

Sparse Data Reconstruction, Missing Value and Multiple Imputation through Matrix Factorization

Sociological Methodology
2023, Vol. 53(1) 72–114
© American Sociological Association 2022
DOI: 10.1177/00811750221125799
<http://sm.sagepub.com>



Nandana Sengupta¹ , Madeleine Udell²,
Nathan Srebro^{3,4}, and James Evans^{5,6} 

Abstract

Social science approaches to missing values predict avoided, unrequested, or lost information from dense data sets, typically surveys. The authors propose a matrix factorization approach to missing data imputation that (1) identifies underlying factors to model similarities across respondents and responses and (2) regularizes across factors to reduce their overinfluence for optimal data reconstruction. This approach may enable social scientists to draw new conclusions from sparse data sets with a large number of features, for example, historical or archival sources, online surveys with high attrition rates, or data sets created from Web scraping, which confound traditional imputation techniques. The authors introduce matrix factorization techniques and detail their probabilistic interpretation, and they demonstrate these techniques' consistency with Rubin's multiple imputation framework. The authors show via simulations using artificial data and data from real-world subsets of the General Social Survey and National Longitudinal Study of Youth cases for which matrix factorization techniques may be preferred. These findings recommend the use of matrix factorization for data reconstruction in several settings, particularly when data are Boolean and categorical and when large proportions of the data are missing.

Keywords

machine learning, matrix factorization, missing value imputation, social science data, sparse data reconstruction

In this article, we introduce the application of matrix factorization methods for missing value imputation into the social science literature. We show that under certain assumptions, matrix factorization satisfies the requirements of the multiple imputation framework, and we demonstrate how to generate multiple imputations using matrix factorization. We compare imputations from LowRankModels, an open-source

¹School of Public Policy at the Indian Institute of Technology, Delhi, India

²Cornell University, Ithaca, NY, USA

³Toyota Technology Institute of Chicago, Chicago, IL, USA

⁴Department of Computer Science and Committee on Computational and Applied Mathematics, University of Chicago, Chicago, IL, USA

⁵Knowledge Lab and Program in Computational Social Science, University of Chicago, Chicago, IL, USA

⁶Santa Fe Institute, Chicago, IL, USA

Corresponding Author:

James Evans, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, USA

Email: jevans@uchicago.edu

software package that implements the matrix factorization techniques, with those obtained from the two most popular missing value imputation packages in social science, Amelia and MICE, using simulation exercises on artificial and real-world data. On the basis of the underlying assumptions of the different frameworks and the results of the simulations, we discuss conditions under which matrix factorization may be preferred over traditional missing value imputation techniques.

Empirical analysis in the social sciences is plagued with the problem of missing values. The origins of missing values are diverse: respondents may choose not to answer sensitive questions from opinion surveys, respondents may move and inadvertently drop out of longitudinal studies, and the vagaries of history—such as the fire and subsequent destruction of much of the 1890 U.S. census—create gaps so wide that missing values outnumber observations (Blake 1996). Although some researchers ignore missing values or resort to unprincipled methods to deal with them, which weakens the robustness and generality of their findings to new contexts, the application of statistical approaches toward missing value imputation has become an essential component in a wide range of empirical social science analyses.

Social science scholars have used missing value imputation techniques to improve their inferences in a number of contexts, including analysis of marriage among cohabiting partners (Sassler and McNally 2003), feminist identities (Harnois 2005), religious service attendance (Brenner 2011), local and global inequality (Solt 2009), the effects of poverty on youth development (Davey, Shanahan, and Schafer 2001) and education (Alexander, Entwisle, and Olson 2007), and employment dynamics with census data (Treiman, Bielby, and Cheng 1988). Missing value imputation has also been used to study public opinion (Linos and West 2003), voting behavior (Ansolabehere, Iyengar, and Simon 1999; Bernhagen and Marsh 2007; Pasek et al. 2009), and the analysis of conflict (Fearon 2005). Missing value imputation is a valuable tool for policy evaluation (Chiozza and Goemans 2004; King et al. 2009; Scheve and Slaughter 2001), survey methodology (Frankel and Hillygus 2014; Si, Reiter, and Hillygus 2015), and the design of social and political institutions (Gay 2002; Gerber and Hopkins 2011; Honaker, Katz, and King 2002; Stasavage 2005).

Within the social sciences, the popular multiple imputation technique developed by Rubin (1987) is the preferred framework for missing value imputation. The framework provides a structured and statistically valid approach for inference in the presence of missing values. The goal of multiple imputation is to ensure that inferences from data reflect the uncertainty inherent in the data's reconstruction. The chief weakness of the framework is that strong parametric assumptions are often required to prove that its requirements are met. A number of authors have criticized the approach and suggested competing imputation techniques (notably Fay 1992; Nielsen 2003; Rao and Shao 1992), but more have defended its value as an overarching framework (see Meng 1994; Rubin 1996; Schafer 1999). The current state of the literature on missing value imputation in the social sciences remains firmly centered around the multiple imputation framework, even incorporating new approaches like Gaussian copula models (Hollenbach et al. 2018). In fact, it is the preferred imputation technique in all articles cited in the preceding paragraph; although not all can prove they satisfy its stringent

requirements. Missing value imputation continues to be a vibrant area of methodological research (see Blackwell, Honaker, and King 2015; Kropko et al. 2014; Lall 2016). This article stays within the multiple imputation framework by demonstrating how a novel machine learning technique, sparse matrix reconstruction using matrix factorization, can be extended to provide multiple imputations. We further suggest specific settings in which matrix factorization methods may improve on traditional social science techniques for multiple imputation.

Sparse matrix reconstruction using matrix factorization has become an active area of computer science research that has fueled successful recommendation engines that predict preferences for movies, songs, and products not experienced previously. In particular, use of low-norm and low-rank approximations of the data matrix have proven successful in predicting a wide range of missing values. Many influential studies have contributed to this research, and Candès and Plan (2009), Candès and Recht (2009), Foygel and Srebro (2011), Negahban and Wainwright (2011), Rubin (1987), and Srebro (2004) presented results on statistical properties of these estimators. Much early work in matrix reconstruction revolved around principal components analysis (PCA), or for non-square data matrices, singular value decomposition (SVD), in which all entries in the data matrix are numerical valued. The literature has since developed to accommodate heterogeneous data types under a unified matrix completion framework (see Collins, Dasgupta, and Schapire 2001; Singh and Gordon 2008; Udell et al. 2014). The matrix factorization framework differs from traditional imputation models in social science on the basis of (1) the ability to handle highly sparse data sets and (2) the use of out-of-sample testing or cross-validation for model selection. In this article, we demonstrate how these approaches compare with popular, state-of-the-art imputation techniques used for social science analysis.

Some recent work has explored the intersection of social science data imputation and machine learning techniques. For instance, causal inference questions may be interpreted as a missing data problem wherein counterfactual outcomes are treated as “missing.” This has led to recent work in economics (Athey et al. 2018) and computer science (Kallus, Mao, and Udell 2018) in which matrix factorization has improved the quality of causal inferences. Yu, Rao, and Dhillon (2016) introduced a temporal regularized matrix factorization framework for data-driven temporal learning and forecasting with applications in demand forecasting and climatology. Ouzienko and Obradovic (2014) and Huisman (2014) discussed imputation methods for estimating missing links in social network data using a variety of data-driven techniques. All but one of these articles, however, were published in computer-science venues. To our knowledge, our article marks the first attempt to introduce matrix factorization techniques to the larger social science community for general imputation and to extend them for the multiple imputation framework.

This article provides two main contributions. First, we introduce the idea of sparse data reconstruction via matrix factorization methods into the social science literature by presenting an overview of the framework, from its relationship to linear regression to its probabilistic and geometric interpretation. We present a menu of loss functions that may be used for common data types arising in social surveys, including loss

functions new to the social science domain, as well as a menu of regularization penalties that may be used to prevent model overfitting. We also show that matrix factorization satisfies the requirements of the multiple imputation framework, and we demonstrate how to generate multiple imputations using matrix factorization. In doing so, we summarize results from the statistical literature that guarantee accurate imputation, along with the assumptions they rely on and recent attempts to relax these assumptions. Finally, we detail mathematically how our approach yields proper imputations according to Rubin's (1996) framework.

Second, we introduce readers to *LowRankModels*, an open-source software package that implements the matrix factorization techniques discussed here. We compare imputations from *LowRankModels* with those obtained from the two most popular missing value imputation packages in social science, *Amelia* and *MICE*, using simulation exercises on artificial and real-world data. Matrix factorization methods have lower root mean squared error (RMSE) and better coverage rates for Boolean variables in the simulations using artificial data, but in the simulations using real-world data, matrix factorization leads to reduced RMSE and high coverage rates across all data types. In both sets of simulations, we find that matrix factorization techniques are robust to increasing sparsity of the data set, whereas the traditional imputation techniques often fail to produce any imputations.

Specifically, our simulation results suggest that matrix factorization may be uniquely profitable for use in settings with substantial sparsity from missing entries, and when data are predominantly qualitative, yielding Boolean and categorical variables. Social science contexts in which such data are particularly prevalent include surveys with substantial attrition, a critical issue in the contemporary era of vanishing response rates (Boyle et al. 2021); text and network data derived from social media sources such as Twitter, Reddit, and Facebook; Web-scraped data from e-commerce or government Web sites; and historical archival data, as from ship manifests, commercial contracts, and letters. Moreover, increasing efforts to impute individual preferences and perspectives indirectly from online behavior almost invariably faces extensive challenges of sparsity.

To be clear, we are not claiming that matrix factorization techniques are the best under all conditions. Rather, our focus here is to introduce the factorization design to the social science literature, and to demonstrate its superiority in some cases, such as those involving very sparse data. In doing so, we provide baseline comparisons with the most popular traditional imputation techniques to demonstrate conditions under which matrix factorization may be preferred. This is an active area of research where a number of related techniques have been explored, including Gaussian copulas (Hollenbach et al. 2018; Zhao and Udell 2020), Gaussian mixture models (Sang, Kim, and Lee 2020), random forests (Xia et al. 2017), and deep neural networks (Che et al. 2018; Choudhury and Pal 2019; Z. Wang et al. 2021). Some of the recent explorations have deep architectural similarity to matrix factorization (e.g., neural network autoencoders are heuristic approximations of matrix factorization [Levy and Goldberg 2014]). *Amelia* too has an option to incorporate priors in the model, and *MICE* has the

option to incorporate random forests (Shah et al. 2014). We demonstrate the power of matrix factorization approaches in comparison with standard methods, show their special application in cases of data sparsity, and suggest their promising, principled relationship with these other emerging methods.

MULTIPLE IMPUTATION FRAMEWORK

The most popular framework among social scientists for dealing with missing data is multiple imputation (Rubin 1987). The framework ensures that the inherent uncertainty in missing data imputation is accounted for in any statistical estimation using imputed data.

Consider a matrix comprising real numbers organized into n rows and p columns, the number of respondents and responses in a survey or objects and features from an archive, which we denote as $\mathbb{R}^{n \times p}$. The complete, observed, and missing data are then denoted by $Y \in \mathbb{R}^{n \times p}$, $Y_{obs} \in \mathbb{R}^{n \times p}$, and $Y_{miss} \in \mathbb{R}^{n \times p}$, respectively. Observed entries in Y_{obs} belong to the set $(i, j) \in \Omega \subset \{1, \dots, n\} \times \{1, \dots, p\}$. Suppose Q denotes the true value of the statistic of interest, and $\hat{Q}$ its estimate from the complete data set. The key assumption in the literature is that data are missing at random (MAR): the missingness depends only on the observed data (Y_{obs}) and is independent of the missing data (Y_{miss}).

The multiple imputation framework estimates Q from imputed data in three steps:

- *Imputation*: m complete data sets are created by imputing missing values in the incomplete data sets m times. Generally, m is chosen to be between 5 and 10.
- *Analysis*: The m complete data sets are used to perform the analysis of interest, leading to m estimates $\hat{Q}^{(l)}$ with estimated variance $R^{(l)}$ where $l = \{1, 2, \dots, m\}$.
- *Pooling*: The m estimates are then averaged to get the final estimate

$$\bar{Q}_m = \frac{1}{m} \sum_{l=1}^m \hat{Q}^{(l)}. \quad (1)$$

Different methods for imputing missing data in the first step result in different estimates $\bar{Q}_m$. To assess uncertainty in this estimate, within-imputation and between-imputation variances corresponding to m estimates are defined, respectively, as

$$R_m = \frac{1}{m} \sum_{l=1}^m \hat{R}^{(l)}, \quad B_m = \frac{1}{m} \sum_{l=1}^m (\hat{Q}^{(l)} - \bar{Q}_m)^2.$$

The total estimated variance of the pooled estimate $\bar{Q}_m$ is

$$T_m = \frac{B_m}{1+m} + R_m. \quad (2)$$

Proper Imputations

When does the multiple imputation procedure result in accurate estimates? Rubin (1996) introduced a number of technical conditions (see equations 2.6–2.8 in Rubin 1996) under which imputations are guaranteed to be “proper.” Imputations are proper when the pooled estimate from imputed data sets approaches the estimate from the complete data set with additional imputations. The pooled estimate $\bar{Q}_m$ obtained from proper imputations is an asymptotically consistent estimator for Q and follows a t -distribution (Rubin 1987). Given MAR data and proper imputations, tests of significance and confidence intervals based on this distribution are statistically valid even in the presence of missing data. The relevant distribution is characterized as

$$(\bar{Q}_m - Q)/\sqrt{T_m} \sim t_\nu \quad \text{with degrees of freedom} \quad \nu = (m-1) \left[1 + \frac{R_m}{(1+m^{-1})B_m} \right]^2.$$

Proper imputations are typically obtained via Bayesian models of the data: (1) a parametric model of the complete data is assumed, typically $Y \sim \mathcal{N}(\mu, \Sigma)$; (2) the unknown parameters are estimated using observed data via maximum likelihood estimation; and (3) m independent draws from the distribution of Y_{miss} conditional on Y_{obs} make up m imputed data sets. Across these imputed data sets, inferences are made with appropriate variance.

Two of the most popular software packages (both available in R) for conducting multiple imputations are Amelia (Honaker, King, and Blackwell 2011) and MICE (Buuren and Groothuis-Oudshoorn 2011). Multiple imputations produced by Amelia follow from a joint modeling (JM) approach, which assumes a multivariate joint normal distribution for the entire data, that is, $Y \stackrel{iid}{\sim} \mathcal{N}(\mu, \Sigma)$, where μ is a $(p \times 1)$ vector of means and Σ refers to a covariance matrix of dimension $(p \times p)$. To estimate parameters given the presence of missing entries, Amelia uses an expectation maximization (EM) algorithm in which the following two steps are repeated iteratively until convergence. Starting with some initialization of the model parameters $(\hat{\mu}, \hat{\Sigma})$, in the first “expectation step,” these estimates of the parameters are used to fill in the missing data. In the second “maximization step,” the filled in data are used to reestimate the parameters (Honaker et al. 2011). MICE does not assume a joint multivariate normal distribution but allows for different models for each variable. This is referred to as a fully conditional specification (FCS). Imputations are produced using a chained equations algorithm that cycles through each column of the data, fitting a single model at a time for each column. This chain is repeated for a number of iterations, at the end of which a single imputation is produced. The modeling and imputation step is repeated m times to generate m imputed data sets (Buuren and Groothuis-Oudshoorn 2011).

MATRIX FACTORIZATION FRAMEWORK

The imputation of missing values is also a central problem in many computer-science applications. The most common application is to online recommendation systems, which predict a consumer’s rating of a new product from previously observed ratings

for other products by other consumers. Some of the most successful imputation models in computer science are based on matrix factorization techniques. Unsupervised deep learning neural network models, called autoencoders, are now widely used for heuristic optimization in the factorization of very large matrices (Goodfellow, Bengio, and Courville 2016). Importantly, and possibly of particular interest to social scientists, matrix factorization can be interpreted as the result of a probabilistic process, as discussed in greater detail in the next subsection.

As the name suggests, matrix factorization techniques factorize tabular data of interest (i.e., $Y \in \mathbb{R}^{n \times p}$) into two component matrices $U \in \mathbb{R}^{n \times k}$ and $V \in \mathbb{R}^{k \times p}$ such that the product UV recovers the original matrix. Traditional imputation methods use a linear regression probability model that exploits variation in items within an observation, or column values within a row, to predict missing values. Matrix factorization methods add another dimension to the imputation solution by additionally using variation in the observations or subjects “within” an item, or rows within a column. In this way, the matrix factorization framework can be understood as a two-dimensional regression model that simultaneously takes into account similarities across respondents, cases, or objects *and* similarities across responses, attributes, or features. A classic method of matrix factorization was proposed by social statisticians Kroonenberg and De Leeuw (1980) through alternating least squares, on items within each case, and then cases within each item (see also Takane, Young, and De Leeuw 1977). Formally, if Y is a matrix we want to factorize, and we are interested in finding factors U and V that minimize the squared difference $\|Y - UV\|^2$, we can interpret this in two ways:

- the matrix U is given and solving for V is a linear regression problem for each column of Y , or
- the matrix V is given and solving for U is a linear regression problem for each row of Y .

