Group-Aware Matrix Estimation and Latent Subspace Recovery

Hamza Golubovic*
Department of Statistics,
Columbia University
hg2723@columbia.edu

Matthew Shen*
Department of Statistics,
Columbia University
ms7079@columbia.edu

Genevera I. Allen
Department of Statistics,
Irving Institute for Cellular Dynamics,
Zuckerman Institute,
Columbia University
genevera.allen@columbia.edu

Tarek M. Zikry†
Division of Data Science and Society,
School of Data and Information Sciences
University of North Carolina at Chapel Hill
tarek@unc.edu

Abstract

Modern matrix completion problems often involve heterogeneous data whose rows simultaneously belong to many meta-categories, such as demographic and age groups in recommendation systems, or region and recording session labels in neural electrophysiological experiments. Standard low-rank estimators impose a single global latent geometry, which can recover average structure but may smooth away subgroup-specific variation, especially when observations are unevenly distributed across groups. We introduce Group-Aware Matrix Estimation (GAME), a convex estimator for overlapping subgroup-wise low-rank matrix estimation. GAME regularizes category-specific submatrices through overlapping nuclear-norm penalties, allowing related groups to borrow information while preserving local latent structure in a shared coordinate system. We provide finite-sample guarantees for both reconstruction error and subgroup-specific subspace recovery, showing how performance depends on sampling density, subgroup rank, and overlap structure. Experiments on synthetic, recommendation, ecological, and neuroscience datasets show that GAME is most beneficial in structured missingness regimes, where subgroup-aware regularization improves both reconstruction accuracy and latent subspace fidelity. Across these benchmarks, GAME is competitive or best among global low-rank, side-information, and modern imputation baselines, with the largest gains when subgroups exhibit distinct low-rank structure. All code is made available on GitHub. [3]

1 Introduction

In recommendation systems, users may belong to overlapping demographic groups such as age, gender, and occupation [Harper and Konstan, 2015]. In biological, experimental, and longitudinal studies, samples may be indexed by subject, batch, session, spatial region, or other partially overlapping sources of variation [Wang and Lock, 2024, Chen et al., 2024]. In such settings, the goal is often not only to reconstruct missing entries accurately on average, but also to preserve

*Equal contribution.
†Corresponding author.
[3]Code: https://github.com/mattyshen/game

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

subgroup-specific latent structure that may be scientifically or operationally important. Standard nuclear-norm estimators impose a single global low-rank penalty [Candes and Recht, 2008], which can smooth away subgroup-specific variation – especially for groups that are unevenly or weakly observed, where faithful estimation often matters most. More broadly, heterogeneous data need not conform to one shared low-rank geometry; different subgroups may exhibit local low-dimensional structure that is related through overlap but not identical [Elhamifar and Vidal, 2013]. This motivates an overlap-aware estimator that shares information across groups while preserving local structure.

In this work, we propose Group-Aware Matrix Estimation (GAME), a convex framework for heterogeneous matrix estimation with overlapping meta-categories (Figure 1). GAME balances faithful reconstruction of the observed entries with nuclear-norm regularization on meta-category-defined submatrices, allowing local low-rank structure to be learned within groups while overlap propagates information across them. This yields a single coherent estimate that borrows strength across groups without forcing all groups into the same latent subspace. GAME is not restricted to settings with an explicit block-sampling model, but its inductive bias is most beneficial when subgroups exhibit distinct low-rank structure. Structured missingness is one important instance: when some subgroups are weakly observed, overlap allows under-represented groups to borrow information without being smoothed into a single global estimate.

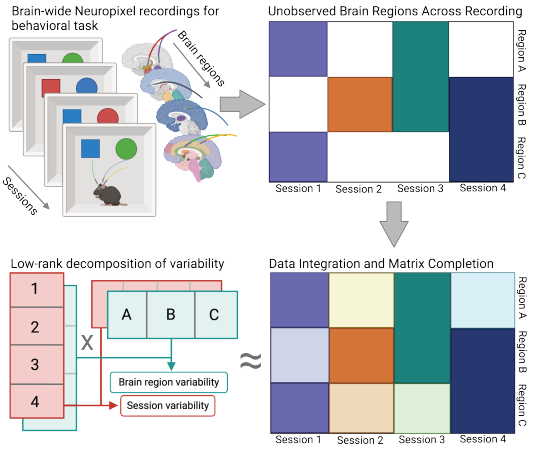


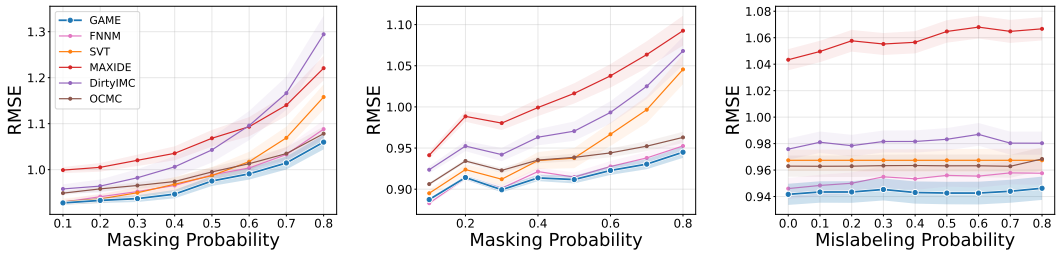
Figure 1: **Neuroscience motivation of GAME.** Brain-wide Neuropixels recordings measure single-neuron temporal responses to stimuli, but only a few brain regions can be recorded at once, leaving large block-missingness across recording sessions and brain regions. GAME imputes missing region dynamics and decomposes variability across meta-categories such as region and session.

To optimize the GAME objective efficiently, we use the proximal average framework [Bauschke et al., 2008, Yu, 2013], which avoids the auxiliary-variable and consensus overhead of ADMM in high-category settings [Boyd et al., 2011]. Beyond scalable optimization, GAME also supports analysis of subgroup-specific latent geometry: by regularizing each subgroup through an overlapping nuclear norm, the estimator supports both reconstruction and local subspace recovery.

We evaluate GAME on synthetic data with hierarchical group structure, MovieLens-100k [Harper and Konstan, 2015] under global and block-wise missingness and meta-category misspecification, multi-label birdsong classification from the BirdSet benchmark [Clapp et al., 2023, Rauch et al., 2025], and a scientific case study on multi-session Neuropixels recordings [Chen et al., 2024].

Contributions Together, these ideas lead to four main contributions:

- **A subgroup-aware convex estimator for heterogeneous matrix estimation.** We introduce GAME, which promotes overlapping subgroup-wise low-rank structure while maintaining a shared coordinate system across the full matrix.
- **Scalable optimization without consensus overhead.** We develop a proximal-averaging solver for the GAME objective that is practical in settings with many overlapping categories.
- **Provable reconstruction and subspace fidelity under overlap.** We provide finite-sample guarantees for both reconstruction error and subgroup-specific subspace recovery, clarifying how performance depends on sampling, rank, and overlap structure.
- **Empirical validation on synthetic and real datasets.** We show that GAME performs competitively or better than benchmark methods across diverse settings, particularly when meta-categories carry distinct low-rank structure and missingness is group-structured. On Neuropixels data, GAME recovers region-specific latent dynamics, illustrating value beyond aggregate imputation accuracy.



(a) Global missingness (b) Block-wise missingness (c) Mislabeled meta-categories

Figure 2: Test RMSE (↓) on MovieLens-100k under global missingness, block-wise missingness targeting users aged $35+$, and mislabeled demographic metadata. GAME achieves the lowest RMSE across nearly all regimes. Shaded regions denote ± 1 standard error across 3 masking realizations.

The remainder of the paper develops the formulation, optimization, and theory, followed by experiments on synthetic and real data.

2 Related Work

2.1 Matrix Completion with Side Information

Classical matrix completion assumes a single global low-rank structure, often estimated by nuclear-norm regularization [Candes and Recht, 2008, Recht et al., 2010] or scalable non-convex methods such as alternating least squares [Hastie et al., 2014, Mazumder et al., 2010]. Side-information methods extend this framework by using row or column features: IMC restricts the target matrix through known feature matrices [Jain and Dhillon, 2013], Maxide estimates within feature-induced subspaces using a lower-dimensional trace norm [Xu et al., 2013], and DirtyIMC adds residual structure when features are imperfect [Chiang et al., 2015]. FNNM combines global low-rank structure with feature-linked sparsity [Yang et al., 2022], while recent work regularizes components outside the observed feature subspaces when side information is incomplete [Chang et al., 2025]. These side-information methods primarily rely on feature-spanned subspaces and can be biased when features are uninformative. GAME instead treats metadata as overlapping group labels and regularizes the corresponding submatrices directly, allowing subgroup-specific low-rank structure without relying on the feature-spanned latent space.

2.2 Low-Rank Structure in Heterogeneous Data

Heterogeneous data are often better described by multiple local low-dimensional structures than by one global subspace, as in subspace clustering, low-rank representation, multi-task learning, and integrative factor analysis [Elhamifar and Vidal, 2013, Liu et al., 2013, Argyriou et al., 2008, Obozinski et al., 2011, Wang and Lock, 2024]. This issue is especially salient in neural population recordings, where low-rank and tensor methods are used to denoise and interpret neural activity [Buchanan et al., 2018, Pellegrino et al., 2024, Williams et al., 2018], and latent-variable models impose temporal structure on neural dynamics [Yu et al., 2009, Pandarinath et al., 2018]. Recent multi-session and multi-animal methods further exploit shared structure across partially observed recordings [Azabou et al., 2025, Zhang et al., 2024, 2026, Xia et al., 2025]. These methods typically impose a shared latent space or require aligned recordings; GAME targets the complementary matrix-estimation setting where overlapping, partially observed groups can each have their own low-rank structure while still borrowing information.

3 Group-Aware Matrix Estimation (GAME)

We introduce Group-Aware Matrix Estimation (GAME), a convex estimator designed to leverage overlapping categorical structures, promoting group-wise low-rank reconstructions while preserving a shared coordinate system. Unlike global estimators that smooth out distinct latent geometries, GAME allows information to be shared across categories while controlling complexity locally via category-specific penalties.

3.1 Problem Formulation

We consider a partially observed matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, with rows as observational units and columns as shared features. Each row $i \in [n]$ may belong to multiple meta-categories $c \in \mathcal{C}$, such as a user being both 35 years old and female in recommendation systems. Let $\Omega \subseteq [n] \times [m]$ denote the observed entries and $\mathcal{P}_\Omega$ the associated projection operator. The meta-categories in $\mathcal{C}$ may overlap: a single row can belong to several categories simultaneously, so the matrix cannot be decomposed into independent blocks.

Rather than assuming that $\mathbf{X}$ admits a single global low-rank explanation, GAME posits that each meta-category $c \in \mathcal{C}$ induces meaningful low-rank structure on its corresponding submatrix, even though these structures overlap and may not align perfectly across categories. The GAME objective, defined in Equation 1, is to estimate a single matrix $\mathbf{W}$ that preserves these category-specific patterns while coherently reconciling them:

$$\widehat{\mathbf{W}} \in \underset{\mathbf{W} \in \mathbb{R}^{n \times m}}{\arg\min} \frac{1}{2} \left\| \mathcal{P}_\Omega(\mathbf{X} - \mathbf{W}) \right\|_F^2 + \lambda \sum_{c \in \mathcal{C}} \alpha_c \left\| \mathbf{W}_c \right\|_*, \tag{1}$$

where $\mathbf{W}_c \in \mathbb{R}^{n_c \times m}$ denotes the submatrix of $\mathbf{W}$ corresponding to rows in meta-category c, $\| \cdot \|_*$ is the nuclear norm, $\lambda > 0$ is a global regularization parameter, and $\{\alpha_c\}_{c \in \mathcal{C}}$ are nonnegative weights satisfying $\sum_{c \in \mathcal{C}} \alpha_c = 1$.

