EM-Based Transfer Learning for Gaussian Causal Models Under Domain Shift: Supplementary Materials

Mohammad Ali Javidian
Computer Science Department
Appalachian State University
Boone, USA
javidianma@appstate.edu

I. Population–EM Operator under a Known Causal DAG

We begin in the infinite–data regime, free of sampling noise, and assume that the true joint law of $X = (X_1, \dots, X_p)^\top$ is linear–Gaussian with covariance $\Sigma^* \in \mathbb{R}^{p \times p}$. Let $T = X_t$ denote the coordinate that is unobserved in the target domain, so that we may write $X = (X_{-t}, T)$.

For a single observation X, the negative complete–data log–likelihood is

$$\ell_{\text{comp}}(X;\Sigma) \;=\; \tfrac{1}{2}\Big[\log\det\Sigma \;+\; \operatorname{trace}(\Sigma^{-1}XX^\top)\Big], \qquad \Sigma \succ 0. \tag{1}$$

When data have missing entries, the EM algorithm maximizes the observed–data likelihood by iteratively maximizing a surrogate function. Given a dataset with missing T, this surrogate is

$$Q_N(\Sigma \mid \Sigma^{(r)}) \;:=\; -\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\Big[\ell_{\text{comp}}\big(X^{(i)};\Sigma\big) \;\Big|\; X_{-t}^{(i)},\, \Sigma^{(r)}\Big],$$

where the inner expectation is taken with respect to the conditional law of the missing coordinate(s) given the observed part $X_{-t}^{(i)}$ under the current iterate $\Sigma^{(r)}$. In the infinite–sample limit, the empirical average is replaced by an expectation over the *true* marginal distribution of the observed block $X_{-t} \sim \mathcal{N}(0, \Sigma^*_{-t,-t})$, leading to

$$Q_{\text{pop}}(\Sigma \mid \Sigma^{(r)}) \;:=\; -\mathbb{E}_{X_{-t}}\Big[\mathbb{E}\Big[\ell_{\text{comp}}(X;\Sigma) \;\Big|\; X_{-t},\, \Sigma^{(r)}\Big]\Big]. \tag{2}$$

The outer expectation averages over what is actually observed (the true X_{-t}), while the inner expectation integrates out the missing coordinate according to the working model induced by the current iterate $\Sigma^{(r)}$.

To evaluate the inner expectation, partition the covariance matrix as

$$\Sigma^{(r)} = \begin{pmatrix} \Sigma^{(r)}_{-t,-t} & \Sigma^{(r)}_{-t,t} \\ \Sigma^{(r)}_{t,-t} & \Sigma^{(r)}_{tt} \end{pmatrix}.$$

For fixed x_{-t}, the conditional law of T given $X_{-t} = x_{-t}$ under $\Sigma^{(r)}$ is Gaussian with mean

$$\mu_t^{(r)}(x_{-t}) \;=\; \Sigma^{(r)}_{t,-t}\big(\Sigma^{(r)}_{-t,-t}\big)^{-1}x_{-t}, \qquad V_t^{(r)} \;=\; \Sigma^{(r)}_{tt} - \Sigma^{(r)}_{t,-t}\big(\Sigma^{(r)}_{-t,-t}\big)^{-1}\Sigma^{(r)}_{-t,t}. \tag{3}$$

Using (1), the conditional second moment of X is then

$$\mathbb{E}\Big[XX^\top \;\Big|\; X_{-t} = x_{-t},\, \Sigma^{(r)}\Big] \;=\; \begin{pmatrix} x_{-t}x_{-t}^\top & x_{-t}\,\mu_t^{(r)}(x_{-t}) \\ \mu_t^{(r)}(x_{-t})\,x_{-t}^\top & V_t^{(r)} + \big(\mu_t^{(r)}(x_{-t})\big)^2 \end{pmatrix}. \tag{4}$$

Plugging this into (2) and exploiting linearity of expectation shows that

$$\mathbb{E}\Big[\ell_{\text{comp}}(X;\Sigma) \;\Big|\; X_{-t}, \Sigma^{(r)}\Big] \;=\; \tfrac{1}{2}\Big[\log\det\Sigma + \operatorname{trace}\Big(\Sigma^{-1}\,\mathbb{E}[XX^\top \mid X_{-t}, \Sigma^{(r)}]\Big)\Big].$$

Averaging over the true marginal of X_{-t} defines the filled–in second moment

$$M^{(r)} \;:=\; \mathbb{E}_{X_{-t}}\Big[\mathbb{E}\Big[XX^\top \;\Big|\; X_{-t},\, \Sigma^{(r)}\Big]\Big]. \tag{5}$$

Letting $a^{(r)} := \big(\Sigma^{(r)}_{-t,-t}\big)^{-1}\Sigma^{(r)}_{-t,t}$ and recalling that $X_{-t} \sim \mathcal{N}(0, \Sigma^*_{-t,-t})$, one computes

$$\mathbb{E}[X_{-t}X_{-t}^\top] = \Sigma^*_{-t,-t}, \quad \mathbb{E}\big[X_{-t}\,\mu_t^{(r)}(X_{-t})\big] = \Sigma^*_{-t,-t}\,a^{(r)}, \quad \mathbb{E}\big[(\mu_t^{(r)}(X_{-t}))^2\big] = a^{(r)\top}\Sigma^*_{-t,-t}a^{(r)},$$

so that

$$M^{(r)} \;=\; \begin{pmatrix} \Sigma^*_{-t,-t} & \Sigma^*_{-t,-t}\, a^{(r)} \\ a^{(r)\top}\Sigma^*_{-t,-t} & V_t^{(r)} + a^{(r)\top}\Sigma^*_{-t,-t}a^{(r)} \end{pmatrix}. \tag{6}$$