We then alternate between the two, reminiscent of the iterative EM method for estimating model parameters. In traditional linear regression, however, it is required that the number of columns in V is strictly less than the number of rows in U , and therefore the second option cannot be feasibly solved. In matrix factorization models, the addition of a regularization penalty (discussed later in this section) allows both problems to be solved simultaneously. In this way, matrix factorization can and should be understood as a probabilistic approach that simultaneously accounts for multiple dimensions of similarity in the data.

To describe matrix factorization mathematically, let us define $u_i \in \mathbb{R}^{1 \times k}$ to be the i th row of U , and $v_j \in \mathbb{R}^{1 \times p}$ to be the j th column of V . Then $u_i v_j = (UV)_{ij} \in \mathbb{R}$ denotes the inner product between the two. Observed entries in Y_{obs} belong to the set $(i, j) \in \Omega \subset \{1, \dots, n\} \times \{1, \dots, p\}$. To learn factors U and V , the simplest approach is to solve

$$\min_{U, V} \sum_{(i, j) \in \Omega} (Y_{ij} - u_i v_j)^2. \quad (3)$$

The value of k determines the dimensionality of the factorization and the rank of the matrix UV . When $k < \min\{n, p\}$, the product UV of the solution (U, V) to (3) forms a low-rank approximation of Y . If all entries in Y are observed, equation (3) is the solution to finding the principal component decomposition of matrix Y . When $k = \min\{n, p\}$, then the problem is easy to solve: the solution simply sets one factor to Y and the other to the identity. In practice, the value of k is typically selected via a data-driven method known as cross-validation, described in the Implementation section.

A regularization penalty is also added to the loss function (equation 3) to simplify and reduce variance of the model to avoid overfitting, without substantially increasing bias. Without a penalty, matrix factorization can predict precise values on the sample from which it was trained, but fail to generalize beyond. One can understand regularization as an enforcement of Occam's razor simplicity heuristic for model selection, or conceptualize it as a smooth Bayesian prior that directly penalizes model complexity. In the specific context of missing values, regularization introduces stability that is critical for imputing from sparse data. Trace norm regularization is one of the most popular choices for matrix factorization and, as demonstrated in the simulations that follow, one of the most successful for missing value imputation. The trace norm Y_Σ , which we list and explain in Table 2, is defined as the sum of the singular values of Y . In trace norm regularization, the norm is added as a regularization penalty term to the loss function

$$\min_{U, V} \sum_{(i,j) \in \Omega} (Y_{ij} - u_i v_j)^2 + \gamma (U_F^2 + V_F^2), \quad (4)$$

where $\gamma \geq 0$ controls the relative weight given to the regularization penalty. A trace norm with $\gamma > 0$ will limit the average size of the entries in U and V to be closer to zero, preferring a smoother factorization. If we solve equation (4) *without* imposing a dimensionality constraint on U and V (i.e., letting $k = \min(n, p)$), then the problem is referred to as a low-norm matrix factorization problem. This formulation is preferred by some analysts because it leads to a convex, more easily soluble optimization problem. These and related mathematical details are explained fully in Udell et al. (2014). In practice, the value of γ is typically selected via a data-driven method known as cross-validation, described in the Implementation section.

Interpretation

Matrix factorization can be profitably interpreted as the result of a probabilistic process. Consider the following data-generating process:

$$Y_{ij} = (UV)_{ij} = (\bar{U} \ \bar{V})_{ij} + \eta_{ij}, \quad \eta_{ij} \sim N(0, \sigma_\eta^2).$$

An observed data matrix Y_{ij} equals the true underlying matrix factorization model $(\bar{U} \ \bar{V})_{ij}$ together with additive Gaussian noise η_{ij} . $\bar{U}$ and $\bar{V}$ are the true latent factors—component matrices whose entries are assumed to be generated by drawing independent samples from a normal distribution with mean 0 and variance γ^{-1} (for

some $\gamma > 0$). The estimator that finds matrices U and V most likely to have generated the data is then the maximum a posteriori estimator (U, V) of $(\bar{U}, \bar{V})$:

$$\max_{U, V} \exp\left(-\frac{\gamma}{2} U_F^2\right) \exp\left(-\frac{\gamma}{2} V_F^2\right) \prod_{(i,j) \in \Omega} \exp\left(-\left(Y_{ij} - (UV)_{ij}\right)^2\right). \quad (5)$$

The solution of equation (4) is equivalent to the solution of equation (5) as the former can be obtained via a negative log transformation of the latter.

Factors U and V admit a number of other useful interpretations (see Udell et al. 2014). A geometric articulation of factors U and V is particularly intuitive and can assist in identifying the appropriate scope of application. Here, U and V represent latent variables, features inferred from inherent structure in the data. Each row of U , u_i , represents a given row in the original data (e.g., a survey respondent) by a vector in $\mathbb{R}^k$, and each column represents its coordinates in a k -dimensional Euclidean space. This allows us to map and measure the distance between each respondent in the space of questions, and each question in the space of respondents. The matrix V maps these representations back into the original feature space in $\mathbb{R}^p$.

Loss Functions. So far, our discussion has focused on real-valued data, but most social science data elicited by surveys is not real-valued. Four of the most common data types in social science surveys¹ are (1) *real-valued*, such as income data; (2) *Boolean*, including responses to “yes/no” questions; (3) *ordinal*, as from a Likert-type scale for ordered responses; and (4) *categorical*, comprising data on membership in different groups (e.g., religions or ethnicities). The data type may be represented as labels on the points in the latent feature space. The minimization problem in equation (4) can be rewritten for general loss functions, $\mathcal{L}(Y_{ij}, u_i v_j)$, and for general regularizers, $r(U, V)$, as

$$\min_{U, V} \sum_{(i,j) \in \Omega} \mathcal{L}(Y_{ij}, u_i v_j) + \gamma \quad r(U, V). \quad (6)$$

The first term in equation (6) refers to the loss function applied to a given column. Table 1 presents a menu of loss function specifications for different data types. Representations of distinct data types and corresponding loss functions in the latent factor space are presented in Appendix Figure A1.

Regularization. As discussed previously, a regularization penalty is appended to the loss function to prevent overfitting. The penalty is usually some function of the norms of factors U and V . Table 2 presents a menu of matrix norm specifications, their relative strengths for regularization, and their geometric interpretation.

An Illustration. Figure 1 illustrates a low-rank matrix factorization. Figure 1a represents matrix Y as approximated by the multiplication of factors U and V . The rows in Y represent n individuals and the columns their p -dimensional characteristics. Note that some entries in Y are missing. The columns of U can be interpreted as the k -dimensional latent factors, and the rows of U the value of these latent factors for a given individual. Rows of V map each individual’s latent factors back to the original

Table 1. A Menu of Loss Functions

Data Type	Loss Function Name	Loss Type	Loss Function Specification ^a	Notes
Real	Quadratic	Generative	$(y - w)^2$.	Most commonly used loss for diverse data types. Prone to overfitting as outliers are penalized heavily. From a probabilistic interpretation, the loss is the negative log-likelihood of the normal distribution with mean w and variance 1.
	Huber	Generative	$\begin{cases} \frac{1}{2}(y - w)^2, & \text{if } y - w; \leq \delta \\ \delta(y - w - \frac{1}{2}\delta), & \text{otherwise.} \end{cases}$	Quadratic loss for small deviations, linear penalty for larger deviations ($> \delta$).
	Logistic Hinge	Generative	$\log(1 + \exp(-wy))$	$P(y = +1 w) = \frac{1}{1 + \exp(-wy)}$
	Ordered probit	Discriminative	$\max(0, 1 - wy)$	Only penalize values of w that would predict the class incorrectly.
Ordinal		Generative	$-\sum_{j=1}^M \mathbf{1}_{\{y=j\}} \{\ln(\phi(\mu_j - w) - \phi(\mu_{j-1} - w))\}$	Thresholds $\mu_1 \cdots \mu_{M-1}$ such that $y = m \Rightarrow \mu_{m-1} < w \leq \mu_m$. Probabilistically,
	All thresholds ordinal		$\sum_{j=1}^{M-1} \log(1 + \exp(-w_j \mathbf{1}_{y>j}))$	$P(y = m) = P(\mu_{m-1} < w \leq \mu_m)$ Where 1 is an indicator function that is 1 if its argument is true and -1 otherwise. In words, the loss fits a logistic regression model to separate all instances with values $> j$ from all instances with values $\leq j$, for each $j = 1, \dots, M - 1$. See Rennie and Srebro (2005).
Categorical	Multinomial logistic	Generative	$\log(\exp(w_y) / (\sum_{j=1}^M \exp(w_j)))$	Extension of logistic loss to $M > 2$ categories where one category is fixed as a baseline. See Agresti (1996).
	One vs. all hinge	Discriminative	$\sum_{j=1}^{M-1} \max(1 - w_j \mathbf{1}_{y=j}, 0)$.	Extension of hinge loss to $M > 2$ categories. See Rifkin and Klautau (2004).

^aLoss function specification: $l(y, w)$ where the true value is y and the predicted value is $w = u_i v_j$.

Table 2. A Menu of Regularization Norms

Norm	Norm Specification	Notes
Maximal entry norm	$\ Y_\infty\ = \max_{i,j} (y_{ij})$	Magnitude of the largest entry in Y where y_{ij} is the entry corresponding to row i and column j .
Spectral norm	$\ Y_2\ = \sigma_{\max}(Y)$	Largest singular value of Y , which is equal to the square root of the largest eigenvalue of $Y'Y$.
Frobenius norm	$\ Y_F\ = \sqrt{\text{tr}(Y'Y)} = \left(\sum_{i=1}^r \sigma_i^2\right)^{1/2}$.	Euclidean or l_2 norm of the vector of singular values for Y .
Trace (or nuclear) norm	$Y_\Sigma = \sum_{j=1}^p \sigma_j$. For matrices with a factorization $Y = UV$, $Y_\Sigma = \min Y = UV \quad \frac{1}{2}(U_F^2 + V_F^2)$	Sum of the singular values of the matrix, σ_i , ($i = 1, \dots, p$). Corresponds to the form of the regularization penalty in equation (4). If a matrix has rank k , then it has exactly k nonzero singular values. Although the rank of a matrix counts the number of nonzero or nonvanishing singular values, the trace norm sums their magnitude. We note that matrix factorization using trace norm regularization without explicitly constraining the rank has been proposed as a convex optimization problem that encourages low-rank solutions. Adding a trace norm regularization penalty encourages smaller entries in U and V , <i>on average</i> . In other words, the trace norm ensures that <i>most</i> entries in U and V are “small.”
Max norm	For matrices with a factorization $Y = UV$, $Y_{\max} = \inf\{U_{2,\infty} V_{2,\infty} : UV = Y\}$ where $U_{2,\infty} = \max_i \left(\sum_{j=1}^p U_{ij}^2\right)^{1/2}$; $V_{2,\infty} = \max_i \left(\sum_{j=1}^p V_{ij}^2\right)^{1/2}$.	Maximum of the L_2 norms of each of its rows y_i , ($i = 1, \dots, n$). Ensures that all entries in U and V are bounded above. Heuristically, max-norm regularization is expected to be effective for uniformly bounded data, such as preferences. Geometrically, adding a max-norm regularization penalty encourages smaller entries in U and V <i>uniformly</i> . In other words, the max norm ensures <i>all</i> entries in U and V are “bounded” within some constant distance of the origin.

Note: For a rectangular matrix $Y \in \mathbb{R}^{n \times p}$, σ_i denotes the i th largest singular value. Denote the true rank of the matrix by r .

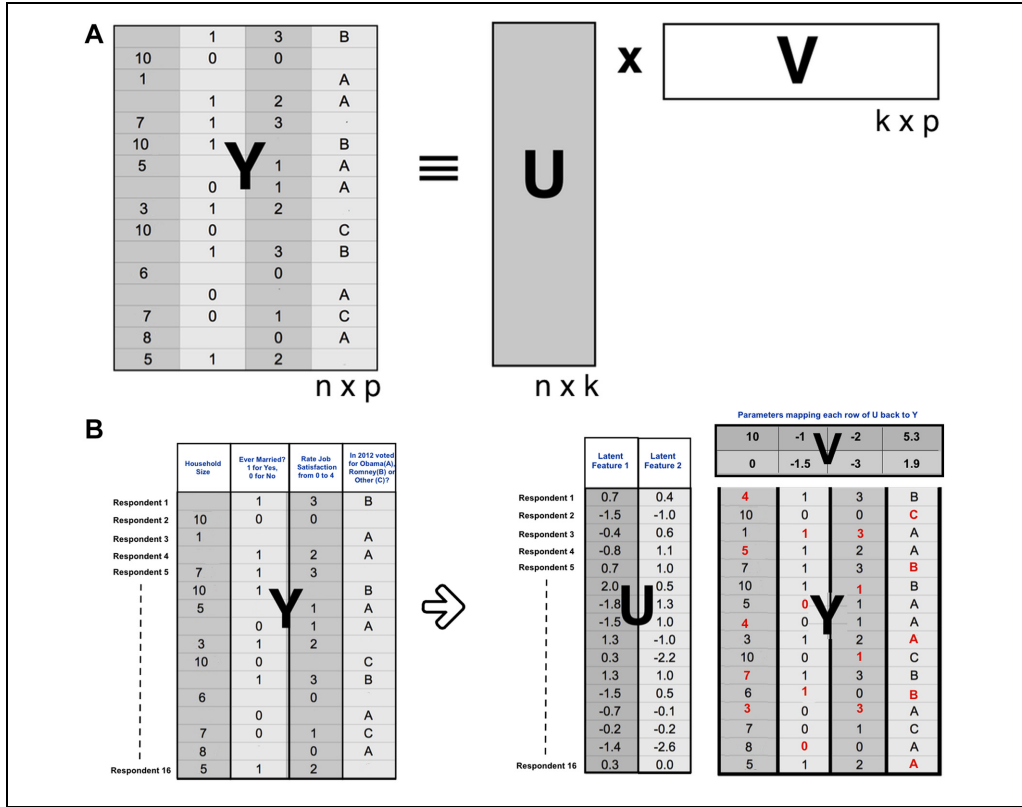


Figure 1. An illustration of low-rank matrix factorization.

Note: Panel (a) represents matrix Y as approximated by the multiplication of factors U and V where some entries in Y are missing. The rows in Y represent n individuals and the columns their p -dimensional characteristics. The columns of U can be interpreted as the k -dimensional latent factors, and the rows of U represent the value of these latent factors for a given individual. Rows of V map each individual's latent factors back to the original p -dimensional characteristic space. In panel (b), we consider a solution to the matrix completion problem when the dimensionality of the factors is constrained to $k=2$. The matrix U generates latent factors of the original matrix (two for each respondent), and the parameters in matrix V map these latent features back to the original matrix space (two parameters per original question) via the generalized loss functions for different data types as described in the "Loss Functions" section. The predicted product columns also contain estimates of the missing entries,

p -dimensional characteristic space. In Figure 1b, we consider a solution to the matrix completion problem when the dimensionality of the factors is constrained to $k=2$. In this stylized example of a social science survey, the original data contain four columns corresponding to questions on (1) household size, (2) ever married, (3) job satisfaction rating, and (4) vote in 2012. Some respondents did not answer certain questions and thus some entries are missing or blank. This matrix is approximated by the product of two matrices, U and V . The matrix U generates latent factors of the original matrix (two for each respondent), and the parameters in matrix V map these latent features back to the original matrix space (two parameters per original question) via the general

loss functions for different data types (as described in the “Loss Functions” section). The predicted columns also contain estimates of the missing entries, which are highlighted. See the online version for a color figure where the missing entries are highlighted in red.

Comparison with Amelia and MICE

We now compare matrix factorization techniques with the two traditional imputation techniques Amelia and MICE.

Underlying Model. As discussed above, imputations from Amelia follow from a JM approach, whereas those from MICE follow from an FCS. Matrix factorization algorithms combine the joint estimation properties of Amelia with the flexibility of MICE. This is possible by having different loss functions for different columns of the data set while solving the minimization problem jointly rather than one column at a time.

