AIMS Mathematics, 11(5): 14641–14654.
DOI: 10.3934/math.2026600
Received: 08 February 2026
Revised: 23 April 2026
Accepted: 29 April 2026
Published: 25 May 2026

https://www.aimspress.com/journal/Math

Research article

Mahalanobis-geometry imputation for multivariate data with missing entries

Alvaro H. Salas S.[1,*], **David L. Ocampo R.**[1] **and Lorenzo J. Martínez H.**[1,2]

[1] Department of Mathematics and Statistics, Universidad Nacional de Colombia, Manizales, Colombia

[2] Universidad de Caldas, Manizales, Colombia

* **Correspondence:** Email: ahsalass@unal.edu.co.

Abstract: Missing entries in multivariate data distort not only marginal summaries but also the covariance geometry that governs scale-adjusted and correlation-aware comparisons between observations. Motivated by covariance-sensitive downstream tasks, this paper develops a deterministic imputation framework driven by Mahalanobis distance. The first stage is a linear frozen-covariance procedure: missing entries are temporarily replaced by simple columnwise values, a fixed covariance matrix is computed, and the sum of the nonconstant squared Mahalanobis distances is minimized with respect to the unknown entries. Since the inverse covariance is fixed at that stage, the objective is quadratic and the first-order optimality conditions reduce to a linear system. The second stage is a nonlinear covariance-updating refinement in which the covariance matrix depends on the imputed values themselves and the optimization is performed locally, using the linear solution as initializer. We derive a compact matrix representation of the linear objective, give a sufficient full-rank condition guaranteeing uniqueness of the stationarity system, discuss the bias induced by freezing the covariance, and provide a regularized fallback for singular or ill-conditioned systems. The framework also clarifies its scope with respect to MCAR, MAR-type, and structured block masks, and uses covariance stabilization only as a numerical safeguard rather than as a determinant-minimization estimator. A repeated-mask experiment on the red wine quality dataset shows that the Mahalanobis method substantially improves on mean imputation at all masking levels and becomes the strongest among the tested methods at the highest missingness level considered. The resulting method is transparent, reproducible, and intended for moderate continuous-data settings in which preserving empirical covariance geometry is more important than fitting a large black-box model.

Keywords: missing data; imputation; Mahalanobis distance; sample covariance matrix; covariance geometry; regularized linear systems; MCAR masking
Mathematics Subject Classification: 62F35, 62H12, 62H25, 62J10, 65F10

1. Introduction

Incomplete multivariate data arise in surveys, laboratory measurements, industrial monitoring, finance, clinical records, and many other data-rich settings. When entries are missing, the damage is not restricted to isolated variables: the empirical covariance structure is altered, and therefore so are procedures that depend on scale and dependence, including covariance-based clustering, anomaly detection, principal component analysis, and Mahalanobis-type classification. A method that fills gaps while preserving multivariate geometry is therefore especially relevant when the downstream analysis is itself covariance-sensitive.

Classical responses to missing data include complete-case analysis, mean substitution, regression imputation, expectation-maximization, and multiple imputation; each addresses a different inferential or predictive goal [3, 6, 9, 12]. Deterministic baselines such as mean, regression, and nearest-neighbor imputation remain important because they are simple and transparent [11]. At the other end of the spectrum, low-rank completion, random-forest methods, and recent deep generative or attention-based models can exploit richer structure, but they also introduce heavier modeling assumptions, larger tuning burdens, or reduced interpretability [1, 8, 10, 13–18]. No single class dominates uniformly across all data regimes.

The present paper studies a narrower but mathematically clear objective: imputing missing entries so that the completed matrix remains geometrically coherent with respect to the covariance metric. Mahalanobis distance is the canonical distance associated with a positive definite covariance matrix [4, 5, 7]. Because it simultaneously accounts for variable scale and cross-correlation, it provides a natural objective whenever the preservation of covariance-aware row geometry is more important than uncertainty quantification or purely predictive black-box accuracy.

