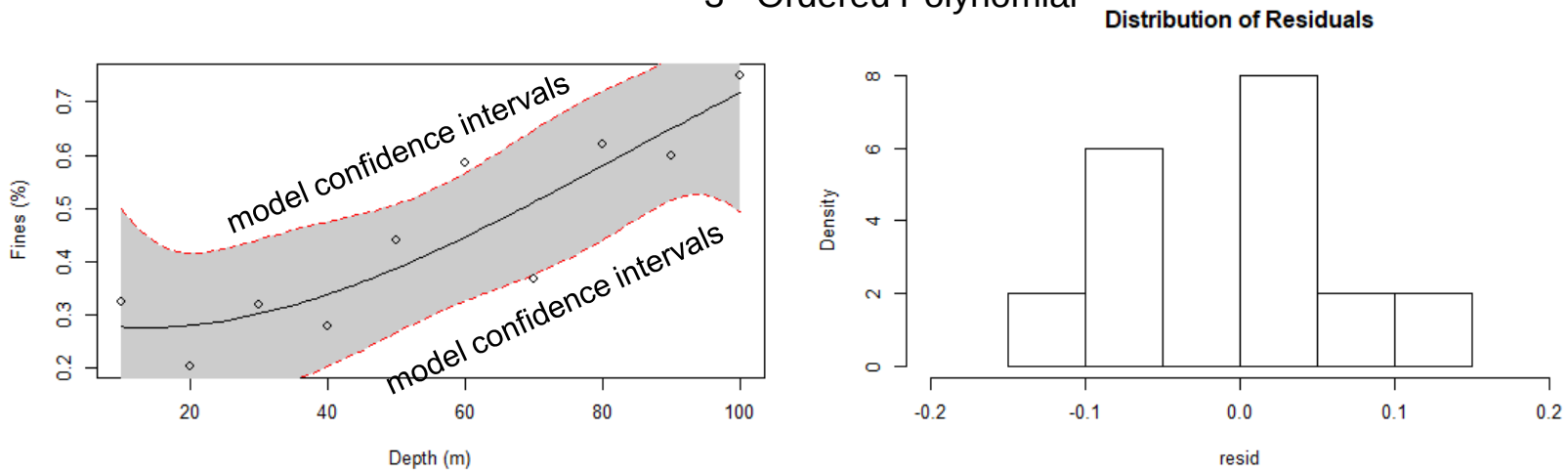
Multivariate Modeling:
Feature Selection

Lecture outline . . .