Stability. The regularization penalty inbuilt into matrix factorization models, which can be interpreted as a Bayesian prior, introduces stability into resulting imputations, particularly when data sets are sparse. Although default models for Amelia and MICE do not take into account regularization or priors, in practice their respective R packages do allow some options for the introduction of priors. However, these packages do not provide users with clear instructions on the choice of associated hyperparameters, unlike the use of cross-validation to select tuning parameters in a principled data-driven manner, which is inbuilt into the matrix factorization framework.

Statistical Guarantees. The underlying models for both Amelia and MICE imply that they require order of p^2 observations to estimate the model parameters. On the other hand, for matrix factorization models, in practice, one may be able to recover a complete matrix even if *very few* of its entries are observed, because in many practical cases the required sample size for exact reconstruction is much smaller than the total number of entries in the original matrix. This is expected to hold true as long as the number of observations for each respondent is greater than $\mathcal{C} \log(p)$ (where p is the total number of questions in the data set and $\mathcal{C}$ is some constant) as shown in Candès and Recht (2009). Such statistical guarantees are discussed in more detail in the Appendix (for more details and results on the exact recovery of sparse matrices, see Candès and Plan 2009; Candès and Recht 2009).

Assumptions. Whereas the standard assumption underlying Amelia and MICE is that data are MAR, the statistical guarantees on exact recovery in matrix factorization models are based on a stronger assumption that the data are sampled uniformly at random, which is equivalent to missing completely at random (MCAR). Despite this, several studies (e.g., Josse, Pagès, and Husson 2011; Shapiro, Xie, and Zhang 2018), including ours, demonstrate how they perform better on data that are MAR than do probabilistic models designed for this task. One possibility derives from the finding that maximizing the observed likelihood of data using the EM algorithm (Dempster, Laird, and Rubin 1977) theoretically satisfies the MAR assumption. Matrix factorization can also be learned using the EM algorithm (Josse and Husson 2016), such that

matrix factorization may approximately satisfy the assumptions proven to be satisfied with MAR data. Another potential reason is that our use of regularization methods and the native exploitation of both axes of data similarity (i.e., column and row) more than compensates for dependent biases in the data.

MATRIX FACTORIZATION FOR MULTIPLE IMPUTATION

In this section, we extend the matrix factorization framework to incorporate multiple imputations. First, using the framework presented in Gunasekar, Ravikumar, and Ghosh (2014), we sketch out the argument that under certain plausible assumptions, matrix factorization models can produce proper imputations, and hence produce asymptotically consistent estimates of matrix values. Second, we discuss in detail the implementation of matrix factorization models with multiple imputation using the open-source software LowRankModels.

Proper Imputations

Fix a low-rank matrix $Y \in \mathbb{R}^{n \times p}$ that satisfies a few additional technical assumptions described in Gunasekar et al. (2014). Suppose, without loss of generality, that $n < p$, and suppose, with loss of generality, that $\Omega \subseteq [1, \dots, n] \times [1, \dots, p]$ is sampled uniformly from the set of matrix entries and at least $|\Omega| > p \log p$ entries have been observed. For these entries, suppose that the values $Y_{obs,ij}$ are sampled according to an exponential family probability distribution P_{ij} for each $(i,j) \in \Omega$. Pick γ sufficiently large and estimate $\hat{Y}$ by solving the regularized maximum likelihood estimation problem

$$\sum_{(i,j) \in \Omega} -\log P_{ij}(Y_{obs,ij} | Y_{ij}) + \gamma \|Y\|_*.$$

Then with high probability,

$$\|\hat{Y} - Y\|_F^2 \leq C \frac{p \log p}{|\Omega|},$$

where C is a constant independent of n and p (for details, see Gunasekar et al. 2014).

Using this result, we can show that $\hat{Y} \rightarrow Y$ in at least two limits:

1. *Constant dimension, repeated samples.* $\hat{Y} \rightarrow Y$ as $|\Omega| \rightarrow \infty$, for any matrix dimensions n and p .
2. *Growing dimensions, constant density.* Fix a constant $\alpha \in (0, 1]$, and sample entries with probability α for a total of $|\Omega| = \alpha np$ observations. Then $\hat{Y} \rightarrow Y$ as $n, p \rightarrow \infty$.

Now for simplicity, suppose the imputation procedure returns $\hat{Y}_{obs} = \hat{Y}$. Then for any estimand Q that is continuous with respect to the data $\hat{Y}_{obs}$, we see that equations 2.6 to 2.8 from Rubin (1996) will hold as either

1. the number of observations $|\Omega| \rightarrow \infty$, fixing n and p , or
2. $n, p \rightarrow \infty$, fixing $|\Omega| = \alpha np$ for any $\alpha \in (0, 1]$.

Implementation

The minimization problem in equation (4) has two parameters, rank k and regularization penalty weight γ , which need to be set. If we set $\gamma=0$, the solution reduces to a pure low-rank approximation. If we set $k \geq \min\{n, p\}$, the solution reduces to a pure low-norm approximation. In all cases, the parameter is selected using hold out sample evaluation or cross-validation. Hold out samples (also called the testing set) are randomly selected samples from the observed values. The training set consists of the remaining observed values. The performance of a given model (i.e., a model with a fixed value of k and/or γ) fit to the training set is evaluated on the testing set. With cross-validation, this is done repeatedly, with distinct, rotating bins of training and testing sets. The idea here is to simulate performance of the model with hold out sample data. Subsamples may be randomized over rows, columns, or indices.

Because both U and V are unknown, the problem in equation (4) is a nonconvex optimization problem. If we solve for one set of factors, while keeping the other fixed, the problem becomes convex and can be solved optimally for most loss functions. The method of “alternating minimization” builds on this principle. One of the benefits of alternating minimization is that it parallelizes easily over factors U and V . Another popular solution method is stochastic gradient descent (see, e.g., Gemulla et al. 2011).

We use the open-source software package LowRankModels written in Julia to compute the matrix factorizations² (an implementation in Python is also available). This package uses the alternating proximal gradient method to fit models, and performs cross-validation by subsampling over indices. A range of loss functions and regularization penalties, including most of those listed in Tables 1 and 2, are implemented in the package. The basic capabilities of the software package include (1) fitting matrix factorization models to data matrices, (2) choosing optimal rank k and regularization parameter γ via cross-validation, and (3) imputing missing values from the optimal model (a) without making uncertainty corrections (maximum likelihood estimate) or (b) after sampling from the resulting distribution (proper imputations).

We fit the low-rank model by finding submatrices U and V that minimize the objective, given values of the (observed) data Y_{ij} for $(i, j) \in \Omega$. Unobserved entries Y_{ij} for $(i, j) \notin \Omega$ are imputed from the fitted model.

For clarity, we will assume the loss functions chosen are *probabilistic*; that is, $L_j(y, z) = -\log \mathbb{P}(y, z)$ is the negative log-likelihood of the probability of observing value y , if observations are drawn from a probability distribution with parameter z . For example, the quadratic loss $L_j(y, z) = (y - z)^2$ is the negative log-likelihood of the normal distribution with mean z and variance 1.

There are two plausible ways to impute missing data in this probabilistic framework.

- *Imputation.* Find the *most likely* value of Y_{ij} , given U and V . This value is the *mode* of the conditional distribution $\mathbb{P}(y|z)$,

$$\hat{Y}_{ij} = \mathbb{P}(y|z).$$

LowRankModels supports this functionality with the command `impute(glrn)`, where the object type `glrm`, or *generalized low-rank model*, in the software refers to a prespecified model that can be invoked for the purpose of matrix factorization and imputation.³

- *Sampling.* Sample a value y from the conditional distribution $\mathbb{P}(y|z)$, $\hat{Y}_{ij} \sim \underset{y}{\operatorname{argmin}} L_j(y, z) = \underset{y}{\operatorname{argmax}} \mathbb{P}(y|z)$.

The output $\hat{Y}$ is a sample from the posterior distribution of Y . LowRankModels supports this functionality with the command `sample_missing(glrn)`. To sample every entry of Y , even those observed, use the command `sample(glrn)`.

For single imputations, which generalizes to simple, effective nonprobabilistic loss functions (e.g., hinge loss), we simply compute the value of y minimizing $L_j(y, z)$ for fixed z .

For multiple imputation, we incorporate observational uncertainty and sampling uncertainty into our imputations (as in Amelia and MICE). For observational uncertainty, we bootstrap with replacement over entries of the matrix. We then incorporate sampling uncertainty into the imputations on the bootstrapped matrices. This produces a *different* $\hat{Y}$ each time new samples are drawn. To extend the approach to losses that do *not* correspond to a probability distribution, we must define a sampling distribution for those losses. The default sampling methodology in LowRankModels is to define

$$\mathbb{P}(y|z) = \frac{\exp(L(y, z))}{\sum_{y'} \exp(L(y', z))},$$

where it sums over all values of y' in the domain of loss L .⁴

SIMULATIONS ON ARTIFICIAL DATA

As a first step to evaluate the relative performance of matrix factorization models, we run a series of simulations on artificial data, along the lines of the simulation implemented by Hollenbach et al. (2018). We compare imputations from Amelia and MICE with two types of matrix factorization models: (1) a possibly low-rank model, Rank, and (2) a full-rank model with trace norm regularization, Trace. The rank k and regularization weight γ are fit via cross-validation. For matrix factorization models, the default loss types are quadratic loss for real-valued columns, logistic loss for Boolean and categorical columns, and all-thresholds loss for ordinal columns.

For each simulation, we first generate a complete matrix consisting of $p = 80$ columns, which is the combination of four submatrices of $p/4 = 20$ columns, each consisting of real, Boolean, categorical valued data. These submatrices have an underlying latent variable structure as follows:

1. $X_{real} = U_{real} V_{real} + \varepsilon$ where $U_{real} \in \mathbb{R}^{n \times k}$, $V_{real} \in \mathbb{R}^{k \times p/4}$, and ε is a zero-mean error.

2. X_{bool} is sampled from a binomial distribution. $Pr(X_{bool}=1)$ corresponds to the logit model on $U_{bool}V_{bool}$ where $U_{bool} \in \mathbb{R}^{n \times k}$ and $V_{bool} \in \mathbb{R}^{k \times p/4}$, that is, $\text{logit}(Pr(X_{bool}=1)) = U_{bool}V_{bool}$.
3. X_{cat} is sampled from a multinomial distribution across D categories. $Pr(X_{cat}=d)$, $d \in \{1, \dots, D\}$ corresponds to the multinomial logistic model on $U_{cat}V_{cat}$ where $U_{cat} \in \mathbb{R}^{n \times k}$, $V_{cat} \in \mathbb{R}^{k \times (D \times p/4)}$.
4. X_{ord} is sampled from a multinomial distribution across D ordered categories, $\{1, \dots, D\}$. Probabilities for each category correspond to the ordered logit model on $U_{ord}V_{ord}$ where $U_{ord} \in \mathbb{R}^{n \times k}$ and $V_{ord} \in \mathbb{R}^{k \times p/4}$. Mathematically, $\text{logit}(Pr(X_{ord} \leq d)) = \alpha_d - UV_{ord}$, where $\alpha_d \in \{\alpha_1, \dots, \alpha_{D-1}\}$ are preset threshold parameters.

The parameter of interest is $\beta = (\beta_1, \beta_2, \dots, \beta_p)$, which is kept fixed throughout our simulations. Using this fixed vector of parameters and the simulated, complete data set $X = \{X_{real}, X_{bool}, X_{cat}, X_{ord}\}$, we simulate $Y = \beta X + \varepsilon$.

The final set of simulations consists of 24 cases corresponding to (1) two different sample sizes, $n=250$ and $n=1,000$; (2) three different levels of MAR missingness, $\rho_{miss} = 20$ percent, $\rho_{miss} = 40$ percent, and $\rho_{miss} = 60$ percent; (3) two different underlying ranks of the various latent variables, a low-rank case of $k=2$ and an intermediate rank case of $k=10$, and (4) two different distributions from which the columns of the latent variable are drawn, a standard normal distribution $\sim N(0, 1)$ and a uniform distribution $\sim \text{Unif}(-1, 1)$.

We create missingness in the X matrix via a MAR mechanism (described in detail in the Appendix) and denote the resulting matrix as X_{obs} . We generate $n_{sim}=500$ such data sets for each of the 24 cases considered and hold these fixed throughout the simulation exercise. For each of the 500 simulated data sets, we obtain $m=5$ imputed data sets from the four different imputation techniques corresponding to Amelia, MICE, Trace, and Rank. Thus $500 \times 20 = 10,000$ imputations are obtained for each of the 24 cases.

The parameter vector $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$ is then estimated by regressing Y on X_m^{imp} where X_m^{imp} is the m th imputed matrix obtained via multiple imputation from the four candidate imputation techniques. The resulting estimates from each technique $\hat{\beta}_{ij}$ corresponding to simulation $i \in (1, \dots, n_{sim})$ and column $j \in (1, \dots, p)$ are compared via the following metrics averaged over all the real, Boolean, categorical, and ordinal valued columns:

1. RMSE: $\sqrt{\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\beta}_{ij} - \beta_j)^2}$.
2. Mean interval length: size of the 95 percent confidence interval associated with $\hat{\beta}_{ij}$, averaged over n_{sim} simulations.
3. Bias: $\left(\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\beta}_{ij} \right) - \beta_j$.
4. Empirical coverage: proportion of times the true parameter value β_j lies in the 95 percent confidence interval associated with $\hat{\beta}_{ij}$.

Results

Comparison of the estimates obtained from different imputation techniques for all the cases considered provides us with the following broad insights. First, traditional techniques consistently perform better than the matrix factorization techniques for coefficients associated with real-valued columns in terms of both RMSE and empirical coverage metrics. This is driven largely by the fact that coefficients from matrix factorization methods have larger biases.

Second, for coefficients associated with Boolean and categorical valued columns, we find that matrix factorization methods have lower RMSE values along with higher or competitive empirical coverage metrics, particularly when the underlying data are not normally distributed or come from a low-rank data-generation process. In particular, the comparatively lower coverage rates for Rank are driven by the higher value of bias associated with the estimates.

Third, for coefficients associated with ordinal valued columns, traditional imputation techniques and matrix factorization techniques are very competitive in terms of both RMSE and empirical coverage metrics.

Fourth, overall, we find that Amelia and MICE perform better in comparison with Trace and Rank in terms of RMSE and empirical coverage probabilities when the underlying data are normally distributed, not low-rank, and for larger sample sizes. Conversely, matrix factorization techniques perform better in comparison with traditional techniques in terms of RMSE and empirical coverage probabilities when the underlying data are not normally distributed, low-rank, and for smaller sample sizes. We also note that empirical coverage rates are affected by the bias-variance trade-off of the estimates. Estimates from matrix factorization techniques tend to have higher coverage rates when they are associated with either lower bias or larger interval lengths or (in some cases) both.

Finally, traditional imputation techniques are sensitive to the level of missingness, often failing to produce imputations when the proportion of missing values is 40 percent and higher, particularly for smaller sample sizes. Detailed numbers are presented in Appendix Table D1.

We present two sets of plots corresponding to small and large sample sizes that demonstrate these insights in greater detail. See the online version for color figures. Average RMSE, interval length, bias, and empirical coverage are presented for each variable type corresponding to different proportions of missingness and the different imputation techniques. Averages are not displayed for cases in which the imputation method crashed repeatedly and failed to produce at least 30 imputations.

