PGE 383
Feature Imputation

Lecture outline . . .

- **Feature Imputation**

Michael Pyrcz, The University of Texas at Austin

Motivation for Feature Imputation

- Most spatial, subsurface datasets are not complete, missing values from the database.

- Data analytics and machine learning require complete data

- Dealing with missing data is an essential part of feature / data engineering, prerequisite for data analytics and machine learning

PGE 383
Feature Imputation

Lecture outline . . .

- **Feature Imputation**

Michael Pyrcz, The University of Texas at Austin

Missing Data Bias

- Missing at random (MAR) is not common and is not evaluated
  - Global random omission may not result in data bias and bias in the resulting models
  - This is typically not the case as missing data often has a confounding feature, e.g. cost, rock rheology, project goals / prioritization, sampling to reduce uncertainty and maximize profitability instead of statistical representativity

- Missing data consequences
  - More than reducing the amount of training and testing data, missing data, if not completely at random will result in:
  - Biased sample statistics resulting in biased model training and testing
  - Biased models with biased predictions with potentially no indication of the bias!

Missing Data on Calculation

- Samples with Missing Features Cannot be Applied in Many Data Analytics and Machine Learning Methods

- Inferential Machine Learning: PCA, MDS, Cluster Analysis require all the features, $x_{1,i}, \ldots, x_{m,i}$ for each of the data samples $i = 1, \ldots, n$.
  - We cannot calculate distance / dissimilarity, projects etc. without placing each sample in the m dimensional space

- Predictive Machine Learning: require all features to train and test the model.

$$\hat{Y} = \hat{f}(X_1, \ldots, X_m)$$

Likewise Deletion

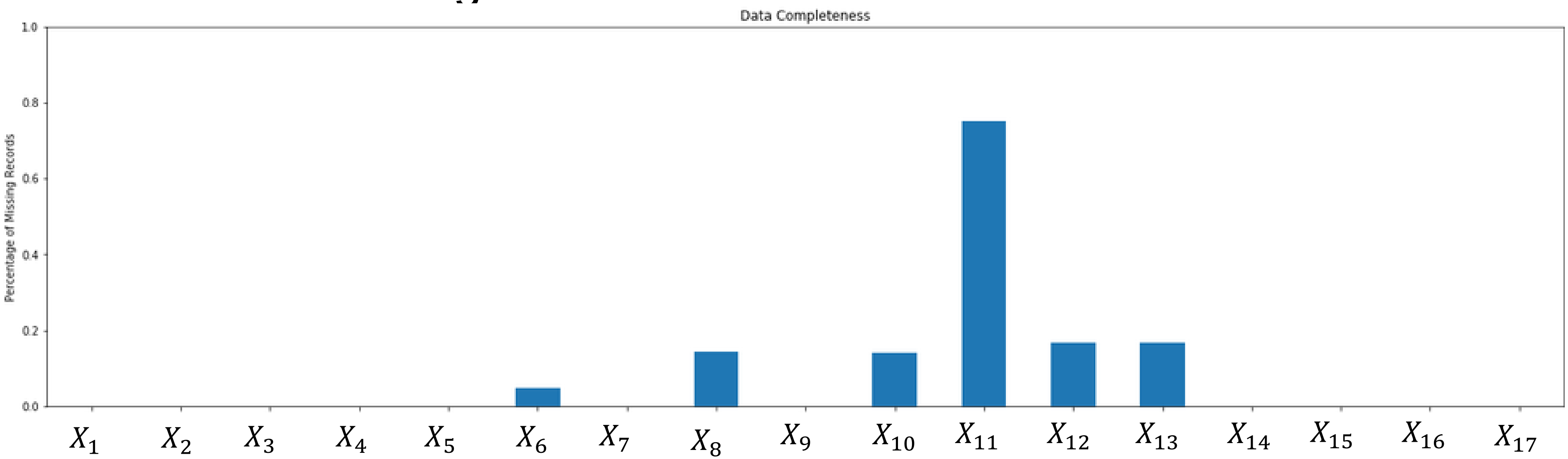
Most Common / Default Approach in Data Analytics and Machine Learning