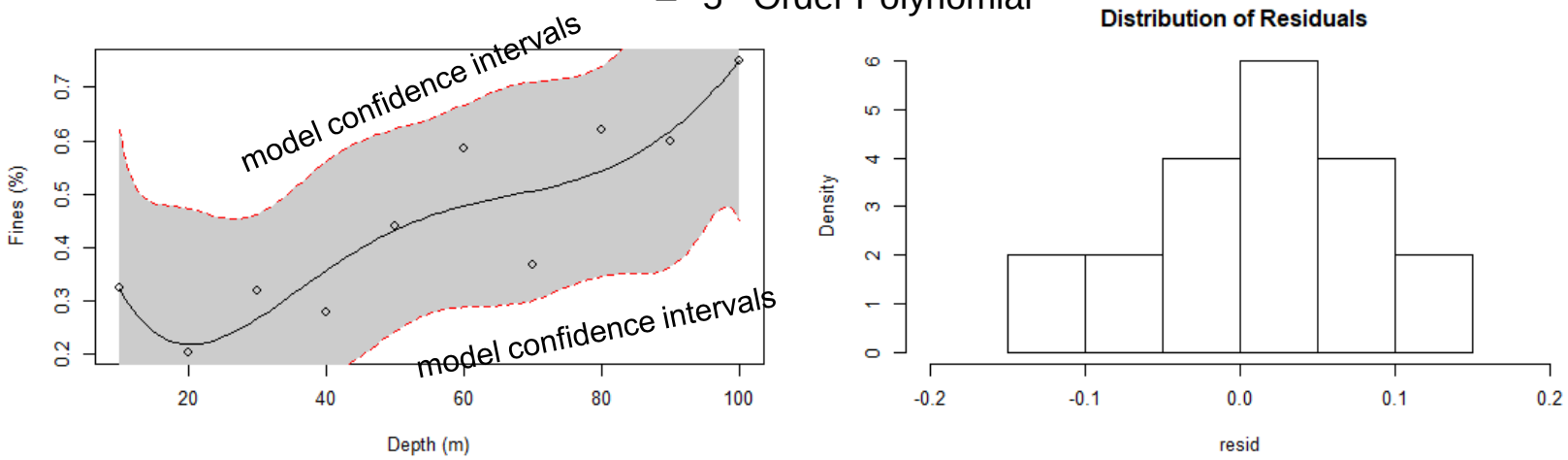
- **Overfit / Model Complexity**

Overfitting

- Example of trend fits:
  - 3rd Ordered Polynomial



  - 5th Order Polynomial

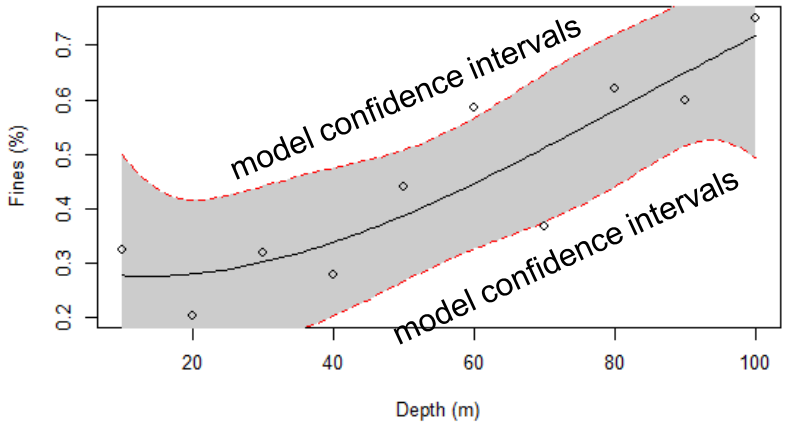


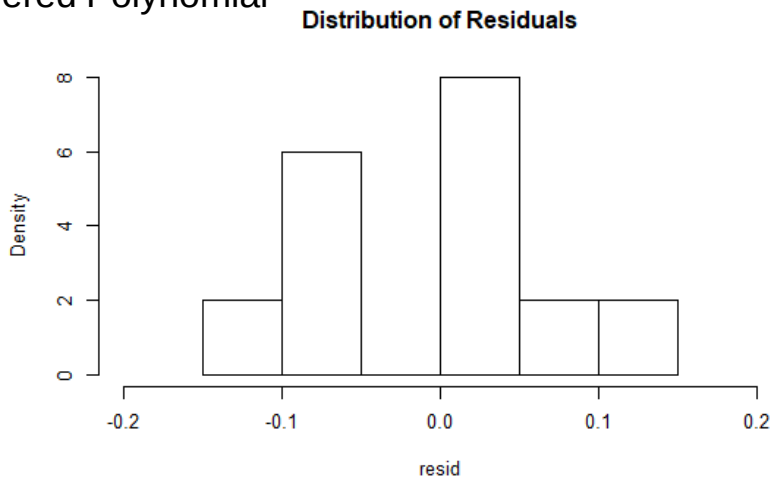
Overfit demonstration in R, code is here:
https://github.com/GeostatsGuy/geostatsr/blob/master/overfit.R

