

Assignment 2: Analyze Challenges in Multivariate Data Analysis

A. Sepúlveda-Jiménez

Data Science Dept., SoTE, CoBET, National University

TIM-8515: Multivariate Analysis

Course Instructor: Y. Karahan, PhD

November 14, 2025

Contents

Introduction	3
Multivariate Data: Definition and Complexity	4
Why Multivariate Data Are More Complex	5
Major Challenges in Multivariate Data Analysis	6
Curse of dimensionality and high-dimensional geometry	6
Description and impact	6
Mitigation strategies	7
Multicollinearity and complex dependence structures	7
Description and impact	7
Mitigation strategies	8
Missing data and outliers in multivariate contexts	9
Missing-data mechanisms	9
Outliers and Mahalanobis distance	9
Mitigation strategies	10
Visualization and interpretability	11
Description and impact	11
Mitigation strategies	11
Strategies for Mitigating Multivariate Challenges	12
Discussion and Conclusion	13
Appendix: Figures	18

List of Figures

- 1 Two-dimensional illustration of the unit ball $\mathbb{B}_2(1)$ inside the hypercube $[-1, 1]^2$. In higher dimensions, the volume of the ball becomes a vanishing fraction of the cube’s volume, one aspect of the curse of dimensionality (Donoho, 2000; Vershynin, 2018). 19
- 2 Graphical representation of a generic missing-data setup. Solid arrows indicate data generation; dashed arrows indicate potential dependencies of the missingness indicator R on observed and unobserved data. MCAR, MAR, and MNAR mechanisms correspond to restrictions on the dashed arrows (Little & Rubin, 2019; Rubin, 1976). 20
- 3 Covariance ellipse of a bivariate normal distribution with mean $\boldsymbol{\mu}$ and Mahalanobis distance $d_M(\mathbf{x}, \boldsymbol{\mu})$ from $\boldsymbol{\mu}$ to an observation \mathbf{x} . Points far outside the ellipse have large Mahalanobis distances and are potential multivariate outliers (Mardia et al., 1979; Rousseeuw & Leroy, 1987). 21

Introduction

Multivariate data—where each observational unit is described by three or more variables—are now ubiquitous in data science, business analytics, and applied research. Compared with univariate or bivariate data, multivariate datasets are more informative but also substantially harder to understand and model. This paper defines multivariate data, explains why multivariate settings are intrinsically more complex, and examines several major challenges: the curse of dimensionality and sparsity, multicollinearity, missing data and outliers, and the difficulty of visualizing and interpreting high-dimensional structures. For each challenge, the discussion highlights how the issue distorts statistical inference and decision-making, and then reviews strategies to mitigate the problem, including dimensionality reduction, feature selection, regularization, robust methods, and advanced visualization. The goal is not to present multivariate methods as a solved problem, but to underscore where routine practice often goes wrong and how a more

deliberate analytic strategy can reduce risk in empirical conclusions.

Modern data science rarely deals with a single variable at a time. Most real datasets describe each observational unit—a person, transaction, patient, customer, a general entity type, or sensor reading—by a vector of measurements. Multivariate statistical methods have therefore become central to empirical work in the social sciences, biomedical research, marketing, and machine learning (Hastie et al., 2009a; Johnson & Wichern, 2007; McQuitty, 2018). At the same time, multivariate analysis is notorious for being fragile: small mistakes in data preparation, model choice, or interpretation can propagate through a complex dependence structure and lead to confident but wrong decisions (Everitt, 1975).

The purpose of this paper is to critically examine core challenges in multivariate data analysis and outline strategies to address them. The focus is conceptual rather than algorithmic. First, multivariate data are defined and contrasted with univariate and bivariate settings. Second, several prominent difficulties are described: (a) the curse of dimensionality and sparsity, (b) multicollinearity, (c) missing data and outliers, and (d) the problems of visualization and interpretability. Third, for each challenge, the paper discusses implications for analysis and decision-making and reviews practical mitigation strategies, drawing on both classical statistics and more recent developments in statistical learning (Dormann et al., 2013; Guyon & Elisseeff, 2003; Hastie et al., 2009a; Schafer & Graham, 2002).