With this notation, the population Q–function (2) simplifies to

$$Q_{\mathrm{pop}}(\Sigma \mid \Sigma^{(r)}) \;=\; -\tfrac{1}{2}\Big[\log\det\Sigma \;+\; \mathrm{trace}\big(\Sigma^{-1}M^{(r)}\big)\Big]. \tag{7}$$

Differentiating shows that its gradient is

$$\nabla_\Sigma Q_{\mathrm{pop}}(\Sigma \mid \Sigma^{(r)}) \;=\; -\tfrac{1}{2}\Big[\Sigma^{-1} - \Sigma^{-1}M^{(r)}\Sigma^{-1}\Big],$$

and setting this to zero yields $\Sigma = M^{(r)}$. Because (7) is strictly convex in $\Sigma \succ 0$, this stationary point is the unique minimizer. Thus the population EM operator is

$$F\big(\Sigma^{(r)}\big) \;:=\; \Sigma^{(r+1)} \;=\; M^{(r)} \;=\; \mathbb{E}_{X_{-t}}\Big[\mathbb{E}\Big[XX^\top \;\Big|\; X_{-t},\, \Sigma^{(r)}\Big]\Big]. \tag{8}$$

In words, each population EM step replaces the current covariance guess $\Sigma^{(r)}$ by the exact conditional second moment of X (computed under $\Sigma^{(r)}$), averaged over the true marginal of the observed block X_{-t}.

Finally, when X has nonzero mean μ^*, the complete–data objective is written in centered form as

$$\ell_{\mathrm{comp}}(X;\mu,\Sigma) = \tfrac{1}{2}\Big[\log\det\Sigma + (X-\mu)^\top\Sigma^{-1}(X-\mu)\Big].$$

With current iterate $(\mu^{(r)}, \Sigma^{(r)})$, the conditional mean becomes

$$\mu_t^{(r)}(x_{-t}) = \mu_t^{(r)} + \Sigma^{(r)}_{t,-t}\big(\Sigma^{(r)}_{-t,-t}\big)^{-1}\big(x_{-t} - \mu^{(r)}_{-t}\big),$$

and the corresponding sufficient statistic is the centered second moment

$$M^{(r)} \;=\; \mathbb{E}_{X_{-t}}\Big[\mathbb{E}\Big[(X-\mu^{(r)})(X-\mu^{(r)})^\top \;\Big|\; X_{-t},\, \Sigma^{(r)}\Big]\Big].$$

The covariance update remains $\Sigma^{(r+1)} = M^{(r)}$, while the mean updates according to

$$\mu^{(r+1)}_{-t} \;=\; \mathbb{E}[X_{-t}] \;=\; \mu^*_{-t}, \qquad \mu^{(r+1)}_t \;=\; \mu^{(r)}_t + \Sigma^{(r)}_{t,-t}\big(\Sigma^{(r)}_{-t,-t}\big)^{-1}\big(\mu^*_{-t} - \mu^{(r)}_{-t}\big),$$

which is simply the conditional–mean rule averaged over the true marginal of X_{-t}.

II. First-Order (Partial) M–Step via Gradient–EM

In the finite–sample setting, the E–step produces the empirical completed covariance

$$\widehat{Q}^{(r)} \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[X^{(i)}X^{(i)\top} \;\Big|\; X^{(i)}_{-t},\, \Sigma^{(r)}\Big], \tag{9}$$

which is the analogue of the population moment $M^{(r)}$ in (5). The M–step then updates the DAG parameters (A, Δ) so that the implied covariance $\Sigma(A, \Delta)$ better matches $\widehat{Q}^{(r)}$.

To recall, a Gaussian DAG model with coefficient matrix A and noise variances $\Delta = (\Delta_1, \ldots, \Delta_p)^\top$ induces the precision

$$K(A,\Delta) = A^\top\Lambda A, \qquad \Lambda = \mathrm{diag}(1/\Delta), \qquad \Sigma(A,\Delta) = K(A,\Delta)^{-1}.$$

The corresponding finite–sample Q–function is

$$\mathcal{L}(A,\Delta \mid \widehat{Q}^{(r)}) = -\tfrac{n}{2}\Big[\log\det K(A,\Delta) + \mathrm{trace}\big(K(A,\Delta)\widehat{Q}^{(r)}\big) + \sum_{i=1}^{p}\log\Delta_i\Big], \tag{10}$$

and maximizing this expression corresponds to solving p generalized least–squares (GLS) regressions, which costs $O(p^3)$ per iteration.

Rather than fully maximizing (10), we adopt a cheaper strategy: take only a single gradient–ascent step in (A, Δ). This reduces the cost to $O(p^2)$ while still ensuring ascent in $\mathcal{L}$, consistent with the general gradient–EM theory of [1]. The gradients take simple forms. For A, we use the facts that $\partial\log\det K/\partial K = K^{-1}$, $\partial\,\mathrm{trace}(K\widehat{Q})/\partial K = \widehat{Q}$, and $\partial K/\partial A = 2\Lambda A$, which yield

$$G_A = n\,\Lambda A\big(K^{-1} + \widehat{Q}^{(r)}\big), \tag{11}$$

restricted to the structural nonzeros $j \in \mathrm{pa}(i)$.