Standard matrix completion assumes that all rows share a common rowspace, whose principal right singular vectors explain the low-rank behavior of every row. But rows belonging to different meta-categories may exhibit partially orthogonal latent structure; for example, distinct age groups or sexes in a recommendation system. Partitioning the regularization by meta-category allows GAME to respect these distinctions, using category membership as a proxy for differing low-rank behavior.

We present a simple motivating example in Appendix B illustrating GAME as a denoiser. Under the model, GAME's combined global and category-wise penalties preserve both globally shared and category-specific components while removing noise from categories where a component is absent. In contrast, global nuclear-norm denoising discards weaker category-specific components, while category-wise denoising misses shared components that require pooling evidence across categories.

3.2 Algorithm and Optimization

Since the meta-categories overlap, the GAME regularizer is non-separable, making consensus-based methods such as ADMM [Glowinski and Marrocco, 1975, Gabay and Mercier, 1976, Boyd et al., 2011] impractical due to the scaling of auxiliary variables with $|\mathcal{C}|$.

Despite this coupling, the GAME regularizer decomposes as a weighted sum of $g_c(\mathbf{W}) := \lambda \|\mathbf{W}_c\|_*$, where each $\mathbf{W}_c$ is obtained from $\mathbf{W}$ via a fixed row-subsetting operator $\mathbf{D}_c$. Specifically, let $\mathbf{W}_c = \mathbf{D}_c \mathbf{W}$, where $\mathbf{D}_c$ selects the rows belonging to category c. The row-selection operator is semi-orthogonal, satisfying $\mathbf{D}_c \mathbf{D}_c^\top = \mathbf{I}$. This implies that each g_c has a simple proximal operator given by singular value thresholding.

Algorithm 1 GAME Algorithm (PA-APG)

Initialize $\mathbf{W}^0 = \mathbf{Y}^1$, $\gamma > 0$, $\eta_1 = 1$
for $k = 1, 2, \ldots$ **do**
  $\mathbf{Z}^k = \mathbf{Y}^k - \gamma \, \mathcal{P}_\Omega(\mathbf{Y}^k - \mathbf{X})$
  $\mathbf{D}_c \mathbf{Z}^k = \mathbf{U}_c^k \mathbf{\Sigma}_c^k \mathbf{V}_c^{k\top}$ **for** $c = 1, \ldots, |\mathcal{C}|$ **(SVD)**
  $\mathbf{W}^k = \sum_{c \in \mathcal{C}} \alpha_c \Big[(\mathbf{I} - \mathbf{D}_c^\top \mathbf{D}_c) \mathbf{Z}^k + \mathbf{D}_c^\top \mathbf{U}_c^k S_{\lambda\gamma}(\mathbf{\Sigma}_c^k) \mathbf{V}_c^{k\top} \Big]$
  $\eta_{k+1} = \frac{1 + \sqrt{1 + 4\eta_k^2}}{2}$
  $\mathbf{Y}^{k+1} = \mathbf{W}^k + \frac{\eta_k - 1}{\eta_{k+1}} (\mathbf{W}^k - \mathbf{W}^{k-1})$
end for
return $\mathbf{W}^k$

This structure naturally motivates the use of the Proximal Average (PA) algorithm [Yu, 2013], which optimizes objectives composed of weighted sums of functions with simple proximal operators. PA combines the category-specific proximal updates directly, enabling efficient optimization without introducing auxiliary variables or dual updates. Each iteration requires one reduced SVD per meta-category, costing $O(n_c m^2)$ for $n_c \geq m$.

We defer proofs and technical details regarding semi-orthogonality and proximal derivations of the GAME algorithm (Algorithm 1) to Appendix A. With these results, we are able to obtain convergence guarantees from [Yu, 2013], highlighted in Theorem 3.1.

We use the abbreviation $\overline{L^2} = \lambda^2 \sum_{c\in\mathcal{C}} \alpha_c \min(n_c, m)$ in the following theorem to denote the weighted sum of per-category Lipschitz constants. We also define $f(\mathbf{W}) := \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{X} - \mathbf{W})\|_F^2$ and $\bar{g}(\mathbf{W}) := \lambda \sum_{c\in\mathcal{C}} \alpha_c \|\mathbf{W}_c\|_*$, so that $f + \bar{g}$ is the GAME objective in Equation (1).

Theorem 3.1 (Theorem 1 of [Yu, 2013])**.** *Let $\mathbf{W}_0$ be the initialization of the PA-APG algorithm. Fix the accuracy $\epsilon > 0$. Let $\gamma = \min\{1, 2\epsilon/\overline{L^2}\}$. After at most $k = \sqrt{\frac{2}{\gamma\epsilon}}\|\mathbf{W}_0 - \widehat{\mathbf{W}}\|_F$ steps, the PA-APG approximation, $\widetilde{\mathbf{W}}_k$, satisfies $f(\widetilde{\mathbf{W}}_k) + \bar{g}(\widetilde{\mathbf{W}}_k) \le f(\widehat{\mathbf{W}}) + \bar{g}(\widehat{\mathbf{W}}) + 2\epsilon$.*

Let $\widetilde{\mathbf{W}}_k$ denote the output of Algorithm 1 after k iterations. Under the conditions of Theorem 3.1, its objective value is at most 2ϵ above the minimum in (1).

3.3 Theoretical Guarantees of GAME

We now establish finite-sample guarantees for the entrywise-constrained GAME estimator (2) under a partially observed noisy matrix model. We observe $\mathbf{X} = \mathbf{W}^\star + \mathbf{E}$ on a random set Ω in which every entry is included independently with probability $q = \overline{N}/(nm)$, so that $\overline{N}$ is the expected number of observed entries. We highlight the uniform Bernoulli sampling across rows in the main text, and defer block-wise sampling extension to Appendix C.9. In developing theoretical results, we analyze the non-spiky restricted estimator

$$\widehat{\mathbf{W}}^{\mathcal{F}} \in \arg\min_{\mathbf{W}\in\mathcal{F}(\alpha^\star)} \tfrac{1}{2}\big\|\mathcal{P}_\Omega(\mathbf{X} - \mathbf{W})\big\|_F^2 + \sum_{c\in\mathcal{C}} \lambda_c \|\mathbf{W}_c\|_*, \tag{2}$$

where $\mathcal{F}(\alpha^\star) := \{\mathbf{W} : \|\mathbf{W}\|_\infty \le \alpha^\star/\sqrt{nm}\}$. For each meta-category $c \in \mathcal{C}$, $\mathbf{W}_c := \mathbf{W}[I_c, :] \in \mathbb{R}^{n_c\times m}$ with $n_c := |I_c|$, $d_c := n_c + m$, and $r_c := \operatorname{rank}(\mathbf{W}_c^\star)$. Row overlap multiplicity is $\kappa(i) := |\{c : i \in I_c\}|$ with extremes $\kappa_{\max}$ and $\kappa_{\min}$, and we write $r_{\text{tot}} := \sum_c r_c$ and $d_0 := \max\{n, m\}$. We make the following assumptions.

Assumption 3.2 (Subexponential Noise)**.** The entries E_{ij} are independent, mean zero, and subexponential with parameters (σ, R), meaning $\mathbb{E}|E_{ij}|^p \le \frac{p!}{2} R^{p-2}\sigma^2$ for every integer $p \ge 2$.

Assumption 3.3 (Non-Spikiness)**.** $\|\mathbf{W}^\star\|_\infty \le \alpha^\star/\sqrt{nm}$, so that $\mathbf{W}^\star$ is feasible for (2).

Assumption 3.4 (Cover)**.** $\bigcup_{c\in\mathcal{C}} I_c = [n]$, i.e. $\kappa_{\min} \ge 1$. Further, $\kappa_{\max} = O(1)$.

The spikiness constraint rules out matrices whose mass is concentrated on a small number of entries, which cannot be identified from partial observations [Negahban and Wainwright, 2011]. Under these assumptions, the restricted GAME estimator (2) satisfies the following theorem.

Theorem 3.5 (GAME Frobenius Error)**.** *Suppose Assumptions 3.2–3.4 hold and $\overline{N} \gtrsim (1 + R^2/\sigma^2)\, d_0 \log(n+m)$. Let*

$$\lambda_c \asymp \sigma\sqrt{\frac{\overline{N}\,(n_c \vee m)\,\log d_c}{nm}} + R\log d_c \qquad \forall\, c \in \mathcal{C}. \tag{3}$$

Then with probability at least $1 - 2\sum_c d_c^{-1} - (n+m)^{-c}$, the restricted GAME estimator $\widehat{\mathbf{W}}^{\mathcal{F}}$ from (2) satisfies

$$\frac{\|\widehat{\mathbf{W}}^{\mathcal{F}} - \mathbf{W}^\star\|_F^2}{nm} \lesssim \underbrace{\frac{\sigma^2\log(n+m)}{\overline{N}}\sum_{c\in\mathcal{C}} r_c\,(n_c \vee m)}_{\textit{noise term}} + \underbrace{(\alpha^\star)^2\,\theta_\lambda^2\, r_{\text{tot}}\cdot\frac{d_0\log(n+m)}{nm\,\overline{N}}}_{\textit{spikiness term}}, \tag{4}$$

where $\theta_\lambda := \max_c \lambda_c / \min_c \lambda_c$ is a constant when the categories are comparably sized.

The proof is given in Appendix C. The first term is the statistical price of estimating each category's low-rank structure from the noise on its observed entries. The second comes from the sampling tolerance of the completion problem and is dominated by the first whenever $\|\mathbf{W}^\star\|_\infty \lesssim \sigma$, the usual noisy completion regime. The Frobenius bound translates into a per-category latent-subspace guarantee via Wedin's $\sin\Theta$ theorem, which is the statement of interest when the goal is to preserve category-specific latent structure.

Corollary 3.6 (Per-Category Right-Subspace Recovery)**.** *Let $\mathcal{B}^2$ denote nm times the right-hand side of* (4)*. For each $c \in \mathcal{C}$, let $\mathbf{Q}_c^\star, \widehat{\mathbf{Q}}_c \in \mathbb{R}^{m \times r_c}$ denote the top r_c right singular vectors of $\mathbf{W}_c^\star$ and $\widehat{\mathbf{W}}_c^{\mathcal{F}}$, and let $\sigma_{r_c}(\mathbf{W}_c^\star)$ be the r_c-th singular value of $\mathbf{W}_c^\star$. On the event of Theorem 3.5, for every c with $\mathcal{B} \le \frac{1}{2}\sigma_{r_c}(\mathbf{W}_c^\star)$,*

$$\min_{\mathbf{O}:\, \mathbf{O}^\top \mathbf{O} = \mathbf{I}_{r_c}} \left\| \widehat{\mathbf{Q}}_c \mathbf{O} - \mathbf{Q}_c^\star \right\|_F^2 \lesssim \frac{\mathcal{B}^2}{\sigma_{r_c}^2(\mathbf{W}_c^\star)}.$$

The proof is given in Appendix C.8. The ratio on the right is error over signal. The subspace of category c is recovered well when the total error is small compared with the weakest latent direction of that category, and every improvement in the entrywise rate transfers directly to subspace recovery.