Multivariate Data: Definition and Complexity

Definition 1. A multivariate observation is a vector

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^\top,$$

where $p \geq 3$ and $i = 1, \dots, n$ index observational units. The collection of such vectors is typically represented as a data matrix $\mathbf{X} = \{\mathbf{x}_i\}_1^p \in \mathbb{R}^{n \times p}$, possibly with a mixture of continuous, ordinal, and categorical variables. Multivariate methods are those that explicitly empirically model, learn, or describe the (theoretical) joint distribution

$\{x_i\}_1^p \mid \boldsymbol{\theta} \sim D_{\boldsymbol{\theta}}$ with parameterization vector $\boldsymbol{\theta}$, of the p variables or their dependence structure (Hazra & Gogtay, 2017; Johnson & Wichern, 2007).

Let $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}_i] \in \mathbb{R}^p$ denote the mean vector and

$$\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X}_i) = \mathbb{E}[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^\top] \in \mathbb{R}^{p \times p}$$

denote the covariance matrix. The pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and, more generally, the full joint distribution of \mathbf{X}_i encode dependence between variables. Many multivariate techniques—such as principal component analysis (PCA), discriminant analysis, and multivariate regression—are built directly on properties of $\boldsymbol{\Sigma}$ and its eigen-decomposition (Jolliffe & Cadima, 2016; Mardia et al., 1979).

By contrast, univariate analysis concerns a single random variable X and its mean, variance, and distribution. Bivariate analysis mainly considers pairs (X, Y) and summarizes their association via covariance or a correlation coefficient. In multivariate settings, even simple tasks such as describing “typical” behavior or detecting anomalies require reasoning in the joint space of all p variables. For example, an observation can be extreme along a combination of variables while appearing unremarkable in any single marginal distribution. (Hastie et al., 2009a; McQuitty, 2018).

Why Multivariate Data Are More Complex

There are several reasons why multivariate analysis is substantially more complex than univariate or bivariate analysis.

High-dimensional geometry behaves quite differently from the familiar intuition in two or three dimensions (Donoho, 2000; Vershynin, 2018). A standard illustration is the relationship between the p -dimensional unit ball $\mathbb{B}_p(1) = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\|_2 \leq 1\}$ and the hypercube $[-1, 1]^p$. The volume of the unit ball is

$$V_p = \frac{\pi^{p/2}}{\Gamma(1 + \frac{p}{2})},$$

where $\Gamma(\cdot)$ is the gamma function. The volume of the hypercube is 2^p . As $p \rightarrow \infty$, the ratio $V_p/2^p \rightarrow 0$, meaning that the unit ball occupies an exponentially vanishing fraction of the

hypercube. In other words, most of the volume of high-dimensional space lies in the corners of the hypercube, not near the origin. See Figure 1.

A related phenomenon is concentration of measure. For example, if $\mathbf{Z} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$, then $\|\mathbf{Z}\|_2^2 \sim \chi_p^2$, so $\mathbb{E}\|\mathbf{Z}\|_2^2 = p$ and $\text{Var}(\|\mathbf{Z}\|_2^2) = 2p$. Standard arguments show that $\|\mathbf{Z}\|_2/\sqrt{p} \rightarrow 1$ in probability as $p \rightarrow \infty$: most of the mass is concentrated in a thin shell of radius approximately \sqrt{p} (Vershynin, 2018). Distances between random points thus become very similar in high dimensions, which undermines algorithms that rely on nearest neighbors or local geometry.

These geometric facts underlie several of the challenges discussed below.

Major Challenges in Multivariate Data Analysis

Curse of dimensionality and high-dimensional geometry

Description and impact

The *curse of dimensionality* refers to a collection of phenomena that occur when analyzing data in high-dimensional spaces, first articulated in the context of dynamic programming by Bellman (1957). As the number of variables p increases, the volume of the feature space grows exponentially, and data become sparse in that space (Donoho, 2000). Combined with norm concentration, this means that (a) most points are roughly the same distance from each other, and (b) regions of space that might be well-sampled in low dimensions become essentially empty when p is large.

From a modeling perspective, consider a linear regression with predictor vector $\mathbf{X}_i \in \mathbb{R}^p$ and response Y_i ,

$$Y_i = \beta_0 + \boldsymbol{\beta}^\top \mathbf{X}_i + \varepsilon_i.$$

The parameter vector $(\beta_0, \boldsymbol{\beta})$ has $p + 1$ unknowns. When p is large relative to n , the matrix $\mathbf{X}^\top \mathbf{X}$ is ill-conditioned or singular, so ordinary least squares becomes unstable or undefined. More flexible models, such as tree ensembles or kernel methods, can overfit: they may interpolate training data while generalizing poorly (Bühlmann & van de Geer, 2011; Hastie et al., 2009b).