For the first set of plots, we consider two extreme cases when sample size is small, $n = 250$. In Figure 2, the underlying data are normally distributed and of intermediate rank, $k = 10$, for each block of $p/4 = 20$ variables corresponding to a given data type. In this case, we find that, especially for lower levels of missingness, matrix factorization techniques have higher RMSE and bias for regression coefficients corresponding to real-valued columns. Although interval lengths are smaller, indicating more precise estimates, this property combined with higher bias leads the matrix factorization techniques to have substantially lower coverage for regression coefficients corresponding

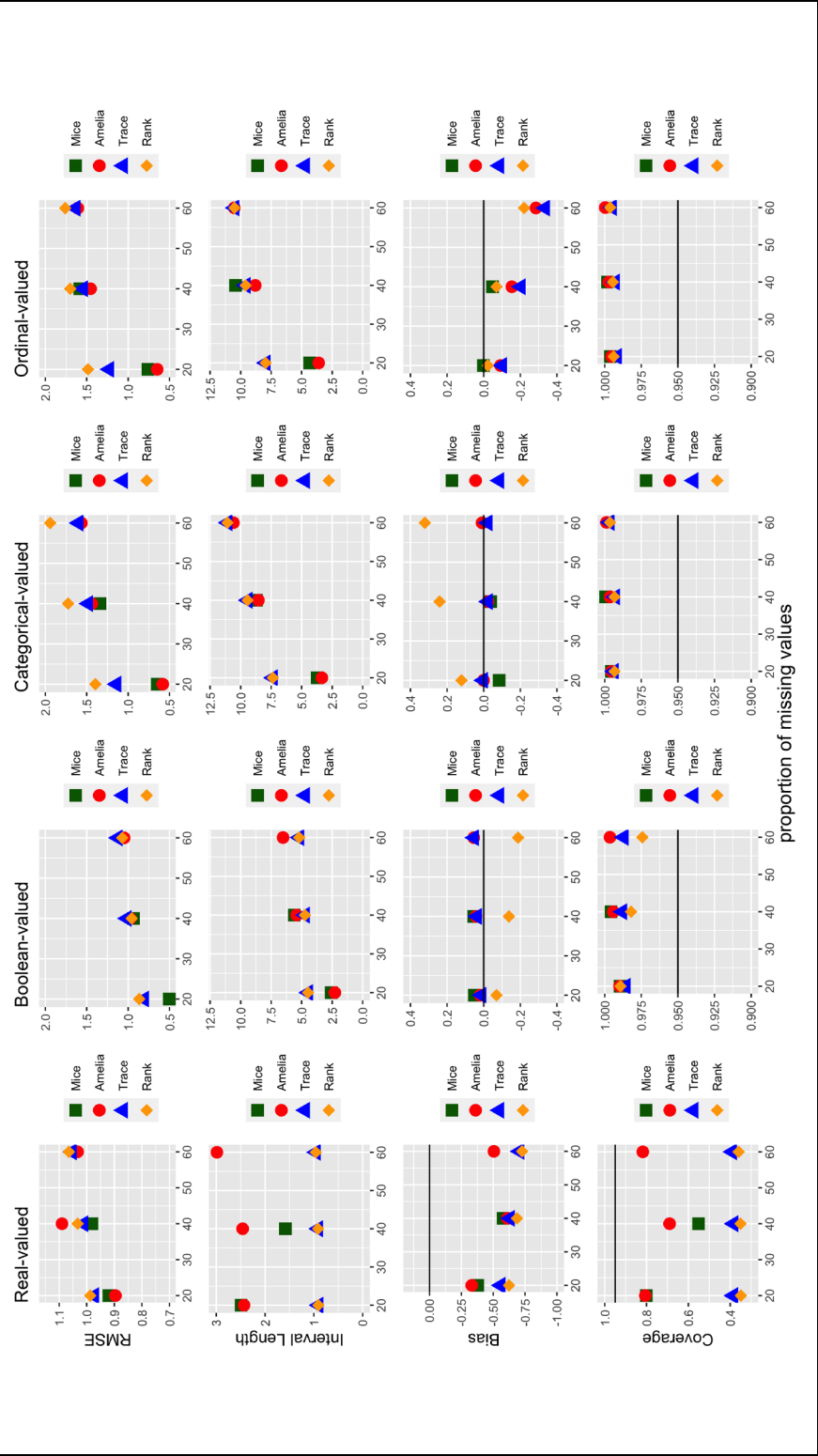


Figure 2. Simulation results on artificial data: normal distribution and intermediate rank.

Note: Simulation results for artificial data when underlying data are normally distributed and the rank of the underlying latent variables corresponding to each submatrix of real, Boolean, categorical, and ordinal valued columns (of dimension 250×20) is set to $k = 10$. Root mean squared error (RMSE), interval length, bias, and coverage for estimated regression coefficients from multiple imputation using the different imputation techniques is plotted for increasing proportion of missing values.

to real-valued data. For the remaining data types too, the average RMSE values are typically higher for the matrix factorization methods, albeit the difference is less pronounced for higher levels of missingness. Empirical coverage levels are above 95 percent for all methods. These results also hold broadly for the larger sample size value of $n = 1,000$. This indicates that even with smaller sample sizes, traditional imputation techniques perform very well when the underlying data are normally distributed and of intermediate rank.

In Figure 3, the underlying data are uniformly distributed and of low-rank, $k = 2$, for each block of $p/4 = 20$ variables corresponding to a given data type. In this case, however, when the underlying data-generating process is low rank and nonnormal, we find that matrix factorization methods produce estimates that have lower RMSE values and competitive coverage rates corresponding to all non-real-valued data types. For Boolean columns in particular, the RMSE values from Trace and Rank are substantially lower, and coverage values substantially higher, than estimates from Amelia and MICE, with the gap widening as the proportion of missing data increases. When sample size is lower ($n = 250$), both Amelia and MICE fail to produce imputations for a number of the simulated data sets, whereas Trace and Rank produce imputations in all cases.

For the second set of plots, we consider two intermediate cases when sample size is large, $n = 1,000$. In Figure 4, the underlying data are normally distributed and have low-rank, $k = 2$. In this case, we find that coefficient estimates corresponding to Boolean valued columns obtained from Trace and Rank have lower RMSE values and substantially higher empirical coverage values, with the gap widening with a greater proportion of missing data. However, for all other data types, despite the low underlying rank, we find that Amelia and MICE produce coefficient estimates that have higher coverage and lower or comparable RMSE values. In Figure 5, the underlying data are uniformly distributed and have intermediate rank, $k = 10$. In this case too, despite the underlying data being very different, we note similar trends for Boolean and real-valued data types. Again, both Amelia and MICE fail to produce imputations for a number of the simulated data sets when the proportion of missing data is 40 percent or more.

In summary, these simulations suggest that matrix factorization techniques may be particularly useful for non-real-valued data when the interest is in minimizing RMSE or prediction errors, especially for higher levels of sparsity. For Boolean data generated from nonnormal underlying data in particular, matrix factorization techniques exhibit systematically better empirical coverage properties and lower bias, along with lower RMSE values, than do Amelia or MICE.

SIMULATIONS ON REAL-WORLD DATA

Simulations with artificial data are useful in providing intuition about the properties of imputation techniques, but their use in real-world scenarios may be somewhat limited, given that the underlying data-generating structure may be far more complex. In this section, we turn our attention to simulations using real-world survey data. The design

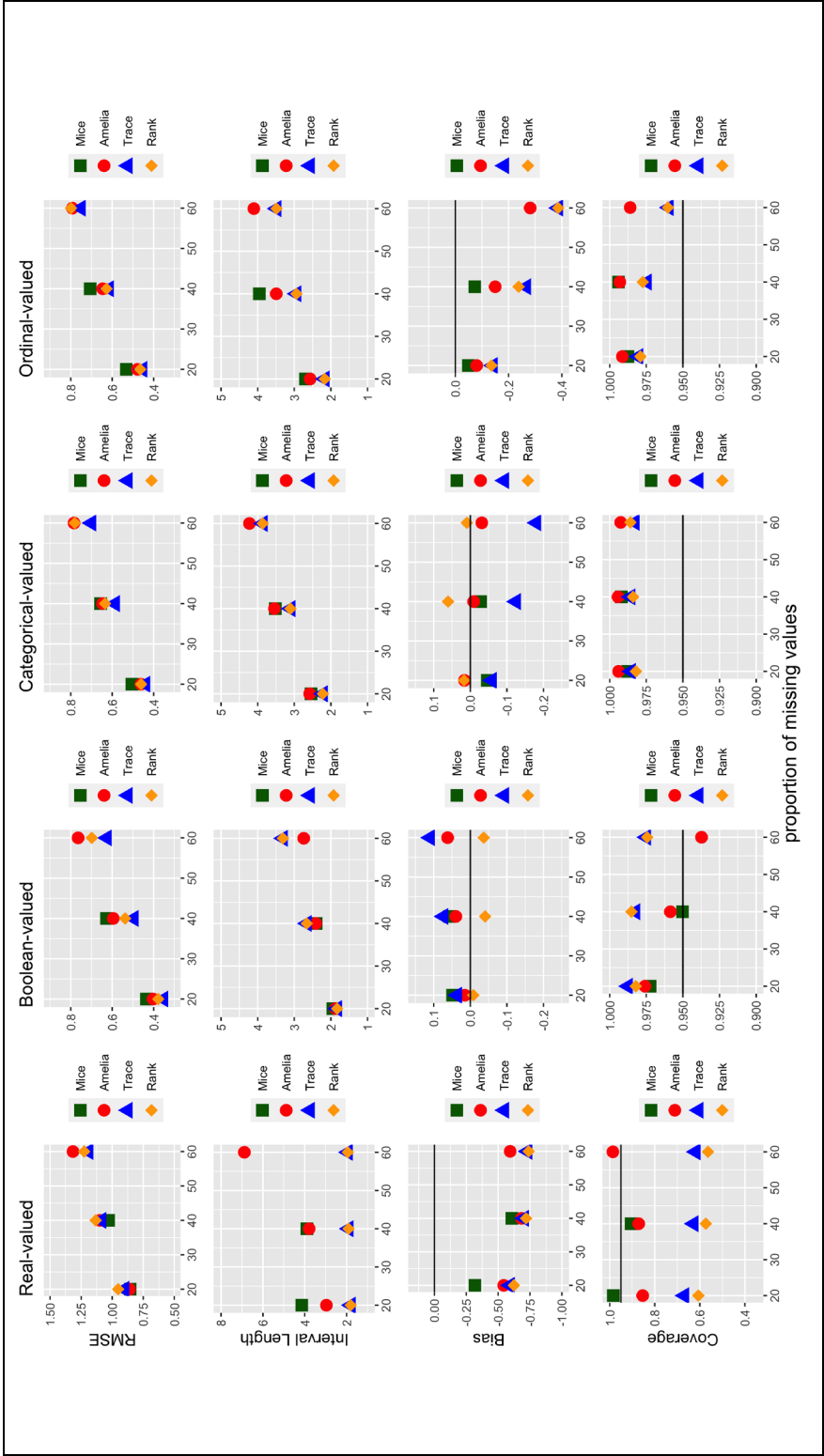


Figure 3. Simulation results on artificial data: uniform distribution and low rank.
Note: Simulation results for artificial data when underlying data are uniformly distributed and the rank of the underlying latent variables corresponding to each submatrix of real, Boolean, categorical, and ordinal valued columns (of dimension 250×20) is set to $k = 2$. Root mean squared error (RMSE), interval length, bias, and coverage for estimated regression coefficients from multiple imputation using the different imputation techniques is plotted for increasing proportion of missing values.

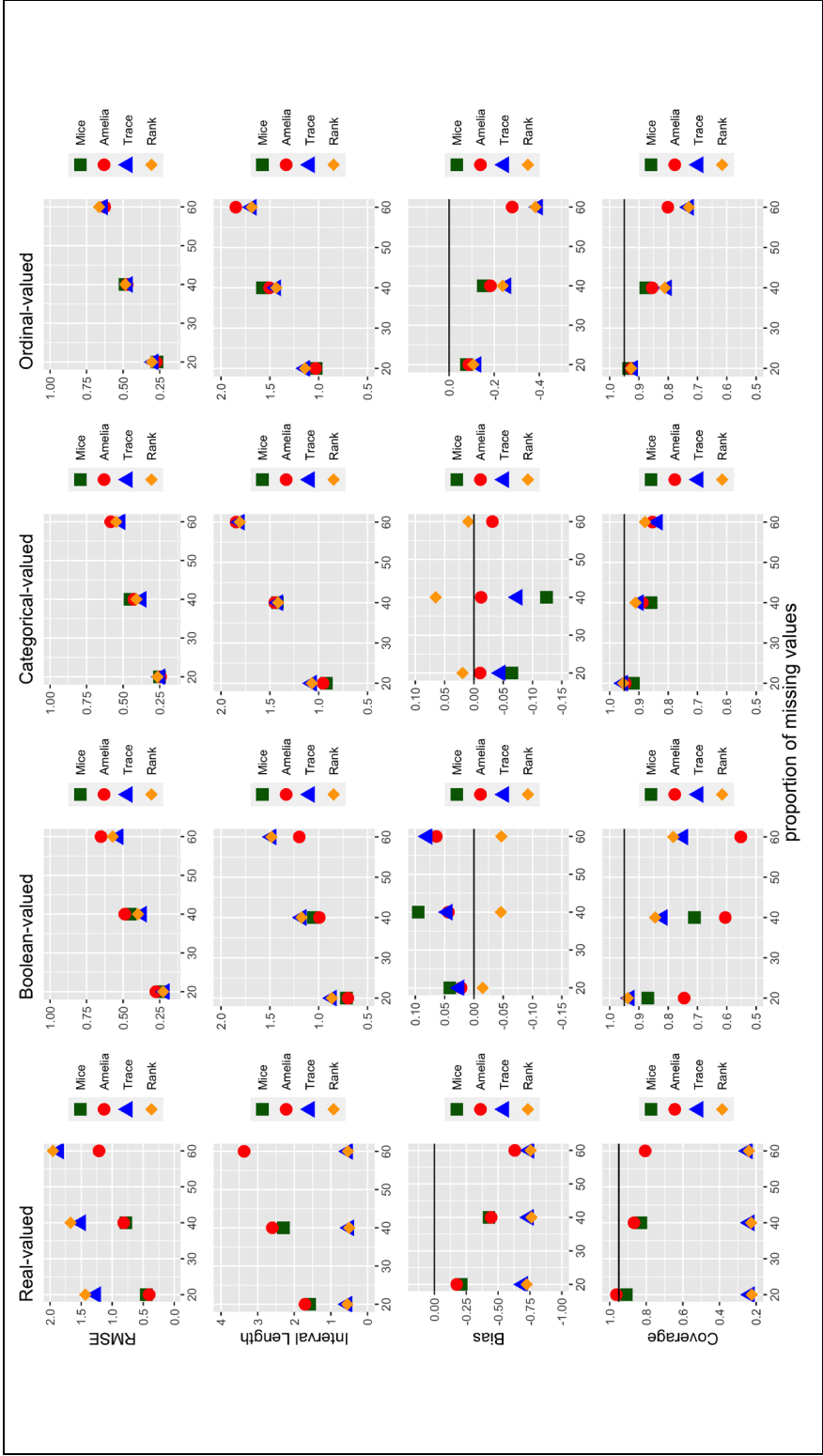


Figure 4. Simulation results on artificial data: normal distribution and low rank.
Note: Simulation results for artificial data when underlying data are normally distributed and the rank of the underlying latent variables corresponding to each submatrix of real, Boolean, categorical, and ordinal valued columns (of dimension $1,000 \times 20$) is set to $k=2$. Root mean squared error (RMSE), interval length, bias, and coverage for estimated regression coefficients from multiple imputation using the different imputation techniques is plotted for increasing proportion of missing values.

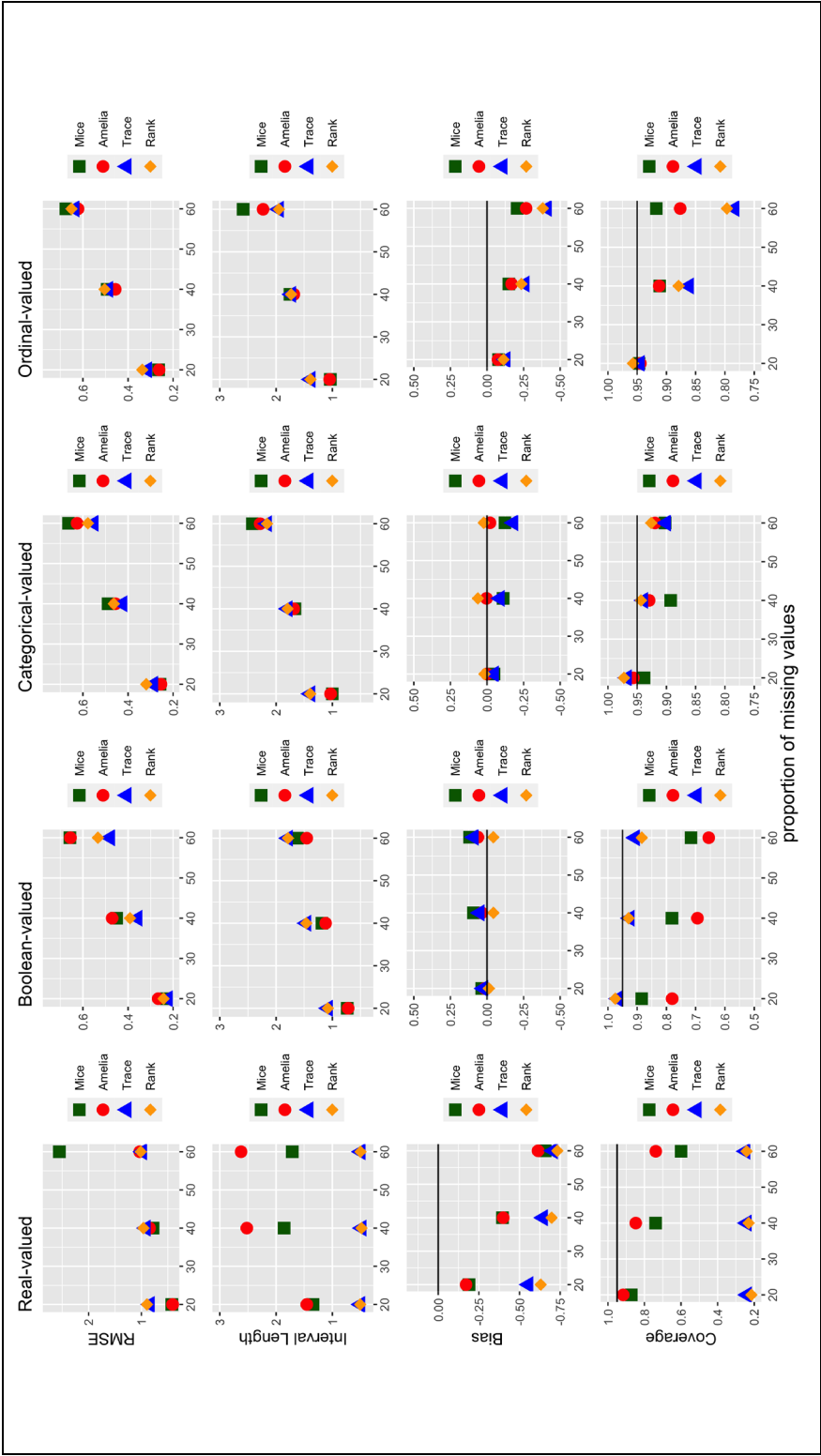


Figure 5. Simulation results on artificial data: uniform distribution and intermediate rank.