For Δ, let $a_i = A_{\cdot,i}$ denote the ith column of A. Then $\partial K/\partial\Delta_i = -\Delta_i^{-2}\, a_i a_i^\top$, which implies

$$\frac{\partial}{\partial\Delta_i}\log\det K = -\Delta_i^{-2}\, a_i^\top K^{-1} a_i, \qquad \frac{\partial}{\partial\Delta_i}\operatorname{trace}(K\widehat{Q}) = -\Delta_i^{-2}\, a_i^\top \widehat{Q}\, a_i,$$

and of course $\partial\sum_j \log\Delta_j/\partial\Delta_i = 1/\Delta_i$. Putting these together gives

$$\frac{\partial\mathcal{L}}{\partial\Delta_i} = -\tfrac{n}{2}\Big[-\Delta_i^{-2}(a_i^\top K^{-1} a_i) - \Delta_i^{-2}(a_i^\top \widehat{Q} a_i) + \tfrac{1}{\Delta_i}\Big]. \tag{12}$$

At this point a key identity simplifies matters: since $K = A^\top \Lambda A$ and $\Sigma = K^{-1} = A^{-1}\Lambda^{-1}A^{-\top}$, one finds

$$a_i^\top \Sigma\, a_i = \Delta_i, \qquad \boxed{a_i^\top K^{-1} a_i = \Delta_i.}$$

Substituting this into (12), the terms involving $\pm\Delta_i^{-1}$ cancel, leaving the compact expression

$$G_{\Delta_i} = \tfrac{n}{2}\,\Delta_i^{-2}\, a_i^\top \widehat{Q}\, a_i. \tag{13}$$

Unlike earlier heuristics, this is not simply $[\widehat{Q}]_{ii}$, but rather a quadratic form of $\widehat{Q}$ along the structural column a_i.

The resulting first–order update rule at iteration r is therefore

$$G_A = N_t\, \Lambda^{(r)} A^{(r)} \big(\Sigma^{(r)} + \widehat{Q}^{(r)}\big), \qquad G_{\Delta_i} = \tfrac{N_t}{2}\,(\Delta_i^{(r)})^{-2}\,(a_i^{(r)})^\top \widehat{Q}^{(r)} a_i^{(r)}.$$

With step sizes $\eta_A, \eta_\Delta > 0$, we set

$$\widetilde{A} = A^{(r)} + \eta_A G_A, \qquad \widetilde{\Delta}_i = \Delta_i^{(r)} + \eta_\Delta G_{\Delta_i}.$$

Sparsity is enforced by zeroing out entries $\widetilde{A}_{ij}$ whenever $j \notin \mathrm{pa}(i)$, and positivity by projecting each $\widetilde{\Delta}_i$ onto $[10^{-6}, \infty)$. The updated covariance is then

$$A^{(r+1)} = \widetilde{A}, \quad \Delta^{(r+1)} = \widetilde{\Delta}, \quad \Sigma^{(r+1)} = \big[A^{(r+1)\top}\operatorname{diag}(1/\Delta^{(r+1)})A^{(r+1)}\big]^{-1}.$$

In summary, this partial M–step performs a cheap $O(p^2)$ update in (A, Δ), in contrast to solving p GLS regressions at $O(p^3)$. General results on gradient–EM [1] guarantee monotone ascent and, under suitable conditions, geometric convergence. In practice, one may occasionally re–solve the full GLS subproblem (e.g. every 50–100 iterations) to re–project onto the DAG covariance manifold and maintain numerical stability.

III. Domain-Adaptive EM

We now describe how to leverage both fully observed *source* data and partially observed *target* data within a unified EM framework. The key idea is to first identify which conditional parameters of the DAG remain invariant across domains (from the source fit), then perform EM updates in the target domain only on the subset of parameters affected by the shift.

Let $\{X_\mathrm{s}^{(i)}\}_{i=1}^{N_s}$ be N_s i.i.d. samples from the source distribution, fully observed on all p variables. The empirical second moment is

$$S_\mathrm{s} = \frac{1}{N_s}\sum_{i=1}^{N_s} X_\mathrm{s}^{(i)} X_\mathrm{s}^{(i)\top}. \tag{14}$$

Using the GLS fitting procedure (Sec. II) we estimate the DAG parameters:

$$(A^{(\mathrm{s})}, \Delta^{(\mathrm{s})}) = \mathrm{FitDAG}(A, S_\mathrm{s}, N_s), \tag{15}$$

which yields structural coefficients $A^{(\mathrm{s})}$ and noise variances $\Delta^{(\mathrm{s})}$ consistent with the known DAG. We then partition these parameters into *invariant* versus *shifted* blocks, depending on the type of domain shift: under **covariate shift**, the structural equations remain identical across domains, i.e. $(A^{(\mathrm{s})}, \Delta^{(\mathrm{s})})$ are frozen in their entirety; under **target shift**, only the conditional distribution of the target node T may differ, so $(A^{(\mathrm{s})}_{-t,-t}, \Delta^{(\mathrm{s})}_{-t})$ are fixed, while the parameters $(A_{t,\cdot}, \Delta_t)$ may adapt to the target.