In decision-making, this produces an illusion of certainty: high-dimensional models can achieve extremely low error on training data, encouraging overconfident deployment. Yet predictions may degrade sharply when applied to new populations or changing environments.

Mitigation strategies

Mitigation typically reduces the effective dimensionality or increases information per parameter:

- **Dimension reduction.** Techniques such as PCA and related factor models construct low-dimensional summaries $\mathbf{Z}_i = \mathbf{W}^\top (\mathbf{X}_i - \boldsymbol{\mu})$, where $\mathbf{W} \in \mathbb{R}^{p \times k}$ with $k \ll p$, chosen to maximize explained variance (Jolliffe & Cadima, 2016; Mardia et al., 1979). Nonlinear methods (e.g., manifold learning) attempt to approximate lower-dimensional manifolds embedded in \mathbb{R}^p (Borsboom et al., 2021).
- **Feature selection and sparsity.** Penalized methods such as the lasso and elastic net constrain $\|\boldsymbol{\beta}\|_1$ or a combination of $\|\boldsymbol{\beta}\|_1$ and $\|\boldsymbol{\beta}\|_2$, yielding sparse or shrinkage-based solutions that reduce variance in high-dimensional settings (Bühlmann & van de Geer, 2011; Hastie et al., 2009b).
- **Sample size and experimental design.** When feasible, increasing n or using experimental designs that maximize information about key contrasts can partially offset high dimensionality.

These techniques help ensure that downstream conclusions—such as which covariates predict risk or which segments of customers behave similarly—are supported by structure in the data rather than high-dimensional noise.

Multicollinearity and complex dependence structures

Description and impact

In multivariate datasets, many variables are often correlated or partially redundant, leading to *multicollinearity*. In matrix terms, the covariance matrix $\boldsymbol{\Sigma}$ has eigenvalues

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$, and the condition number

$$\kappa(\mathbf{\Sigma}) = \frac{\lambda_1}{\lambda_p}$$

quantifies the degree of ill-conditioning. Large $\kappa(\mathbf{\Sigma})$ implies that some directions in predictor space are weakly identified by the data; regressions along those directions will have highly variable coefficient estimates (Bühlmann & van de Geer, 2011; Mardia et al., 1979).

In a linear model $Y = \beta_0 + \boldsymbol{\beta}^\top \mathbf{X} + \varepsilon$, multicollinearity inflates the variance of the least-squares estimator $\hat{\boldsymbol{\beta}}$, makes individual t -tests unreliable, and can even flip coefficient signs under small perturbations of the data or model specification (Hastie et al., 2009b). Although predictions may remain reasonable, interpretation of individual predictors becomes unstable, which is problematic when decisions hinge on identifying specific levers (e.g., pricing, dosage, or risk factors).

Mitigation strategies

Several strategies can help manage multicollinearity and complex dependence:

- **Orthogonal transformations.** PCA replaces \mathbf{X} by uncorrelated components $\mathbf{Z} = \mathbf{V}^\top (\mathbf{X} - \boldsymbol{\mu})$, where columns of \mathbf{V} are eigenvectors of $\mathbf{\Sigma}$ (Jolliffe & Cadima, 2016). Regressing on a subset of principal components can reduce variance and stabilize estimates.
- **Regularization.** Ridge regression adds an ℓ_2 penalty $\lambda \|\boldsymbol{\beta}\|_2^2$ to the loss, effectively replacing $\mathbf{X}^\top \mathbf{X}$ by $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ and improving conditioning. The lasso and elastic net combine sparsity and shrinkage (Bühlmann & van de Geer, 2011; Hastie et al., 2009b).
- **Domain-informed grouping.** Aggregating strongly related indicators into composite scores or indices reduces redundancy and can align statistical structure with substantive constructs.

- **Graphical models and networks.** Representing conditional dependencies via sparse precision matrices (inverse covariance) or graphical models can reveal clusters of variables that behave as units (Borsboom et al., 2021).

These approaches treat collinearity as a property to be accommodated rather than eliminated, improving stability without discarding information.