Discussion. Theorem 3.5 is prescriptive, in that it fixes the relative weights α_c from category-level statistics, leaving one global scale to tune, and comparative, in that it identifies the regime where GAME's guarantee sharpens that of global nuclear-norm regularization. We discuss each in turn.

(i) Category-specific regularization. The choice (3) is the noise level σ times the square root of the category's expected observation count divided by its smaller block dimension, up to a log factor. Since the observed signal scales with the sampling rate but λ_c only with its square root, sparsely observed categories are regularized more strongly relative to their signal.

(ii) Category-specific sample budgets. Category c receives $\overline{N} n_c / n$ observations in expectation and contributes $r_c(n_c \vee m)$ to the noise term of (4). This contribution is smaller for lower-rank categories and decreases with category size when $n_c > m$. For bounded R/σ, the theorem's sample-size requirement is $\overline{N} \gtrsim (n \vee m)\log(n+m)$.

(iii) Comparison with global nuclear-norm regularization. With a single category Theorem 3.5 recovers the bound of Negahban and Wainwright [2011, Cor. 1], whose noise term $\sigma^2 r^\star (n \vee m)\log(n+m)/\overline{N}$, with $r^\star = \operatorname{rank}(\mathbf{W}^\star)$, is minimax optimal up to a logarithm. Dividing noise terms,

$$\frac{\text{GAME noise term}}{\text{global noise term}} \asymp \frac{\sum_c r_c(n_c \vee m)}{r^\star(n \vee m)}. \tag{5}$$

The global bound pays rank times dimension once. GAME pays it once per category, with that category's rank and block size. When each row has categories that are comparably sized and tall with common local rank, and their row spaces are largely orthogonal, the ratio scales as $O(1/|\mathcal{C}|)$: a guarantee sharper by a factor of order $|\mathcal{C}|$. Conversely, when $n \le m$ every category is wide and (44) is at least one, so the bound favors GAME only when categories are tall (Appendix C.7). The gain vanishes when all categories share one row space, since then the labels carry no rank information. By Corollary 3.6 the same factor transfers to per-category subspace recovery.

Appendix C gives the proofs and a fuller reading of the results, with an additional extension of the above results to nonuniform sampling. Sharper rates under separation of the category-wise row spaces are left to future work.

4 Experiments

We evaluate GAME on four experiments probing (1) preservation of local latent structure that global methods smooth away (Section 4.1), (2) reconstruction under uniform and structured missingness (Section 4.2), and (3) downstream classification (Section 4.3) and (4) latent dynamics recovery (Section 4.4).

Across all experiments, we compare GAME to singular value thresholding [SVT; Cai et al., 2010], alternating least squares [ALS; Jain et al., 2012], dirty inductive matrix completion [DirtyIMC; Chiang et al., 2015], MAXIDE [Xu et al., 2013], FNNM [Yang et al., 2022], and OCMC [Chang et al., 2025]. We restrict DirtyIMC and MAXIDE to using only row-wise side information, stored in matrix $\mathbf{A} \in \mathbb{R}^{n \times d_r}$: the one-hot encoding of each row's meta-categories. Section 4.3 additionally baselines against two modern matrix imputation methods: MissForest [Stekhoven and Bühlmann, 2012] and TabImpute [Feitelberg et al., 2026].

Hyperparameters for baseline methods are tuned via cross-validation; for the GAME weights α_c, the heuristic in Theorem 3.5 is used as initialization, with additional scaling selected by cross-validation. Additionally, when executing the iterative proximal updates for GAME, we utilized a

reduced SVD, which discards only zero singular values and so computes the exact proximal update. For most experiments, completing the matrices with GAME is computationally feasible to run locally. However, for the Svoboda Lab Neuropixels dataset we resorted to implementing a GPU-compatible version of GAME using a truncated SVD for GAME and all SVD-based baselines, since exact SVDs were computationally infeasible at that scale. Even with a trade-off for computational efficiency, GAME demonstrated robust performance against other methods.

4.1 Synthetic Clustering Under Subgroup Overlap

Setup. We construct a synthetic matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ with crossed group structure. Each row x_i belongs to an observed group $g \in \mathcal{G}$ and an unobserved subcluster $s \in \mathcal{S}$. The observed groups define the meta-categories available to GAME and the side-information baselines, while the hidden subclusters define the latent structure we aim to recover after matrix completion.

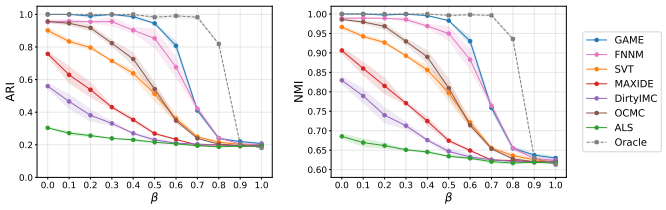
We set $|\mathcal{G}| = 10$, $|\mathcal{S}| = 5$, $n = 1000$, and $m = 500$. Each row is generated as

$$x_i = \mu_g + \nu_s + u_i V_g^\top + (1 - \beta) z_i W_s^\top,$$

where μ_g is a group-specific mean vector, ν_s is a subcluster-specific mean vector, V_g is a low-rank basis associated with the observed group g, and W_s is a low-rank basis associated with the hidden subcluster s. The vectors u_i and z_i are normalized Gaussian latent scores that control row-level variation within the group and subcluster components, respectively.

The parameter $\beta \in [0, 1]$ controls the strength of the hidden subcluster signal. When $\beta = 0$, the subcluster-specific low-rank component is strongest; as $\beta \to 1$, this component vanishes and the hidden subclusters become harder to recover.

After generating $\mathbf{X}$, we uniformly mask 60% of its entries and apply each matrix completion method. We then run K-means on the completed matrix and evaluate recovery of the hidden subcluster labels s. This experiment tests whether using observed group information during completion helps preserve latent row geometry, rather than merely improving entrywise reconstruction. This simulation can be viewed through an effects-decomposition lens: the observed group labels act as nuisance factors, and GAME uses them to regularize group-specific variation so that the completed matrix better preserves the crossed latent subcluster geometry.



(a) ARI (b) NMI

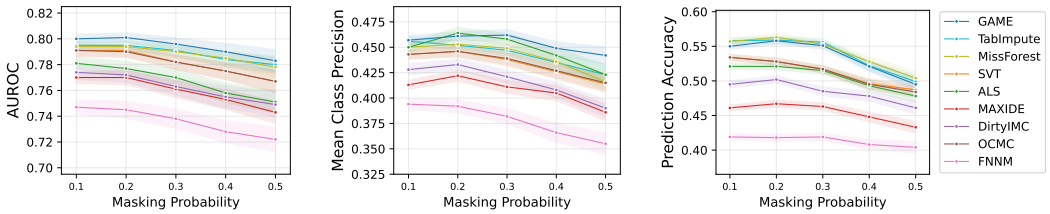
Figure 3: Synthetic clustering recovery of hidden subclusters under varying subcluster signal strength β, evaluated by ARI (↑) and NMI (↑). GAME nearly matches the fully observed oracle at small and moderate β, outperforming all baselines. Shaded regions denote ± 1 standard error across replications.

Results. Figure 3 shows that GAME most accurately recovers the hidden subclusters across small and moderate values of β, nearly matching the fully observed oracle that clusters X_{full}. In contrast, global completion methods and side-information baselines degrade earlier as the subcluster signal weakens, indicating that they preserve less of the local row geometry after completion. As $\beta \to 1$, the subcluster-specific component vanishes, the oracle performance drops, and all methods converge to similar clustering performance.

These results support the central intuition of GAME: when observed meta-categories carry low-rank structure, group-aware regularization can account for group-specific variation while preserving the crossed latent directions that define hidden subclusters.

4.2 MovieLens-100k Dataset: Missingness and Meta-Category Robustness

Setup. The MovieLens-100k dataset [Harper and Konstan, 2015] contains 100,000 ratings from 943 users on 1,682 movies, together with user metadata including gender, age, occupation, and ZIP code. We use age and gender as meta-categories and omit occupation and ZIP code. We



(a) AUROC (b) Mean Class Precision (c) Prediction Accuracy

Figure 4: Downstream species-classification performance on BirdSet HSN after imputing masked MFCC features under label-structured missingness, evaluated by AUROC (↑), cmAP (↑), and top-1 prediction accuracy (↑). GAME-completed features preserve classification signal across masking probabilities, remaining competitive or best among all methods. Shaded regions denote ± 1 standard error across 3 masking realizations.

evaluate matrix completion by holding out observed ratings and reporting test $\mathrm{RMSE}(\widehat{\mathbf{W}}) = \sqrt{\frac{1}{|\Omega_{\mathrm{test}}|}\sum_{(i,j)\in\Omega_{\mathrm{test}}}\left(\widehat{W}_{ij} - X_{ij}\right)^2}$, where Ω_{test} is the held-out set of observed ratings.

We consider three regimes. First, in the *global missingness* setting, we uniformly mask a proportion p_{global} of observed ratings. Second, in the *block-wise missingness* setting, we first hold out 10% of ratings globally, then additionally mask a proportion p_{block} of ratings from users aged $35+$. This setting tests whether methods can recover ratings for a systematically under-observed subgroup by borrowing information through overlapping demographic structure.

Third, we evaluate robustness to mislabeled metadata. We impose 50% global missingness and mislabel the user-level meta-categories in the training set. For each user, we randomly select either gender or age with equal probability, then flip the selected label with probability p_{swap}. We sweep p_{swap} from 0 to 0.8 to assess sensitivity to unreliable demographic metadata.

Results. Figure 2 shows GAME is consistently among the strongest methods across the three MovieLens settings. We omit ALS because it yields substantially weaker reconstruction performance. Under global missingness, GAME achieves the lowest RMSE across all masking probabilities and remains stable as sparsity increases. Its advantage is largest under block-wise missingness, where GAME attains the lowest RMSE across nearly all values of p_{block} and the gap widens as subgroup missingness increases. Under meta-category mislabeling, GAME remains stable as p_{swap} increases and outperforms the side-information baselines. FNNM is the strongest competing method.

Together, these results suggest that overlapping demographic structure is most useful when missingness is itself structured by subgroup. The mislabeling experiment further indicates that GAME does not rely blindly on metadata: mislabeled rows weaken the alignment between demographic groups and rating patterns, but the estimator can still borrow strength from the observed ratings and the remaining correctly specified group memberships.

4.3 Multi-label BirdSet: Downstream Classification

Setup. BirdSet [Rauch et al., 2025] is a large-scale benchmark for birdsong classification. We use the High Sierras Nevada (HSN) subset [Clapp et al., 2023], which contains 5460 focal training recordings across 21 bird species, and 12,000 soundscape test recordings that may contain multiple bird species. We exclude rare species, unlabeled samples, and low-quality recordings.

Because GAME and the baselines operate on matrices, we do not apply them directly to raw audio waveforms. Instead, we convert each audio clip into a fixed-length feature vector using Mel-frequency cepstral coefficients (MFCCs), producing a feature matrix $\mathbf{X} \in \mathbb{R}^{n\times 40}$. Rows correspond to audio clips and columns correspond to MFCC features. Here, we use bird species and recording location as meta-categorical information when applying GAME and other baselines.

We use species labels to define observed group structure for matrix completion, but the labels themselves are not imputed. We then impose structured block missingness by masking a fraction of MFCC entries among selected species. This creates a label-structured missingness pattern. After

completing the feature matrix, we evaluate whether the imputed features preserve information useful for downstream multi-label species classification.

Results. Figure 4 evaluates whether the completed MFCC features preserve information useful for downstream multi-label species classification. For each completed feature matrix, we train a random forest classifier and report AUROC, class mean average precision (cmAP), and top-1 prediction accuracy on the multi-label soundscape test set.

