Efficient Algorithms for Nearest Correlation Matrix Computation with Missing Data

Ibrahim Oyeyinka

Department of Mathematical Sciences
Northern Illinois University

Thesis Director: Prof. Nathan Krislock — October 13, 2025

Masters Thesis Defense

Abstract Overview

This thesis investigates *efficient algorithms* for computing the nearest correlation matrix (NCM) with missing data. Empirical estimates often violate PSD and unit-diagonal constraints. We compare **Modified Alternating Projections (MAP)** and **Anderson Acceleration (AA)** and analyze how **MCAR** and **NMAR** mechanisms affect performance and reliability. The results provide practical guidance for financial practitioners.

Introduction and Context

Background and Context

Role in Finance

- Correlation matrices underpin risk management, portfolio optimization, derivative pricing, fraud detection, and regulatory compliance.

Modern Portfolio Theory (MPT)

- Low/negative correlations reduce portfolio risk without reducing expected return.

Data Challenge

- Missing/asynchronous observations (non-trading days, delistings, reporting gaps, corporate actions) yield invalid empirical matrices.

Problem Statement: Invalid Correlation Matrices

Goal: Reconstruct a valid correlation matrix X from imperfect A.

- Validity: $X = X^\top$, $X \succeq 0$, $X_{ii} = 1$.
- Naive fixes can still violate these, leading to poor risk estimates and decisions.
- Seek the nearest valid X to A under a chosen norm.

Research Objectives

Objective 1: Algorithm Comparison

- **MAP:** alternating projections onto $\mathcal{S}_+^n$ and $\mathcal{U}^n$ with Dykstra's correction.
- **AA:** accelerates convergence using prior iterates/residuals.
- Metrics: efficiency, speed, accuracy, robustness, scalability.

Objective 2: Missing Data Analysis

- Examine MCAR (scattered) vs. NMAR (structured) on estimation and algorithm behavior.
- Provide practical recommendations for finance.

Core Definitions: Correlation Matrix

For a, b:

$$\rho_{ab} = \frac{\mathrm{Cov}(a,b)}{\sigma_a \sigma_b} \in [-1, 1].$$

Correlation matrix $X \in \mathbb{R}^{n \times n}$:

1. $X = X^\top$ (symmetric)
2. $X \succeq 0$ (PSD)
3. $X_{ii} = 1$ (unit diagonal)

Feasible sets:

$$\mathcal{S}_+^n = \{Y = Y^\top : \; Y \succeq 0\}, \quad \mathcal{U}^n = \{Y = Y^\top : \; \mathrm{diag}(Y) = e\}.$$

Nearest Correlation Matrix (NCM) Problem

$$\min_{X \in \mathcal{S}_+^n \cap \mathcal{U}^n} \|A - X\|_F, \qquad \|A\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2.$$

Compute the minimal distance $\gamma(A)$ and the minimizer X in $\mathcal{S}_+^n \cap \mathcal{U}^n$.

Weighted Norms for NCM

W**-norm:**

$$\|A\|_W = \left\|W^{1/2} A W^{1/2}\right\|_F, \quad W \succ 0$$

Preserves inertia under congruence; emphasizes rows/cols via W; fits dense financial data. (*Adopted in this thesis.*)

H**-norm:**

$$\|A\|_H = \|H \circ A\|_F, \quad H \geq 0$$

Entrywise weighting (Hadamard); useful for sparse/partial/confidence-weighted observations.

Literature: NCM Methods & Acceleration

MAP (Higham): alternating projections onto $\mathcal{S}_+^n$ and $\mathcal{U}^n$; converges in Frobenius norm; *linear* rate can be slow for large/ill-conditioned cases. Dykstra's correction stabilizes by compensating projection bias.

Newton-type: Qi & Sun; refined by Borsdorf & Higham offers faster convergence, higher complexity.

Anderson Acceleration (AA): leverages past iterates/residuals for fixed-point acceleration; often cuts iterations by factors of 2–3+ with minimal overhead. Widely effective in practice.

Theoretical Background

Convex Characterization of the NCM Problem

Minimize $\|A - X\|_W$ over $X \in \mathcal{S}_+^n \cap \mathcal{U}^n$, where

$$\|A\|_W = \|W^{1/2} A W^{1/2}\|_F, \quad W \succ 0.$$

Optimality condition:

$$\langle Z - X, A - X \rangle_W \leq 0, \quad \forall Z \in \mathcal{S}_+^n \cap \mathcal{U}^n,$$

which implies

$$A - X \in \partial(\mathcal{S}_+^n \cap \mathcal{U}^n)(X).$$

Normal cone sum rule:

$$\partial(\mathcal{S}_+^n \cap \mathcal{U}^n)(X) = \partial\mathcal{S}_+^n(X) + \partial\mathcal{U}^n(X),$$

valid because $\mathrm{ri}(\mathcal{S}_+^n) \cap \mathrm{ri}(\mathcal{U}^n) \neq \emptyset$.