- Removal of any sample with any missing feature

- Missing at Random (MAR)
  - Should not result in biased (or increased bias)
  - Caution: MAR is rare
  - Will result in a decrease in the effective data size and increase in model uncertainty

Likewise Deletion

Most Common / Default Approach in Data Analytics and Machine Learning



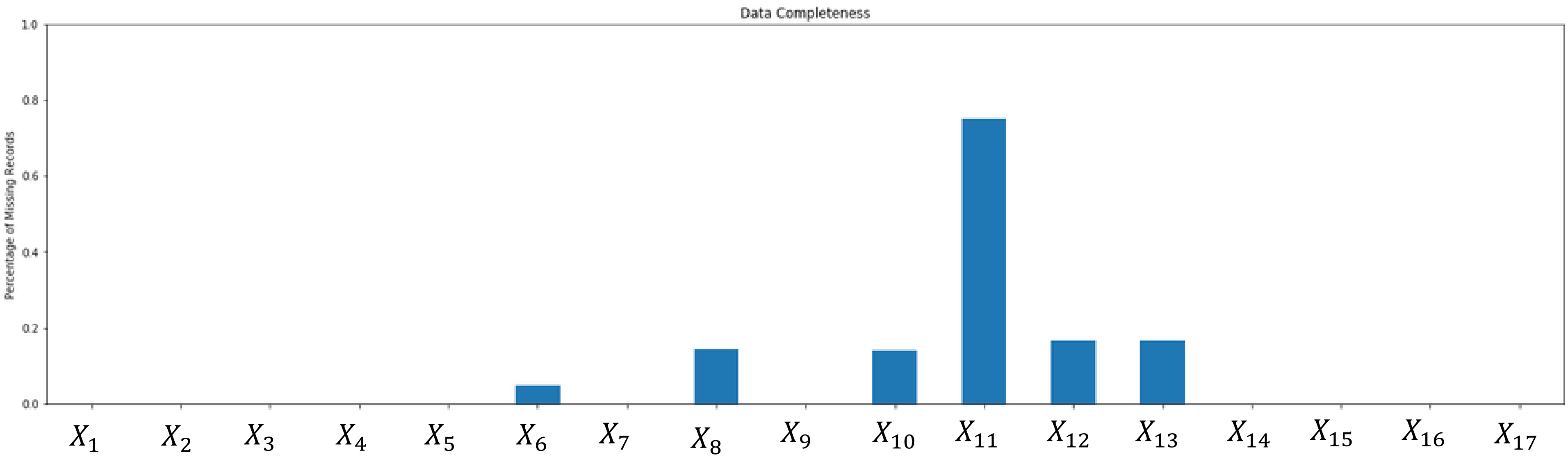
Proportion of missing records for each feature.

- Data completeness, coverage for each feature
  - Missing records in X_{10} may not all be in X_{11} etc.
  - May result in loss of much more than the largest proportion of missing
- If missing not at random (MNAR), sample bias is increased
  - Missing data diagnosis – best method fill in missing data, practical method is to evaluate the conditional statistics of missing samples over other features.

Feature Selection

Removal of features low data completeness

- Reduces missing data severity, treat data completeness as feature reliability for feature selection



Proportion of missing records for each feature.

- Removing the features with low coverage
  – Removal of X_{11} and likewise deletion fortunately resulted in a 18% reduction in samples, fortunately missing X_6, X_8, X_{10}, X_{12} and X_{13} were coincidental in this case

Traditional Alternatives to Likewise Deletion

Substitute the Global Mean

- Optimum estimate (minimizes the L2 loss function) given no other information
- Do not do this:
  - Cause conditional bias in the model in the presence of other features, systematic shift in the expectation of the substituted predictor feature over combinatorials of the other features.
  - Reduce variance of the substituted predictor feature limiting the training and testing data coverage

What is a Feature Imputation?