Let the missing entries be collected into a vector of unknowns $x \in \mathbb{R}^m$. Once a covariance matrix S is available, each squared Mahalanobis distance between rows becomes a quadratic form in the corresponding row difference, and summing the terms that actually depend on x produces an optimization criterion. Two versions of that criterion are especially natural:

(i) In the linear frozen-covariance method, missing entries are first replaced by provisional values, yielding a numeric matrix Y and hence a fixed covariance matrix $S(Y)$. Once $S(Y)^{-1}$ is frozen, the Mahalanobis objective is quadratic in x, and the first-order conditions reduce to a linear algebraic system.

(ii) In the nonlinear covariance-updating method, the covariance matrix is recomputed from the symbolic completion itself. Then both $S(x)$ and $S(x)^{-1}$ depend on the unknown entries, leading to a nonlinear local optimization problem initialized by the linear solution.

The contribution is not a generalized-variance minimization principle. Reducing the determinant of a covariance matrix does not, by itself, imply accurate recovery of masked entries. For this reason, determinant information is used only as a diagnostic of numerical conditioning, whereas the estimation principle remains the preservation of Mahalanobis geometry.

The paper has four main goals:

G1. to position the method clearly relative to classical missing-data methods and recent imputation developments;

G2. to derive the linear Mahalanobis system in a compact matrix form and provide a usable sufficient condition for uniqueness;

G3. to state explicitly the method's limitations, especially rank deficiency, covariance instability, and the absence of uncertainty quantification;

G4. to strengthen empirical validation through repeated masking, fair mask sharing across methods, and a transparent experimental protocol.

The paper is organized as follows. Section 2 reviews the main imputation paradigms and explains where the proposed method fits. Section 3 introduces the notation, missingness scope, linear and nonlinear Mahalanobis imputers, and stabilization strategy. Section 4 describes the experimental protocol. Section 5 reports the repeated-mask wine experiment. Section 6 discusses strengths, limitations, and interpretation, and Section 7 concludes.

2. Related work

It is useful to distinguish the proposed framework from the main families of imputation methods that appear in the missing-data literature.

2.1. Classical statistical approaches

Complete-case analysis discards incomplete rows and is often simple to implement, but it may be wasteful or biased when the missingness mechanism is not benign. Parametric likelihood-based methods, especially expectation–maximization, are central when one is willing to posit an incomplete-data model and focus on estimation under that model [3, 6]. Multiple imputation extends this philosophy by propagating uncertainty across several completed datasets rather than committing to a single deterministic fill-in [9, 12]. These approaches are essential when inferential validity is the primary goal.

2.2. Deterministic baselines and machine-learning comparators

Mean imputation remains a basic reference point, even though it shrinks empirical variability and attenuates covariances. Regression imputation uses the observed coordinates to predict missing ones, but it is directional and may become unstable under multicollinearity. Neighbor-based approaches such as k-NN exploit local similarity in the observed space and have long served as practical baselines in biological and tabular applications [11]. MissForest extends this spirit through iterative random-forest prediction and is a particularly relevant nonlinear comparator for mixed-type tables [10].

2.3. Low-rank and matrix-completion viewpoints

When the data matrix is governed by a latent low-rank structure, matrix-completion methods based on nuclear-norm regularization or related spectral ideas can be effective [1, 8]. Their target, however, is low-rank recovery rather than explicit preservation of the empirical covariance metric of the observed table.

2.4. Recent deep and generative imputers

The recent literature includes GAN-based imputers such as GAIN [13], variational approaches such as GP-VAE [14], diffusion-based imputers such as CSDI [15], self-attention architectures such as SAITS [16], and masked-autoencoding transformer approaches for tabular data such as ReMasker [17]. A recent survey by Wang et al. emphasizes both the breadth of this landscape and the importance of matching the imputation method to the data regime and downstream objective [18]. These modern methods are powerful, but they are usually more computationally demanding and less transparent than simple covariance-driven rules.

2.5. Position of the present work

The method proposed here is deterministic like mean, regression, and single-imputation neighbor methods, but it is explicitly multivariate because its objective is built from the covariance-induced Mahalanobis metric. It is not intended to replace multiple imputation when uncertainty quantification is required, nor to compete head-to-head with large deep architectures on broad benchmark suites. Rather, it offers a mathematically transparent alternative for moderate continuous-data settings in which preserving covariance geometry is itself the main structural priority.