Lemma: Normal Cone of $\mathcal{U}^n$

For $A \in \mathcal{U}^n$ and $W \succ 0$:

$$\boxed{\partial\mathcal{U}^n(A) = \{\, W^{-1}\,\mathrm{Diag}(\theta)\,W^{-1} :\ \theta \in \mathbb{R}^n \,\}.}$$

Idea:

- If WYW had any nonzero off-diagonal entry, then $\langle Z - A, Y\rangle_W$ could become unbounded for feasible $Z \in \mathcal{U}^n$.
- To keep the supremum finite, WYW must therefore be diagonal.
- Hence every normal element has the form $Y = W^{-1}\,\mathrm{Diag}(\theta)\,W^{-1}$ for some $\theta \in \mathbb{R}^n$.

Lemma: Normal Cone of $\mathcal{S}_+^n$

For $A \in \mathcal{S}_+^n$:

$$\boxed{\partial \mathcal{S}_+^n(A) = \{\, Y = Y^\top : \; Y \preceq 0, \; \langle Y, A \rangle_W = 0 \,\}.}$$

Idea:

- If $Y \not\preceq 0$, then

$$\sup_{Z \succeq 0} \langle Y, Z \rangle_W = +\infty$$

  (by scaling along a positive eigenvector of Y), so Y cannot lie in the normal cone.

- If $Y \preceq 0$, the supremum is finite and $\langle Y, A \rangle_W = 0$ follows from orthogonality between the supports of Y and A.

Corollary: Structure of $\partial\mathcal{S}_+^n(A)$

Statement: For $A \in \mathcal{S}_+^n$, the elements of the normal cone are

$$\partial\mathcal{S}_+^n(A) = \{\, Y : \; WYW = -VDV^\top,\; D = \mathrm{Diag}(d_i) \succeq 0 \,\},$$

where $V \in \mathbb{R}^{n\times p}$ has orthonormal columns spanning $\mathrm{null}(A)$.

Idea: Let $A = Q\Lambda Q^\top$ with $Q = [Q_1, Q_2]$ partitioned so that Q_1 corresponds to positive eigenvalues and Q_2 spans $\mathrm{null}(A)$. From $\langle Y, A\rangle_W = 0$ and $Y \preceq 0$,

$$Q^\top(WYW)Q = \begin{bmatrix} G & H \\ H^\top & M \end{bmatrix} \preceq 0, \quad \text{with } G = 0,\; H = 0,\; M \preceq 0.$$

Hence $WYW = Q_2 M Q_2^\top = -VDV^\top$ for some diagonal $D \succeq 0$.

Theorem: Solution Characterization

X solves

$$\min_{X \in \mathcal{S}_+^n \cap \mathcal{U}^n} \|A - X\|_W \quad \Longleftrightarrow \quad X = A + W^{-1}\big(VDV^\top + \mathrm{Diag}(\theta)\big)\, W^{-1},$$

where V spans null(X), $D \succeq 0$, and $\theta \in \mathbb{R}^n$.

Proof Sketch:

- From optimality: $A - X = Y_1 + Y_2$ with $Y_1 \in \partial\mathcal{S}_+^n(X)$ and $Y_2 \in \partial\mathcal{U}^n(X)$.
- Substitute the normal cone expressions: $Y_1 = W^{-1}V(-D)V^\top W^{-1}$ and $Y_2 = W^{-1}\mathrm{Diag}(\theta)\, W^{-1}$.
- Combine terms and rearrange to obtain the stated form of X.

Constraints and Eigenvalue Analysis

Assume $A = A^\top$, $a_{ii} \geq 1$, and W is diagonal.

- **Diagonal correction:** From the unit-diagonal constraint, the diagonal multipliers θ_i must satisfy

$$\theta_i \leq 0,$$

  ensuring that x_{ii} is reduced (or unchanged) to enforce $\mathrm{diag}(X) = \mathbf{e}$.
- **Eigenvalue structure:** If A has t nonpositive eigenvalues, each iterate $R_k = A + \Delta_k$ and the solution X retain at least t zero eigenvalues. Negative directions cannot be "recovered" under the PSD constraint.

Theorem: Projection onto $\mathcal{S}^n_+$

For symmetric A and $W \succ 0$,

$$\boxed{P_{\mathcal{S}^n_+}(A) = W^{-1/2}\Big((W^{1/2}AW^{1/2})_+\Big)W^{-1/2}.}$$

Optimality Conditions:

$$A - X \preceq 0, \qquad \mathrm{trace}\big((A-X)WXW\big) = 0.$$

Proof Sketch:

- Let $B = W^{1/2}AW^{1/2} = B_+ - B_-$ with $B_\pm \succeq 0$.
- Set $X = W^{-1/2}B_+W^{-1/2}$.
- Then $W^{1/2}(A-X)W^{1/2} = -B_- \preceq 0$, and $B_-B_+ = 0$ ensures orthogonality.