Note: Simulation results for artificial data when underlying data are uniformly distributed and the rank of the underlying latent variables corresponding to each submatrix of real, Boolean, categorical, and ordinal valued columns (of dimension $1,000 \times 20$) is set to $k = 10$. Root mean squared error (RMSE), interval length, bias, and coverage for estimated regression coefficients from multiple imputation using the different imputation techniques is plotted for increasing proportion of missing values.

of the evaluation closely mirrors the previous section. The main difference is that rather than generating new data with specific underlying characteristics, we use fully observed⁵ subsets of data from the General Social Survey (GSS) (Smith et al. 2019) and the National Longitudinal Survey of Youth (NLSY) 1997 (U.S. Department of Labor, Bureau of Labor Statistics 2015), following the design suggested by Z. Wang et al. (2021).⁶

We use a subsample of the GSS 2014 data set. The GSS is a public opinion survey that collects data on attitudes, opinions, and demographic characteristics for adults from randomly selected U.S. households. The data set used in the analysis comprises 399 rows and 83 columns, of which we use three columns corresponding to respondent's age, gender, and race to generate MAR missingness. Of the remaining 80 columns, 7 are real-valued, 16 Boolean, 21 categorical, and 36 ordinal.

For the NLSY 1997 cohort data set, we use a subset from 2011. The NLSY is a longitudinal social survey that follows the lives of a sample of U.S. youth born between 1980 and 1984. The data set used in the analysis comprises 666 rows and 43 columns, of which we use three columns corresponding to respondent's age, gender, and race to generate MAR missingness. Of the remaining 40 columns, 8 are real-valued, 5 Boolean, 9 categorical, and 18 ordinal.

The final set of simulations consists of six cases corresponding to (1) the two fully observed subsamples of GSS and NLSY data and (2) three different levels of MAR missingness $\rho_{miss} = 20$ percent, $\rho_{miss} = 40$ percent, and $\rho_{miss} = 60$ percent. Starting with the full data set X , we create missingness via an MAR mechanism where we set columns corresponding to respondent age, gender, and race as fixed reference columns (the detailed MAR mechanism is described in the Appendix) and denote the resulting matrix as X_{obs} . We generate $n_{sim} = 500$ such data sets for each of the six cases considered and hold these fixed throughout the simulation exercise. For each of the 500 simulated data sets, we obtain $m = 5$ imputed data sets from the four different imputation techniques corresponding to Amelia, MICE, Trace, and Rank. Thus, $500 \times 20 = 10,000$ imputations are obtained for each of the six cases.

The parameters of interest $\beta_1, \dots, \beta_{p-3}$ (where the three columns used to generate MAR missingness are not included) are summary statistics corresponding to the fully observed data. For a given real or ordinal valued column, β_j equals the arithmetic mean of the column on the fully observed data (i.e., $\beta_j = \frac{\sum_i x_{i,j}}{N}$). For a given Boolean and categorical valued column, β_j equals the proportion of times the most frequent category of response, $X_{maxfreq_j}$, is recorded in the fully observed data (i.e., $\beta_j = \frac{\sum_i \mathcal{I}\{x_{i,j} = X_{maxfreq_j}\}}{N}$). The estimated parameter vector $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{p-3})$ is obtained from the imputed matrix via multiple imputation from the four candidate imputation techniques. As before, the resulting estimates $\hat{\beta}_{ij}$ corresponding to simulation $i \in (1, \dots, n_{sim})$ and column $j \in (1, \dots, p)$ from each of the imputation techniques are compared via the following metrics averaged over all the real, Boolean, categorical, and ordinal valued columns⁷:

1. RMSE: $\sqrt{\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\beta}_{ij} - \beta_j)^2}$.
2. Mean interval length: size of the 95 percent confidence interval associated with $\hat{\beta}_{ij}$, averaged over n_{sim} simulations.
3. Bias: $\left(\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\beta}_{ij} \right) - \beta_j$.
4. Empirical coverage: proportion of times the true parameter value β_j lies in the 95 percent confidence interval associated with $\hat{\beta}_{ij}$.

Results

Comparison of the estimates obtained from different imputation techniques for the six cases provides us with the following broad insights. First, unexpectedly, for real-valued columns, Trace performs particularly well in terms of lower RMSE and higher empirical coverage, especially when compared with MICE for all levels of missingness. A possible explanation for this deviation from the results with artificial data may be that the underlying distribution generating the real-valued data is nonnormal or low rank.

Second, for coefficients associated with Boolean and categorical valued columns, we find that matrix factorization methods have substantially lower RMSE values and higher empirical coverage metrics than do Amelia and MICE. Although this mirrors the results from the simulations using artificial data, the gap in performance is much more pronounced for the real-world subsets. The higher coverage rates for estimates using matrix factorization methods are driven here mainly by their substantially lower biases.

Third, for coefficients associated with ordinal valued columns, the results are mixed, with different techniques leading to lower RMSE values in different scenarios. In terms of coverage, all methods have similar performance. These results are in line with the simulations using artificial data.

Finally, traditional imputation techniques, particularly MICE, are sensitive to the level of missingness, often failing to produce imputations when the proportion of missing values is 40 percent or higher. This pattern is broadly consistent with the simulations using artificial data. Detailed numbers are presented in Appendix Table D2.

Next, we present two sets of plots corresponding to the GSS and NLSY subsamples that demonstrate these insights in greater detail. See the online version for a color figure. Average RMSE, interval length, bias, and coverage are presented for each variable type corresponding to different proportions of missingness and the different imputation techniques. Averages are not displayed for cases in which the imputation method crashed repeatedly and failed to produce at least 30 imputations.

Figure 6 presents detailed results for all cases corresponding to the GSS subsample. Imputations from Trace produce consistently low RMSE values and high coverage rates compared with the other three techniques across all data types and for all levels of missingness. Imputation results from Rank are similar to those from Trace for Boolean and categorical data types, but mixed for real-valued columns, where Rank

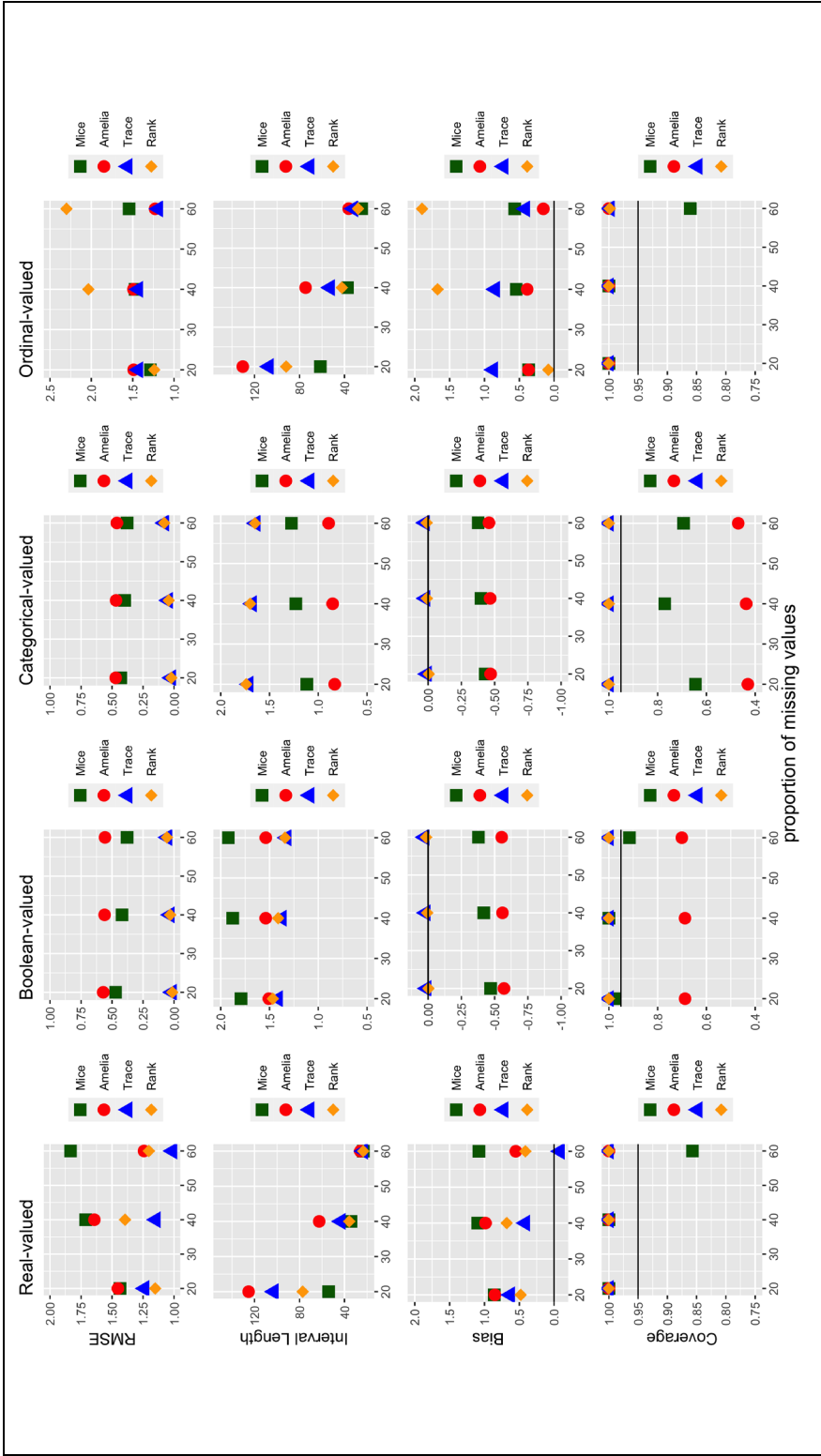


Figure 6. Simulation results for General Social Survey subsample.

Note: Root mean squared error (RMSE), interval length, bias, and coverage for estimated regression coefficients from multiple imputation using the different imputation techniques is plotted for increasing proportion of missing values.

produces lower RMSE values at low levels of missingness, and for ordinal valued columns, where Rank performance is substantially worse than all other techniques at higher levels of missingness. Imputations from Amelia are competitive with the matrix factorization techniques in terms of empirical coverage rates for both ordinal and real-valued columns. Imputations from MICE are competitive with the matrix factorization techniques in terms of empirical coverage rates for both ordinal and real-valued columns at lower levels of missingness. However, for Boolean and categorical valued data types, both traditional imputation techniques lead to substantially higher RMSE values and substantially lower coverage rates at all levels of missingness. Amelia and MICE also fail to produce imputations for a number of the simulated data sets when the proportion of missing data is 40 percent or more, whereas Trace and Rank produce imputations in all cases.

Figure 7 presents detailed results for all cases corresponding to the NLSY subsample. Here too, we find that imputations from Trace produce consistently high coverage rates compared with the other three techniques across all data types and for all levels of missingness. In terms of RMSE values, Trace exhibits better performance, except for ordinal valued columns. Imputation results from Rank are almost identical to those from Trace for Boolean and categorical data types, but substantially poorer in terms of RMSE for both real and ordinal valued columns, where Rank produces consistently higher RMSE values at all levels of missingness (coverage rates, however, are the same). Imputations from Amelia are competitive with Trace in terms of empirical coverage rates and RMSE values for ordinal and real-valued columns, but consistently worse for Boolean and categorical valued columns. Imputations from MICE are competitive with the matrix factorization techniques in terms of empirical coverage rates for ordinal and real-valued columns. MICE produces higher RMSE values than Trace and Amelia for real-valued columns, but consistently lower RMSE compared with all methods for ordinal columns. For Boolean and categorical valued columns, however, MICE is associated with substantially higher RMSE values and substantially lower coverage rates at almost all levels of missingness. Again, MICE fails to produce imputations for a number of the simulated data sets across all proportions of missing data, whereas Amelia, Trace, and Rank produce imputations in almost all cases.

In summary, these simulations provide further evidence that matrix factorization can be a powerful tool for the imputation of missing data in social science surveys. The methods perform particularly well for Boolean and categorical valued data, which comprise a large proportion of typical social science surveys, particularly those that elicit respondents' subjective opinions and self-reported characteristics. In contrast to simulations from artificial data, we find that for real and ordinal valued data, matrix factorization techniques perform very well. Furthermore, the methods are more robust to increasing data sparsity. The main reason for the improved performance in sparse data sets is the regularization penalty in equation (4) that explicitly penalizes overfitting. We recommend the use of matrix factorization using Trace norm regularization across a variety of settings, but specifically for data sets with high sparsity or with predominantly categorical or Boolean valued columns.⁸

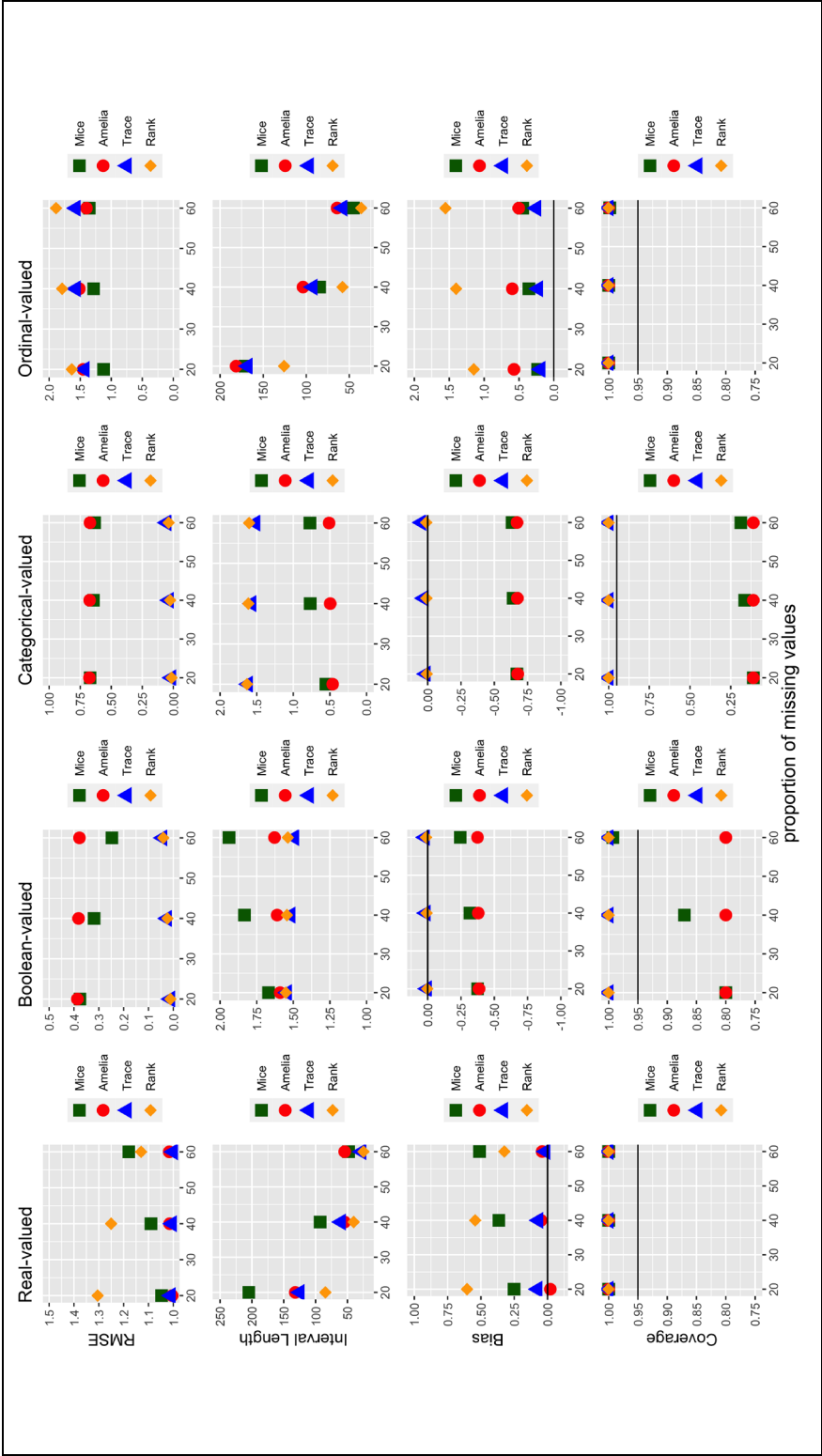


Figure 7. Simulation results for National Longitudinal Survey of Youth subsample.
Note: RMSE, interval length, bias, and coverage for estimated regression coefficients from multiple imputation using the different imputation techniques is plotted for increasing proportion of missing values.