GAME achieves strong downstream performance across all three metrics as the masking probability increases. It is consistently among the best methods for AUROC and cmAP, indicating that the completed features preserve species-discriminative ranking information. For top-1 prediction accuracy, GAME remains competitive, although the modern tabular imputation baselines perform slightly better at some masking levels. Overall, these results suggest that group-aware completion can preserve downstream classification signal under label-structured missingness, even when the evaluation is not based directly on reconstruction error.

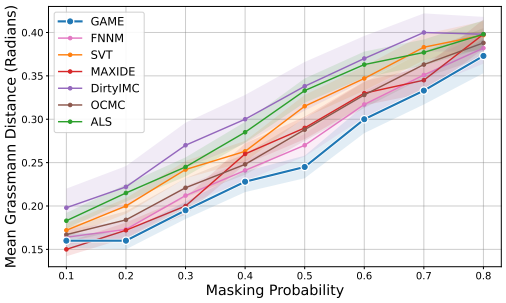


Figure 5: **Latent subspace recovery on the Svoboda Lab Neuropixels dataset**, measured by mean Grassmann distance (↓) between recovered and fully observed striatal subspaces. GAME attains the lowest distance across masking probabilities, indicating best recovery of latent neural dynamics. Shaded regions denote ± 1 standard error across 3 masking realizations.

4.4 Svoboda Lab Neuropixels Dataset

The Svoboda dataset [Chen et al., 2024] consists of large-scale, simultaneous neuronal spike activity of mice performing standardized behavioral tasks (memory-guided directional licking stimulated at $t = 50$), repeated across many recording sessions and spanning prespecified brain regions (Figure 1). For a single trial, we observe a matrix $\mathbf{X} \in \mathbb{R}^{N \times T}$, where N is the number of neurons recorded in the selected regions and sessions, and T is the number of time bins. Matrix entries are 1 if the neuron spiked at time t and 0 otherwise. Figure 9 shows the structured missingness of the dataset for regions measured in 20 of the most recorded sessions. We select recording sessions 123, 146, and 134, and within these sessions, select only regions 124 (midbrain reticular nucleus), 248 (striatum), and 258 (superior colliculus). This results in $N = 50,000$ and $T = 400$. As we observe in the region-session dynamics from PCA (Figure 10) and in the mean neuron firing rates in Figure 11, the same region can have differing temporal dynamics even when controlling for task, since here we select only from a single trial type.

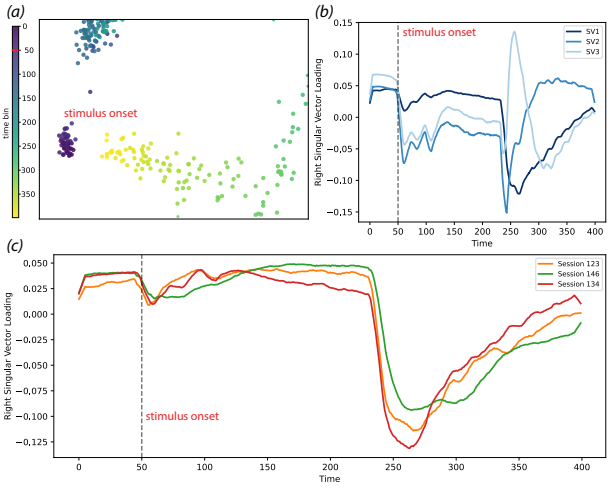


Figure 6: **GAME identifies meaningful neural dynamics while accounting for time-dependent batch effects from different recording sessions.** *(a)* 2D singular vectors from GAME from the striatum over time. *(b)* Top 3 singular vectors show clear neural dynamic response to stimulus and from the onset of mouse movement ($\sim t = 250$). *(c)* Temporal session-specific batch effects from GAME.

In this analysis, region and session latent dynamics overlap significantly, and we cannot expect $V_r \perp V_s$. Generally, region dynamics are of greater interest, where sessions are treated as nuisance

parameters. We therefore include sessions as meta-categories alongside regions and evaluate recovery of the striatal latent dynamics with the Grassmann distance [Ye and Lim, 2016]. Given orthonormal bases $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{T\times k}$ spanning the top-k right singular vectors of the striatum block in the completed and fully observed matrices, the principal angles θ_i satisfy $\cos\theta_i = \sigma_i(\mathbf{U}^\top\mathbf{V})$, and the Grassmann distance is $\left(\sum_{i=1}^k \theta_i^2\right)^{1/2}$, with smaller values indicating closer alignment.

Results. We evaluate matrix completion when masking the striatum, a region with distinct temporal dynamics from other experimentally measured regions [Chen et al., 2024]. As shown in Figure 5, GAME consistently outperforms comparison methods in Grassmann distance, indicating more accurate recovery of latent neural dynamics. In the setting of Neuropixels data, true latent dynamics are of much greater scientific interest [Pandarinath et al., 2018, Churchland et al., 2012, Vyas et al., 2020] than the imputation of individual neuron spike trains which are noisy and highly varying [Boussard et al., 2021]. In Figure 6(a), the top two singular vectors show a shared latent dynamic learned across sessions after adjusting for session-specific effects; compared with the striatum column in Figure 9 and firing rates in Figure 11, these dynamics separate pre-stimulus activity ($0 < t < 50$), pre-movement activity, and movement onset ($\sim 250 < t < 400$). Figure 6(b) further shows the top three singular vectors, revealing stimulus-locked activity near $t = 50$ and additional movement-related dynamics near $t = 250$, while Figure 6(c) shows that session-specific effects are minimal through stimulus onset but become more pronounced during movement. Together, these results recapitulate the striatal dynamics identified in [Chen et al., 2024], but do so directly from the learned decomposition without requiring extensive electrophysiological validation or domain-specific annotation.

5 Discussion

In this work, we presented Group-Aware Matrix Estimation (GAME), a convex framework for heterogeneous matrix completion that leverages overlapping meta-categorical information to promote subgroup-wise low-rank structure. We developed finite-sample theoretical guarantees for both Frobenius reconstruction error and per-category subspace recovery, showing that GAME's statistical complexity depends on the sum of local category ranks rather than the global rank, yielding improved guarantees over standard nuclear-norm regularization when subgroups exhibit distinct low-rank geometry. We further introduced an efficient proximal-averaging optimization procedure that avoids the consensus overhead of ADMM, making GAME practical even when the number of overlapping categories is large. From collaborative filtering on user-movie ratings, to multi-label classification from acoustic features in ecological monitoring, to recovery of latent neural dynamics from large-scale electrophysiological recordings, GAME consistently performed competitively or best, particularly when subgroups carried distinct low-rank structure and missingness was non-uniform. In neural electrophysiological data, GAME further recovered low-dimensional neural dynamics despite substantial session-specific batch effects.

A current limitation is that GAME's advantages depend on the meta-categories capturing meaningful subgroup-specific low-rank structure; when metadata are weakly informative, frequently mislabeled, or unrelated to the missingness pattern, the estimator may offer limited improvement over global methods. In practice, the choice of meta-categories should be guided by domain knowledge of the problem. GAME's use of domain-specified groups keeps the practitioner in the loop. However, with limited domain knowledge, one can keep groups that maximize the orthogonality of the right-singular vectors of the submatrices they induce, as we did when binning MovieLens ages and selecting Neuropixels regions and sessions, and merge groups too small to satisfy the per-category sample requirement of Theorem 3.5 into groups with non-orthogonal right-singular vectors. When no metadata are available, clustering a naively imputed matrix can *potentially* supply group labels (Appendix D.1). Our guarantees assume exact proximal updates, so analyzing GAME with truncated SVDs, as used in our Neuropixels study, is an avenue for future work. On the experimental side, future work on Neuropixels data will incorporate meta-categories for different trials and different mice, further testing GAME's capacity to disentangle overlapping sources of neural variability. The Svoboda data in full contains 293 neural regions and 173 recording sessions [Chen et al., 2024]; big-data extensions could incorporate other publicly available Neuropixels data from the International Brain Laboratory [Angelaki et al., 2025] and include site or laboratory as a meta-category.

Acknowledgements

G.I.A. acknowledges funding from NSF DMS-1554821. We thank Ji Xia for her help and expertise in processing the Svoboda Neuropixels data.

References

Dora Angelaki, Brandon Benson, Julius Benson, Daniel Birman, Niccolò Bonacchi, Kcénia Bougrova, Sebastian A Bruijns, Matteo Carandini, Joana A Catarino, Gaelle A Chapuis, Anne K Churchland, Yang Dan, Felicia Davatolhagh, Peter Dayan, Eric Ej DeWitt, Tatiana A Engel, Michele Fabbri, Mayo Faulkner, Ila Rani Fiete, Charles Findling, Laura Freitas-Silva, Berk Gerçek, Kenneth D Harris, Michael Häusser, Sonja B Hofer, Fei Hu, Félix Hubert, Julia M Huntenburg, Anup Khanal, Christopher S Krasniak, Christopher Langdon, Christopher Langfield, Petrina Y P Lau, Zachary F Mainen, Guido T Meijer, Nathaniel J Miska, Thomas D Mrsic-Flogel, Jean-Paul Noel, Kai Nylund, Alejandro Pan-Vazquez, Liam Paninski, Alexandre Pouget, Cyrille Rossant, Noam Roth, Rylan Schaeffer, Michael Schartner, Yanliang Shi, Karolina Z Socha, Nicholas A Steinmetz, Karel Svoboda, Anne E Urai, Miles J Wells, Steven J West, Matthew R Whiteway, Olivier Winter, Ilana B Witten, and International Brain Laboratory. A brain-wide map of neural activity during complex behaviour. *Nature*, 645(8079):177–191, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09235-0.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008. doi: 10.1007/s10994-007-5040-8. URL https://doi.org/10.1007/s10994-007-5040-8.

Mehdi Azabou, Krystal Xuejing Pan, Vinam Arora, Ian Jarratt Knight, Eva L Dyer, and Blake Aaron Richards. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IuU0wcO0mo.

Heinz H Bauschke, Rafal Goebel, Yves Lucet, and Xianfu Wang. The proximal average: basic theory. *SIAM Journal on Optimization*, 19(2):766–785, 2008.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542. URL https://doi.org/10.1137/080716542.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

Julien Boussard, Erdem Varol, Hyun Dong Lee, Nishchal Dethe, and Liam Paninski. Three-dimensional spike localization and improved motion correction for neuropixels recordings. *Advances in Neural Information Processing Systems*, 34:22095–22105, 2021.

Stephen Boyd, Neal Parikh, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3:1–122, 07 2011. doi: 10.1561/2200000016.

E. Kelly Buchanan, Ian Kinsella, Ding Zhou, Rong Zhu, Pengcheng Zhou, Felipe Gerhard, John Ferrante, Ying Ma, Sharon Kim, Mohammed Shaik, Yajie Liang, Rongwen Lu, Jacob Reimer, Paul Fahey, Taliah Muhammad, Graham Dempsey, Elizabeth Hillman, Na Ji, Andreas Tolias, and Liam Paninski. Penalized matrix decomposition for denoising, compression, and improved demixing of functional imaging data, 2018. URL https://arxiv.org/abs/1807.06203.

Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010. doi: 10.1137/080738970. URL https://doi.org/10.1137/080738970.

Emmanuel J. Candes and Benjamin Recht. Exact matrix completion via convex optimization, 2008. URL https://arxiv.org/abs/0805.4471.