3. Materials and methods

3.1. Problem setting and notation

Let

$$X = [a_{ij}] \in \mathbb{R}^{n\times p}$$

be a data matrix with n observations and p variables. Let

$$\Omega \subset \{1,\dots,n\} \times \{1,\dots,p\}$$

be the set of observed indices and let Ω^c denote the missing set. Write $m := |\Omega^c|$.

For each $(i, j) \in \Omega^c$ introduce a variable c_{ij}, collect these variables into a vector

$$x \in \mathbb{R}^m,$$

and define the symbolic completion

$$\widetilde{X}(x) = \left[\widetilde{a}_{ij}(x)\right]_{n\times p}, \qquad \widetilde{a}_{ij}(x) = \begin{cases} a_{ij}, & (i,j) \in \Omega, \\ c_{ij}, & (i,j) \in \Omega^c. \end{cases}$$

The ith completed observation is denoted by $\widetilde{x_i}(x) \in \mathbb{R}^p$ and is understood as a column vector. Given any complete matrix $Y \in \mathbb{R}^{n\times p}$ with observations $y_i \in \mathbb{R}^p$ and column mean vector $\bar{y} \in \mathbb{R}^p$, its sample covariance matrix is

$$S(Y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T. \tag{3.1}$$

Whenever $S(Y)$ is nonsingular, the squared Mahalanobis distance between observations y_i and y_h is

$$\rho_{ih}^2(Y) = (y_i - y_h)^T S(Y)^{-1}(y_i - y_h). \tag{3.2}$$

3.2. Scope with respect to missingness mechanisms

The method is deterministic and geometry-based; it is not a stochastic model of the missingness process itself. In the experimental protocol, performance is evaluated under synthetic missing completely at random (MCAR) masking, because the original complete matrix makes the true masked entries available for direct error computation. The method can also be stress-tested under missing at random (MAR) schemes when masking probabilities depend only on observed variables. It is not presented as a principled solution for missing not at random (MNAR) settings, where missingness depends on unobserved values and should be modeled jointly with the data-generation mechanism.

Structured blocks of missing entries are admissible in principle, but their success depends on how much multivariate information remains available outside the block. In the linear frozen-covariance stage, the essential issue is whether the active Mahalanobis-difference design retains enough information to identify the missing degrees of freedom. Large contiguous blocks may reduce the number of informative row contrasts and can drive the stationarity system toward rank deficiency. Thus block masking is a natural stress regime in which ridge stabilization or minimum-norm regularization may become necessary.

3.3. Linear frozen-covariance method

The linear method begins with a provisional complete matrix Y obtained by replacing each missing entry in column j by the mean of the observed values in that column:

$$\mu_j = \frac{1}{|\{i : (i, j) \in \Omega\}|} \sum_{(i,j)\in\Omega} a_{ij}, \qquad j = 1, \ldots, p.$$

From this numeric matrix we compute the fixed covariance matrix

$$S_0 := S(Y),$$

and assume first that S_0 is positive definite. Write

$$Q := S_0^{-1}.$$

The linear Mahalanobis objective is defined as the sum of the nonconstant squared Mahalanobis distances,

$$\Phi_L(x) = \sum_{(i,h)\in\mathcal{A}} (\widetilde{x_i}(x) - \widetilde{x_h}(x))^T Q(\widetilde{x_i}(x) - \widetilde{x_h}(x)), \tag{3.3}$$

where $\mathcal{A}$ denotes the set of active pairs, that is, row pairs whose distance actually depends on at least one missing variable. Distances between two fully observed rows are constant and have no effect on the minimizer.

The next proposition makes the quadratic structure explicit.