Turning to the target domain, let $\{X_\mathrm{t}^{(j)}\}_{j=1}^{N_t}$ be N_t i.i.d. samples in which T is missing and X_{-t} is observed. At iteration r, we alternate an E–step and an M–step that exploit the known DAG structure and the invariances identified from the source. In the E–step, as in Sec. I, we impute the completed covariance by conditioning on $X_{-t,\mathrm{t}}^{(j)}$ under $\Sigma^{(r)}$:

$$Q_\mathrm{tgt}^{(r)} = \frac{1}{N_t}\sum_{j=1}^{N_t}\mathbb{E}\Big[X_\mathrm{t}^{(j)} X_\mathrm{t}^{(j)\top} \;\Big|\; X_{-t,\mathrm{t}}^{(j)},\, \Sigma^{(r)}\Big], \tag{16}$$

thereby "filling in" the missing T entries with their conditional second moments. In the M–step, we maximize a hybrid Q–function using both source and target information, but only over the parameters affected by the shift; the invariant parameters remain locked to their source estimates. Under **covariate shift**, all conditionals except $P(T \mid \mathrm{pa}(T))$ are invariant, so only the block $(A_{t,-t}, \Delta_t)$ is updated. Concretely, we regress X_t on its parents $X_{\mathrm{pa}(t)}$ using the block rows and columns of $Q^{(r)}_{\mathrm{tgt}}$:

$$A^{(r+1)}_{t,\mathrm{pa}(t)} = -Q^{-1}_{\mathrm{pa}(t),\mathrm{pa}(t)}\, Q_{\mathrm{pa}(t),t}, \quad \Delta^{(r+1)}_t = Q_{tt} - Q_{t,\mathrm{pa}(t)} Q^{-1}_{\mathrm{pa}(t),\mathrm{pa}(t)} Q_{\mathrm{pa}(t),t}.$$

Under **target shift**, the structural equation for T is unchanged but its marginal variance may differ; thus $A^{(\mathrm{s})}$ is frozen entirely, and we update only

$$\Delta^{(r+1)}_t = Q^{(r)}_{tt}.$$

For clarity, the resulting updates can be summarized as

$$\boxed{\begin{aligned} \text{Covariate shift:}\quad & A^{(r+1)}_{t,\mathrm{pa}(t)} = -Q^{-1}_{\mathrm{pa}(t),\mathrm{pa}(t)}\, Q_{\mathrm{pa}(t),t}, \\ & \Delta^{(r+1)}_t = Q_{tt} - Q_{t,\mathrm{pa}(t)} Q^{-1}_{\mathrm{pa}(t),\mathrm{pa}(t)} Q_{\mathrm{pa}(t),t}; \\ \text{Target shift:}\quad & A^{(r+1)} = A^{(\mathrm{s})}, \\ & \Delta^{(r+1)}_t = Q^{(r)}_{tt}. \end{aligned}} \tag{17}$$

After updating the shifted block, we reconstruct the precision and covariance via

$$K^{(r+1)} = A^{(r+1)\top}\, \mathrm{diag}(1/\Delta^{(r+1)})\, A^{(r+1)}, \qquad \Sigma^{(r+1)} = (K^{(r+1)})^{-1},$$

and we stop when the Frobenius gap falls below tolerance,

$$\|\Sigma^{(r+1)} - \Sigma^{(r)}\|_F < \varepsilon.$$

In practical terms, each target M–step now reduces to a block regression of X_t on its parents (an $O(p^2)$ update) or even a single variance update (an $O(1)$ step) rather than a full DAG refit at $O(p^3)$. Under covariate shift, all parameters except $(A_{t,-t}, \Delta_t)$ remain tied to the source estimates; under target shift, only Δ_t is updated. This separation exploits the invariances guaranteed by the causal DAG, allowing the EM algorithm to focus its statistical effort exclusively on the non-invariant parameters.

IV. Population-Level Contraction in a Neighborhood of Σ^*

Let Σ^* denote the true target covariance, the unique fixed point of the population EM operator F defined in (8). Our goal is to show that F acts as a contraction mapping in a Frobenius neighborhood of Σ^*, which immediately implies geometric convergence of iterates $\Sigma^{(k+1)} = F(\Sigma^{(k)})$. To set the stage, recall that $F : \mathcal{M}^p_+ \to \mathcal{M}^p_+$ is defined by

$$F(\Theta) = \mathbb{E}_{X_{-t}}\big[\mathbb{E}(XX^\top \mid X_{-t}, \Theta)\big]\,, \tag{18}$$

where $\mathcal{M}^p_+$ denotes the cone of $p \times p$ positive definite matrices equipped with the Frobenius norm $\|M\|_F = \sqrt{\mathrm{trace}(M^\top M)}$. By definition, $F(\Sigma^*) = \Sigma^*$. We seek constants $r > 0$ and $0 \le \kappa < 1$ such that

$$\|F(\Sigma) - \Sigma^*\|_F \;\le\; \kappa\, \|\Sigma - \Sigma^*\|_F, \quad \forall \Sigma \in \mathcal{B}_F(\Sigma^*; r), \tag{19}$$

where $\mathcal{B}_F(\Sigma^*; r) = \{\Sigma \succ 0 : \|\Sigma - \Sigma^*\|_F \le r\}$.

The EM surrogate curvature admits a natural decomposition into a stabilizing complete–data term and a destabilizing missing–data term. On the stabilizing side, consider the complete–data negative log–likelihood

$$\ell_{\mathrm{comp}}(X; \Sigma) = \tfrac{1}{2}\big[\log\det(\Sigma) + X^\top \Sigma^{-1} X\big].$$

Its population expectation is strongly convex in Σ^{-1}, which we formalize as follows.