R code at Code/Overfit.R

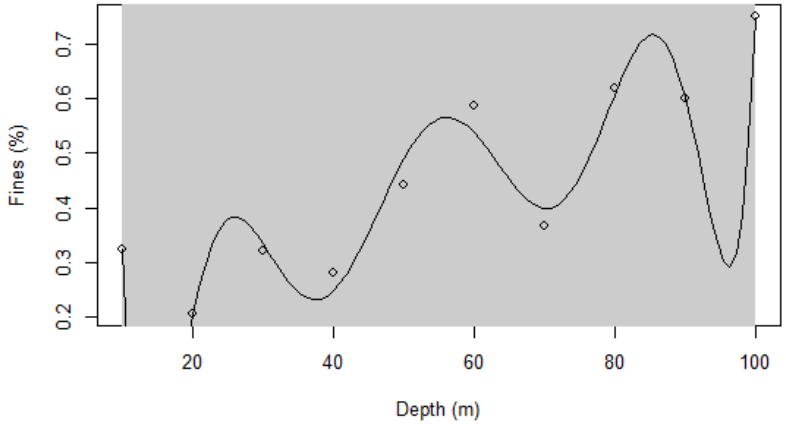
Overfitting

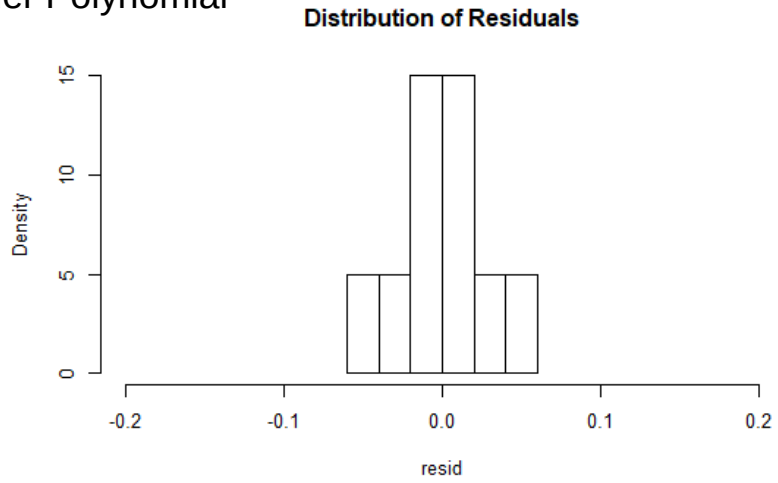
- Example of trend fits:
  - 3rd Ordered Polynomial





  - 8th Order Polynomial





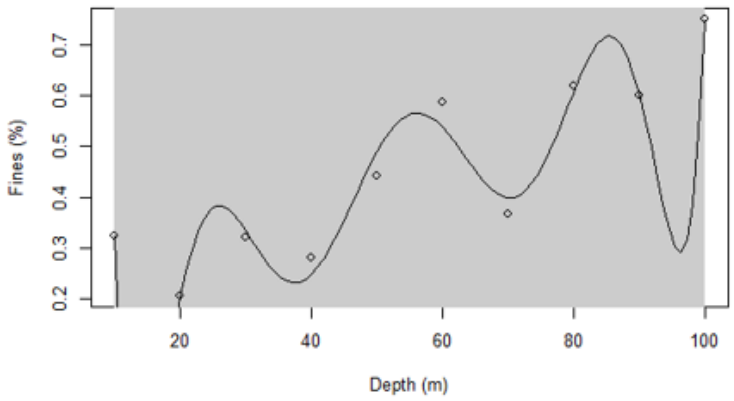
Overfit demonstration in R, code is here:
https://github.com/GeostatsGuy/geostatsr/blob/master/overfit.R

R code at Code/Overfit.R

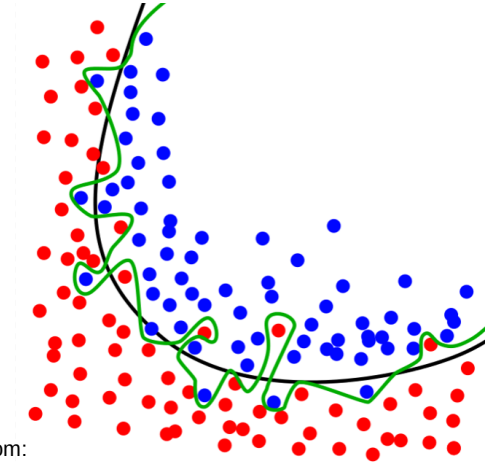
Definition of Overfitting

- Overly complicated model to explain “idiosyncrasies” of the data, capturing data noise in the model
- More parameters than can be justified with the data

- Results in likely very high error away from the data
- But, results in low residual variance!
- High training R^2
- Very accurate at the data! - Claim you know more than you actually do!