Proposition 1. *(Quadratic form) For each active pair $(i, h) \in \mathcal{A}$, there exist a vector $d_{ih} \in \mathbb{R}^p$ and a matrix $B_{ih} \in \mathbb{R}^{p\times m}$ such that*

$$\widetilde{x_i}(x) - \widetilde{x_h}(x) = d_{ih} + B_{ih}x.$$

If B denotes the block matrix obtained by stacking the matrices B_{ih} and d denotes the stacked vector of the vectors d_{ih}, then

$$\Phi_L(x) = c + 2g^T x + x^T H x, \qquad H = 2B^T(I_{|\mathcal{A}|} \otimes Q)B, \tag{3.4}$$

for a constant $c \in \mathbb{R}$ and a vector $g \in \mathbb{R}^m$. Consequently,

$$\nabla\Phi_L(x) = 2Hx + 2g, \tag{3.5}$$

and the stationarity equations reduce to the linear system

$$Hx = -g. \tag{3.6}$$

Proof. The map $x \mapsto \widetilde{x}_i(x) - \widetilde{x}_h(x)$ is affine because each entry of $\widetilde{X}(x)$ is either observed or equal to one coordinate of x. Expanding each quadratic term in (3.3) and summing over $\mathcal{A}$ gives (3.4). Differentiating the quadratic form gives (3.5) and hence (3.6). □

The block representation yields a simple sufficient condition for uniqueness.

Theorem 1. *(Sufficient condition for nonsingularity) Assume that S_0 is positive definite, so $Q = S_0^{-1} \succ 0$. If the stacked design matrix B in* (3.4) *has full column rank m, then the Hessian matrix H is symmetric positive definite. In particular, the stationarity system* (3.6) *has a unique solution.*

Proof. Because $Q \succ 0$, the Kronecker product $I_{|\mathcal{A}|} \otimes Q$ is also positive definite. For any nonzero vector $z \in \mathbb{R}^m$,

$$z^T H z = 2(Bz)^T(I_{|\mathcal{A}|} \otimes Q)(Bz).$$

If B has full column rank, then $Bz \neq 0$ whenever $z \neq 0$, and hence the right-hand side is strictly positive. Therefore $H \succ 0$. □

Remark 1. *(Interpretation of the rank condition) The matrix B encodes how each missing variable enters the active row differences. Full column rank means that no nontrivial linear combination of missing variables is invisible to all active Mahalanobis differences. In practical terms, each missing degree of freedom must be geometrically identifiable through the chosen distance structure.*

Remark 2. *(Failure mode) If two missing variables always occur through exactly the same row-difference pattern, then two columns of B coincide and B is rank-deficient. In that case H is singular and the stationarity equations do not determine a unique completion. This can happen when a large missing block leaves only a small number of effective contrasts between rows.*

3.4. Regularized fallback for the linear system

When H is singular or nearly singular, we solve the Tikhonov-stabilized system

$$(H + \tau I)x_\tau = -g, \qquad \tau > 0, \tag{3.7}$$

or compute the minimum-norm solution via the Moore-Penrose pseudoinverse. This fallback prevents the linear stage from failing numerically. However, rank deficiency should not be hidden: it signals underidentification of the completion by the available Mahalanobis-difference information.

3.5. Algorithmic summary of the linear stage

The linear frozen-covariance stage can be summarized as follows:

L1. Replace missing entries by observed column means to build Y.

L2. Compute $S_0 = S(Y)$ and set $Q = (S_0 + \eta I)^{-1}$, with a small $\eta \geq 0$ if stabilization is needed.

L3. Form the active-pair design matrix B and vector g.

L4. Solve either (3.6) or the stabilized system (3.7).

L5. Insert the resulting vector of missing entries into $\widetilde{X}(x)$ to obtain the completed matrix.

3.6. Bias induced by freezing the covariance

Freezing the covariance matrix makes the first stage transparent and computationally simple, but it also introduces an approximation. The matrix S_0 is computed from a provisional completion rather than from the final imputed matrix. Consequently, the linear objective measures geometry in the metric induced by Y, not in the metric induced by $\widetilde{X}(x)$. If the provisional fill-in severely distorts the covariance, then the linear solution may inherit this distortion. This bias is expected to be small when the missing fraction is moderate and columnwise means provide a reasonable covariance initializer, but it may become important under heavy missingness, structured masks, or highly nonlinear dependence patterns. The nonlinear refinement below is designed precisely to reduce this frozen-metric bias.

