Topic:

Missing Value Imputation in a Data Matrix Using the Regularized Singular Value Decomposition

NLA 2023

Team:

1) Daniyal Asif
2) Hasaan Maqsood
3) Anwar Shamim

19 December 2023

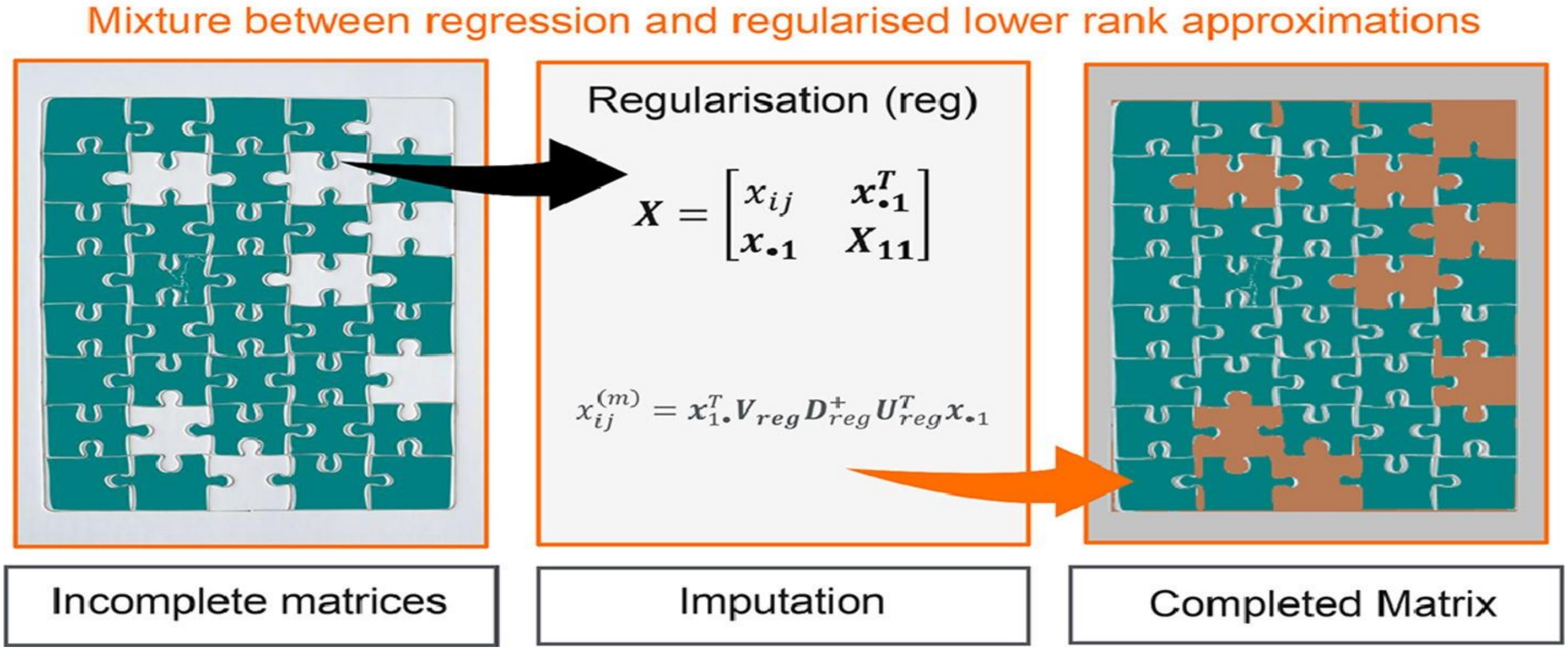
Problem Statement:

- **Issue:** Prevalence of incomplete datasets in statistical analysis.
- **Challenges:** Limitations of traditional imputation techniques.
- **Consequences:** Inadequate imputation affects data integrity.
- **Demand:** Need for an advanced, reliable imputation method.