Gengshuo Chang, Wei Zhang, and Lehan Zhang. Matrix completion with incomplete side information via orthogonal complement projection. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 7393–7416. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chang25e.html.

Susu Chen, Yi Liu, Ziyue Aiden Wang, Jennifer Colonell, Liu D Liu, Han Hou, Nai-Wen Tien, Tim Wang, Timothy Harris, Shaul Druckmann, Nuo Li, and Karel Svoboda. Brain-wide neural activity underlying memory-guided movement. *Cell*, 187(3):676–691.e16, 2024. ISSN 1097-4172. doi: 10.1016/j.cell.2023.12.035.

Kai-Yang Chiang, Cho-Jui Hsieh, and Inderjit S Dhillon. Matrix completion with noisy side information. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/0609154fa35b3194026346c9cac2a248-Paper.pdf.

Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487 (7405):51–56, 2012.

Mary Clapp, Stefan Kahl, Erik Meyer, Megan McKenna, Holger Klinck, and Gail Patricelli. A collection of fully-annotated soundscape recordings from the southern Sierra Nevada mountain range. Dataset, 2023. URL https://zenodo.org/records/7525805.

Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013. doi: 10.1109/TPAMI.2013.57.

Jacob Feitelberg, Dwaipayan Saha, Kyuseong Choi, Zaid Ahmad, Anish Agarwal, and Raaz Dwivedi. One pipeline, many transformers: Pattern-specific imputation specialists for tabular missing data, 2026. URL https://arxiv.org/abs/2510.02625.

Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2 (1):17–40, 1976. ISSN 0898-1221. doi: https://doi.org/10.1016/0898-1221(76)90003-1. URL https://www.sciencedirect.com/science/article/pii/0898122176900031.

R. Glowinski and A. Marrocco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires. *Revue Française d'Automatique, Informatique, et Recherche Opérationnelle*, 9:41–76, 1975.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), December 2015. ISSN 2160-6455. doi: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

Trevor Hastie, Rahul Mazumder, Jason Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares, 2014. URL https://arxiv.org/abs/1410.2596.

Prateek Jain and Inderjit S. Dhillon. Provable inductive matrix completion, 2013. URL https://arxiv.org/abs/1306.0626.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization, 2012. URL https://arxiv.org/abs/1212.0467.

Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12(67):2297–2334, 2011. URL http://jmlr.org/papers/v12/jenatton11a.html.

T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, May 2005. ISSN 2168-894X. doi: 10.1214/009117905000000044. URL http://dx.doi.org/10.1214/009117905000000044.

Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013. doi: 10.1109/TPAMI.2012.88.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80):2287–2322, 2010. URL http://jmlr.org/papers/v11/mazumder10a.html.

Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise, 2011. URL https://arxiv.org/abs/1009.2118.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, November 2012. ISSN 2168-8745. doi: 10.1214/12-sts400. URL http://dx.doi.org/10.1214/12-STS400.

Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach, 2011. URL https://arxiv.org/abs/1110.0413.

Chethan Pandarinath, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, Jonathan C. Kao, Eric M. Trautmann, Matthew T. Kaufman, Stephen I. Ryu, Leigh R. Hochberg, Jaimie M. Henderson, Krishna V. Shenoy, L. F. Abbott, and David Sussillo. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15 (10):805–815, 2018. doi: 10.1038/s41592-018-0109-9. URL https://doi.org/10.1038/s41592-018-0109-9.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, January 2014. ISSN 2167-3888. doi: 10.1561/2400000003. URL https://doi.org/10.1561/2400000003.

Arthur Pellegrino, Heike Stein, and N. Alex Cayco-Gajic. Dimensionality reduction beyond neural subspaces with slice tensor component analysis. *Nature Neuroscience*, 27(6):1199–1210, 2024. doi: 10.1038/s41593-024-01626-2. URL https://doi.org/10.1038/s41593-024-01626-2.

Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, and Christoph Scholz. Birdset: A large-scale dataset for audio classification in avian bioacoustics, 2025. URL https://arxiv.org/abs/2403.10380.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, January 2010. ISSN 1095-7200. doi: 10.1137/070697835. URL http://dx.doi.org/10.1137/070697835.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4): 389–434, August 2012. ISSN 1615-3375.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596. URL https://www.cambridge.org/core/books/highdimensional-probability/797C466DA29743D2C8213493BD2D2102.

Saurabh Vyas, Matthew D Golub, David Sussillo, and Krishna V Shenoy. Computation through neural population dynamics. *Annual review of neuroscience*, 43(1):249–275, 2020.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Jiuzhou Wang and Eric F Lock. Multiple augmented reduced rank regression for pan-cancer analysis. *Biometrics*, 80(1):ujad002, 03 2024. ISSN 1541-0420. doi: 10.1093/biomtc/ujad002. URL https://doi.org/10.1093/biomtc/ujad002.

Alex H. Williams, Tony Hyun Kim, Forea Wang, Saurabh Vyas, Stephen I. Ryu, Krishna V. Shenoy, Mark Schnitzer, Tamara G. Kolda, and Surya Ganguli. Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. *Neuron*, 98(6):1099–1115.e8, 2018. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2018.05.015. URL https://www.sciencedirect.com/science/article/pii/S0896627318303878.

Ji Xia, Yizi Zhang, Shuqi Wang, Genevera I. Allen, Liam Paninski, Cole Lincoln Hurwitz, and Kenneth D. Miller. Inpainting the neural picture: Inferring unrecorded brain area dynamics from multi-animal datasets, 2025. URL https://arxiv.org/abs/2510.11924.

Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/e58cc5ca94270acaceed13bc82dfedf7-Paper.pdf.

Mengyun Yang, Yaohang Li, and Jianxin Wang. Feature and nuclear norm minimization for matrix completion. *IEEE Transactions on Knowledge and Data Engineering*, 34(5):2190–2199, 2022. doi: 10.1109/TKDE.2020.3005978.

Ke Ye and Lek-Heng Lim. Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1176–1197, 2016.

Byron M Yu, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102(1):614–635, 2009. ISSN 0022-3077. doi: 10.1152/jn.90941.2008.

Y. YU, T. WANG, and R. J. SAMWORTH. A useful variant of the davis—kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. ISSN 00063444. URL http://www.jstor.org/stable/43908537.

Yao-Liang Yu. Better approximation and faster algorithm using the proximal average. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/49182f81e6a13cf5eaa496d51fea6406-Paper.pdf.

Yizi Zhang, Yanchen Wang, Donato M. Jiménez-Benetó, Zixuan Wang, Mehdi Azabou, Blake Richards, Renee Tung, Olivier Winter, The International Brain Laboratory, Eva Dyer, Liam Paninski, and Cole Hurwitz. Towards a "universal translator" for neural dynamics at single-cell, single-spike resolution. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 80495–80521. Curran Associates, Inc., 2024. doi: 10.52202/079017-2559. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/934eb45b99eff8f16b5cb8e4d3cb5641-Paper-Conference.pdf.

Yizi Zhang, Hanrui Lyu, Cole Hurwitz, Shuqi Wang, Charles Findling, Yanchen Wang, Felix Hubert, Alexandre Pouget, Erdem Varol, and Liam Paninski. Exploiting correlations across trials and behavioral sessions to improve neural decoding. *Neuron*, 114(3):536–551, 2026.

A Theory for Proximal Averaging Convergence Guarantees

A.1 Proximal Averaging Background

Definition A.1 (Proximal Operator)**.** Let $x, z \in \mathbb{X}$, where $\mathbb{X}$ is a real Hilbert space. We define the *proximal operator* of a function f as a mapping

$$\text{prox}_{\gamma f}(x) = \arg\min_z \frac{1}{2\gamma}||x - z||_2^2 + f(z).$$

A function is *simple* if it has a closed form proximal operator. Proximal operators exhibit many desirable algebraic properties that prove to be extremely useful in solving convex optimization problems. A particularly nice property of proximal operators is utilized in constructing the GAME solution. More generally, an optimization procedure known as *proximal gradient descent* is successful if the objective is separable into a convex and differentiable component f, and another convex, but not necessarily differentiable, component g [Parikh and Boyd, 2014]:

$$\arg\min_{x \in \mathbb{X}} f(x) + g(x).$$

To minimize the following general multi-term, nonsmooth convex minimization problem

$$\arg\min_{x \in \mathbb{X}} F(x) = f(x) + \bar{g}(x), \quad \text{where } \bar{g}(x) = \sum_{k=1}^{K} \alpha_k g_k(x) \tag{6}$$

Here,

$$\alpha_k \geq 0, \qquad \sum_{k=1}^{K} \alpha_k = 1,$$

each $g_k : \mathbb{X} \to [-\infty, \infty]$ is a proper closed convex function, and $f : \mathbb{X} \to (-\infty, \infty)$ is a continuously differentiable and gradient L_f Lipschitz convex function. This is exactly the minimization problem [Yu, 2013] considered, relying on proximal averaging approximations to solve the (6).

Definition A.2 (Proximal Average, Bauschke et al. 2008 [Bauschke et al., 2008])**.** We define the *Proximal Average function* of $\bar{g}(x) = \sum_{k=1}^{K} \alpha_k g_k(x)$ with parameter $\gamma > 0$ is defined by the optimization problem

$$g_\gamma(x) := \min_{y_k} \left\{ \sum_{k=1}^{K} \alpha_k g_k(y_k) + \frac{1}{2\gamma} \left(\sum_{k=1}^{K} \alpha_k \|y_k\|^2 - \|x\|^2 \right) \quad : \quad \sum_{k=1}^{K} \alpha_k y_k = x \right\}$$

The proximal operator of the Proximal Average function satisfies the nice property

$$\text{prox}_{g_\gamma, \gamma} = \alpha_1 \text{prox}_{g_1, \gamma} + \cdots + \alpha_K \text{prox}_{g_K, \gamma}$$

This result says that the proximal operator of g_γ with respect to the same step size γ is additive in the individual proximal operators of each g_k. The power of this formulation stems from the fact that the proximal operator of the function $\bar{g}(x)$ might not have a nice closed form equation, but each of the individual g_k's do, allowing us to construct an approximation with convergence guarantees. With the construction of a Proximal Average, [Yu, 2013] solve (6) via the approximation

$$\arg\min_{x \in \mathbb{X}} f(x) + g_\gamma(x) \tag{7}$$

Observe that this problem is of the same form as proximal gradient descent. [Yu, 2013] makes this exact connection, plugging the approximate objective function into the updates provided in proximal gradient descent. [Yu, 2013] also provides theoretical justifications for the aforementioned algorithm, under a few assumptions.

In particular, their proposed algorithm takes the form of the Fast Iterative Shrinkage–Thresholding Algorithm (FISTA) [Beck and Teboulle, 2009] outlined in Algorithm 2, as well as a non-accelerated

Algorithm 2 PA-APG [Yu, 2013].

Initialize $x_0 = y_1$, γ, $\eta_1 = 1$.
for $t = 1, 2, \ldots$ **do**
  $z_t = y_t - \gamma \nabla f(y_t)$,
  $x_t = \sum_k \alpha_k \cdot \text{prox}_{\gamma g_k}(z_t)$,
  $\eta_{t+1} = \dfrac{1 + \sqrt{1 + 4\eta_t^2}}{2}$,
  $y_{t+1} = x_t + \dfrac{\eta_t - 1}{\eta_{t+1}}(x_t - x_{t-1})$.
end for

Algorithm 3 PA-PG [Yu, 2013].