3.7. Nonlinear covariance-updating refinement

The nonlinear refinement removes the frozen-covariance approximation by allowing the covariance matrix to depend on the unknown entries. Define

$$S(x) := S(\widetilde{X}(x)).$$

Whenever $S(x)$ is invertible, the nonlinear Mahalanobis objective is

$$\Phi_N(x) = \sum_{(i,h)\in\mathcal{A}} (\widetilde{x}_i(x) - \widetilde{x}_h(x))^T S(x)^{-1} (\widetilde{x}_i(x) - \widetilde{x}_h(x)). \tag{3.8}$$

This objective is generally nonconvex because both the covariance matrix and its inverse depend on x.

A practical and reproducible strategy is to use the linear solution x_L as initializer and minimize (3.8) only on a local admissible set,

$$\mathcal{B}(x_L; \delta) = \{x \in \mathbb{R}^m : |x_k - (x_L)_k| \leq \delta_k,\ k = 1, \ldots, m\}, \tag{3.9}$$

with fixed user-chosen radii $\delta_k > 0$. The nonlinear estimate is then defined by

$$x_N \in \arg\min_{x\in\mathcal{B}(x_L;\delta)} \Phi_N(x), \tag{3.10}$$

subject to numerical feasibility conditions ensuring that $S(x)$ remains safely invertible. The local formulation is intentional: the linear solution already captures the dominant covariance geometry, and the nonlinear step is meant to correct the metric rather than replace the method by an unconstrained global search.

3.8. Covariance stabilization

Because Mahalanobis distances require inversion of a covariance matrix, numerical stability must be handled explicitly. Whenever a covariance matrix S is ill-conditioned, we replace it by

$$S_\eta = S + \eta I, \qquad \eta > 0 \text{ small}, \tag{3.11}$$

before inversion. This is the default stabilization device in both the linear and nonlinear stages. If a covariance matrix becomes singular despite stabilization, the Moore-Penrose pseudoinverse can be used for diagnostic purposes, but the corresponding completion should be interpreted as a regularized solution rather than a strict Mahalanobis optimum. The determinant of the covariance matrix is monitored only as an indicator of ill-conditioning; it is not minimized as a surrogate for true missing-value recovery.

4. Experimental protocol

The numerical protocol is organized around repeated masking, fair comparison, and transparent stabilization. The objective is not to benchmark against every modern imputer, but to determine whether Mahalanobis geometry adds value beyond simple deterministic baselines on moderate continuous data.

4.1. Baselines

For continuous data, the natural deterministic and semi-parametric baselines are:

B1. columnwise mean imputation;

B2. ordinary-least-squares or ridge-stabilized regression imputation;

B3. k-NN imputation with a small set of k values;

B4. EM-based Gaussian imputation when a Gaussian working model is acceptable;

B5. MICE as a standard chained-equation baseline.

The experiment reported below compares mean imputation, ridge-stabilized regression imputation, and the proposed Mahalanobis method. MissForest and deep-learning imputers are important related methods, but they are not the most immediate comparators for the present scope, which is a deterministic covariance-geometry principle for continuous tabular data.

4.2. Masking mechanisms

The primary benchmarking regime is MCAR masking. Since the benchmark matrix is complete before masking, the true hidden values remain available for error computation. MAR-type masks and structured block masks are compatible with the same framework as stress tests, but direct claims about MNAR recovery are avoided.

4.3. Evaluation metrics

When the true matrix is available before masking, the imputed values are evaluated only at the masked positions. We report

$$\mathrm{RMSE} = \left(\frac{1}{m} \sum_{(i,j)\in\Omega^c} (\widehat{a}_{ij} - a_{ij}^{\mathrm{true}})^2 \right)^{1/2}, \qquad \mathrm{MAE} = \frac{1}{m} \sum_{(i,j)\in\Omega^c} |\widehat{a}_{ij} - a_{ij}^{\mathrm{true}}|, \tag{4.1}$$

and, when informative, the minimum and maximum values over repeated masks.

4.4. Repeated masking and fair comparison

For each masking percentage, the experiment is repeated over independent random masks. At a fixed masking rate, the same mask bank is used for all compared methods. This prevents the comparison from being contaminated by different random missingness patterns across methods.