Assumption 1 (Strong concavity of complete data): There exists $\lambda > 0$ such that

$$\nabla^2_{\Sigma^{-1}}\, \mathbb{E}[\ell_{\mathrm{comp}}(X; \Sigma)]\big|_{\Sigma=\Sigma^*} \;\succeq\; \lambda\, I_{p(p+1)/2}.$$

On the destabilizing side, the missing–data Hessian

$$\Delta(\Sigma) := \mathbb{E}_{X_{-t}}\big[\nabla^2_{\Sigma^{-1}} \log p(X_{-t}; \Sigma)\big]$$

may vary with Σ. We assume this dependence is Lipschitz, so that

Assumption 2 (Lipschitz missing-data term): There exists $\gamma \ge 0$ such that for all $\Sigma_1, \Sigma_2 \in \mathcal{B}_F(\Sigma^*; r)$,

$$\|\Delta(\Sigma_1) - \Delta(\Sigma_2)\|_F \;\le\; \gamma\, \|\Sigma_1 - \Sigma_2\|_F.$$

Heuristically, λ quantifies the stabilizing curvature of the complete–data likelihood, while γ measures how sensitively the missing–data curvature changes. Throughout, we assume $\gamma < \lambda$.

To analyze this setting we invoke a standard convex optimization fact, adapted to Frobenius geometry.

Lemma 1 (Gradient growth under strong convexity): Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be a finite-dimensional Hilbert space. Suppose $f : \mathcal{H} \to \mathbb{R}$ is differentiable and μ–strongly convex on $\mathcal{D} \subseteq \mathcal{H}$, i.e.,

$$f(y) \;\geq\; f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{\mu}{2}\|y - x\|^2 \quad \forall x, y \in \mathcal{D}.$$

Then

$$\mu\|y - x\| \;\leq\; \|\nabla f(y) - \nabla f(x)\|, \quad \forall x, y \in \mathcal{D}.$$

Corollary 1 (Matrix version): Let $f : \mathbb{S}^p \to \mathbb{R}$ be differentiable and μ–strongly convex on $\mathcal{D} \subseteq \mathbb{S}^p$ w.r.t. the Frobenius inner product. Then

$$\mu\|\Sigma_2 - \Sigma_1\|_F \;\leq\; \|\nabla f(\Sigma_2) - \nabla f(\Sigma_1)\|_F, \quad \forall \Sigma_1, \Sigma_2 \in \mathcal{D}.$$

With these ingredients in place, we can state the main result.

Theorem 1 (Population EM Contraction): Suppose Assumptions 1–2 hold with $\gamma < \lambda$. Then there exists $r > 0$ such that for all $\Theta \in \mathcal{B}_F(\Sigma^*; r)$,

$$\|F(\Theta) - \Sigma^*\|_F \;\leq\; \frac{\gamma}{\lambda}\|\Theta - \Sigma^*\|_F.$$

Thus Σ^* is the unique fixed point in $\mathcal{B}_F(\Sigma^*; r)$ and iterates $\Sigma^{(k+1)} = F(\Sigma^{(k)})$ converge geometrically.

Proof: Fix $\Theta \in \mathcal{B}_F(\Sigma^*; r)$ and let $f_\Theta(\Sigma) := Q_{\text{pop}}(\Sigma \mid \Theta)$. By the Louis decomposition (Sec. V) and A1–A2, f_Θ is $(\lambda - \gamma)$–strongly convex in Σ near Σ^*. Since $F(\Theta) = \arg\min_\Sigma f_\Theta(\Sigma)$ and $F(\Sigma^*) = \Sigma^*$, Lemma 1 with $x = \Sigma^*$ and $y = F(\Theta)$ yields

$$(\lambda - \gamma)\|F(\Theta) - \Sigma^*\|_F \;\leq\; \|\nabla_1 G(\Sigma^*; \Theta) - \nabla_1 G(\Sigma^*; \Sigma^*)\|_F.$$

By A2, the right-hand side is at most $\gamma\|\Theta - \Sigma^*\|_F$. Hence

$$\|F(\Theta) - \Sigma^*\|_F \;\leq\; \frac{\gamma}{\lambda - \gamma}\|\Theta - \Sigma^*\|_F \;\leq\; \frac{\gamma}{\lambda}\|\Theta - \Sigma^*\|_F,$$

since $\gamma < \lambda$. This proves contraction with factor $\kappa = \gamma/\lambda < 1$. ■

The radius of this contraction region can be quantified as follows. Let L bound the third derivative of $\mathbb{E}[\ell_{\text{comp}}]$ in a neighborhood of Σ^*. Choosing

$$r = \frac{\lambda - \gamma}{L}$$

ensures uniform strong convexity and Lipschitz bounds, making the argument valid throughout $\mathcal{B}_F(\Sigma^*; r)$. In practice, a source-domain initialization or a covariate-aware pilot estimate is sufficient to guarantee $\|\Sigma^{(0)} - \Sigma^*\|_F \leq r$.

Finally, under *covariate shift*, the complete–data curvature λ is unaffected, while γ depends only on the observed block $\Sigma^*_{-t,-t}$; hence if $\lambda > \gamma$ in the source it persists in the target. Under *target shift*, γ may increase since $P(T)$ changes, but as long as Σ^* remains well conditioned ($\lambda_{\min}(\Sigma^*) > 0$), the inequality $\gamma < \lambda$ continues to hold and contraction remains valid.