Skoltech

Proposed Solution:

- **Solution:** GabrielEigen Imputation System
- **Method:** Combines regression with lower rank approximations
- **Innovation:** Regularised Singular Value Decomposition
- **Benefits:** Improves imputation quality, reduces overfitting
- **Applicability:** Suitable for various multivariate data matrices



Singular Value Decomposition

Singular Value Decomposition is a mathematical method where any matrix $A \in \mathbb{R}^{m \times n}$ can be decomposed into three matrices:

$$A = U\Sigma V^*$$

where:

- U is an $n \times K$ unitary matrix.
- V is an $m \times K$ unitary matrix, where $K = \min(m, n)$.
- Σ is a diagonal matrix with non-negative elements $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K$ on the diagonal.
- If A has rank r, then $\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_K = 0$.

GabrielEigen Method

García-Peña et al. introduced GabrielEigen, an imputation method that combines regression with lower rank approximations for matrix structured datasets, without relying on distributional or structural assumptions. It leverages Gabriel's cross-validation approach and SVD eigenvectors and eigenvalues to derive lower rank approximations.

Algorithm

Given a data matrix $X \in \mathbb{R}^{n \times p}$ with missing elements x_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$, the following algorithm is used for imputation:

Step 1: Fill each missing entry with the mean of its respective column:

$$\hat{x}_{ij} = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \text{ for missing } x_{ij}$$

Step 2: Standardize the columns of the completed matrix:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

where $\bar{x}_j$ is the mean and s_j is the standard deviation of the j-th column.

Step 3: For each originally missing entry x_{ij}, replace with:

$$\hat{x}_{ij} = \mathbf{U}_{(i)}\mathbf{D}^{+}\mathbf{V}_{(j)}^{T}$$

where $\mathbf{D}^{+}$ is the generalized inverse of $\mathbf{D}$, and $\mathbf{U}_{(i)}$, $\mathbf{V}_{(j)}$, and $\mathbf{D}$ are obtained from the SVD of X_{11}:

$$X_{11} = \sum_{k=1}^{m} \sigma_k \mathbf{u}_k \mathbf{v}_k^T$$

with $m \leq \min\{n-1, p-1\}$.

Step 4: Choose m to be the smallest value satisfying:

$$\frac{\sum_{k=1}^{m} \sigma_k^2}{\sum_{k=1}^{\min\{n-1,p-1\}} \sigma_k^2} \geq 0.75$$

Step 5: Convert the imputed values $\hat{x}_{ij}$ back to their original scale:

$$x_{ij}^{\text{imputed}} = \hat{x}_{ij} \cdot s_j + \bar{x}_j$$

Step 6: Repeat steps 2 to 5 until the imputations achieve stability. If $n \leq p$, transpose the matrix before conducting the iterations.

Skoltech

Regularized Version of the GabrielEigen

Regularization is employed to prevent overfitting, ensuring higher quality imputations and more reliable parameter estimation. To enhance the original GabrielEigen imputation system, a regularized Singular Value Decomposition is used, effectively creating a regularized version of the method.

Algorithm

Let X_{11} denote the matrix $(n-1 \times p-1)$ and m the desired rank. The algorithm is as follows:

Step 1: Initially, a V $(p-1 \times m)$ matrix is obtained with random entries from a uniform distribution $(0,1)$.

Step 2: The matrix U $(n-1 \times m)$ is calculated as:

$$U = X_{11} V (V^T V + \lambda I_m)^+$$

where I_m represents the identity matrix $(m \times m)$ and $(+)$ represents a generalized inverse and λ is the regularization parameter.

Step 3: The matrix V is updated through:

$$V = X_{11}^T U (U^T U + \lambda I_m)^+$$

Step 4: The value of the regularized objective function is calculated by:

$$J = \|X_{11} - UV^T\|_F^2 + \lambda(\|U\|_F^2 + \|V\|_F^2)$$

.

Step 5: Steps 2, 3 and 4 are repeated iteratively until reaching convergence in the value of J.

Step 6: The standard SVD is calculated over UV^T to obtain the corresponding regularized eigenvalues and eigenvectors.