DISCUSSION

In this article we introduced into the empirical social science literature the use of matrix factorization techniques for missing values imputation. Such techniques are now ubiquitous in computer-science applications, such as recommender systems, and can be profitably used in social science applications. Many popular neural networks methods (e.g., unsupervised autoencoders) can also be understood as solutions to very large matrix factorization problems through heuristic optimization (Levy and Goldberg 2014).

We provided an overview of the matrix factorization literature, including its relationship with linear regression, its probabilistic and geometric interpretation, some common parametric and nonparametric loss functions corresponding to distinct data types, as well as regularization penalties readily implementable within this framework. In using matrix factorization models, we used the notion of *cross-validation*, wherein optimal parameters are selected on the basis of their predictive performance on a held out sample. We first presented results from a set of simulations on artificial data comparing the performance of matrix factorization methods with standard social science techniques, and we found that matrix factorization techniques tend to have lower RMSE values for regression estimates on non-real-valued columns, and in the case of Boolean data higher coverage when the underlying data-generating process is low rank and uniformly distributed. We also present results from a set of simulations on real-world data where we find these methods lead to lower RMSE values and higher or comparable coverage rates for parameter estimates corresponding to all data types, particularly as sparsity increases. Importantly, we find that matrix factorization methods are particularly robust to increasing levels of sparsity, in contrast to traditional imputation techniques that fail to produce imputations in a number of cases. On the other hand, our simulation results suggest that traditional imputation techniques are a good choice when underlying data follow an approximately normal distribution, especially when the ratio of rows to columns is larger and for lower levels of sparsity. Finally, we point the reader to the freely available software LowRankModels, written in Julia and Python, for implementing matrix factorization models on social science data.

Why do multiple imputations using matrix factorization approaches improve over standard approaches in a variety of settings, particularly for simulations using real-world data? First, matrix factorization explicitly models multiple dimensions of variation in data. By simultaneously modeling similarity between respondents and between questions in a survey, factorization approaches can exploit more real variation relevant to sparse data reconstruction. Second, claims that standard approaches produce optimal estimates are based on strong parametric assumptions, which we argue may not hold in realistic scenarios. For example, columns are assumed to follow a multivariate normal distribution. This assumption fails to hold in real-world surveys where a large majority of columns are Boolean or ordinal (e.g., Likert-type scales). Finally, machine learning methods like matrix factorization are built on the principle of minimizing bias *and* variance, which produce estimates with better predictive properties and robustness to sparsity. This is accomplished with the addition of a penalty term in the loss function

to discourage overfitting, which acts as a heuristic method of model-selection. For example, in multiple imputation, researchers typically pare down their data to include only variables they assume will be related to a missing variable, ignoring others with limited justification. Matrix factorization methods do not drop any variables explicitly, but cross-validation and the regularization penalty do the job of model selection explicitly and automatically. Given all of these factors, it is unsurprising that matrix factorization methods do a better job of predicting missing data in large and sparse data sets.

This article adds to the growing list of techniques for missing data imputations using machine learning and related methods, including Gaussian copulas (Hollenbach et al. 2018; Zhao and Udell 2020), random forests (Xia et al. 2017), and deep neural networks (Che et al. 2018; Choudhury and Pal 2019; Z. Wang et al. 2021). Although an empirical comparison across each of these emerging methods with imputations produced via matrix factorization is beyond the purview of this article, there are theoretical arguments for why matrix factorization may be more attractive for use by social scientists. The Gaussian copula approach is particularly interesting, and it can work at large scale using ideas similar to the low-rank approach presented here, but it may be extremely sensitive to the MCAR assumption, because estimates of the marginal distributions are drawn directly from the empirical distribution. Deep learning methods can produce excellent imputations, but they lack theoretical foundation and tend to be very slow and expensive to fit. These models require resources beyond the means of many social science users of imputation software. They also lack theoretical guarantees, and their complexity may hide hidden bias. Finally, random forest imputation methods are an excellent general purpose framework for imputation and work especially well for categorical data. However, the performance of the random forest method *MissForest* (and its analog in the *MICE* package) deteriorates for large-scale data, especially as the number of columns increases, because the number of models scales with the number of columns. Zhao and Udell (2020) demonstrated cases in which *LowRankModels* and a copula EM method outperform forest-based methods on social science survey data. We do not claim that the techniques discussed here are superior to all other methods under all circumstances, but we do make the case that matrix factorization (and related) methods maybe particularly attractive, especially in contexts where data are sparse and responses are more qualitative in nature. Such contexts are becoming increasingly common in social science research and are expected to grow with the increase of non-standard data collection, such as Web scraping, crowdsourcing, and online surveys conducted at scale on social media and other online platforms.

Our findings suggest a number of future directions for research in this area. A natural next step is the extension of these techniques to longitudinal archival or panel data. The related field of tensor factorization is a vibrant area of research in computer science (see, e.g., Acar et al. 2011; Kolda and Bader 2009) and is a natural candidate for missing value imputation in longitudinal analysis of archival data like government documents, censuses, or panel surveys. In this setting, individual respondents, their responses, and time represent three critical dimensions of the tensor whose factorization would reconstruct missing entries. Moreover, although we introduced a number of matrix norms in this article, we used only the trace norm in our empirical evaluation.

Ongoing research in computer science and discrete math on other norms and their implementation promises to lead to further improvements in the quality of imputation.

Blackwell et al. (2015) introduced multiple overimputation as a means to check and correct for measurement errors in fully observed social science data. This is closely related to the use of matrix factorization techniques in signal processing as a “denoising” technique to extract original information from corrupted data. This connection opens up the possibility of applying matrix and tensor factorization techniques to deal with measurement error issues in the social sciences.

Finally, adaptive survey design is an important new area of research that uses results from matrix factorization models by dynamically targeting a subset of survey questions to individual respondents (see Blumenstock, Cadamuro, and On 2015; Salganik and Heckathorn 2004; Salganik and Levy 2015). This technique is actively used in recommendation systems to make accurate predictions while minimizing the number of questions asked. We believe this is a promising area of future investigation for sociological research design and questionnaire construction, enabling field researchers to significantly reduce respondent and surveyor fatigue as well as respondent drop-out rates. Coming full circle, this would lessen the demand for missing value imputation.

We believe matrix and tensor factorization approaches have the potential for wide application and adoption in the social sciences. We hope that our easy-to-use software LowRankModels will help realize these hopes. Here, we demonstrate the promise of matrix factorization for missing value imputation and reveal where and how it improves on existing approaches while fitting within the useful “multiple imputation” framework. This creates new opportunities for the reconstruction and systematic analysis of social data containing many missing values, as from historical archives or lengthy surveys with high drop-out rates. The impressive performance of matrix factorization approaches across our evaluations recommend their use for robust missing value imputation, and suggest their potential to make a difference for social analysis, enabling inferences in sparse data that are impossible without them.

APPENDIX

Relationship to SVD

In the main text, we discuss how matrix factorization models are related to linear regression. Here we elaborate on their relationship to another popular statistical technique of SVD. The reduced SVD of a matrix $Y \in \mathbb{R}^{n \times p}$ with true rank k can be written as

$$Y = \tilde{U} \Sigma \tilde{V}' \quad (7)$$

$$= \sum_{r=1}^k \sigma_r \tilde{u}_r \tilde{v}_r', \quad (8)$$

where $\tilde{U} \in \mathbb{R}^{n \times k}$, $\tilde{V}' \in \mathbb{R}^{k \times p}$, $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_k) \in \mathbb{R}^{k \times k}$. The columns of $\tilde{U}$ and $\tilde{V}$, $\tilde{u}_r \in \mathbb{R}^{n \times 1}$ and $\tilde{v}_r \in \mathbb{R}^{1 \times p}$, respectively, represent an orthonormal basis. Note that setting

$$U = \tilde{U} \Sigma^{1/2} \quad \text{and} \quad V = \Sigma^{1/2} \tilde{V}'$$

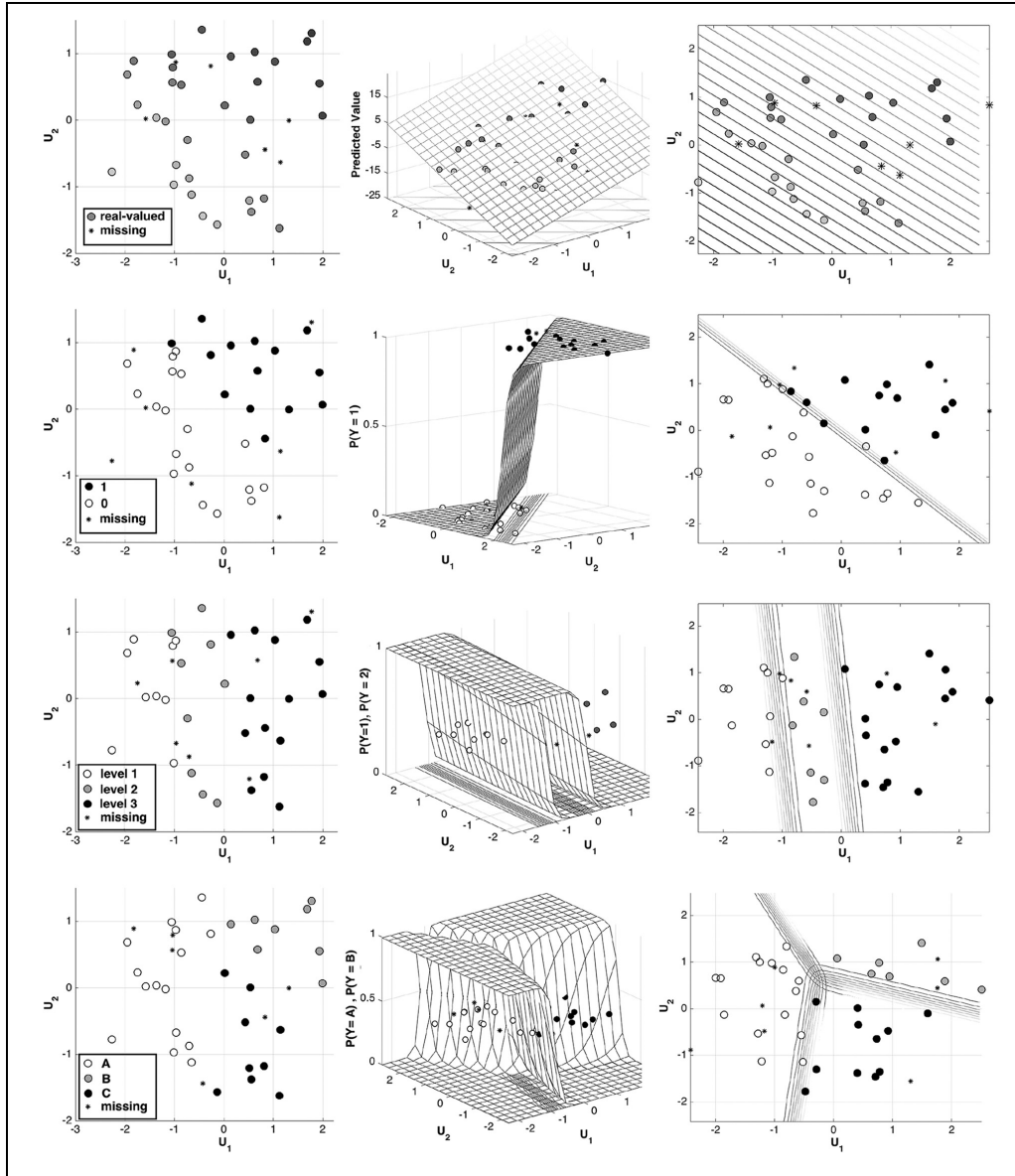


Figure A1. Distinct data types and corresponding loss functions in the latent factor space.

Note: The horizontal slices of this figure correspond to columns Y of four different data types (real, Boolean, ordinal, categorical). The first vertical slice represents the distribution of individuals in the latent factor space $U \in \mathbb{R}^2$. Note that for a given factorization, the location of the points in the latent space for each individual is fixed across all horizontal slices. The labels of the points represent the value corresponding to a single column of the original data set Y ; missing values are represented by asterisks. The second vertical slice presents the loss functions (quadratic, logistic, ordered probit, multinomial logit) fitted to the observed values of the columns of Y . Contour plots in the third vertical slice visualize how the fitted loss functions can be used to predict the missing Y values corresponding to asterisks.

is a solution to the low-rank approximation problem in equation (3). This formulation will be useful in the next section, “Statistical Guarantees,” for which we summarize the most commonly used assumptions from the literature.

Statistical Guarantees

Matrix factorization techniques can often guarantee the accuracy of imputed values. In particular, statistical guarantees showing the exact or near-exact recovery of matrices have received considerable attention. In this section we briefly discuss the assumptions required for these guarantees to hold. We also direct interested readers to recent theoretical work that relaxes some of these assumptions.

A typical result is that one can recover the original matrix accurately with high probability if the number of observed entries is greater than some constant. In case the samples are noisy, then the error in the recovered matrix is proportional to the noise level.

For instance, Candès and Recht (2009) proved that given uniform sampling, incoherence, and nonspikiness, a matrix $Y \in \mathbb{R}^{n \times n}$ of low-rank k may be perfectly reconstructed with high probability via a matrix factorization program, as long as the observed sample size s is larger than a positive quantity:

$$s \geq Cn^{1.2}k \log n,$$

where C is some positive numerical constant. In practice, this result means one may be able to recover a complete matrix even if *very few* of its entries are observed, because in many practical cases the required sample size for exact reconstruction is much smaller than the total number of entries in the original matrix, that is, $s \ll n^2$. For more details and results on exact recovery, see Candès and Plan (2009) and Candès and Recht (2009).

Assumptions

Assumption of uniform sampling. Some assumption on the sampling pattern of observed matrix entries Y_{obs} is required to make guarantees on the accuracy of imputed values. The most common assumption is that the sampling pattern is uniformly random. The assumption is unlikely to hold in a number of realistic settings. For instance, in a survey we may expect many individuals to systematically avoid certain sensitive questions.

Much of the statistical literature on matrix factorization relies on this assumption. There are exceptions, however; for example, Srebro and Salakhutdinov (2010) showed that matrix completion with trace norm regularization can work significantly less well when entries of the matrix are sampled nonuniformly; they proposed a weighted version of trace norm regularization to address this problem.

Assumption of inherent low-rank structure. Accuracy guarantees also require an assumption on the underlying structure of the complete matrix. One common assumption is that $Y \in \mathbb{R}^{n \times p}$ is approximately low-rank $k < \min\{n, p\}$. In other words, the $(n \times p)$ matrix can be embedded into a lower dimensional space. This assumption holds when only a few (latent) factors are sufficient to determine all entries in the matrix. Somewhat surprisingly, it is possible to show that this assumption approximately holds for all sufficiently large data sets drawn from a natural class of distributions (Udell and Townsend 2017).

The maximum likelihood estimation problem corresponding to this assumption is nonconvex; without making further assumptions, it is not possible to guarantee that an efficient optimization method for this problem finds the global optimum. As a result, much of the statistical literature focuses on the convex relaxation of this problem, which is to minimize the trace norm, and which generally recovers a low-rank solution with low norm. Some authors argue that low-norm methods produce more accurate results as well. For more details and results, see Negahban and Wainwright (2011).