Initialize x_0, γ.
for $t = 1, 2, \ldots$ **do**
  $z_t = x_{t-1} - \gamma \nabla f(x_{t-1})$,
  $x_t = \sum_k \alpha_k \cdot \text{prox}_{\gamma g_k}(z_t)$.
end for

ISTA in Algorithm 3. The accelerated version is referred to as "accelerated proximal gradient" (PA-APG), whereas the non-accelerated counterpart is PA-PG.

Denote $\bar{L}^2 = \sum_{k=1}^K \alpha_k L_k^2$ as a convex combination of each Lipschitz constant L_k.

Theorem A.3 (Theorem 1 of [Yu, 2013]). *Assuming the following:*

1. *Each g_k is convex and L_k-Lipschitz continuous w.r.t. the Euclidean norm.*
2. *Each g_k is simple.*
3. *f is convex with L_f-Lipschitz continuous gradient w.r.t. the Euclidean norm.*

Let x_0 be the initialization of the PA-APG algorithm. Fix the accuracy $\epsilon > 0$. Under assumptions 1-3, let $\gamma = \min\{1/L_f, 2\epsilon/\bar{L}^2\}$. After at most $\sqrt{\frac{2}{\gamma\epsilon}}\|x_0 - x\|_2$ steps, the PA-APG approximation, $\tilde{x}$, satisfies

$$f(\tilde{x}) + \bar{g}(\tilde{x}) \leq f(x) + \bar{g}(x) + 2\epsilon.$$

The same guarantee holds for the non-accelerated Proximal Average approximation after at most $\dfrac{1}{2\gamma\epsilon}\|x_0 - x\|_F^2$ steps.

A.2 Solving GAME via Proximal Averaging

The Proximal Average algorithm [Yu, 2013] requires $\sum_{c \in \mathcal{C}} \alpha_c = 1$. This yields us our final GAME optimization problem

$$\widehat{\mathbf{W}} \in \arg\min_{\mathbf{W}} \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{X} - \mathbf{W})\|_F^2 + \lambda \sum_{c \in \mathcal{C}} \alpha_c \|\mathbf{D}_c \mathbf{W}\|_*. \tag{8}$$

Here, $\mathbf{D}_c \in \{0, 1\}^{n_c \times n}$ is the row-selection matrix corresponding to category c, so that

$$\mathbf{W}_c = \mathbf{D}_c \mathbf{W}.$$

We define

$$\lambda_c := \lambda \alpha_c$$

to represent the amount of regularization assigned to category c. Conversely, any collection of positive category regularization parameters $\{\lambda_c\}_{c \in \mathcal{C}}$ can be represented by setting

$$\lambda = \sum_{c \in \mathcal{C}} \lambda_c, \qquad \alpha_c = \frac{\lambda_c}{\lambda}.$$

As discussed in the previous subsection, to leverage proximal averaging techniques for solving GAME (8), we first require $f(\mathbf{W}) := \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{X} - \mathbf{W})\|_F^2$ to be convex with L_f-Lipschitz continuous gradient (with respect to the Euclidean norm), and each $g_c(\mathbf{W}) := \lambda ||\mathbf{D_c W}||_*$ to be convex and L_c-Lipschitz continuous with respect to the Euclidean norm. Secondly, we require that each g_c is

simple (i.e. has a defined closed-form proximal operator). We claim that the assumptions hold for our GAME formulation (8) through the following propositions, constructing an analogous theorem to A.3.

It should be noted that although Yu [2013] states the results for $x \in \mathbb{R}^d$ with the ℓ_2 norm, they still apply to $\mathbb{R}^{n\times m}$ with the Frobenius norm: vectorization is an isometry, $\|\mathrm{vec}(\mathbf{X})\|_2 = \|\mathbf{X}\|_F$, and every argument in Yu [2013] is carried out in a real Hilbert space.

We recognize the GAME formulation (8) as a Proximal Average problem [Yu, 2013], where g_c are simple (easy to compute proximal operator). Thus, we aim to reconstruct the objective function into a composition of simple functions, but this formulation appears to be complex due to the penalties that depend on slices of $\mathbf{W}$ that are not disjoint (i.e. $I_c \cap I_{c'} \neq \emptyset$ for some $c \neq c'$). Recall that $\mathbf{W}_c := \mathbf{D}_c\mathbf{W}$ is the row-wise "slice" of $\mathbf{W}$ that corresponds to the c meta-category block.

The key observation to make at this stage is that $\mathbf{D_c}\mathbf{D_c}^\top = \mathbf{I}$, i.e. $\mathbf{D_c}$ is semi-orthogonal. Using Table 10.1 from [Combettes and Pesquet, 2011], we may rewrite the proximal operator of g_c in terms of linear operations of $\mathbf{D_c}$ and the proximal operator of the nuclear-norm, given in Proposition A.4.

Proposition A.4 (Table 10.1 of [Combettes and Pesquet, 2011])**.** *Define* $g_c(\mathbf{W}) := \lambda||\mathbf{D_c}\mathbf{W}||_*$. *Then, for every* $\gamma > 0$,

$$prox_{\gamma g_c}(\mathbf{A}) = prox_{\gamma\lambda||\mathbf{D_c}\mathbf{A}||_*}(\mathbf{A}) = \mathbf{A} - \mathbf{D}_c^\top(\mathbf{D}_c\mathbf{A} - prox_{\gamma\lambda||\cdot||_*}(\mathbf{D}_c\mathbf{A}))$$

From Proposition A.4, we see the nuclear-norm having a simple proximal operator is sufficient for g_c to have a simple proximal operator. The closed form proximal operator of the nuclear-norm is a well established fact [Parikh and Boyd, 2014].

Proposition A.5 (Nuclear-Norm Proximal Operator [Parikh and Boyd, 2014] Section 6.7.3)**.** *Let* $\tau > 0$ *and let* $\mathbf{Y} \in \mathbb{R}^{m\times n}$ *have the singular value decomposition of rank* r

$$\mathbf{Y} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^\top,\ \mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r),\ \sigma_1 \geq \cdots \geq \sigma_r > 0.$$

Then the proximal operator of the nuclear-norm,

$$\mathrm{prox}_{\tau\|\cdot\|_*}(\mathbf{Y}) \ := \ \underset{\mathbf{Z}\in\mathbb{R}^{m\times n}}{\arg\min}\left\{\tfrac{1}{2}\|\mathbf{Z}-\mathbf{Y}\|_F^2 + \tau\|\mathbf{Z}\|_*\right\},$$

is given by singular value soft-thresholding:

$$\mathrm{prox}_{\tau\|\cdot\|_*}(\mathbf{Y}) \ = \ \mathbf{U}\ \mathrm{diag}((\sigma_i - \tau)_+)\mathbf{V}^\top,$$

where $(t)_+ := \max\{t, 0\}$ *is the positive part applied entrywise to the singular values.*

The following series of propositions demonstrate that assumptions made by [Yu, 2013] (found in Theorem A.3 to establish convergence guarantees) are met by the GAME estimator.

Proposition A.6. *Let* $\mathbf{X} \in \mathbb{R}^{n\times m}$ *be fixed and let* $\mathbf{M} \in \{0,1\}^{n\times m}$ *be a binary mask encoding the observed entries, i.e.* $\mathbf{M}_{ij} = 1$ *if* (i,j) *is observed and* $\mathbf{M}_{ij} = 0$ *otherwise. Define the orthogonal projection* $\mathcal{P}_\Omega : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ *by*

$$\mathcal{P}_\Omega(\mathbf{Z}) \ := \ \mathbf{M} \odot \mathbf{Z},$$

where $\odot$ *denotes the Hadamard (entrywise) product. Consider the function*

$$f(\mathbf{W}) \ := \ \frac{1}{2}\left\|\mathcal{P}_\Omega(\mathbf{X}-\mathbf{W})\right\|_F^2, \qquad \mathbf{W} \in \mathbb{R}^{n\times m}.$$

Then f *is convex and differentiable, with gradient*

$$\nabla f(\mathbf{W}) \ = \ \mathcal{P}_\Omega(\mathbf{W}-\mathbf{X}) \ = \ \mathbf{M}\odot(\mathbf{W}-\mathbf{X}).$$

Moreover, ∇f *is* L_f*–Lipschitz continuous with respect to the Frobenius norm with constant* $L_f = 1$, *that is, for all* $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{n\times m}$,

$$\begin{aligned}\left\|\nabla f(\mathbf{W}_1) - \nabla f(\mathbf{W}_2)\right\|_F \ &\leq \ L_f\left\|\mathbf{W}_1 - \mathbf{W}_2\right\|_F \\ &= \ \left\|\mathbf{W}_1 - \mathbf{W}_2\right\|_F.\end{aligned}$$

Proof. We first rewrite f using the definition of $\mathcal{P}_\Omega$:

$$f(\mathbf{W}) = \frac{1}{2}\big\|\mathcal{P}_\Omega(\mathbf{X} - \mathbf{W})\big\|_F^2 = \frac{1}{2}\sum_{i,j}\big(\mathbf{M}_{ij}(\mathbf{X}_{ij} - \mathbf{W}_{ij})\big)^2.$$

Since each term $\frac{1}{2}\big(\mathbf{M}_{ij}(\mathbf{X}_{ij} - \mathbf{W}_{ij})\big)^2$ is a convex function of $\mathbf{W}_{ij}$ (a nonnegative scalar multiple of a quadratic), and f is a finite sum of convex functions, f is convex.

Next, f is differentiable, and we can compute its gradient entrywise. For each (i, j),

$$f(\mathbf{W}) = \frac{1}{2}\sum_{i,j}\mathbf{M}_{ij}^2(\mathbf{W}_{ij} - \mathbf{X}_{ij})^2 = \frac{1}{2}\sum_{i,j}\mathbf{M}_{ij}(\mathbf{W}_{ij} - \mathbf{X}_{ij})^2,$$

since $\mathbf{M}_{ij}^2 = \mathbf{M}_{ij}$ for $\mathbf{M}_{ij} \in \{0, 1\}$. Differentiating with respect to $\mathbf{W}_{ij}$,

$$\frac{\partial f}{\partial \mathbf{W}_{ij}} = \mathbf{M}_{ij}(\mathbf{W}_{ij} - \mathbf{X}_{ij}).$$

Thus the gradient matrix has entries

$$\big[\nabla f(\mathbf{W})\big]_{ij} = \mathbf{M}_{ij}(\mathbf{W}_{ij} - \mathbf{X}_{ij}),$$

which we can write compactly as

$$\nabla f(\mathbf{W}) = \mathbf{M} \odot (\mathbf{W} - \mathbf{X}) = \mathcal{P}_\Omega(\mathbf{W} - \mathbf{X}).$$

It remains to show that ∇f is 1–Lipschitz with respect to the Frobenius norm. Let $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{n\times m}$. Then

$$\begin{aligned}\nabla f(\mathbf{W}_1) - \nabla f(\mathbf{W}_2) &= \mathcal{P}_\Omega(\mathbf{W}_1 - \mathbf{X}) - \mathcal{P}_\Omega(\mathbf{W}_2 - \mathbf{X}) \\ &= \mathcal{P}_\Omega(\mathbf{W}_1 - \mathbf{W}_2) \\ &= \mathbf{M} \odot (\mathbf{W}_1 - \mathbf{W}_2).\end{aligned}$$