4.5. Hyperparameters

Any local-box width, ridge level, or other tuning parameter is chosen independently of the evaluation mask. A transparent default is

$$\eta = 10^{-6} \frac{\mathrm{tr}(S_0)}{p}$$

for covariance stabilization, together with a local radius of the form

$$\delta_k = \alpha \max\{1, |(x_L)_k|\},$$

where $\alpha > 0$ is fixed in advance. Ridge stabilization is also used in regression imputation to avoid unstable normal equations.

5. Results

5.1. Repeated MCAR imputation experiment on the red wine dataset

The proposed method was evaluated on the red wine quality dataset of Cortez et al. [2]. After removing the response variable "quality", the data matrix contained

$$X \in \mathbb{R}^{1599\times 11},$$

that is, 1599 observations and 11 physicochemical variables. Since the original matrix is complete, artificial missingness was introduced under an MCAR mechanism.

Three masking rates were considered:

$$5\%, \qquad 10\%, \qquad 20\%.$$

Since the matrix contains $1599 \times 11 = 17589$ entries, these percentages correspond to

$$879, \qquad 1759, \qquad 3518$$

masked values, respectively. For each masking level, 20 independent repetitions were performed. In every repetition, the same mask was used for all compared methods, ensuring a fair comparison.

The compared imputers were columnwise mean imputation, ridge-stabilized regression imputation, and the proposed Mahalanobis-based imputation method. For each repetition, the imputed values at the masked positions were compared against the true values using RMSE and MAE. Table 1 reports the average errors, standard deviations, and the minimum and maximum values over the 20 repetitions.

Table 1. Comparison of imputation methods on the red wine dataset under repeated MCAR masking.

Method	Rate	Mean RMSE	Std RMSE	Mean MAE	Std MAE	Min RMSE	Max RMSE	Min MAE	Max MAE
Mean	0.05	10.8632	1.3572	3.46164	0.308527	8.40325	13.1118	2.88395	3.88712
Regression	0.05	7.50771	1.04101	2.25273	0.218825	5.55434	9.20947	1.87577	2.53632
Mahalanobis	0.05	7.58160	1.10951	2.27471	0.227211	5.45821	9.33568	1.88708	2.59551
Mean	0.10	10.4437	0.932977	3.42256	0.247555	8.48634	11.9067	2.94430	3.89350
Regression	0.10	7.56230	0.679759	2.30068	0.154387	6.16593	8.67329	1.99225	2.55631
Mahalanobis	0.10	7.56837	0.806996	2.32573	0.171856	5.80464	8.88118	1.95330	2.60345
Mean	0.20	10.5069	0.564140	3.39329	0.116127	9.55654	11.2904	3.21773	3.52966
Regression	0.20	8.48272	0.543052	2.62145	0.154393	7.31429	9.40239	2.23514	2.84381
Mahalanobis	0.20	8.10188	0.505182	2.49399	0.083015	7.19624	8.94252	2.30402	2.61985

The results show that both regression and Mahalanobis imputation substantially outperform simple mean imputation at all masking levels. At 5% and 10% missingness, regression yields slightly lower RMSE and MAE values than the Mahalanobis method, although the two methods are close. At 20% missingness, however, the Mahalanobis method becomes the best-performing method among the three tested procedures, attaining both the lowest mean RMSE and the lowest mean MAE.

At the highest masking level, the Mahalanobis method reduces the mean RMSE from 10.5069 for mean imputation and 8.48272 for regression imputation to 8.10188. The mean MAE decreases from 3.39329 and 2.62145 to 2.49399, respectively. This suggests that covariance-aware geometric structure becomes increasingly useful as the proportion of missing entries grows.

5.2. Interpretive scope of the experiment

The repeated-mask wine experiment should be read as a controlled MCAR benchmark rather than as a claim of universal superiority across all missingness regimes. Its role is to show that, on a real continuous dataset of moderate dimension, the proposed Mahalanobis method is not merely algebraically well defined but also empirically competitive. In particular, the method remains close to regression imputation at low masking levels and becomes strongest among the tested methods at the largest masking level considered.