Overfit demonstration in R, code is here:
https://github.com/GeostatsGuy/geostatsr/blob/master/overfit.R



Overfit classification model example from:
https://en.wikipedia.org/wiki/Overfitting#/media/File:Overfitting.svg

Complexity / Flexibility

Complexity / Flexibility are Closely Related

- Consider these potential polynomials $\hat{f}$ to predict $\hat{Y}$

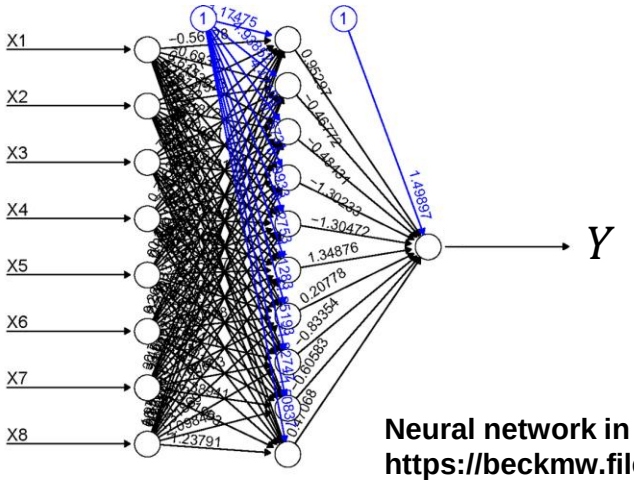
$$Y = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \beta_6 X^6$$

- The 6th order polynomial is more complicated and more flexible to fit the relationship between feature, X, and response, Y
- Now, what if we use 8 bins on X and 10 nodes in a hidden layer of a neural net?:

Indicator Code X into Bins

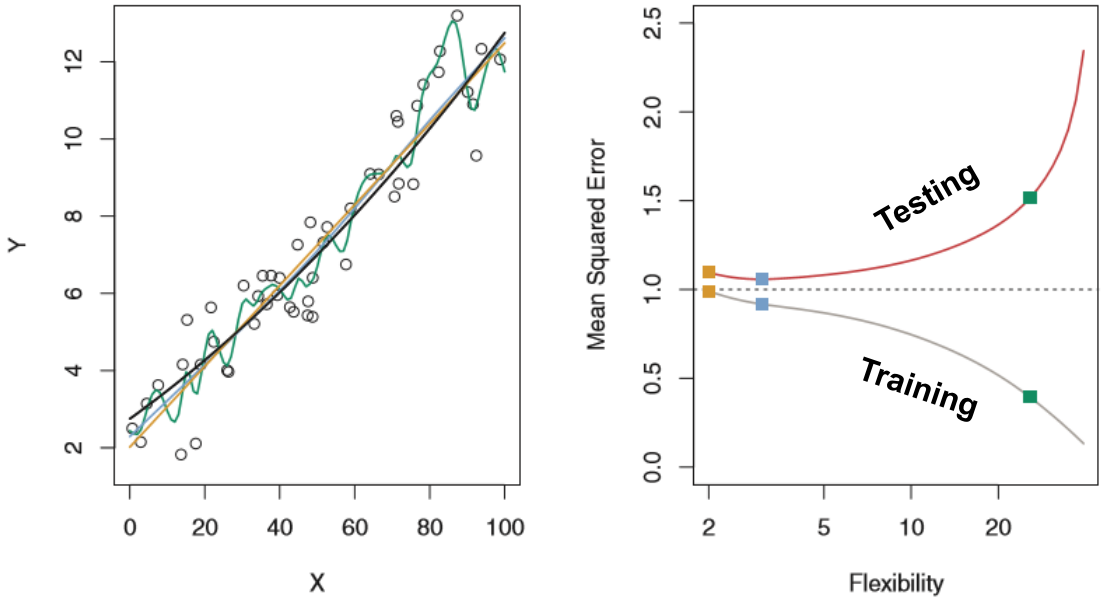
$$I(x; x_k) = \begin{cases} \mathbf{1}, if\ x \in X_k \\ \mathbf{0}, otherwise \end{cases}$$