Missing data and outliers in multivariate contexts

Missing-data mechanisms

Let Y denote the (hypothetical) complete data matrix and R a matrix of missingness indicators with entries $R_{ij} = 1$ if Y_{ij} is observed and 0 otherwise. Following Rubin (1976), missing-data mechanisms are often classified via the conditional distribution $p(R | Y)$:

- **Missing completely at random (MCAR):** $p(R | Y) = p(R)$, i.e., missingness is independent of the data.
- **Missing at random (MAR):** $p(R | Y) = p(R | Y_{\text{obs}})$, i.e., missingness may depend on observed values but not on unobserved ones.
- **Missing not at random (MNAR):** Any situation where the above conditions fail, so missingness depends on unobserved values.

In multivariate settings, different variables can have different mechanisms, and the pattern of missingness can be high-dimensional. Naive complete-case analysis discards all rows with any missing value, which can severely reduce effective sample size and induce bias when data are not MCAR (Hughes et al., 2019; Little & Rubin, 2019). See Figure 2.

Outliers and Mahalanobis distance

In multivariate settings, an observation can be an outlier with respect to the joint distribution even if its coordinates look ordinary marginally. A common diagnostic is the

Mahalanobis distance from a mean vector $\boldsymbol{\mu}$ with covariance matrix $\boldsymbol{\Sigma}$:

$$d_M(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}.$$

Under multivariate normality, $d_M(\mathbf{X}_i, \boldsymbol{\mu})^2 \sim \chi_p^2$, which allows the use of χ^2 -quantiles as cutoffs for potential outliers (Mardia et al., 1979). In practice, classical estimates $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ can be distorted by outliers themselves, motivating robust alternatives (Rousseeuw & Leroy, 1987; Zimek et al., 2012). See Figure 3.

Mitigation strategies

Best practices for missing data and outliers in multivariate settings include:

- **Principled missing-data methods.** Multiple imputation, maximum likelihood under MAR, and inverse probability weighting can yield approximately unbiased estimates under explicit assumptions (Austin et al., 2021; Hughes et al., 2019; Little & Rubin, 2019; Rubin, 1976).
- **Sensitivity analysis.** Because the true missingness mechanism is often uncertain, sensitivity analyses examine how conclusions change under plausible deviations from MAR (Little & Rubin, 2019).
- **Robust estimation and outlier detection.** Robust estimators of location and scatter, as well as high-dimensional outlier detection methods, reduce sensitivity to a small fraction of anomalous points (Rousseeuw & Leroy, 1987; Zimek et al., 2012).
- **Diagnostic visualization.** Scatterplot matrices, parallel coordinates, and distance plots of $d_M(\mathbf{x}_i, \hat{\boldsymbol{\mu}})$ versus observation index can highlight unusual observations.

These strategies directly affect the reliability of inference and decisions in fields where missing data and outliers are common.

Visualization and interpretability

Description and impact

Because humans reason visually in at most three dimensions, visualizing multivariate data requires projection or more abstract encodings. Let $\mathbf{X}_i \in \mathbb{R}^p$. A linear projection $\mathbf{a}^\top \mathbf{X}_i$ onto direction $\mathbf{a} \in \mathbb{R}^p$ can be plotted univariately; a pair of directions (\mathbf{a}, \mathbf{b}) provides a two-dimensional scatterplot. PCA chooses directions that maximize projected variance, solving

$$\mathbf{v}_1 = \arg \max_{\|\mathbf{v}\|_2=1} \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v},$$

and yields orthogonal eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_p$ (Jolliffe & Cadima, 2016; Mardia et al., 1979). Nonlinear methods such as t-SNE and related techniques instead aim to preserve neighborhood relationships by minimizing a Kullback–Leibler divergence between high- and low-dimensional similarity distributions (Borsboom et al., 2021; van der Maaten & Hinton, 2008).

Each visualization emphasizes certain aspects of the data and can hide others. A low-variance direction might be crucial for separating classes but nearly invisible in a PCA plot; a t-SNE map may exaggerate cluster separation while obscuring global distances. For decision-makers, these limitations can foster overconfidence in apparent patterns.

Mitigation strategies

Mitigating visualization and interpretability challenges involves:

- **Multiple complementary views.** Using several visualization methods with different projections and encodings reduces the risk of drawing strong conclusions from a single view (Cleveland, 1993; Jolliffe & Cadima, 2016).
- **Dimension reduction for visualization.** PCA, non-linear embeddings, and clustering-based maps can reveal structure that is difficult to see in raw coordinates, provided their objectives and limitations are clearly explained (Borsboom et al., 2021; van der Maaten & Hinton, 2008).