Assumption of “nonspikiness.” The next two assumptions of “nonspikiness” and “incoherence” are conceptually more subtle. Recall that the aim in low-rank matrix factorization is to find a k -dimensional subspace that approximates the $(n \times p)$ matrix well. For expositional purposes, say the matrix represents the answers of n individuals to p survey questions. The “nonspikiness” assumption implies that there is some information overlap among all individuals and among all questions, so that when some entries are missing, there is enough information in the observed entries to be able to predict the missing values reasonably well. Mathematically, consider the SVD of the complete matrix Y from, for example, equation (7):

$$Y = \sum_{r=1}^k \sigma_r \tilde{u}_r \tilde{v}_r.$$

The nonspikiness assumption requires that the maximum entry in each of the basis vectors $\tilde{u}_r \in \mathbb{R}^{n \times 1}$ and $\tilde{v}_r \in \mathbb{R}^{1 \times p}$ is bounded above. Because each $\tilde{u}_r$ is a unit vector, this implies that the entries of $\tilde{u}_r$ are not very large for any one of the n coordinates but are spread across all n coordinates. Intuitively, this means each of the basis vectors in U contains information on most individuals. Similarly, each of the basis vectors in V should contain information on most questions. In terms of the matrix of interest, Y , this means there is some information overlap among all individuals and among all questions. The nonspikiness assumption also implies that all entries in the matrix of interest Y are bounded above.

Assumption of incoherence. The incoherence condition common in the literature is closely related to the notion of nonspikiness. The condition states that no single individual (or no single question) is so different from the remaining individuals (or questions) that one of the orthonormal basis vectors $\tilde{u}_r$ (or $\tilde{v}_r$) consists solely of information on that individual (or question).

Mathematically, we explain the condition with respect to n individuals corresponding to the latent factors U . The same logic and condition applies for the p questions corresponding to the latent factors V . There exists a standard basis vector, $e_i \in \mathbb{R}^n, i = (1, \dots, n)$, corresponding to each row of the matrix. The incoherence condition requires that the subspace spanned by the latent features U is not parallel to any of these standard basis vectors, that is, the following quantity is bounded above:

$$\max_{1 \leq i \leq k} \|P_U e_i\|^2,$$

where P_U represents the orthogonal projection onto the subspace spanned by U . Intuitively, this condition does not admit a latent factor whose role is purely to explain the behavior of a specific individual.

An illustration of the issue is presented next, borrowing from Candès and Recht (2009). Suppose a matrix $Y \in \mathbb{R}^{n \times p}$ is of low-rank $k=1$ where $U \in \mathbb{R}^{n \times 1} = e_1$ and $V \in \mathbb{R}^{1 \times p} = [x_1, x_2, x_3, \dots, x_p]$, then

$$Y = UV = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} (x_1 \quad x_2 \quad x_3 \quad \dots \quad x_p) = \begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_p \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

In this example, we see that the first row corresponding to individual 1 is completely different from all other rows of the matrix. Therefore, observing entries from row 2 to row n is not at all informative for predicting entries in the first row. This is the kind of situation the incoherence assumption rules out.

Recent theoretical work on statistical results has relaxed the assumption of incoherence, required by the exact recovery results. This strand of work shows that *approximate recovery* of the matrix of interest is possible without the strong incoherence assumptions. Such approximate recovery results are relative to some measure of the *scale* of the matrix Y , which may be measured as $scale(M) = \|Y\|_\infty^2$ or $scale(M) = \frac{1}{np} \|Y\|_F^2$. For more details and results on approximate recovery, see Foygel and Srebro (2011).

Simulations: MAR Mechanisms

Artificial Data.

1. We start with a full data set $X = [X_{real}, X_{bool}, X_{cat}, X_{ord}]$ where dimension X is $n \times 80$ and each of the submatrices dimensions is $n \times 20$.
2. We randomly set 10 percent of the columns as complete columns, implying these eight columns (two columns each from $X_{real}, X_{bool}, X_{cat}, X_{ord}$) contain no missing values. We denote these columns as $X_{comp,1}, \dots, X_{comp,8}$ and the remaining 72 columns as $X_{miss,1}, \dots, X_{miss,72}$.
3. We then loop over these 72 columns to generate MAR missingness patterns. Consider a generic column j :
 - We first set eight random coefficients $\gamma_1, \dots, \gamma_8$ to create a regression function $L = \gamma_1 X_{comp,1} + \dots + \gamma_8 X_{comp,8} + \epsilon$, where ϵ is a zero-mean error term.
 - We set a threshold T such that entries in column $X_{miss,j}$ where the regression function value falls below the threshold ($L < T$) are assigned NA values.
 - The threshold is adjusted on the basis of the proportion of missing values required overall.
4. The resulting matrix is denoted as X_{obs} and saved.
5. We save $n_{sim} = 500$ versions of the X_{obs} matrix for each of the 24 cases we consider.

Real-World Data.

1. We start with the full data set $X = [X_{real}, X_{bool}, X_{cat}, X_{ord}]$ where dimension X is $n \times p$.
2. We set columns corresponding to respondent age, gender, and race as fixed reference columns and denote these as $X_{age}, X_{gender}, X_{race}$ columns. The remaining columns are denoted $X_{miss,1}, \dots, X_{miss,p-3}$.
3. We then loop over the X_{miss} columns. Consider a generic column j :

- We first set three random coefficients $\gamma_1, \gamma_2, \gamma_3$ to create a regression function $L = \gamma_1 X_{age} + \gamma_2 X_{gender} + \gamma_3 X_{age} + \varepsilon$, where ε is a zero-mean error term.
 - We set a threshold T such that entries in column $X_{miss,j}$ where the regression function value falls below the threshold ($L < T$) are assigned NA values.
 - The threshold is adjusted based on the proportion of missing values required overall.
4. The resulting matrix is denoted as X_{obs} and saved.
5. We save $n_{sim} = 500$ versions of the X_{obs} matrix for each of the six cases we consider.

Details of Imputation Numbers

Table D1. Number of Times Each Technique Produced a Set of Five Imputations (out of 500) for Each of 24 Cases Considered for Simulations Using Artificial Data

<i>n</i>	<i>k</i>	Distribution	Proportion Missing	Trace	Rank	MICE	Amelia
<i>n</i> = 250	<i>k</i> = 2	$\sim N(0, 1)$	20%	500	500	500	253
			40%	500	500	500	436
			60%	500	500	1	256
		$\sim U(-1, 1)$	20%	500	500	500	174
			40%	500	500	500	492
			60%	500	500	28	271
	<i>k</i> = 10	$\sim N(0, 1)$	20%	500	500	500	401
			40%	500	500	500	392
			60%	500	500	0	127
		$\sim U(-1, 1)$	20%	500	500	500	110
			40%	500	500	500	491
			60%	500	500	19	243
<i>n</i> = 1000	<i>k</i> = 2	$\sim N(0, 1)$	20%	500	500	500	500
			40%	500	500	500	223
			60%	500	500	8	178
		$\sim U(-1, 1)$	20%	500	500	500	500
			40%	500	500	500	221
			60%	500	500	500	181
	<i>k</i> = 10	$\sim N(0, 1)$	20%	500	500	500	500
			40%	500	500	500	212
			60%	500	500	3	139
		$\sim U(-1, 1)$	20%	500	500	500	500
			40%	500	500	500	224
			60%	500	500	500	187

Table D2. Number of Times Each Technique Produced a Set of Five Imputations (out of 500) for Each of Six Cases Considered for Simulations Using Real-World Data

Dataset	(<i>n</i> , <i>p</i>)	Proportion Missing	Trace	Rank	MICE	Amelia
GSS subsample	<i>n</i> = 399, <i>p</i> = 83	20%	500	500	454	491
		40%	500	500	329	201
		60%	500	500	85	215
NLSY subsample	<i>n</i> = 666, <i>p</i> = 43	20%	500	500	363	480
		40%	500	500	260	496
		60%	500	500	84	497

Note: GSS = General Social Survey; NLSY = National Longitudinal Survey of Youth.

Acknowledgments

We thank participants at seminars at the University of Chicago and the Wisconsin Institute of Discovery for helpful comments.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We acknowledge support from the Air Force Office of Scientific Research (FA9550-19-1-0354), the Defense Advanced Research Projects Agency (DARPA) (HR00111820006), titled “Social MIND: Social Machine Intelligence for Novel Discovery” in the program “Ground Truth” (2,000,000). 2017–2020 Principal Investigator (with Gary Lupyan) on Sloan Foundation Grant “The Impact of Programming Languages and Data Science Frameworks on Thinking, Software, and Science” (750,000). 2016 Principal Investigator for “International Symposium on the Science of Science,” March 22–23, 2016; National Science Foundation #1623809 (35,000). 2015–2019 Co-Principal Investigator (with Adrian Johns and James Chandler) on Mellon Foundation Grant “Algorithms, Models and Formalisms” (797,000). 2015–2019 Co-Principal Investigator (with Alberto Cambrosio, PI) “Oncology’s Metaknowledge Networks: An Investigation of the Collective Dynamics of Bio-clinical Innovation”; Canadian Institutes of Health Research (640,000). 2015–2019 Co-Principal Investigator (with Boaz Keysar, principal investigator) “The Impact of Foreign Languages on Problem Solving: An Interdisciplinary Study of Language and Society”; Interdisciplinary Behavioral and Social Sciences Research grant from National Science Foundation (NSF) (900,000). 2015–2017 Principal Investigator (with Dashun Wang) “Discovering the Extent of Estimable Prediction (DEEP) in Science and Technology”; Air Force Office of Scientific Research (FA9550-15-1-0162) (767,811). 2014–2017 Co-Investigator (with Jason Owen-Smith, Julia Lane and Bruce Weinberg) “The Institute for Research on Innovation & Science (IRIS)” from the Sloan (1,200,000) and Kauffman (1,497,751) foundations. 2014–2017 Senior Personnel on NIH BEST (Broadening Experience in Scientific Training), “myCHOICE (Chicago Options-In-Careers Empowerment).” 2014–16 Principal Investigator (with Julia Lane and Jordan Boyd-Graber) on NSF National Center for Science and Engineering Statistics grant #1422902, “Collaborative Research: Scaling Insight into Science: Assessing the Value and Effectiveness of Machine Assisted Classification within a Statistical System” (500,000). 2014–2017 Co-Investigator on DARPA grant #14145043, “Big Machine Science” in the Big Mechanism program (4,500,000). 2013–2016 Principal Investigator on John Templeton Foundation “Metaknowledge Network” (5,300,000). Financial support for this project was received from the NSF SciSIP Science of Science and Innovation Policy Grant #1158803, “Tracing Influence & Predicting Impact in Science.”

ORCID iDs

Nandana Sengupta  <https://orcid.org/0000-0002-3874-7308>

James A. Evans  <https://orcid.org/0000-0001-9838-0707>

Supplemental Material

Supplemental material for this article is available online.

Notes

1. A limitation of the current discussion is that not all missing data are equivalent, and data are often missing for a reason (e.g., “won’t say,” “refuse to answer”), complications we gloss here and supplement as deviations from the missing completely at random assumption, either MAR or missing not at random. Incorporation of complex dependencies underlying missingness is a challenge beyond the scope of this article, although we demonstrate performance robust to some of these, as shown in later sections and in the online supplement. Within the context of our framework, columns with a mix of “won’t say” or “refuse to answer” and other data types may be treated as mixed Boolean, categorical,

- and/or cardinal data. Such complex data types may be treated as predictable on the basis of other covariates (e.g., treated with a Heckman correction [Heckman 1974, 1976; see also Christopher and Mare 1992]) and incorporated into a more fulsome estimate.
2. Software available at <https://github.com/madeleineudell/LowRankModels.jl/tree/sample>. An R interface for basic LowRankModels functionalities is available at: <https://docs.h2o.ai/h2o-tutorials/latest-stable/tutorials/glm/glm-tutorial.html>. We are working on providing an R wrapper for the multiple imputation functionality.
 3. Note that the output $\hat{Y}$ is also the maximum a posteriori estimator of Y .
 4. This methodology defines a valid probability distribution. However, because the normalizing constant $\sum_{y'} \exp(L(y', z))$ is not constant in z , it is *not* true that, fixing y , minimizing $L(y, z)$ over z is the same as minimizing $-\log \mathbb{P}(y|z)$ over z .
 5. Subsets were generated on the basis of the following criteria: (1) removing columns with identifying information (e.g., interviewer characteristics, interview description, sampling details); (2) removing columns with nonvarying entries; (3) keeping only one column from pairs of very highly correlated columns (correlation >0.70); (4) removing columns with more than 25 percent missing entries for GSS and columns with more than 50 percent missing entries for National Longitudinal Survey of Youth; and (5) removing rows with any missing entries in the resulting data set. The choice of threshold for removing columns with some p proportion of missing entries was done by trial and error to achieve fully observed subsets with 300 to 1,000 rows and 40 to 100 columns. We would ideally have liked to keep the correlated columns in the data; however, we removed these because column correlation showed up as a common reason for failure of imputation by MICE and Amelia. To compute correlations across columns with diverse data types, we used the R command `hetcor` within the package `polycor`, which “computes a heterogenous correlation matrix, consisting of Pearson product-moment correlations between numeric variables, polyserial correlations between numeric and ordinal variables, and polychoric correlations between ordinal variables” (Fox and Fox 2016).
 6. Such real-world data sets typically come with population weights that are not considered in the demonstration simulations in this article. However, it is straightforward to incorporate a weighting approach into the LowRankModels methodology: each term in the loss function (corresponding to the observation of the $(i, j)^{th}$ element in the array) can be premultiplied by a weight w_{ij} . For example, column weights $w_{ij} = w_j$ can be used to indicate the importance of fitting different questions, and row weights $w_{ij} = w_i$ can be used to reweight a survey population toward a reference population. The LowRankModels software includes support for column weights. We are working on an extension of the software to allow for row weights. In the meantime, one can simulate row weights by up-sampling rows that should have higher weights by replicating the row. LowRankModels is able to fit large data sets quickly, so this approach is of practical utility.
 7. These metrics are simply averaged for the Boolean and categorical columns. For the real and ordinal columns, the metrics (other than empirical coverage) are averaged after scaling by the standard deviation of the columns.
 8. Our results demonstrate that the performance of Trace and Rank methods are mostly similar overall, but Trace typically performs better for categorical and Boolean valued columns. The similarity in performance of Trace and Rank methods is unsurprising given that Trace norm regularization is widely used as an approach for learning low-rank matrices (see, e.g., Ciliberto, Stamos, and Pontil 2017; Wang, Zhang, and Wang 2021).

References

- Acar, Evrim, Daniel M. Dunlavy, Tamara G. Kolda, and Morten Mørup. 2011. “Scalable Tensor Factorizations for Incomplete Data.” *Chemometrics and Intelligent Laboratory Systems* 106:41–56.
- Agresti, Alan. 1996. *An Introduction to Categorical Data Analysis*. New York: John Wiley.
- Alexander, Karl L., Doris R. Entwisle, and Linda Steffel Olson. 2007. “Lasting Consequences of the Summer Learning Gap.” *American Sociological Review* 72(2):167–80.