Neural network in R image from:
https://beckmw.files.wordpress.com/2013/11/neuralnet_plot.jpg

Assessing Model Accuracy

- Flexibility vs. Accuracy
  - Increased flexibility will generally decrease MSE on the **training dataset**
  - May result in increase MSE with **testing data**
  - Not generally a good idea to select method only to minimize training MSE



Data and model fits (left) and MSE for training and testing (right) from James et al. (2013).

  - High flexibility + minimize MSE = likely overfit.

Multivariate Modeling:
Feature Selection

Lecture outline . . .

- **Feature Ranking**

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

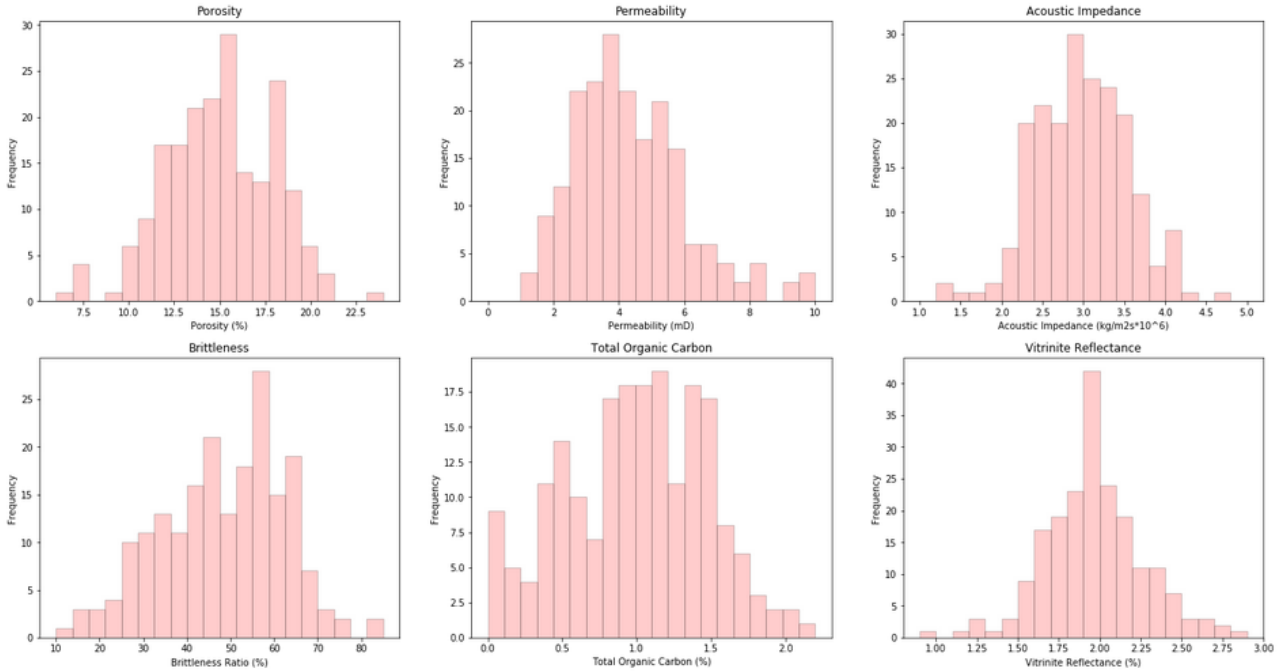
- the number of valid (non-null) values for each feature
- general behaviors such as central tendency, mean, and dispersion, variance.
- issues with negative values, extreme values, and values that are outside the range of plausible values for each property.