6. Discussion

6.1. What the method does well

The framework is deterministic, transparent, and strongly tied to the multivariate geometry induced by the covariance matrix. The linear stage has a clean algebraic core, and the nonlinear stage provides a natural refinement rather than a completely different principle. This makes the method attractive when

the user wants an interpretable completion rather than a black-box predictor.

6.2. What the method does not do

The method does not produce multiple completed datasets, posterior intervals, or a model for MNAR mechanisms. It is not intended to dominate deep-learning or ensemble methods on large heterogeneous benchmarks. Its intended domain is moderate continuous data with meaningful covariance structure.

6.3. Why the comparison class is intentionally classical

The numerical section emphasizes mean, regression, and covariance-based baselines because these methods are methodologically aligned with the proposed deterministic framework. The present paper is not a broad benchmark of all modern imputation algorithms; it is an analysis of a specific covariance-geometry principle. Broader comparisons with MissForest, MICE, EM, and deep generative methods are meaningful, but they belong to a larger benchmarking study.

6.4. Why the rank condition matters

Theorem 1 shows that the real mathematical issue is not merely the invertibility of the covariance matrix but the identifiability of the missing degrees of freedom through active Mahalanobis differences. The design matrix B makes this explicit. When B is rank-deficient, the observed pattern does not contain enough geometric information to determine a unique completion without regularization.

6.5. Noise sensitivity and structured masks

Noise and structured missingness can affect the method through the covariance estimate and the active-pair design. Large missing blocks reduce the number of informative contrasts between rows and can drive B toward rank deficiency. In such cases, the ridge-stabilized linear solution should be viewed as a regularized estimator rather than an exact geometry-preserving optimum.

6.6. Relation to the broader missing-data literature

Mean and regression remain important baselines because they test whether the covariance-geometry principle adds value beyond simpler deterministic rules. EM and MICE are natural bridges to the broader missing-data literature, especially when uncertainty quantification matters. The present method occupies a complementary position: it is a transparent structure-preserving tool for covariance-sensitive continuous data.

7. Conclusions

This paper developed a deterministic missing-value imputation framework based on preserving covariance geometry through Mahalanobis distances. The method contains two complementary components: a frozen-covariance linear imputer with a transparent quadratic structure, and a covariance-updating nonlinear refinement initialized by the linear solution. The compact matrix derivation shows that the linear stationarity system is well posed when the active-pair design matrix

has full column rank; otherwise, ridge or minimum-norm stabilization provides a practical fallback while revealing the underlying underidentification.

The main conceptual point is that covariance-based imputation should be treated as a structure-preserving deterministic tool, not as a universal replacement for probabilistic or machine-learning imputers. Its strength lies in moderate continuous data problems where empirical covariance geometry is itself the object one wants to preserve. Its limitations lie in underidentified masks, small-sample covariance instability, frozen-covariance bias, structured missing blocks, and the lack of uncertainty quantification.

The repeated MCAR experiment on the red wine dataset shows that the method is not only algebraically motivated but also empirically competitive. The Mahalanobis imputer substantially improves on mean imputation at all masking levels and becomes the strongest among the tested methods in the highest missingness regime considered.

Future work should expand the benchmarking study to stronger MAR scenarios, structured block masks, mixed data, EM and MICE comparisons, MissForest, and recent deep-learning imputers. Another important direction is to combine the deterministic Mahalanobis completion with uncertainty-aware procedures, thereby retaining geometric interpretability while adding inferential information.

Author contributions

A. H. S. S.: conceptualization, methodology, formal analysis, writing—original draft, writing—review and editing, supervision. D. L. O. R.: methodology, numerical implementation, validation, data curation, writing—review and editing. L. J. M. H.: formal analysis, validation, discussion of results, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors thank Universidad Nacional de Colombia and Universidad de Caldas for academic support. No specific external funding was received for this study.

Conflict of interest

The authors declare that they have no conflicts of interest.

Data availability

The numerical experiment uses the red wine quality dataset described by Cortez et al. [2]. The implementation details and generated masks can be made available by the corresponding author upon reasonable request.