Theorem: Projection onto $\mathcal{U}^n$

$$\boxed{P_{\mathcal{U}^n}(A) = A - W^{-1}\mathrm{Diag}(\theta)\, W^{-1}, \qquad (W^{-1} \circ W^{-1})\,\theta = \mathrm{diag}(A - I).}$$

Proof Sketch:

- From optimality, $A - X = W^{-1}\mathrm{Diag}(\theta)\, W^{-1}$ for some $\theta \in \mathbb{R}^n$.
- Enforcing the unit-diagonal constraint $x_{ii} = 1$ gives $(W^{-1} \circ W^{-1})\,\theta = \mathrm{diag}(A - I)$.
- Since $W \succ 0 \Rightarrow W^{-1} \succ 0$, the Hadamard product $W^{-1} \circ W^{-1} \succ 0$, so the linear system admits a unique solution θ.

Methodology and Algorithms

Modified Alternating Projections (MAP)

Algorithm:

1. **Input:** Matrix $A \in \mathbb{R}^{n \times n}$, tolerance.
2. **Correction:** $R_k \leftarrow Y_{k-1} - \Delta S_{k-1}$ (Dykstra term)
3. **Projection 1:** $X_k \leftarrow P_{\mathcal{S}_+^n}(R_k)$
4. **Update ΔS:** $\Delta S_k \leftarrow X_k - R_k$
5. **Projection 2:** $Y_k \leftarrow P_{\mathcal{U}^n}(X_k)$
6. Stop when $\|Y_k - X_k\| / \|Y_k\| \leq \texttt{tolerance}$.

Converges to the unique nearest correlation matrix in $\mathcal{S}_+^n \cap \mathcal{U}^n$ (weighted Frobenius).

MAP: Convergence and Efficiency

Assume $A = A^\top$, $a_{ii} \geq 1$, and W is diagonal.

Iteration structure:

$$R_k = A + \Delta_k, \qquad \Delta_k = \sum_{i=1}^{k-1} D_i \preceq 0 \text{ (diagonal)}.$$

Eigenvalue persistence: If A has t nonpositive eigenvalues, then each R_k has at least t, so $P_{\mathcal{S}_+^n}(R_k)$ has at least t zero eigenvalues.

Computational note: For large-scale settings, compute only the top $n-t$ positive eigenpairs of $W^{1/2} R_k W^{1/2}$ (via Lanczos or tridiagonalization) to reduce the cost of the PSD projection.

Anderson Acceleration (AA)

View: MAP as a fixed-point:

$$Z_{k+1} = g(Z_k), \qquad f(z) = \mathrm{vec}(\tilde{g}(Z)) - z,$$

where $Z_k = [Y_k, \Delta S_k]$, $z_k = \mathrm{vec}(Z_k)$.
$g(Z)$ performs one full MAP step; $\tilde{g}(Z)$ is its vectorized form.

Algorithm:

1. Compute residual $f_k = f(z_k)$.
2. Form $X_k = [\Delta z_{k-m_k}, \ldots, \Delta z_{k-1}]$, $F_k = [\Delta f_{k-m_k}, \ldots, \Delta f_{k-1}]$.
3. Solve LS: $\gamma^{(k)} = \arg\min_\gamma \|f_k - F_k\gamma\|_2$.
4. Update: $z_{k+1} = z_k - X_k\gamma^{(k)} + f_k - F_k\gamma^{(k)}$.
5. Stop when $\|Y_k - X_k\|_2 / \|Y_k\|_2 \leq \texttt{tol}$.

Notes: Small history $(m \leq 5) \Rightarrow \mathcal{O}(n^2)$ cost vs. $\mathcal{O}(n^3)$ eigensolve. Cuts MAP iterations by $5-10\times$ with same accuracy.

Missing Data Mechanisms (Rubin, 1976)

Mechanism	**Definition**	**Financial Example**
MCAR	Missingness is independent of $(Y_{\text{obs}}, Y_{\text{mis}})$	Random outages or technical errors
MAR	Missingness depends only on Y_{obs} (after conditioning)	Reporting tied to observed firm attributes
NMAR	Missingness depends on Y_{mis} even after conditioning	Suppression of extreme correlations

Focus here: MCAR and NMAR.

- **Stock Selection:**
  - Dataset of 550 global equities (2020–2025), containing approximately 768,900 observations.
  - Balanced sector representation: 50 stocks randomly selected from each of 11 GICS sectors.
  - Equal weighting scheme applied (no market-cap bias).