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

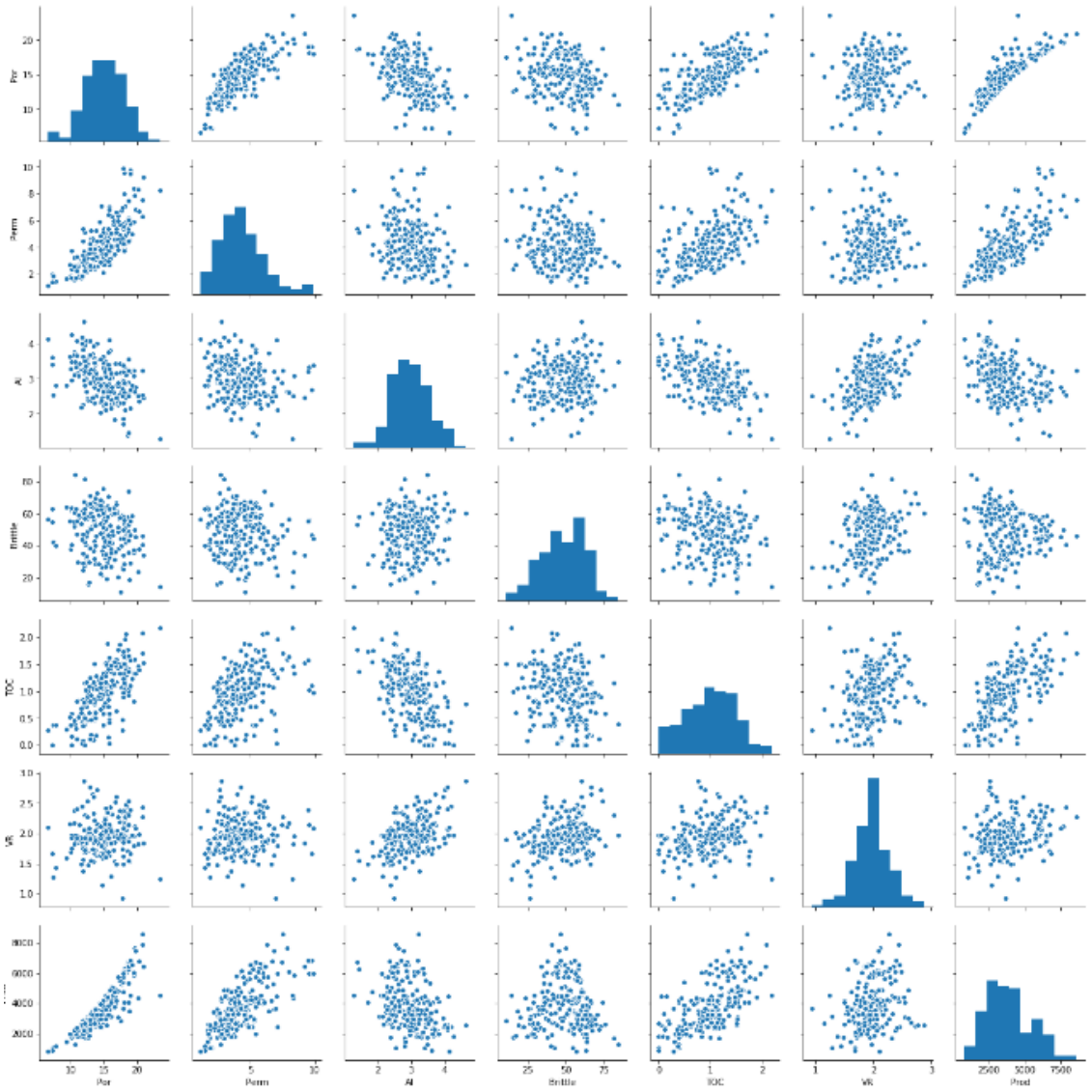
Metric #3: Bivariate

- 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 permeaiblity in Darcy vs. milliDarcy. We can show that if we apply a constant multiplier, cc, to a variable, XX, 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 $Z_{1,\ldots,m-2}$.
2. calculate the residuals in Step #1, $X - X^*$
3. perform linear, least-squares regression to predict Y from $Z_{1,\ldots,m-2}$.
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, $Z_{1,\ldots,m-2}$.
- 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.