References

1. E. J. Candès, B. Recht, Exact matrix completion via convex optimization, *Found. Comput. Math.*, **9** (2009), 717–772. http://doi.org/10.1007/s10208-009-9045-5
2. P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, Modeling wine preferences by data mining from physicochemical properties, *Decis. Support Syst.*, **47** (2009), 547–553. http://doi.org/10.1016/j.dss.2009.05.016
3. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc. B*, **39** (1977), 1–22. http://doi.org/10.1111/j.2517-6161.1977.tb01600.x
4. R. Gnanadesikan, *Methods for statistical data analysis of multivariate observations*, John Wiley & Sons, 1997. http://doi.org/10.1002/9781118032671
5. R. A. Johnson, D. W. Wichern, Applied multivariate statistical analysis, *Biometrics*, **44** (1988), 920.
6. R. Little, D. Rubin, *Statistical analysis with missing data*, John Wiley & Sons, 2019. http://doi.org/10.1002/9781119482260
7. P. C. Mahalanobis, On the generalized distance in statistics, *Proc. Natl. Inst. Sci. India*, **2** (1936), 49–55.
8. R. Mazumder, T. Hastie, R. Tibshirani, Spectral regularization algorithms for learning large incomplete matrices, *J. Mach. Learn. Res.*, **11** (2010), 2287–2322.
9. D. B. Rubin, *Multiple imputation for nonresponse in surveys*, John Wiley & Sons, 1987. http://doi.org/10.1002/9780470316696
10. D. J. Stekhoven, P. Bühlmann, MissForest-non-parametric missing value imputation for mixed-type data, *Bioinformatics*, **28** (2012), 112–118. http://doi.org/10.1093/bioinformatics/btr597
11. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, et al., Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17** (2001), 520–525. http://doi.org/10.1093/bioinformatics/17.6.520
12. S. van Buuren, *Flexible imputation of missing data*, New York: Chapman & Hall/CRC, 2018. http://doi.org/10.1201/9780429492259
13. J. Yoon, J. Jordon, M. van der Schaar, GAIN: Missing data imputation using generative adversarial nets, In: *Proceedings of the 35th International Conference on Machine Learning*, **80** (2018), 5689–5698.
14. V. Fortuin, D. Baranchuk, G. Rätsch, S. Mandt, GP-VAE: Deep probabilistic multivariate time series imputation, In: *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, **108** (2020), 1651–1661.
15. Y. Tashiro, J. Song, Y. Song, S. Ermon, CSDI: Conditional score-based diffusion models for probabilistic time series imputation, In: *Advances in Neural Information Processing Systems*, **34** (2021), 24804–24816.
16. W. Du, D. Côté, Y. Liu, SAITS: Self-attention-based imputation for time series, *Expert Syst. Appl.*, **219** (2023), 119619. http://doi.org/10.1016/j.eswa.2023.119619

17. T. Du, L. Melis, T. Wang, ReMasker: Imputing tabular data with masked autoencoding, In: *The Twelfth International Conference on Learning Representations*, 2024.

18. J. Wang, W. Du, Y. Yang, L. Qian, W. Cao, K. Zhang, et al., Deep learning for multivariate time series imputation: A survey, In: *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025, 10696–10704. http://doi.org/10.24963/ijcai.2025/1187

Appendix

A. Compact derivation of the linear system

For each active pair (i, h) define the affine row difference

$$z_{ih}(x) = d_{ih} + B_{ih}x.$$

Then

$$\Phi_L(x) = \sum_{(i,h)\in\mathcal{A}} z_{ih}(x)^T Q z_{ih}(x).$$

Stacking all active pairs gives

$$z(x) = d + Bx,$$

and therefore

$$\Phi_L(x) = (d + Bx)^T (I_{|\mathcal{A}|} \otimes Q)(d + Bx).$$

Expanding yields

$$\Phi_L(x) = d^T (I \otimes Q)d + 2d^T (I \otimes Q)Bx + x^T B^T (I \otimes Q)Bx,$$

which is exactly the quadratic form used in Proposition 1 after collecting the constant, linear, and quadratic terms.