Step 7: The imputation equation of the regularized GabrielEigen becomes:

$$x_{ij}^{(m)} = x_i^T U_{\text{reg}} D_{\text{reg}}^{+} V_{\text{reg}}^T x_j$$

where $U_{\text{reg}} D_{\text{reg}} V_{\text{reg}}^T$ represents the regSVD of X_{11}.

Validation

- Utilized Kaggle's datasets for Breast and Prostate cancer.
- Introduced missing values randomly at 5%, 15%, and 30% ratios.
- Varied lambda values for comprehensive analysis.
- Evaluation Metrics including mean absolute error and correlation coefficient used..

Skoltech

Mean Absolute Error

The Mean Absolute Error (MAE) is a metric used to quantify the average magnitude of errors between predicted and actual values. The formula for calculating MAE is as follows:

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \hat{x}_i|$$

where x_i and $\hat{x}_i$ represents the actual and predicted value of i^{th} observation respectively.

MAE is a straightforward and interpretable metric. A smaller MAE indicates better agreement between the predicted and actual values, with each absolute difference contributing equally to the overall error.

Correlation Coefficient

The correlation coefficient between two variables, denoted as r, is a statistical measure of the strength and direction of their linear relationship. The formula for calculating the correlation coefficient between two variables X and Y is given by:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

A correlation close to 1 indicates a strong positive linear relationship, while a correlation close to -1 indicates a strong negative linear relationship. A correlation close to 0 suggests a weak or no linear relationship.

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. λ=0 represents the original GabrielEigen.
In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. λ=0 represents the original GabrielEigen.

Summary of study on the breast cancer dataset

	Missing **Ratio**	λ = 0	λ = 0.1	λ = 0.2	λ = 0.3	λ = 0.4	λ = 0.5
Mean absolute error	5%	0.3748	0.3376	0.3255	0.52993	0.4000	0.2791
	15%	1.0597	1.3509	0.9963	0.9978	0.9927	1.0414
	30%	2.9354	2.6785	3.0909	3.2415	2.8780	3.4083
Correlation	5%	0.9994	0.9996	0.9994	0.9981	0.9991	0.9997
	15%	0.9972	0.9956	0.9985	0.9987	0.9986	0.9985
	30%	0.9922	0.9941	0.9921	0.9902	0.9930	0.9885

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. $\lambda = 0$ represents the original GabrielEigen.

Summary of study on the prostate cancer

	Missing **Ratio**	λ = 0	λ = 0.1	λ = 0.2	λ = 0.3	λ = 0.4	λ = 0.5
Mean absolute error	5%	0.92694	0.7487	1.9829	1.5997	0.8035	1.2669
	15%	3.6482	3.9415	1.6656	2.1878	2.9213	3.1672
	30%	5.6748	6.1363	5.1481	2.0581	6.4935	6.5073
Correlation	5%	0.9989	0.9984	0.9963	0.9972	0.9991	0.9981
	15%	0.9929	0.9924	0.9986	0.9965	0.9951	0.9963
	30%	0.9910	0.9912	0.9899	0.9806	0.9907	0.9905

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. $\lambda = 0$ represents the original GabrielEigen.

Conclusion

- A generalization of the GabrielEigen imputation method has been proposed using regularized SVD.
- The regularized version is flexible and can be applied to any data matrix, making it suitable for non-parametric imputation for multivariate data.
- The method proved to be quite adaptable across different types of interaction, matrix dimensions, and percentages of missing data.
- The method has potential for use with methodologies for obtaining robust and multiple imputations.

Skoltech

Future Research

- Exploring different mechanisms of data absence and their impact on the proposed method.
- Investigating the effects of various probability distributions on the efficiency and effectiveness of the method.
- Further research on the optimal choice of the regularization parameter, particularly in different data scenarios.
- Integrating the proposed methodology with existing robust and multiple imputation methods to enhance its applicability and effectiveness.

Skoltech

Thank you for your attention

Daniyal Asif; conceptualization, software, validations, project administration, supervision, visualization, writing—original draft.
Hassan Maqsood; conceptualization, software, visualization, and writing—original draft.
Anwar Shamim; conceptualization, software and writing—original draft.

Questions?