Therefore,

$$\begin{aligned}\big\|\nabla f(\mathbf{W}_1) - \nabla f(\mathbf{W}_2)\big\|_F^2 &= \big\|\mathbf{M} \odot (\mathbf{W}_1 - \mathbf{W}_2)\big\|_F^2 \\ &= \sum_{i,j}\big(\mathbf{M}_{ij}(\mathbf{W}_1 - \mathbf{W}_2)_{ij}\big)^2 \\ &\le \sum_{i,j}\big((\mathbf{W}_1 - \mathbf{W}_2)_{ij}\big)^2 \\ &= \|\mathbf{W}_1 - \mathbf{W}_2\|_F^2,\end{aligned}$$

where the inequality uses $\mathbf{M}_{ij} \in \{0, 1\}$, so $\mathbf{M}_{ij}^2 \le 1$ entrywise. Taking square roots yields

$$\big\|\nabla f(\mathbf{W}_1) - \nabla f(\mathbf{W}_2)\big\|_F \;\le\; \|\mathbf{W}_1 - \mathbf{W}_2\|_F,$$

showing that ∇f is 1–Lipschitz with respect to the Frobenius norm. This proves the claim with $L_f = 1$. □

Proposition A.7. *$g_c(\mathbf{W}) = \lambda\|\mathbf{D_c}\mathbf{W}\|_*$ is Lipschitz continuous with respect to the Frobenius norm on $\mathbb{R}^{n\times m}$ with Lipschitz constant*

$$L_c = \lambda\sqrt{r}, \qquad r := \min\{n_c, m\}.$$

That is, for all $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{n\times m}$,

$$\big|g_c(\mathbf{W}_1) - g_c(\mathbf{W}_2)\big| \;\le\; L_c\,\|\mathbf{W}_1 - \mathbf{W}_2\|_F.$$

Proof. We first show that the nuclear-norm is $\sqrt{r}$-Lipschitz with respect to the Frobenius norm, where $r = \min\{p, q\}$. Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times q}$ and let $\sigma_1(\mathbf{A}) \ge \cdots \ge \sigma_r(\mathbf{A})$ and $\sigma_1(\mathbf{B}) \ge \cdots \ge \sigma_r(\mathbf{B})$

denote their singular values. Then

$$\begin{aligned}
\big|\|\mathbf{A}\|_* - \|\mathbf{B}\|_*\big| &= \left|\sum_{j=1}^{r}\sigma_j(\mathbf{A}) - \sum_{j=1}^{r}\sigma_j(\mathbf{B})\right| \\
&= \left|\sum_{j=1}^{r}\big(\sigma_j(\mathbf{A}) - \sigma_j(\mathbf{B})\big)\right| \\
&\le \left(\sum_{j=1}^{r}1^2\right)^{1/2}\left(\sum_{j=1}^{r}\big(\sigma_j(\mathbf{A}) - \sigma_j(\mathbf{B})\big)^2\right)^{1/2} \qquad \text{(Cauchy–Schwarz)} \\
&= \sqrt{r}\left(\sum_{j=1}^{r}\big(\sigma_j(\mathbf{A}) - \sigma_j(\mathbf{B})\big)^2\right)^{1/2}.
\end{aligned}$$

By the Hoffman–Wielandt inequality for singular values,

$$\sum_{j=1}^{r}\big(\sigma_j(\mathbf{A}) - \sigma_j(\mathbf{B})\big)^2 \;\le\; \|\mathbf{A}-\mathbf{B}\|_F^2,$$

so we obtain

$$\big|\|\mathbf{A}\|_* - \|\mathbf{B}\|_*\big| \;\le\; \sqrt{r}\,\|\mathbf{A}-\mathbf{B}\|_F.$$

This shows that the nuclear-norm is $\sqrt{r}$-Lipschitz with respect to the Frobenius norm.

Now apply this with $\mathbf{A} = \mathbf{D}_c\mathbf{W}_1$ and $\mathbf{B} = \mathbf{D}_c\mathbf{W}_2$ for arbitrary $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{n\times m}$. We obtain

$$\begin{aligned}
|g_c(\mathbf{W}_1) - g_c(\mathbf{W}_2)| &= \lambda\,\big|\|\mathbf{D}_c\mathbf{W}_1\|_* - \|\mathbf{D}_c\mathbf{W}_2\|_*\big| \\
&\le \lambda\sqrt{r}\,\|\mathbf{D}_c(\mathbf{W}_1 - \mathbf{W}_2)\|_F \\
&\le \lambda\sqrt{r}\,\|\mathbf{D}_c\|_{\mathrm{op}}\|\mathbf{W}_1 - \mathbf{W}_2\|_F \\
&= \lambda\sqrt{r}\,\|\mathbf{W}_1 - \mathbf{W}_2\|_F.
\end{aligned}$$

where in the last line we used submultiplicativity of the operator norm. We further note that since $\mathbf{D}_c$ is a row-selection matrix, $\|\mathbf{D}_c\|_{\mathrm{op}} = 1$. Thus g_c is Lipschitz with constant $L_c = \lambda\sqrt{r}\|\mathbf{D}_c\|_{\mathrm{op}} = \lambda\sqrt{r}$, as claimed. □

With the previous lemmas, we are able to give an analogous statement of Theorem A.3. We use the abbreviation

$$\overline{L^2} = \lambda^2\sum_{c\in\mathcal{C}}\alpha_c\min(n_c, m).$$

in the following theorem, and observe that the masked loss has $L_f = 1$, as previously shown.

Theorem A.8 (Theorem 1 of [Yu, 2013])**.** *Let $\mathbf{W}_0$ be the initialization of the PA-APG algorithm. Fix the accuracy $\epsilon > 0$. Let $\gamma = \min\{1, 2\epsilon/\overline{L^2}\}$. After at most $k = \sqrt{\frac{2}{\gamma\epsilon}}\|\mathbf{W}_0 - \widehat{\mathbf{W}}\|_F$ steps, the PA-APG approximation, $\widetilde{\mathbf{W}}_k$, satisfies*

$$f(\widetilde{\mathbf{W}}_k) + \bar{g}(\widetilde{\mathbf{W}}_k) \le f(\widehat{\mathbf{W}}) + \bar{g}(\widehat{\mathbf{W}}) + 2\epsilon.$$

B Motivating Example

Consider movie ratings from two categories of users, forming row blocks $\mathbf{W}_1, \mathbf{W}_2$ of $n/2$ users each, with all entries observed under Gaussian noise of variance σ^2. Some latent directions are *shared*: both categories contribute signal. Others are *local*: one category contributes signal and the other only noise, which we call the *leak*. A good denoiser should pool evidence for shared directions, keep local directions, and discard the leak.

Two noise levels. Gaussian noise on an $N \times m$ matrix has operator norm of order $\sigma(\sqrt{N} + \sqrt{m})$, so the noise seen by the whole matrix and by one category are

$$\sigma_{\mathrm{g}} \approx \sigma(\sqrt{n} + \sqrt{m}), \qquad \sigma_{\mathrm{c}} \approx \sigma(\sqrt{n/2} + \sqrt{m}), \qquad \text{so} \quad \sigma_{\mathrm{c}} < \sigma_{\mathrm{g}} < \sqrt{2}\,\sigma_{\mathrm{c}}. \tag{9}$$

A category has half the rows but more than $1/\sqrt{2}$ of the noise, because it pays the column term $\sigma\sqrt{m}$ in full. This gap is what separates the three estimators. A global penalty $\lambda\|W\|_*$ must be set to σ_{g} to remove noise, which is too large for a local direction whose signal lives in one block. A group-wise penalty $\lambda(\|\mathbf{W}_1\|_* + \|\mathbf{W}_2\|_*)$ can be set to the smaller σ_{c}, but each block sees only $a/\sqrt{2}$ of a shared direction of strength a.

GAME uses both penalties. Consider GAME (8) with $\mathcal{C} = \{[n], I_1, I_2\}$:

$$\min_W \tfrac{1}{2}\|X - W\|_F^2 + \lambda_0\|W\|_* + \lambda_1\big(\|W_1\|_* + \|W_2\|_*\big). \tag{10}$$

Lemma B.1 below shows that, direction by direction, GAME first removes weak coefficients within each category (the λ_1 penalty) and then shrinks what remains jointly (the λ_0 penalty). A shared direction therefore survives when $a > \lambda_0 + \sqrt{2}\lambda_1$, and a local direction of strength b with leak at most λ_1 survives when $b > \lambda_0 + \lambda_1$, with the leak removed. Two thresholds can be matched to two noise levels, which a single penalty cannot do: $\lambda_0 = \sigma\sqrt{m}$ and $\lambda_1 = \sigma\sqrt{n/2}$ give $\lambda_0 + \lambda_1 = \sigma_{\mathrm{c}}$ and $\lambda_0 + \sqrt{2}\lambda_1 = \sigma_{\mathrm{g}}$.

Proposition B.1 (Detection limits)**.** *Let X have the separable form of Appendix B.1, and tune each estimator to the thresholds of Appendix B.3. Consider a weak shared direction of strength $a \in (\sigma_{\mathrm{g}}, \sqrt{2}\sigma_{\mathrm{c}})$, or a weak local direction of strength $b \in (\sigma_{\mathrm{c}}, \sigma_{\mathrm{g}})$ whose leak is negligible. Then*

	Weak shared signal	*Weak local signal*
Global	***Retained***	*Missed*
Group-wise	*Missed*	***Retained***
GAME	***Retained***	***Retained***

Moreover, along any local direction that both keep, GAME removes the leak and global denoising retains it.

Interpreting the proposition, global denoising can only see a shared direction once its pooled strength beats the whole-matrix noise, and it cannot see a local direction until that single block beats the whole-matrix noise. Group-wise denoising has the opposite blind spot. GAME has neither, because its two penalties are matched to the two noise levels.

B.1 One Direction at a Time

Assume $n/2 \geq m$ and that both blocks share right singular vectors: $X_c = \sum_j x_{j,c}\, u_{j,c} v_j^\top$ with $v_1, \ldots, v_m$ orthonormal in $\mathbb{R}^m$ and, within each block, $u_{1,c}, \ldots, u_{m,c}$ orthonormal in $\mathbb{R}^{n/2}$. Direction v_j is then described by a pair $x_j = (x_{j,1}, x_{j,2})$ with one coordinate per group: its singular value in the whole matrix is $\|x_j\|_2$ and in block c it is $|x_{j,c}|$.

Lemma B.1 (Separation and nested thresholding)**.** *The minimizer of* (10) *is $\widehat{W}_c = \sum_j \hat{w}_{j,c}\, u_{j,c} v_j^\top$, where each pair solves*

$$\hat{w}_j = \arg\min_{w \in \mathbb{R}^2} \Big\{ \tfrac{1}{2}\|x_j - w\|_2^2 + \lambda_0\|w\|_2 + \lambda_1\|w\|_1 \Big\} = T_{\lambda_0}\big(S_{\lambda_1}(x_j)\big), \tag{11}$$

with $[S_t(x)]_i = \mathrm{sign}(x_i)(|x_i| - t)_+$ and $T_t(x) = (1 - t/\|x\|_2)_+ x$.

The matrix problem reduces to one two-dimensional problem per direction. S_t shrinks each coordinate separately by t, so a coordinate smaller than λ_1 (such as a leak) is set to zero; T_t then shrinks the surviving pair toward the origin by t, treating it as one direction.

