

Bibliography

- Aeberhard, Stefan and M. Forina (1991), “Wine.” UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5PC7J>.
- Cattell, R.B. (1966), “The scree test for the number of factors.” *Multivariate Behavioral Research*, 1, 629–637.
- Cook, R. D. and S. Weisberg (1982), *Residuals and Influence in Regression*. Chapman and Hall Ltd, 11 New Fetter Lane, London EC4P 4EE.
- Dunteman, G. H. (1989), *Principal Components Analysis*, volume 69 of *Quantitative Applications in the Social Sciences*. SAGE UNIVERSITY.
- Ferré, L. (1995), “Selection of components in principal component analysis: A comparison of methods.” *Computational Statistics & Data Analysis*, 19, 669–682.
- Gwet, K. L. (2020), *Beginner’s Guide to Principal Components: Applications with Microsoft Excel*. AgreeStat Analytics, MD: USA.
- Gwet, K. L. (2022), *Using R for Excel Analysts: Tools for Increasing Productivity*. AgreeStat Analytics, MD: USA.
- Hotelling, H. (1933), “Analysis of a complex of statistical variables into principal components.” *Journal of Educational Psychology*, 24, 417–441, 498–520.
- Jackson, D. (1993), “Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches.” *Ecology*, 74, 2204–2214.
- Jolliffe, I.T. (1986), *Principal Component Analysis*. Springer-Verlag, New York.
- Jolliffe, I.T. (2002), *Principal Component Analysis*. Springer, New York.
- Kaiser, H.F. (1960), “The application of electronic computers to factor analysis.” *Educational and Psychological Measurement*, 20, 141–151.

- McCabe, G. P. (1984), “Principal variables.” *Technometrics*, 26, 137–144.
- Mevik, B.-H. and H.R. Cederkvist (2004), “Mean squared error of prediction (mse) estimates for principal component regression (pcr) and partial least squares regression (pls).” *Journal of Chemometrics*, 18, 422–429.
- Neter, John, William Wasserman, and Michael H. Kutner (1985), *Applied Linear Statistical Models*, 2 edition. Richard D. Irwin, Inc., Homewood, Illinois 60430.
- Pearson, K. (1901), “On lines and planes of closest fit to systems of points in space.” *Phil. Mag.*, 6, 559–572.
- Rencher, A. C. (1992), “Interpretation of canonical discriminant functions, canonical variates, and principal components.” *The American Statistician*, 46, 217–225.
- Singh, K. (2014), *Linear Algebra: Step by Step*, 1 edition. Oxford University Press, Great Clarendon Street, Oxford, OX2 6DP.
- Wickham, H. and J. Bryan (2023), *R Packages: Organize, Test, Document, and Share Your Code*, 2 edition. O’Reilly Media.
- Wickham, H., M. Çetinkaya Rundel, and G. Grolemund (2023), *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, 2 edition. O’Reilly Media.
- Zwick, W. and W. Velicer (1986), “Comparison of five rules for determining the number of components to retain.” *Psychological Bulletin*, 99, 432–446.

Author Index

Cattell, R.B., 119
Cederkvist, H.R., 179, 181
Cook, R.D., 165

Dunteman, G.H., v

Ferré, L., 118

Gwet, KL, 18, 88

Hotelling, H., v

Jackson, D., 118
Jolliffe, IT, vi, 118, 119, 121, 145,
151

Kaiser, H.F., 118

McCabe, G.P., 145
Mevik, B.-H., 179, 181

Neter, J. and Wasserman, W. and
Kutner, M.H., 47

Pearson, K., v

Singh, K., 18

Velicer, W., 118

Weisberg, S., 165

Zwick W., 118

Subject Index

- `str()`, 59
- Anomaly Detection, 129
- Biplot, 103, 112
- Breast Cancer Wisconsin (BCW) Data, 84
- Cartesian, 3
- Centering, 73
- Cluster, 58
- Comprehensive R Archive Network, 25
- Confidence Ellipse, 102
- Contrast, 153
- CRAN, 25
- Cross-Validation, 160
- Cross-validation, 179
- Cumulative Variance Plot, 90
- Data
 - Centering, 73
 - Standardization, 73, 74
- Data Frame, 34, 39
- Data Reconstruction, 129
- Dataset
 - Testing, 122
 - Training, 122
- Eigenvalues, 62
- Elbow Method, 119
- Elimination Method, 146
- Hat Matrix, 165
- Hat Value, 165
- IDE, 20
- Inclusion Method, 145
- Integrated Development Environment, 20
- Internal Studentized Residuals, 165
- List, 34
- Loading, 11, 66, 145
- Loading Plot, 107, 108, 110
- Loadings Plot, 108
- Loadings Plots, 103
- Matrix, 34, 44
 - Orthogonal, 138
- Mean Squared Error of Prediction, 178
- MSE, 178
- Multicollinearity, 159
- Outliers, 151
- Overfitting, 160
- Partial Least Squares Regression, 161
- PC Coefficients, 6
- PC Space, 138
- Posit
 - Cloud, 20

- predict(), 124
- Principal Component
 - Coefficients, 6, 60
 - Loading, 66
 - Scores, 6, 8, 61
- Principal Component Regression, 159
- Principal Component Scores, 4
- Principal components, 3
- Q-Q line, 163
- Q-Q Plot, 162
- Quality of Representation, 110
- Reconstruction Errors, 129
- RMSEP, 178
- Root Mean Squared Error of Prediction, 178
- Root Mean Squared Errors for Prediction, 182
- Rotation, 60
- rstandard(), 165
- RStudio, 20
 - Cloud, 20
 - Desktop, 20
- Score Plot, 98
- Scores, 4, 6
- Scree Plot, 90
- Screeplot, 89, 90
- Singular Value Decomposition, 57
- Standardization, 74
- Standardized Residuals, 165
- Statistical Noise, 130
- SVD, 57
- Synthetic Variable, 18, 61
- Testing dataset, 122
- Training dataset, 122
- Variable Contribution Plot, 110
- Variable Correlation Circle, 108
- Vector, 32–34
 - Atomic Vector, 34

Introductory Principal Component Analysis Using R

A Practical Guide with RStudio

In this book, you will learn how to perform Principal Component Analysis (PCA) using the R software. R is used through an Integrated Development Environment (IDE), and RStudio is the choice used in this book.

Although having some knowledge of R will help, this book assumes no specific prerequisite, except for good computer and analytical skills. PCA is a collection of advanced statistical techniques. Therefore, some familiarity with statistical thinking is essential. However, this book is a non-mathematical presentation of Principal Component Analysis.

PCA is one of the essential skills that you must have to become a productive analyst in the new world of data science.

www.agreestat.com/books/pca4r/

Author

Kilem L. Gwet, Ph.D.

Statistical consultant, mathematical statistician, researcher, and instructor. Over 20 years of experience in various industries, and several publications in peer-reviewed journals.

AgreeStat Analytics

Maryland – USA