



Contents

1. Introduction	1
2. Methodological Concepts for Missing Data	9
2.1. Missing Values	9
2.1.1. Missing Data Mechanism	10
2.2. An Overview of Traditional Missing Data Techniques	12
2.2.1. Deletion Methods	12
2.2.2. Substitution Methods	12
2.2.3. Imputation	12
2.3. Nearest Neighbors Methods	14
2.3.1. Traditional k Nearest Neighbors Imputation	16
2.4. Modification of the Traditional kNN Imputation	17
2.5. Nearest Neighbors for High-Dimensional Data	18
2.5.1. Practical Issues in High-Dimensional Data ($n \ll p$)	19
2.6. Extensions to Binary and Multi-Categorical Data	20
2.6.1. Selection of Attributes by Weighted Distances	21
2.6.2. Using Nearest Neighbors to Impute Missing Values	22
2.6.3. A Pearson Correlation Strategy	23
2.7. Extension to Mixed Type Data	24
2.7.1. Available Distances for Mixed Type Data	24
2.7.2. Weighted Distance for Mixed Type Data	25
2.7.3. Weighted Imputation by Nearest Neighbors	26
2.8. Bootstrap Inference	27
2.9. Missing Data and Classification	28
2.10. Multiple Imputation in High-Dimensional Data	29
2.10.1. MI using Nearest Neighbors	30
3. Improved Methods for the Imputation of Missing Data by Nearest Neighbor Methods	33
3.1. Introduction	34
3.2. Weighted Neighbors	35
3.2.1. Distances and Computation of Nearest Neighbors	35
3.2.2. Imputation Procedure	36
3.2.3. Choice of Tuning Parameters by Cross-Validation	37
3.2.4. Performance Measures	38
3.2.5. Evaluation of Weighted Neighbors	39
3.3. Weighted Neighbors Including the Selection of Predictors	42
3.3.1. Selection of Dimensions	42
3.3.2. Evaluation of the Method with Selected Weighted Neighbors	44



3.4. Case Studies	54
3.4.1. Gene Expression Data	54
3.4.2. Non-gene Expression Data	55
3.5. Concluding Remarks	56
4. Missing Value Imputation for Gene Expression Data by Tailored Nearest Neighbors	59
4.1. Introduction	60
4.2. Materials and Methods	61
4.2.1. Nearest Neighbors Approaches to Imputation	61
4.2.2. Nearest Neighbor Based on Selected Genes	62
4.2.3. Choice of Tuning Parameters	64
4.2.4. Overview of Competing Methods	64
4.3. Results and Discussion	66
4.3.1. Data	66
4.3.2. Simulation and Evaluation	67
4.3.3. Real Data Sets	69
4.4. Concluding Remarks	71
5. Nearest Neighbor Imputation for Categorical Data by Weighting of Attributes	73
5.1. Introduction	73
5.2. Methods	75
5.2.1. Distances for Categorical Variables	75
5.2.2. Selection of Attributes by Weighted Distances	77
5.2.3. Measuring Association Among Attributes	78
5.3. Using Nearest Neighbors to Impute Missing Values	80
5.3.1. A Pearson Correlation Strategy	81
5.3.2. Cross Validation	81
5.4. Evaluation of Performance	82
5.4.1. Existing Methods	82
5.5. Simulation Studies	83
5.5.1. Binary Variables	84
5.5.2. Multi-Categorical Variables	84
5.5.3. Mixed (Binary and Multi-Categorical) Variables	87
5.6. Applications	89
5.7. Concluding Remarks	90
6. Imputation Methods for High-Dimensional Mixed-Type Datasets by Nearest Neighbors	93
6.1. Introduction	93
6.2. Distances for Mixed-Type Data	95
6.2.1. Available Methods	95
6.2.2. Weighted Distance	96
6.2.3. Weighted Distance With Selection of Variables	97
6.2.4. Weighted Imputation by Nearest Neighbors	98
6.2.5. Choice of Tuning Parameters by Cross-Validation	99
6.2.6. Measuring Association Among Mixed Variables	100
6.2.7. A Pearson Correlation Strategy	101



6.3. Existing Approaches for Comparison	102
6.3.1. Performance Measures	103
6.4. Simulation Studies	103
6.5. Real Data Applications	106
6.6. Concluding Remarks	113
7. Bootstrap Inference for Weighted Nearest Neighbors Imputation	115
7.1. Introduction	115
7.2. Nearest Neighbors Imputation of Missing Values	117
7.3. Bootstrap Sampling and Missing Data	118
7.4. Proposed Bootstrap Imputation Procedure	119
7.4.1. Bootstrap confidence interval	120
7.4.2. Bootstrap Percentile Interval	120
7.4.3. Coverage Probability	120
7.5. Simulation Studies	121
7.5.1. Simulation Study 1	121
7.5.2. Simulation Study 2	123
7.6. Concluding Remarks	126
8. Missing Values in Classification: Improved Imputation Methods for High-Dimensional Settings	129
8.1. Introduction	129
8.2. Missing Data Treatment for Classification	131
8.3. Imputation Methods for Missing Values	132
8.3.1. Mean imputation	133
8.3.2. NN Imputation	133
8.3.3. Random Forests	133
8.3.4. wNN Imputation with Selected Features (wNNSel)	134
8.4. Performance Measures	135
8.5. Experimental Evaluations	136
8.5.1. Datasets	136
8.5.2. Study Design	136
8.5.3. Results	137
8.6. Concluding Remarks	142
9. Multiple Imputation Using Nearest Neighbor Methods	143
9.1. Introduction	143
9.2. Multiple Imputation	146
9.3. Multiple Imputation Based on Nearest Neighbors	148
9.3.1. Sequential Multiple Imputation	150
9.3.2. Multiple Imputation using Bootstrap	150
9.3.3. Existing Approaches for Multiple Imputation	151
9.3.4. Evaluation of Performance	152
9.4. Simulation Studies	152
9.4.1. Simulation Study ($n > p$)	153
9.4.2. Simulation Study for High-Dimensional Data ($n < p$)	158



9.5. Real Data	161
9.5.1. Results for Imputation Error	162
9.5.2. Simulation Using Real Data	163
9.6. Concluding Remarks	165
10. Conclusion and Outlook	167
Appendices	169
A. Additional Results for Chapter 5	171
B. Appendix for Chapter 7	173
C. Appendix for Chapter 8	179
D. Appendix for Chapter 9	181
References	187