V. EM Curvature Decomposition

To verify the contraction constant $\kappa = \gamma/\lambda$ explicitly, it is convenient to decompose the observed–data curvature into a stabilizing complete–data term minus a missing–data correction (cf. [2]). Let $X = (X_1, \ldots, X_p)^\top$ and define the negative complete–data log–likelihood

$$\ell_{\text{comp}}(X; \Sigma) = \tfrac{1}{2}\big[\log\det(\Sigma) + X^\top \Sigma^{-1} X\big].$$

Differentiating with respect to Σ^{-1} yields

$$\nabla^2_{\Sigma^{-1}} \ell_{\text{comp}}(X; \Sigma) \;=\; \tfrac{1}{2}\,(\Sigma \otimes \Sigma),$$

where $\otimes$ denotes the Kronecker product. Taking expectation over $X \sim \mathcal{N}(0, \Sigma)$ therefore gives the complete–data (population) curvature

$$H_{\text{comp}} \;=\; \mathbb{E}_X\big[\nabla^2_{\Sigma^{-1}} \ell_{\text{comp}}(X; \Sigma)\big] \;=\; \tfrac{1}{2}\,(\Sigma \otimes \Sigma).$$

When the target node T is unobserved, the observed–data negative log–likelihood is

$$\ell_{\text{obs}}(X_{-t}; \Sigma) \;=\; -\log \int \exp\!\big(-\ell_{\text{comp}}(X; \Sigma)\big)\, dX_t,$$

and its Hessian at the truth $\Sigma = \Sigma^*$ decomposes as

$$\nabla^2_{\Sigma^{-1}}\, \ell_{\text{obs}}(X_{-t}; \Sigma^*) \;=\; H_{\text{comp}} - H_{\text{miss}}. \tag{20}$$

The next lemma identifies the closed form of the missing–data term. Partition $\Sigma = \begin{pmatrix} \Sigma_{-t,-t} & \Sigma_{-t,t} \\ \Sigma_{t,-t} & \Sigma_{t,t} \end{pmatrix}$ and let $E_{-t,-t}$ denote the selection matrix that extracts the rows and columns indexed by $\{1, \ldots, p\} \setminus \{t\}$.

Lemma 2 (Form of the missing–data curvature): With the notation above,

$$H_{\mathrm{miss}} \;=\; \tfrac{1}{2}\, E_{-t,-t}\, \Sigma^{*}_{-t,-t}\, E^{\top}_{-t,-t},$$

which is positive semidefinite of rank at most $p-1$.

Proof: The marginal density of X_{-t} is

$$p(X_{-t};\Sigma) = \frac{\exp\!\big(-\tfrac{1}{2}X_{-t}^{\top}\Sigma_{-t,-t}^{-1}X_{-t}\big)}{(2\pi)^{(p-1)/2}\sqrt{\det(\Sigma_{-t,-t})}},$$

so the corresponding negative log–density is

$$\ell_{\mathrm{miss}}(X_{-t};\Sigma) = \tfrac{1}{2}\left[\log\det(\Sigma_{-t,-t}) + X_{-t}^{\top}\Sigma_{-t,-t}^{-1}X_{-t}\right].$$

Differentiating twice with respect to $\Psi = \Sigma^{-1}$ gives

$$\nabla^2_{\Psi}\log p(X_{-t};\Sigma) = \tfrac{1}{2}\Big[E_{-t,-t}\Sigma_{-t,-t}E^{\top}_{-t,-t} \;-\; E_{-t,-t}X_{-t}X_{-t}^{\top}E^{\top}_{-t,-t}\Big].$$

Taking expectation over $X_{-t} \sim \mathcal{N}(0,\Sigma^{*}_{-t,-t})$ cancels the quadratic term and leaves the stated expression. ■

Substituting Lemma 2 into (20) yields

$$\nabla^2_{\Sigma^{-1}}\,\ell_{\mathrm{obs}}(X_{-t};\Sigma^{*}) \;=\; \tfrac{1}{2}\Big[\Sigma^{*}\otimes\Sigma^{*} \;-\; E_{-t,-t}\,\Sigma^{*}_{-t,-t}\,E^{\top}_{-t,-t}\Big].$$

A direct spectral comparison now isolates the stabilizing and destabilizing components. Writing

$$\lambda \;=\; \tfrac{1}{2}\,[\lambda_{\min}(\Sigma^{*})]^2, \qquad \gamma \;=\; \tfrac{1}{2}\,\lambda_{\max}(\Sigma^{*}_{-t,-t}),$$

the contraction ratio

$$\kappa \;=\; \frac{\gamma}{\lambda} \;=\; \frac{\lambda_{\max}(\Sigma^{*}_{-t,-t})}{[\lambda_{\min}(\Sigma^{*})]^2} \tag{21}$$

emerges from bounding the smallest eigenvalue of the observed–data Hessian below by $\tfrac{1}{2}[\lambda_{\min}(\Sigma^{*})]^2 - \tfrac{1}{2}\lambda_{\max}(\Sigma^{*}_{-t,-t})$. Whenever $\kappa < 1$, the observed–data curvature is strictly positive definite in Σ^{-1}, and the population EM operator F is locally contractive around Σ^{*}.

This spectral–gap condition provides transparent implications under domain shift. In a pure covariate shift, only $\Sigma^{*}_{-t,-t}$ changes while the cross–block structure that governs $\lambda_{\min}(\Sigma^{*})$ remains controlled; contraction therefore holds so long as $\lambda_{\max}(\Sigma^{*}_{-t,-t}) < [\lambda_{\min}(\Sigma^{*})]^2$. In a pure target shift, the observed block $\Sigma^{*}_{-t,-t}$ is fixed but $\lambda_{\min}(\Sigma^{*})$ may shrink through changes in the T row/column; maintaining $\lambda_{\min}(\Sigma^{*}) \geq \sqrt{\lambda_{\max}(\Sigma^{*}_{-t,-t})}$ preserves contractivity. Under combined shifts both blocks may vary, yet the same inequality in (21) remains the decisive criterion. In short, EM contracts whenever

$$\lambda_{\max}(\Sigma^{*}_{-t,-t}) \;<\; [\lambda_{\min}(\Sigma^{*})]^2,$$

a concrete and easily checkable link between the magnitude of covariate and target shifts and the curvature needed for geometric convergence.