Estimating missing values in the data set / DataFrame

2 Primary Goals

- Maximize model accuracy
- Avoid model bias
- Provide fair measure of model uncertainty

Hot and Cold Deck Methods

Hot Deck Imputation

- Random selection from a similar record in the current dataset
- One implementation is last observation carried forward (LOCF). After sorting the dataset over features of interest (ordering to maximize similarity of adjacent records)

Cold Deck Imputation

- Like hot deck, but from another, analog dataset

Issues:

- Likely introduce bias, disrupt correlations

Mean Value Methods

Mean Value Imputation

- Replace the missing value with the global mean of the feature

$$x_i = E\{X_i\}$$

- Designed to avoid global bias in the specific feature

Conditional Mean Value Imputation

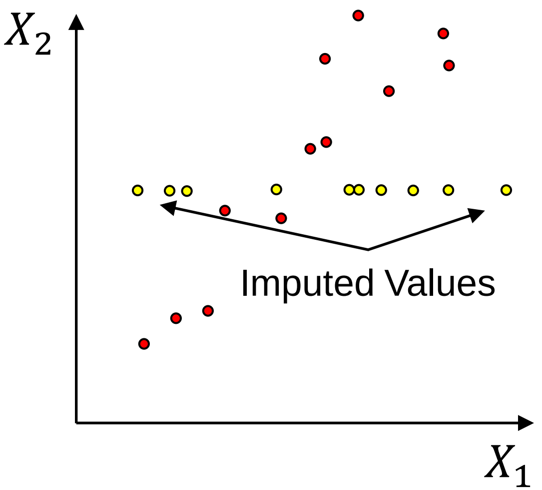
- Replace the missing value with the conditional mean of the feature

$$x_i = E\{X_i | X_{j=1,\ldots,m, j \neq i}\}$$

- Designed to avoid global and conditional bias

Issues:

- This method will attenuate correlations
- Inflating uncertainty models



Scatter plot with imputed values by mean.

Estimation / Regression Methods

Regression Imputation

- Replace the missing value with a model-based estimate of the feature

$$x_i = \hat{f}(X_{j=1,\ldots,m,j\neq i})$$

Issues:

- The imputed values are represented as hard data and fail to represent the uncertainty associated with their estimation
- This method will underestimate the uncertainty models

Alternatives:

Geo-imputation / Geographical Imputation: by spatial analog, similar locations

General Interpolation: a wide variate of interpolation methods

Censoring / Indicator Coding: include a bound / constraint on the missing value, for subsequent methods that integrate soft data

Multiple Imputation

Multiple Imputation

- Replace the missing value with a suite of realizations a multiple model-based estimate (and even scenarios) of the feature

$$x_i^{\ell} = \hat{f}^{\ell}(X_{j=1,\dots,m,j\neq i})$$

- Subsequent workflows must now integrate data realizations to integrate uncertainty

Multiple Imputation by Chained Equations (MICE) Approach:

1. Substitute random values from $F_{X_{i=1,\dots,m}}(X_{i=1,\dots,m})$ for missing values
2. Sequentially predict missing values for a feature with others
3. Iterative until convergence criteria, usually multivariate statistics
4. Repeat for multiple realizations of the dataset

Alternatives:

- Bootstrap, Geostatistics / Spatial Bootstrap

Multivariate and Spatial Imputation

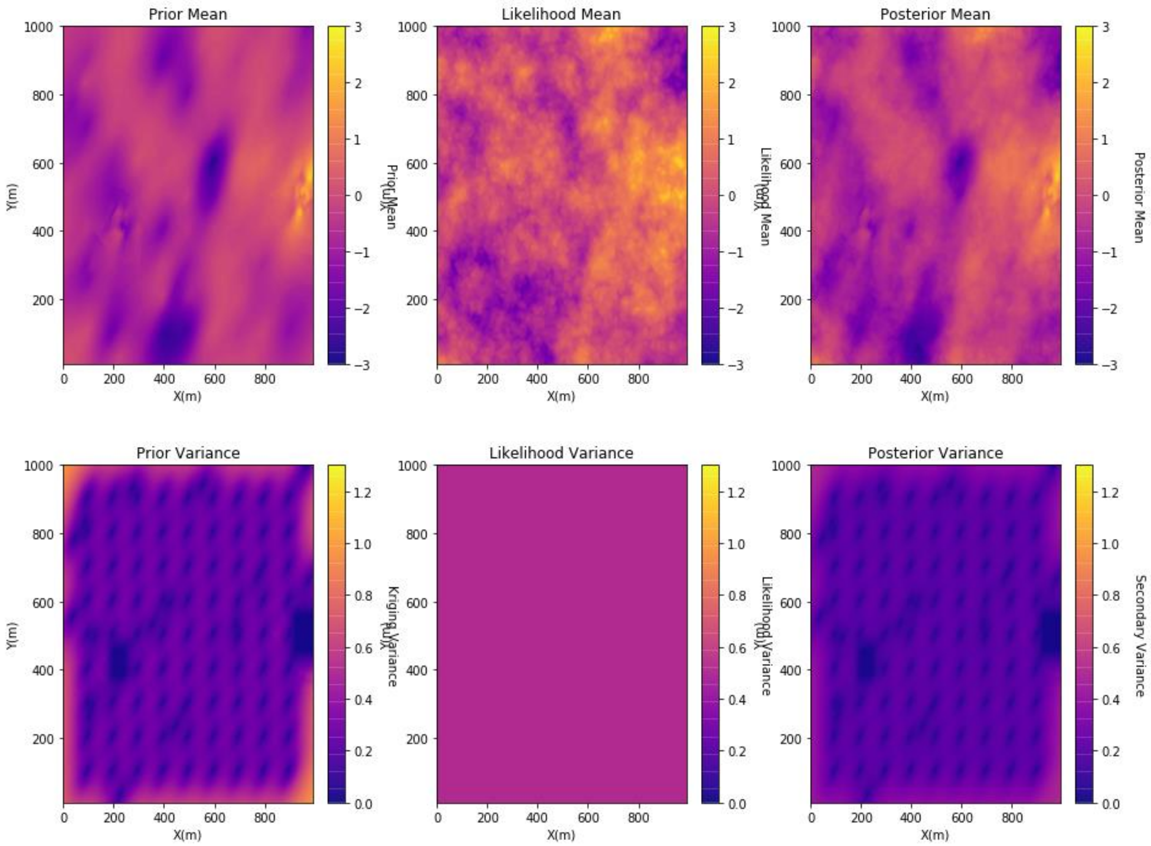
Super Secondary Approach (Deutsch and Zanon, 2004)

1. Transform the selected property to Gaussian
2. **Spatial Primary Information**: Calculate prior through kriging estimate and variance and the Gaussian assumption
3. **Multivariate Secondary Information**: Calculate the likelihood through multivariate relationship with other collocated features
4. **Bayesian Updating to Combine Spatial and Multivariate**: Update to calculate the Gaussian distributed posterior
5. Back transform the property to Gaussian
6. Visualize diagnostics on the impact of the spatial and multivariate on informing the local estimate.

Super Secondary Demonstration

Super Secondary Approach Demonstration – 2D Map

- Prior from well data primary feature, likelihood from multivariate mapped features and posterior.



Example of multivariate and spatial estimation of uncertainty distributions.

PGE 383
Feature Imputation

Lecture outline . . .

- **Feature Imputation**

Michael Pyrcz, The University of Texas at Austin