- **Model interpretability techniques.** Partial dependence plots, accumulated local effect plots, and Shapley-based methods approximate how features influence predictions, even for complex models (Bühlmann & van de Geer, 2011; Hastie et al., 2009b).
- **Explicit uncertainty communication.** Including confidence intervals, prediction intervals, and distributional summaries in visualizations helps audiences appreciate uncertainty.

Visualization is thus both a mathematical and communication problem: the choice of projection and explanation materially affects decisions.

Strategies for Mitigating Multivariate Challenges

Although the challenges above are conceptually distinct, successful practice in multivariate data analysis typically involves an integrated strategy. A practical workflow might include:

1. **Initial data audit.** Use numerical summaries and multivariate diagnostics to examine missingness, outliers, and basic dependence structure.
2. **Exploratory dependence analysis.** Estimate $\hat{\Sigma}$, examine its eigenvalues $\hat{\lambda}_1, \dots, \hat{\lambda}_p$, and compute $\kappa(\hat{\Sigma})$ to assess collinearity (Jolliffe & Cadima, 2016; Mardia et al., 1979).
3. **Dimension reduction and feature selection.** Use PCA, domain-informed variable grouping, and regularized models to reduce effective dimensionality and mitigate the curse of dimensionality and multicollinearity (Bühlmann & van de Geer, 2011; Hastie et al., 2009b; Vershynin, 2018).
4. **Robust modeling and inference.** Apply missing-data methods consistent with plausible mechanisms and robust or high-dimensional outlier-detection techniques (Little & Rubin, 2019; Rousseeuw & Leroy, 1987; Rubin, 1976; Zimek et al., 2012).

5. **Visualization and communication.** Present multiple, complementary visualizations; clarify the objectives and limitations of each; and explicitly communicate uncertainty (Borsboom et al., 2021; Cleveland, 1993).

This workflow improves both statistical performance and the credibility of data-driven decisions.

Discussion and Conclusion

Multivariate data analysis is indispensable in modern data science because most real problems involve multiple interacting variables. However, the shift from univariate or bivariate settings to multivariate and high-dimensional contexts introduces distinct challenges: the curse of dimensionality and counterintuitive geometry, multicollinearity and complex dependence structures, missing data and multivariate outliers, and difficulties in visualization and interpretability (Bellman, 1957; Donoho, 2000; Vershynin, 2018; Zimek et al., 2012). These challenges affect not only the statistical properties of estimators and predictions but also the substantive decisions that analyses are meant to inform.

The literature provides a rich toolkit for mitigating these issues. Dimension reduction and feature selection help control complexity and stabilize models (Bühlmann & van de Geer, 2011; Hastie et al., 2009b; Jolliffe & Cadima, 2016). Robust methods and principled missing-data techniques preserve inferential validity under realistic data imperfections (Austin et al., 2021; Hughes et al., 2019; Little & Rubin, 2019; Rousseeuw & Leroy, 1987; Rubin, 1976). Sophisticated visualization and interpretability tools enable analysts to explore high-dimensional structure while still communicating results in a manner accessible to decision-makers (Borsboom et al., 2021; Cleveland, 1993; van der Maaten & Hinton, 2008). Importantly, these strategies are most effective when combined with careful problem formulation and substantive domain knowledge.

For students and practitioners, the implication is that multivariate data analysis requires both technical skill and methodological skepticism. Default methods that work well in low dimensions can fail dramatically when naively extended to complex multivariate

data. A disciplined workflow that explicitly addresses dimensionality, dependence, missingness, outliers, and visualization is therefore essential. Adopting such a workflow improves not only statistical performance but also the transparency and trustworthiness of data-driven decisions.

References

- Austin, P. C., et al. (2021). Missing data in clinical research: A tutorial on multiple imputation [Tutorial on practical use of multiple imputation in clinical studies]. *Statistics in Medicine*, 40(15), missing–pageinfo. <https://doi.org/10.1002/sim.9020>
- Bellman, R. (1957). *Dynamic programming*. Princeton University Press.
- Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., et al. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, 1(58), 1–18. <https://doi.org/10.1038/s43586-021-00055-w>
- Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer. <https://doi.org/10.1007/978-3-642-20192-9>
- Cleveland, W. S. (1993). *Visualizing data*. Hobart Press.
- Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality [American Mathematical Society lectures]. *Proceedings of the AMS Conference on Mathematical Challenges of the 21st Century*.
- Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J. R., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27–46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Everitt, B. S. (1975). Multivariate analysis: The need for data, and other problems. *British Journal of Psychiatry*, 126(3), 237–240. <https://doi.org/10.1192/bjp.126.3.237>
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009a). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>