Missing Data Simulation: MCAR

Deletion probability $p \in [0.05, 0.50]$. Randomly remove off-diagonals symmetrically.

Example (6×6**,** $p{=}0.4$**; 12 missing off-diagonals):**

$$\begin{bmatrix} 1 & \text{NaN} & 0.12 & 0.08 & \text{NaN} & 0.31 \\ \text{NaN} & 1 & \text{NaN} & 0.15 & 0.26 & \text{NaN} \\ 0.12 & \text{NaN} & 1 & 0.17 & \text{NaN} & 0.22 \\ 0.08 & 0.15 & 0.17 & 1 & 0.29 & \text{NaN} \\ \text{NaN} & 0.26 & \text{NaN} & 0.29 & 1 & 0.19 \\ 0.31 & \text{NaN} & 0.22 & \text{NaN} & 0.19 & 1 \end{bmatrix}$$

Scattered/noisy missingness pattern.

Missing Data Simulation: NMAR

Target large $|A_{ij}|$ until deletion rate p is met; set symmetric pairs missing.

Example (6×6, threshold τ=0.2, 12 missing):

$$\begin{bmatrix} 1 & \text{NaN} & 0.12 & 0.08 & \text{NaN} & \text{NaN} \\ \text{NaN} & 1 & \text{NaN} & 0.15 & \text{NaN} & \text{NaN} \\ 0.12 & \text{NaN} & 1 & 0.17 & \text{NaN} & 0.22 \\ 0.08 & 0.15 & 0.17 & 1 & \text{NaN} & \text{NaN} \\ \text{NaN} & \text{NaN} & \text{NaN} & \text{NaN} & 1 & 0.19 \\ \text{NaN} & \text{NaN} & 0.22 & \text{NaN} & 0.19 & 1 \end{bmatrix}$$

Structured/clustered pattern near stronger entries.

Results, Conclusions, and Future Work

Results — MAP vs. AA (Table 6.2)

Example	**Method**	**Iter**	**Time (s)**	$\|\boldsymbol{A}-\boldsymbol{X}\|_F$
Higham 4×4	MAP	19	0.00104	2.133
	AA	3	0.00601	2.251
Toeplitz 6×6	MAP	27	0.00159	0.369
	AA	4	0.00477	0.402
Real 550×550 (1% MCAR)	MAP	13	0.656	0.116
	AA	3	0.686	0.124
Real 550×550 (1% NMAR)	MAP	19	0.914	0.0392
	AA	3	1.18	0.0412
Real 550×550 (25% MCAR)	MAP	84	6.18	4.832
	AA	**10**	**2.26**	5.072
Real 550×550 (25% NMAR)	MAP	72	4.75	3.558
	AA	**7**	**1.88**	3.734

Convergence: AA slashes iterations (88% fewer at 25% MCAR; 90% fewer at 25% NMAR).
Scalability: At higher missingness on large n, AA is 60–63% faster than MAP.

Results — AA under MCAR vs. NMAR (Table 6.3)

Mechanism	**Missing** (%)	**Iterations**	**Time**(s)	$\|A - X\|_F$
MCAR	5 %	4	1.07	0.827
MCAR	10 %	7	1.41	1.647
MCAR	20 %	7	1.50	3.369
MCAR	**30** %	**30**	**5.46**	**6.342**
MCAR	40 %	46	7.87	9.998
MCAR	50 %	55	9.06	14.845
NMAR	5 %	5	1.18	0.921
NMAR	10 %	5	1.28	1.703
NMAR	20 %	7	1.40	2.717
NMAR	**30** %	**10**	**1.83**	**4.961**
NMAR	40 %	38	6.51	8.211
NMAR	50 %	72	12.8	12.881

Analysis: AA converges markedly faster under NMAR at moderate rates (e.g., 30%: 10 vs 30 iter). Outputs are typically well-conditioned (min eigenvalues near 0^+). At 40–50% missingness, both runtime and reconstruction error rise for obvious reasons (information loss).

Implications & Future Research

Implications for Finance:

- Anderson Acceleration (AA) yields faster and more robust NCM reconstructions as both dimension and missingness increase.
- Improved structural integrity (preserved null eigenvalues) enhances portfolio stability and risk estimation.

Future Directions:

- Incorporate **MAR** mechanisms through explicit dependency modeling.
- Analyze how the *eigenvalue spectrum* affects portfolio optimization and the shape of the efficient frontier.
- Validate at scale (millions of observations, thousands of assets) to test frontier robustness under missing data.

Thank You / Questions

Ibrahim Oyeyinka
Masters Thesis Defense
Department of Mathematical Sciences
Northern Illinois University

Thesis Director: Prof. Nathan Krislock

Date: October 13, 2025

GitHub Repository:

Scan to view the code & data