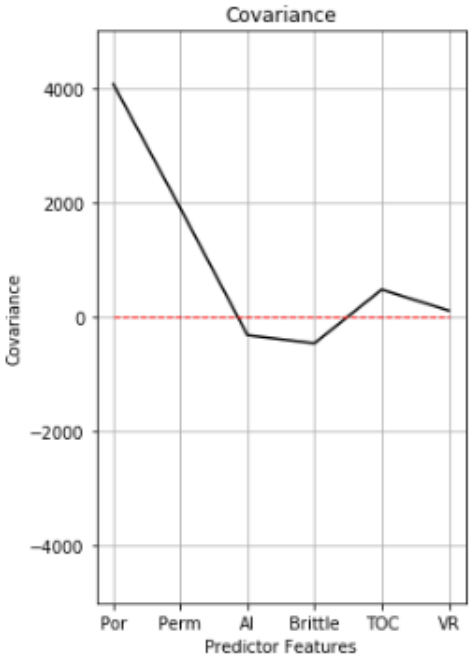
- We apply a Gaussian transform in the demonstration

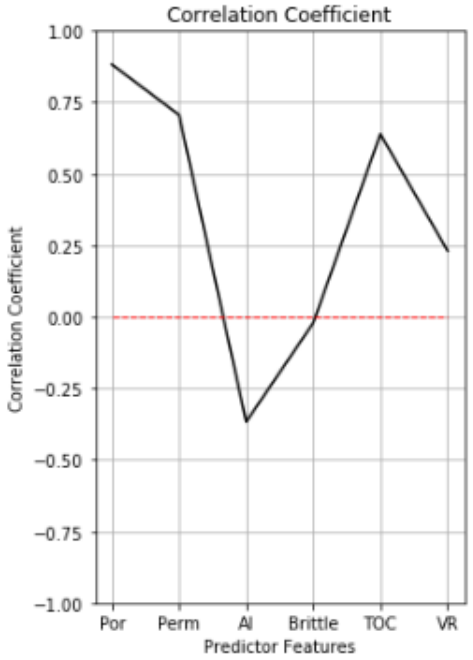
Feature Ranking Metrics

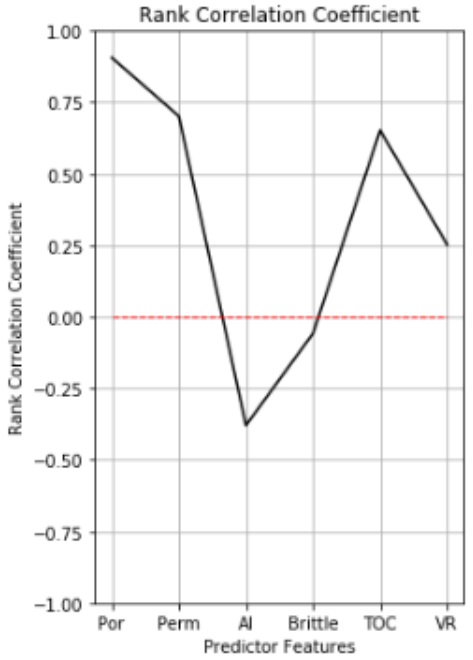
Ranking Methods #4 - #7 – Results

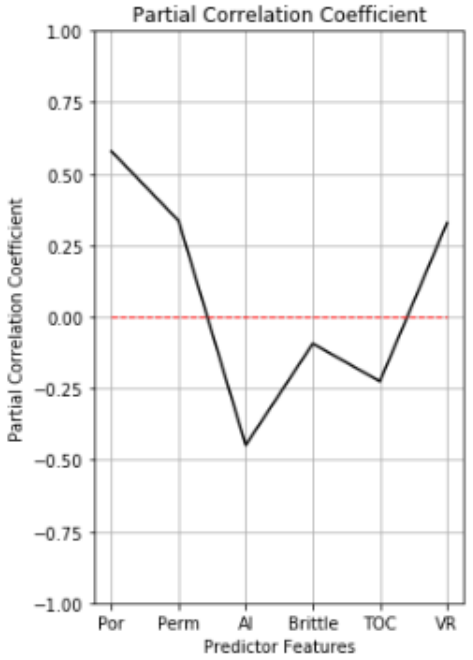
Are we converging on porosity, permeability and vitrinite reflectance as the most important variables with respect to linear relationships with the production?