VI. High-Probability Sample-Level Concentration and Final Error Bound

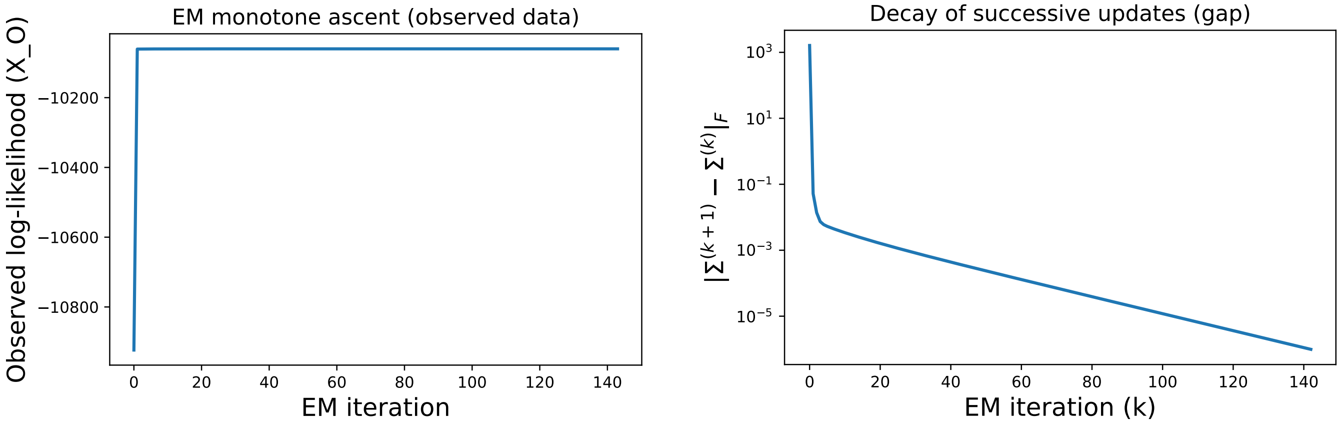
Having established that the population EM operator F is a contraction with factor $\kappa < 1$ around Σ^{*}, we now turn to the finite-sample setting. The sample operator $\widehat{F}$, defined by replacing population moments with empirical covariances from both source and target data, is itself a random perturbation of F. Standard matrix concentration inequalities (e.g. matrix Bernstein for sub-Gaussian vectors) together with Weyl's eigenvalue inequality show that, with probability at least $1-\exp(-\Omega(\min\{N_{\mathrm{s}},N_{\mathrm{t}}\}))$, this deviation is uniformly small in the contraction neighborhood. More precisely, there exists a deviation radius $\delta = \delta(N_{\mathrm{s}},N_{\mathrm{t}})$, vanishing as $N_{\mathrm{s}},N_{\mathrm{t}} \to \infty$, such that

$$\big\|\widehat{F}(\Sigma) - F(\Sigma)\big\|_F \;\leq\; \delta(N_{\mathrm{s}},N_{\mathrm{t}}), \qquad \forall \Sigma \in \mathcal{B}_F(\Sigma^{*};r).$$

The intuition is straightforward: empirical covariances concentrate around their expectations at the rate $O\Big(\sqrt{\frac{p}{N_{\mathrm{s}}+N_{\mathrm{t}}}}\Big)$ under sub-Gaussian tails, and since F is Lipschitz in these moments, the same rate carries over to $\widehat{F}$. Weyl's inequality then controls perturbations in eigenvalues, ensuring that the Frobenius deviation is at most δ.

This uniform concentration, combined with the contraction property of F, yields a finite-sample error bound for EM. Suppose the initialization $\Sigma^{(0)}$ lies within the contraction ball $\mathcal{B}_F(\Sigma^{*};r)$. Then, with the same high probability, the iterates of the sample EM update $\Sigma^{(r+1)} = \widehat{F}(\Sigma^{(r)})$ satisfy

$$\|\Sigma^{(r)} - \Sigma^{*}\|_F \;\leq\; \kappa^{r}\|\Sigma^{(0)} - \Sigma^{*}\|_F + \frac{\delta}{1-\kappa}.$$



(a) Observed-data log-likelihood $\mathcal{L}_{\text{obs}}(\Sigma_k; X_O)$ increases monotonically under first-order EM (GEM).

(b) Successive gap $g_k = \|\Sigma_{k+1} - \Sigma_k\|_F$ decays geometrically (log scale).

Fig. 1. Empirical convergence diagnostics: monotone ascent of the observed likelihood and geometric decay of successive updates.

The proof is a simple recurrence: each step's error is bounded by a contractive population term $\kappa\|\Sigma^{(r)} - \Sigma^*\|_F$ plus an additive deviation δ, and iterating this inequality yields the stated bound.