- Ansolabehere, Stephen D., Shanto Iyengar, and Adam Simon. 1999. "Replicating Experiments Using Aggregate and Survey Data: The Case of Negative Advertising and Turnout." *American Political Science Review* 93(4):901–909.
- Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2018. "Matrix Completion Methods for Causal Panel Data Models." NBER Working Paper No. 25132. Cambridge, MA: National Bureau of Economic Research.
- Bernhagen, Patrick, and Michael Marsh. 2007. "The Partisan Effects of Low Turnout: Analyzing Vote Abstention as a Missing Data Problem." *Electoral Studies* 26(3):548–60.
- Blackwell, Matthew, James Honaker, and Gary King. 2015. "A Unified Approach to Measurement Error and Missing Data Details and Extensions." *Sociological Methods & Research* 46(3):303–41.
- Blake, Kellee. 1996. "First in the Path of the Firemen: The Fate of the 1890 Population Census, Part 1." *Prologue Magazine* 28.
- Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350(6264):1073–76.
- Boyle, John, Lewis Berman, James Dayton, Ronaldo Iachan, Matt Jans, and Randy ZuWallack. 2021. "Physical Measures and Biomarker Collection in Health Surveys: Propensity to Participate." *Research in Social and Administrative Pharmacy* 17(5):921–29.
- Brenner, Philip S. 2011. "Identity Importance and the Overreporting of Religious Service Attendance: Multiple Imputation of Religious Attendance Using the American Time Use Study and the General Social Survey." *Journal for the Scientific Study of Religion* 50(1):103–15.
- Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45(3):1–67.
- Candès, Emmanuel J., and Yaniv Plan. 2009. "Near-Ideal Model Selection by ℓ_1 Minimization." *Annals of Statistics* 37(5A):2145–77.
- Candès, Emmanuel J., and Benjamin Recht. 2009. "Exact Matrix Completion via Convex Optimization." *Foundations of Computational Mathematics* 9:717.
- Che, Zhengping, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. "Recurrent Neural Networks for Multivariate Time Series with Missing Values." *Scientific Reports* 8(1):1–12.
- Chiozza, Giacomo, and Hein E. Goemans. 2004. "International Conflict and the Tenure of Leaders: Is War Still Ex Post Inefficient?" *American Journal of Political Science* 48(3):604–19.
- Choudhury, Suvra Jyoti, and Nikhil R. Pal. 2019. "Imputation of Missing Data with Neural Networks for Classification." *Knowledge-Based Systems* 182:104838.
- Christopher, Winship, and Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18:327–50.
- Ciliberto, Carlo, Dimitris Stamos, and Massimiliano Pontil. 2017. "Reexamining Low Rank Matrix Factorization for Trace Norm Regularization." arXiv. Retrieved September 8, 2022. <https://arxiv.org/abs/1706.08934>.
- Collins, Michael, Sanjoy Dasgupta, and Robert E. Schapire. 2001. "A Generalization of Principal Components Analysis to the Exponential Family." Pp. 617–24 in *Advances in Neural Information Processing Systems*, edited by M. I. Jordan, Y. LeCun, and S. A. Solla. Cambridge, MA: MIT Press.
- Davey, Adam, Michael J. Shanahan, and Joseph L. Schafer. 2001. "Correcting for Selective Nonresponse in the National Longitudinal Survey of Youth Using Multiple Imputation." *Journal of Human Resources* 36(3):500–19.
- Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39: 1–22.
- Fay, Robert E. 1992. "When Are Inferences from Multiple Imputation Valid." Pp. 227–32 in *Proceedings of the Survey Research Methods Section of the American Statistical Association, Vol. 81*. Alexandria, VA: American Statistical Association.
- Fearon, James D. 2005. "Primary Commodity Exports and Civil War." *Journal of Conflict Resolution* 49(4):483–507.

- Fox, J., and M. J. Fox. 2016. "Package 'polycor.'" Retrieved September 8, 2022. <https://cran.r-project.org/web/packages/polycor/index.html>.
- Foygel, Rina, and Nathan Srebro. 2011. "Concentration-Based Guarantees for Low-Rank Matrix Reconstruction." Pp. 315–40 in *Proceedings of the 24th Annual Conference on Learning Theory*, edited by S. M. Kakade and U. von Luxburg. Machine Learning Research Press.
- Frankel, Laura Lazarus, and D. Sunshine Hillygus. 2014. "Looking beyond Demographics: Panel Attrition in the ANES and GSS." *Political Analysis* 22(3):336–53.
- Gay, Claudine. 2002. "Spirals of Trust? The Effect of Descriptive Representation on the Relationship between Citizens and Their Government." *American Journal of Political Science* 46(4):717–32.
- Gemulla, Rainer, Erik Nijkamp, Peter J. Haas, and Yannis Sismanis. 2011. "Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent." Pp. 69–77 in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery.
- Gerber, Elisabeth R., and Daniel J. Hopkins. 2011. "When Mayors Matter: Estimating the Impact of Mayoral Partisanship on City Policy." *American Journal of Political Science* 55(2):326–39.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning, Vol. 1*. Cambridge, MA: MIT Press.
- Gunasekar, Suriya, Pradeep Ravikumar, and Joydeep Ghosh. 2014. "Exponential Family Matrix Completion under Structural Constraints." Pp. 1917–25 in *International Conference on Machine Learning*. Machine Learning Research Press.
- Harnois, Catherine E. 2005. "Different Paths to Different Feminisms? Bridging Multiracial Feminist Theory and Quantitative Sociological Gender Research." *Gender & Society* 19(6):809–28.
- Heckman, James. 1974. "Shadow Prices, Market Wages and Labor Supply." *Econometrica* 42(4):679–94.
- Heckman, James J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5(4):475–92.
- Hollenbach, Florian M., Iavor Bojinov, Shahryar Minhas, Nils W. Metternich, Michael D. Ward, and Alexander Volfovsky. 2018. "Multiple Imputation Using Gaussian Copulas." *Sociological Methods & Research* 50(3):1259–83.
- Honaker, James, Jonathan N. Katz, and Gary King. 2002. "A Fast, Easy, and Efficient Estimator for Multiparty Electoral Data." *Political Analysis* 10(1):84–100.
- Honaker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45(1):1–47.
- Huisman, Mark. 2014. "Imputation of Missing Network Data: Some Simple Procedures." *Journal of Social Structure* 10(1):1–29.
- Josse, Julie, and François Husson. 2016. "missMDA: A Package for Handling Missing Values in Multivariate Data Analysis." *Journal of Statistical Software* 70(1):1–31.
- Josse, Julie, Jérôme Pagès, and François Husson. 2011. "Multiple Imputation in Principal Component Analysis." *Advances in Data Analysis and Classification* 5:231–46.
- Kallus, Nathan, Xiaojie Mao, and Madeleine Udell. 2018. "Causal Inference with Noisy and Missing Covariates via Matrix Factorization." In *Advances in Neural Information Processing Systems, Vol. 31*. Cambridge, MA: MIT Press.
- King, Gary, Emmanuela Gakidou, Kosuke Imai, Jason Lakin, Ryan T. Moore, Clayton Nall, Nirmala Ravishankar, et al. 2009. "Public Policy for the Poor? A Randomised Assessment of the Mexican Universal Health Insurance Programme." *The Lancet* 373(9673):1447–54.
- Kolda, Tamara G., and Brett W. Bader. 2009. "Tensor Decompositions and Applications." *SIAM Review* 51:455–500.
- Kroonenberg, Pieter M., and Jan De Leeuw. 1980. "Principal Component Analysis of Three-Mode Data by Means of Alternating Least Squares Algorithms." *Psychometrika* 45(1):69–97.
- Kropko, Jonathan, Ben Goodrich, Andrew Gelman, and Jennifer Hill. 2014. "Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches." *Political Analysis* 22(4):497–519.

- Lall, Ranjit. 2016. "How Multiple Imputation Makes a Difference." *Political Analysis* 24(4):414–33.
- Levy, Omer, and Yoav Goldberg. 2014. "Neural Word Embedding as Implicit Matrix Factorization." *Advances in Neural Information Processing Systems*, Vol. 27. Cambridge, MA: MIT Press.
- Linos, Katerina, and Martin West. 2003. "Self-Interest, Social Beliefs, and Attitudes to Redistribution: Re-addressing the Issue of Cross-National Variation." *European Sociological Review* 19(4):393–409.
- Meng, Xiao-Li. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science* 9(4):538–58.
- Negahban, Sahand, and Martin J. Wainwright. 2011. "Estimation of (Near) Low-Rank Matrices with Noise and High-Dimensional Scaling." *Annals of Statistics* 39(2):1069–97.
- Nielsen, Søren Feodor. 2003. "Proper and Improper Multiple Imputation." *International Statistical Review* 71(3):593–607.
- Ouzienko, Vladimir, and Zoran Obradovic. 2014. "Imputation of Missing Links and Attributes in Longitudinal Social Surveys." *Machine Learning* 95:329–56.
- Pasek, Josh, Alexander Tahk, Yphtach Lelkes, Jon A. Krosnick, B. Keith Payne, Omair Akhtar, and Trevor Thompson. 2009. "Determinants of Turnout and Candidate Choice in the 2008 U.S. Presidential Election Illuminating the Impact of Racial Prejudice and Other Considerations." *Public Opinion Quarterly* 73(5):943–94.
- Rao, Jon N. K., and Jun Shao. 1992. "Jackknife Variance Estimation with Survey Data under Hot Deck Imputation." *Biometrika* 79(4):811–22.
- Rennie, Jason D. M., and Nathan Srebro. 2005. "Loss Functions for Preference Levels: Regression with Discrete Ordered Labels." Pp. 180–86 in *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*. Norwell, MA: Kluwer.
- Rifkin, Ryan, and Aldebaro Klautau. 2004. "In Defense of One-vs-All Classification." *Journal of Machine Learning Research* 5:101–41.
- Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.
- Rubin, Donald B. 1996. "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association* 91:473–89.
- Salganik, Matthew J., and Douglas D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34:193–240.
- Salganik, Matthew J., and Karen E. C. Levy. 2015. "Wiki Surveys: Open and Quantifiable Social Data Collection." *PLoS ONE* 10:e0123483.
- Sang, Hejian, Jae Kwang Kim, and Danhyang Lee. 2020. "Semiparametric Fractional Imputation Using Gaussian Mixture Models for Handling Multivariate Missing Data." *Journal of the American Statistical Association* 117(538):654–63.
- Sassler, Sharon, and James McNally. 2003. "Cohabiting Couples' Economic Circumstances and Union Transitions: A Re-examination Using Multiple Imputation Techniques." *Social Science Research* 32: 553–78.
- Schafer, Joseph L. 1999. "Multiple Imputation: A Primer." *Statistical Methods in Medical Research* 8(1): 3–15.
- Scheve, Kenneth F., and Matthew J. Slaughter. 2001. "What Determines Individual Trade-Policy Preferences?" *Journal of International Economics* 54(2):267–92.
- Shah, Anoop D., Jonathan W. Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. 2014. "Comparison of Random forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study." *American Journal of Epidemiology* 179(6):764–74.
- Shapiro, Alexander, Yao Xie, and Rui Zhang. 2018. "Matrix Completion with Deterministic Pattern: A Geometric Perspective." *IEEE Transactions on Signal Processing* 67:1088–1103.
- Si, Yajuan, Jerome P. Reiter, and D. Sunshine Hillygus. 2015. "Semi-parametric Selection Models for Potentially Non-ignorable Attrition in Panel Studies with Refreshment Samples." *Political Analysis* 23(1):92–112.
- Singh, Ajit P., and Geoffrey J. Gordon. 2008. "Relational Learning via Collective Matrix Factorization." Pp. 650–58 in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery.

- Smith, Tom W., Michael Davern, Jeremy Freese, and Stephen L. Morgan. 2019. "General Social Surveys, 1972–2018" [Machine-readable data file]. Chicago: NORC.
- Solt, Frederick. 2009. "Standardizing the World Income Inequality Database." *Social Science Quarterly* 90(2):231–42.
- Srebro, Nathan. 2004. "Learning with Matrix Factorizations." PhD dissertation, Massachusetts Institute of Technology.
- Srebro, Nathan, and Ruslan R. Salakhutdinov. 2010. "Collaborative Filtering in a Non-uniform World: Learning with the Weighted Trace Norm." Pp. 2056–64 in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press.
- Stasavage, David. 2005. "Democracy and Education Spending in Africa." *American Journal of Political Science* 49(2):343–58.
- Takane, Yoshio, Forrest W. Young, and Jan De Leeuw. 1977. "Nonmetric Individual Differences Multidimensional Scaling: An Alternating Least Squares Method with Optimal Scaling Features." *Psychometrika* 42(1):7–67.
- Treiman, Donald J., William T. Bielby, and Man-Tsun Cheng. 1988. "Evaluating a Multiple-Imputation Method for Recalibrating 1970 U.S. Census Detailed Industry Codes to the 1980 Standard." *Sociological Methodology* 18:309–45.
- Udell, Madeleine, Corinne Horn, Reza Zadeh, and Stephen Boyd. 2014. "Generalized Low Rank Models." arXiv. Retrieved September 8, 2022. <https://arxiv.org/abs/1410.0342>.
- Udell, Madeleine, and Alex Townsend. 2017. "Nice Latent Variable Models Have Log-Rank." arXiv. Retrieved September 8, 2022. <https://arxiv.org/abs/1705.07474v1>.
- U.S. Department of Labor, Bureau of Labor Statistics. 2015. "National Longitudinal Survey of Youth 1997 Cohort, 1997–2013 (Rounds 1–16)." Columbus: Center for Human Resource Research, The Ohio State University.
- Wang, Wendong, Feng Zhang, and Jianjun Wang. 2021. "Low-Rank Matrix Recovery via Regularized Nuclear Norm Minimization." *Applied and Computational Harmonic Analysis* 54:1–19.
- Wang, Zhenhua, Olanrewaju Akande, Jason Poulos, and Fan Li. 2021. "Are Deep Learning Models Superior for Missing Data Imputation in Large Surveys? Evidence from an Empirical Comparison." arXiv. Retrieved September 8, 2022. <https://arxiv.org/abs/2103.09316>.
- Xia, Jing, Shengyu Zhang, Guolong Cai, Li Li, Qing Pan, Jing Yan, and Gangmin Ning. 2017. "Adjusted Eight Voting Algorithm for Random Forests in Handling Missing Values." *Pattern Recognition* 69(1): 52–60.
- Yu, Hsiang-Fu, Nikhil Rao, and Inderjit S. Dhillon. 2016. "Temporal Regularized Matrix Factorization for High-Dimensional Time Series Prediction." Pp. 847–55 in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press.
- Zhao, Yuxuan, and Madeleine Udell. 2020. "Missing Value Imputation for Mixed Data via Gaussian Copula." Pp. 636–46 in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York: Association for Computing Machinery.

Author Biographies

Nandana Sengupta is an assistant professor in the School of Public Policy and an affiliate faculty of the Economics Unit at the Department of Humanities and Social Sciences at IIT Delhi. Her research lies at the intersection of applied machine learning and economics with a particular focus on novel survey design techniques, algorithmic bias and Indian labor markets. Prior to joining IIT Delhi she was assistant professor at the School of Policy and Governance in Azim Premji University, India and postdoctoral scholar at the Knowledge Lab in University of Chicago. Her current work is supported by Google Research's Award for Inclusion Research and Co-Impact's Gender Fund. Nandana obtained her PhD in economics at the Tepper School of Business at Carnegie Mellon University in 2015.

Madeleine Udell is assistant professor of management science and engineering at Stanford University, with an affiliation with the Institute for Computational and Mathematical Engineering (ICME) and

courtesy appointment in electrical engineering, and associate professor with tenure (on leave) of operations research and information engineering and Richard and Sybil Smith Sesquicentennial fellow at Cornell University. She studies optimization and machine learning for large scale data analysis and control, with applications in marketing, demographic modeling, medical informatics, engineering system design, and automated machine learning. She has received several awards, including an Alfred P. Sloan Research Fellowship (2021), a National Science Foundation CAREER award (2020), an Office of Naval Research (ONR) Young Investigator Award (2020), a Cornell Engineering Research Excellence Award (2020), an INFORMS Optimization Society Best Student Paper Award (as advisor) (2019), and INFORMS Doing Good with Good OR (2018). Her work is supported by grants from the NSF, ONR, DARPA, the Canadian Institutes of Health, and Capital One.

Nati (Nathan) Srebro is a professor at the Toyota Technological Institute at Chicago, with cross-appointments at the University of Chicago's Department of Computer Science, and Committee on Computational and Applied Mathematics. He obtained his PhD from the Massachusetts Institute of Technology in 2004, and previously was a postdoctoral fellow at the University of Toronto, a visiting scientist at IBM, and an associate professor at the Technion. Dr. Srebro's research encompasses methodological, statistical and computational aspects of machine learning, as well as related problems in optimization. Some of Srebro's significant contributions include work on learning "wider" Markov networks, introducing the use of the nuclear norm for machine learning and matrix reconstruction, work on fast optimization techniques for machine learning, and on the relationship between learning and optimization. His current interests include understanding deep learning through a detailed understanding of optimization, distributed and federated learning, algorithmic fairness and practical adaptive data analysis.

James Evans is the Max Palevsky Professor of Sociology, director of Knowledge Lab, and faculty director of Computational Social Science at the University of Chicago. His research uses large-scale data, machine learning and generative models to understand how collectives think and what they know. This involves inquiry into the emergence of ideas, shared patterns of search and inference, and processes of attention, communication, agreement, and certainty, with a special focus on innovation. Evans designs observatories that fuse data from text, networks, images, and structured data with other sensors and results from interactive crowdsourcing and online experiments. Evans' work has investigated modern science and technology to understand the nature of creative teams, identify collective biases and the benefits of diversity, generate new leads taking these into account, and imagine alternative discovery regimes and complementary artificial intelligence. He has identified R&D institutions that generate more and less novelty, disruption, accuracy, and robustness. Evans also explores thinking and knowing in other domains ranging from political ideology to civic discourse and popular culture.