- Hastie, T., Tibshirani, R., & Friedman, J. (2009b). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
<https://doi.org/10.1007/978-0-387-84858-7>
- Hazra, A., & Gogtay, N. (2017). Biostatistics series module 10: Brief overview of multivariate methods. *Indian Journal of Dermatology*, 62(4), 358–366.
https://doi.org/10.4103/ijd.IJD_296_17
- Hughes, R. A., Heron, J., Sterne, J. A. C., & Tilling, K. (2019). Accounting for missing data in statistical analyses: Guidance for practitioners. *International Journal of Epidemiology*, 48(4), 1294–1304. <https://doi.org/10.1093/ije/dyz032>
- Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Pearson Prentice Hall.
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley.
- Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. Academic Press.
- McQuitty, S. (2018). The purposes of multivariate data analysis methods: An applied commentary. *Journal of African Business*, 19(1), 124–142.
<https://doi.org/10.1080/15228916.2017.1374816>
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. John Wiley & Sons. <https://doi.org/10.1002/0471725382>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
<https://doi.org/10.1093/biomet/63.3.581>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>

- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9, 2579–2605.
- Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press. <https://doi.org/10.1017/9781108231596>
- Zimek, A., Schubert, E., & Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5), 363–387. <https://doi.org/10.1002/sam.11161>

Appendix: Figures

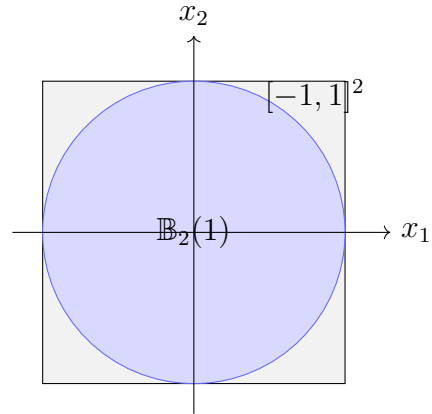


Figure 1

Two-dimensional illustration of the unit ball $\mathbb{B}_2(1)$ inside the hypercube $[-1, 1]^2$. In higher dimensions, the volume of the ball becomes a vanishing fraction of the cube's volume, one aspect of the curse of dimensionality (Donoho, 2000; Vershynin, 2018).

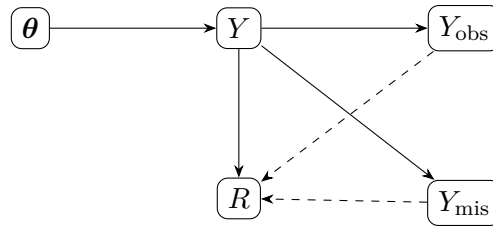


Figure 2

Graphical representation of a generic missing-data setup. Solid arrows indicate data generation; dashed arrows indicate potential dependencies of the missingness indicator R on observed and unobserved data. MCAR, MAR, and MNAR mechanisms correspond to restrictions on the dashed arrows (Little & Rubin, 2019; Rubin, 1976).

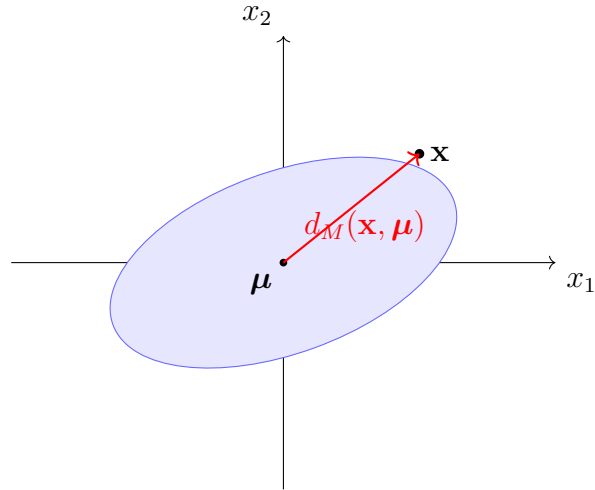


Figure 3

Covariance ellipse of a bivariate normal distribution with mean μ and Mahalanobis distance $d_M(\mathbf{x}, \mu)$ from μ to an observation \mathbf{x} . Points far outside the ellipse have large Mahalanobis distances and are potential multivariate outliers (Mardia et al., 1979; Rousseeuw & Leroy, 1987).