The interpretation is clear. With a suitable initialization, the EM iterates converge geometrically at rate κ, but statistical fluctuations introduce a terminal error plateau of order $\delta/(1-\kappa)$. As the number of source and target samples grows, $\delta \to 0$, so the plateau vanishes and the iterates converge all the way to Σ^*. In this sense, the finite-sample theory mirrors the classical analysis of EM for mixtures and regressions [1], but here it is adapted to Gaussian DAGs under domain shift. Crucially, both contraction and concentration can be checked explicitly via the spectral-gap conditions derived in Sec. V.

VII. Convergence diagnostics for first-order EM

This section documents empirical convergence behavior of our *first-order EM* (mean-aware/GEM) procedure on the motivating example used throughout the paper. We report (i) the observed-data log-likelihood trajectory, (ii) the decay of successive parameter updates, and (iii) convergence of the estimated covariance to a high-accuracy reference ("near-oracle") solution.

a) Setup.: Let Σ_k denote the model covariance after iteration k, with O the set of observed variables and T fully missing in the target domain. Our implementation performs a single (projected) ascent step per iteration on the DAG-constrained parameterization, starting from a mean-aware initialization. We monitor:

$$\begin{aligned}
&\text{(Obs. LL)} && \mathcal{L}_{\text{obs}}(\Sigma_k; X_O) \\
&\text{(Successive gap)} && g_k := \|\Sigma_{k+1} - \Sigma_k\|_F \\
&\text{(Distance to reference)} && d_k := \|\Sigma_k - \Sigma^\star\|_F, \quad \tilde{d}_k := \|\Sigma_k^{1/2}(\Sigma^\star)^{-1/2} - I\|_F,
\end{aligned}$$

where $\Sigma^\star$ is a high-precision solution obtained by running the algorithm to tight tolerance (and cross-checked for stationarity).

b) Monotone ascent of the observed likelihood.: Figure 1 (left) shows a strictly increasing $\mathcal{L}_{\text{obs}}(\Sigma_k; X_O)$ over iterations, as expected for a GEM procedure that guarantees non-decrease of the observed-data objective. The trajectory exhibits rapid rise followed by a clean plateau at convergence.

c) Geometric decay of successive updates.: Figure 1 (right) plots $g_k = \|\Sigma_{k+1} - \Sigma_k\|_F$ on a log scale; the curve is approximately linear, indicating geometric (linear-rate) decay until it reaches the numerical noise floor. This is consistent with local contractivity near a nondegenerate fixed point.

d) Convergence to a reference solution.: Figure 2 (left) reports $d_k = \|\Sigma_k - \Sigma^\star\|_F$, which decays rapidly and then plateaus at solver precision; Figure 2 (right) shows the scale-invariant version $\tilde{d}_k$, confirming the same qualitative behavior when normalizing by the target scale. Together, these diagnostics support convergence to a stable fixed point with near-oracle accuracy for downstream imputation of T.

e) Takeaway.: Across all diagnostics—monotone $\mathcal{L}_{\text{obs}}$, geometric gap decay, and shrinkage of d_k and $\tilde{d}_k$—the first-order EM displays textbook convergence toward a stable fixed point. In our motivating example, this fixed point delivers *near-oracle* predictive accuracy for imputing the fully missing target T (see main text metrics tables/plots).

References

[1] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *The Annals of Statistics*, vol. 45, no. 1, pp. 77–120, 2017.

[2] T. A. Louis, "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 44, no. 2, pp. 226–233, 1982.

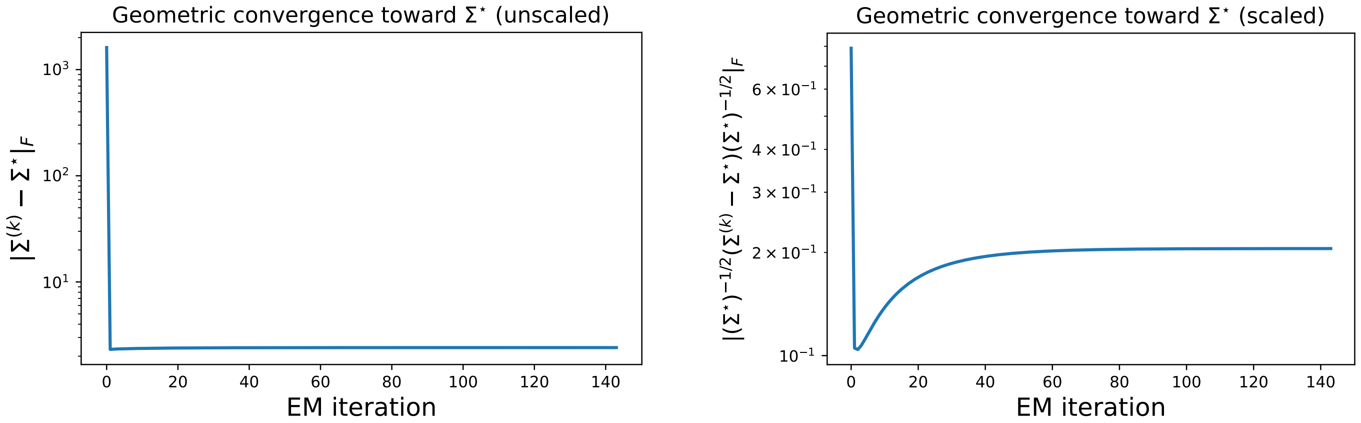


Fig. 2. Convergence of Σ_k to a high-accuracy reference $\Sigma^\star$: (left) unscaled distance $d_k = \|\Sigma_k - \Sigma^\star\|_F$; (right) scale-invariant distance $\tilde{d}_k = \|\Sigma_k^{1/2}(\Sigma^\star)^{-1/2} - I\|_F$.