- 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

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

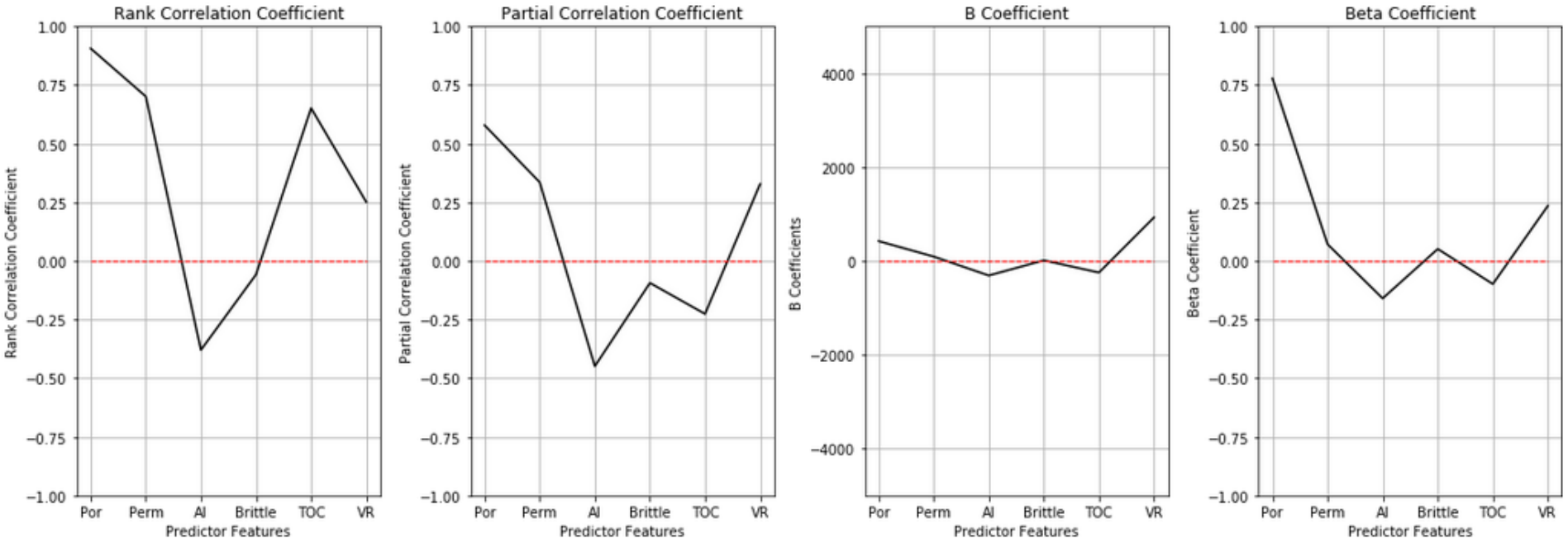
- These are the linear regression coefficients with standardization of the variables, X_i^s and Y^{s*} (variance = 1)
- Not sensitive to variance of the features
- We are capturing interactions between variables.

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 $1, \ldots, m$ 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 Subsuface 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

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 Selection	Apply wide array methods to explore the importance of each predictor feature with respect to the response feature. *Partial correlation reveals that rock type provides little additional information to acoustic impedance.*

Multivariate Modeling:
Feature Selection

Lecture outline . . .

- **Curse of Dimensionality**
- **Overfit / Model Complexity**
- **Feature Ranking**

Introduction

Prerequisites

Probability

Multivariate Analysis

Spatial Estimation

Feature Selection

Multivariate Modeling

Conclusions