Proof. The objective is strongly convex, so the minimizer is unique. Flipping the signs of v_j, $u_{j,1}$ and $u_{j,2}$ together fixes X, preserves the row blocks, and leaves every nuclear norm unchanged, so the minimizer is invariant under each such flip. This zeroes every coefficient of $\widehat{W}_c$ pairing $u_{j,c}$ with

v_k, $k \neq j$, or v_j with a vector orthogonal to $u_{j,c}$. For the resulting W the stacked columns Wv_j are orthogonal across j with norms $\|w_j\|_2$, so $\|W\|_* = \sum_j \|w_j\|_2$ and $\|W_1\|_* + \|W_2\|_* = \sum_j \|w_j\|_1$, and the problem separates. For the closed form, let $y = S_{\lambda_1}(x)$ and $z = T_{\lambda_0}(y)$; z is a nonnegative multiple of y, so $x - y \in \lambda_1 \partial \|z\|_1$ and $y - z \in \lambda_0 \partial \|z\|_2$ (also when $z = 0$), and their sum is the optimality condition [Jenatton et al., 2011]. □

B.2 Dead Zones and Detection Limits

An estimator *removes* direction v_j when $\hat{w}_j = 0$ and *keeps* it otherwise. By (11), the pairs that global ($\lambda_1 = 0$), group-wise ($\lambda_0 = 0$) and GAME denoising remove are

$$\mathcal{D}_{\mathrm{g}} = \{x : \|x\|_2 \leq \tau\}, \qquad \mathcal{D}_{\mathrm{c}} = \{x : \|x\|_\infty \leq \tau\}, \qquad \mathcal{D}_{\mathrm{H}} = [-\lambda_1, \lambda_1]^2 + \lambda_0 B_2, \tag{12}$$

a disk, a square, and a square thickened by a disk (B_2 is the unit disk; the last follows because $\|S_{\lambda_1}(x)\|_2$ is the distance from x to the square $[-\lambda_1, \lambda_1]^2$). GAME's dead zone is a square of half-width $\lambda_0 + \lambda_1$ with corners rounded to radius λ_0: it reaches $\lambda_0 + \lambda_1$ along each axis and $\lambda_0 + \sqrt{2}\lambda_1$ along the diagonal (Figure 7).

Two kinds of direction matter. A *shared* direction of strength a has pair $(a/\sqrt{2}, a/\sqrt{2})$, on the diagonal. A direction *local to group 1* of strength b has pair (b, ε), where the *leak* $\varepsilon \geq 0$ is the noise that group 2 carries along it. Reading off where these points fall relative to the three dead zones gives the detection limits.

Lemma B.2 (Detection limits)**.**

(a) Global: a shared direction is kept iff $a > \tau$; a local direction is kept iff $\sqrt{b^2 + \varepsilon^2} > \tau$, and whenever it is kept its leak is kept with it.

(b) Group-wise: a shared direction is kept iff $a/\sqrt{2} > \tau$; a local direction with $\varepsilon \leq \tau$ is kept iff $b > \tau$, and its leak is removed.

(c) GAME: a shared direction is kept iff $a > \lambda_0 + \sqrt{2}\lambda_1$, with $\hat{w} = \frac{a - \lambda_0 - \sqrt{2}\lambda_1}{\sqrt{2}}(1, 1)$; a local direction with $\varepsilon \leq \lambda_1$ is kept iff $b > \lambda_0 + \lambda_1$, with $\hat{w} = (b - \lambda_0 - \lambda_1, 0)$, so its leak is removed.

Proof. (a) and (b) are (11) with one penalty zero. For (c), the inner step maps $(a/\sqrt{2})(1, 1)$ to $(a/\sqrt{2} - \lambda_1)_+(1, 1)$, of norm $(a - \sqrt{2}\lambda_1)_+$, which the outer step keeps iff this exceeds λ_0; and it maps (b, ε) with $\varepsilon \leq \lambda_1$ to $((b - \lambda_1)_+, 0)$, kept iff $b - \lambda_1 > \lambda_0$. The displayed values follow from T. □

The lemma says three things. A global penalty measures a direction by its pooled strength, so it is the most sensitive to shared directions but the least sensitive to local ones. A group-wise penalty measures each block on its own, so it is the reverse. And the leak is handled by the group penalty alone, which zeroes a second coordinate of size at most λ_1 whatever the first coordinate is: this is the flat side of the square and of the rounded square in the figure. The disk has no flat side, so global denoising keeps the leak whenever it keeps the direction.

B.3 Tuning to the Noise

Let $X = W^\star + E$ with E Gaussian of variance σ^2. The operator norm of an $N \times m$ Gaussian matrix concentrates at $\sigma(\sqrt{N} + \sqrt{m})$ [Vershynin, 2018], which gives the two noise scales in (9). Along a fixed direction v, a block carries noise $\|E_c v\|_2 \approx \sigma\sqrt{n/2}$; this is the scale of the leak. (The noise does not literally have the separable form of Appendix B.1, so the calibration below is a heuristic.)

Each estimator is tuned to the smallest penalty that zeroes noise. Treating whole-matrix noise as a shared pair of strength σ_{g} and within-group noise as a local pair of strength σ_{c}, Lemma B.2 gives $\tau = \sigma_{\mathrm{g}}$ for global denoising, $\tau = \sigma_{\mathrm{c}}$ for group-wise denoising, and for GAME the two equations

$$\begin{gathered} \lambda_0 + \sqrt{2}\lambda_1 = \sigma_{\mathrm{g}}, \qquad \lambda_0 + \lambda_1 = \sigma_{\mathrm{c}}, \\ \text{whence} \qquad \lambda_0^H = \frac{\sqrt{2}\sigma_{\mathrm{c}} - \sigma_{\mathrm{g}}}{\sqrt{2} - 1} = \sigma\sqrt{m}, \qquad \lambda_1^H = \frac{\sigma_{\mathrm{g}} - \sigma_{\mathrm{c}}}{\sqrt{2} - 1} = \sigma\sqrt{n/2}, \end{gathered} \tag{13}$$

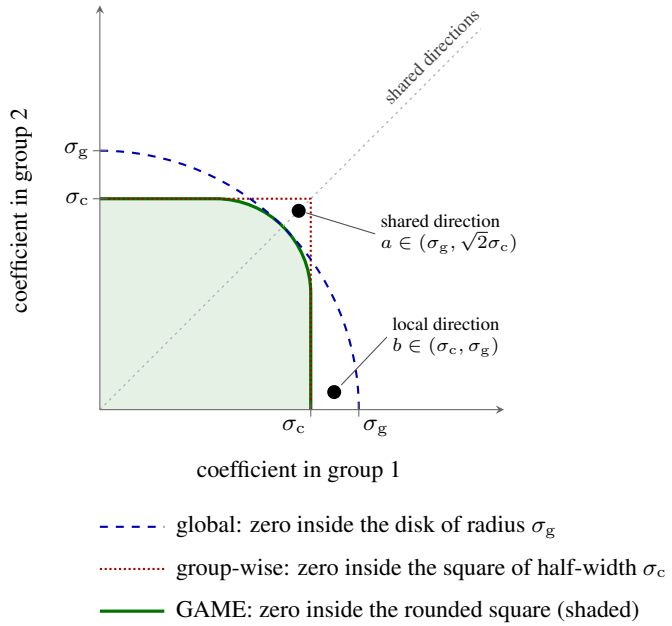


Figure 7: Each latent direction is a point with one coordinate per group (first quadrant shown; the regions are symmetric). An estimator zeroes a point inside its dead zone and keeps one outside it. With the tuning of Appendix B.3, GAME's rounded square reaches σ_c on the axes and σ_g on the diagonal, so it lies inside both the disk and the square. The marked points are the windows of Proposition B.1: the shared point is outside the disk but inside the square; the local point is outside the square but inside the disk. Only the ordering $\sigma_c < \sigma_g < \sqrt{2}\sigma_c$ matters.

where the last equalities use (9). Both penalties are positive. The group penalty λ_1^H is exactly the leak scale, so a leak of Gaussian size is removed up to constants; the global penalty λ_0^H pays only the column term of the noise, since the row term is already paid within each block.

With these penalties, $\mathcal{D}_H \subseteq \mathcal{D}_c \cap \mathcal{D}_g$: the rounded square has the same half-width σ_c as the square, and its farthest point from the origin is at distance $\sqrt{2}\lambda_1^H + \lambda_0^H = \sigma_g$, the radius of the disk. So GAME keeps every direction that either baseline keeps. Both inclusions are strict where it matters: the disk extends past the square along the axes ($\sigma_g > \sigma_c$), and the square past the disk along the diagonal ($\sqrt{2}\sigma_c > \sigma_g$). These two slivers are the windows of the proposition.

Proof of Proposition B.1. Insert the tuned penalties into Lemma B.2. A shared direction is kept by global denoising iff $a > \sigma_g$, by group-wise denoising iff $a > \sqrt{2}\sigma_c$, and by GAME iff $a > \sigma_g$. A local direction with leak $\varepsilon \leq \lambda_1^H$ is kept by group-wise denoising and by GAME iff $b > \sigma_c$, both removing the leak, and by global denoising iff $\sqrt{b^2 + \varepsilon^2} > \sigma_g$, keeping the leak. By (9) both windows $a \in (\sigma_g, \sqrt{2}\sigma_c)$ and $b \in (\sigma_c, \sigma_g)$ are nonempty, which gives the table. The entry "Missed" for global denoising uses the negligible leak: it requires $\varepsilon^2 < \sigma_g^2 - b^2$, which is stronger than the condition $\varepsilon \leq \lambda_1^H$ needed for GAME to remove the leak. □

C Theory for Group-Aware Matrix Estimation with Overlapping Meta-Categories

This appendix establishes finite-sample guarantees for the GAME estimator. Each matrix entry is observed independently with the same probability, so the observation set is a binary mask whose size is random but concentrates around a prescribed expected sample size. Appendix C.9 extends the results to sampling probabilities that vary across meta-category intersections, with the proofs carrying over with mechanical substitutions, and the bounds acquire two explicit factors measuring sampling imbalance.

In developing theoretical results for GAME, we follow the matrix completion proof techniques of [Negahban and Wainwright, 2011, Negahban et al., 2012], who develop arguments based on restricted strong convexity (RSC). This framework is convenient, since it separates the completion bound in two parts: the *curvature* supplied by the random sampling operator, which does not depend on the regularizer, and the *error geometry* induced by the regularizer, which is where GAME differs from the global nuclear norm. Existing matrix-completion theorems for the ordinary nuclear norm generally formulate this geometry in terms of the tangent space of a single low-rank matrix and therefore do not apply directly to the overlapping GAME regularizer.

We do not invoke the with-replacement theorem of [Negahban and Wainwright, 2011] verbatim. Instead, we adopt their spikiness and low-rankness approach of the RSC argument, prove the required curvature result directly for the independent binary mask used here, and combine it with GAME specific decomposability, overlap, and compatibility arguments.

Throughout, let $n, m \geq 1$, $d_0 := \max\{n, m\}$, and let $Z_{ij} := \mathbf{1}\{(i,j) \in \Omega\}$ be the observation indicator. We observe $\mathbf{X} = \mathbf{W}^\star + \mathbf{E}$ on $\Omega := \{(i,j) : Z_{ij} = 1\}$, where $\mathbf{E}$ has independent mean-zero entries independent of the mask. We study the restricted GAME estimator

$$\widehat{\mathbf{W}}^{\mathcal{F}} \in \arg\min_{\mathbf{W} \in \mathbb{R}^{n \times m}} \left\{ \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{X} - \mathbf{W})\|_F^2 + \sum_{c \in \mathcal{C}} \lambda_c \|\mathbf{W}_c\|_* \right\} \quad \text{s.t.} \quad \|\mathbf{W}\|_\infty \leq \frac{\alpha^\star}{\sqrt{nm}}, \tag{14}$$

where

$$\mathbf{W}_c := \mathbf{W}[I_c, :] \in \mathbb{R}^{n_c \times m}, \qquad n_c := |I_c|, \qquad d_c := n_c + m, \qquad r_c := \operatorname{rank}(\mathbf{W}_c^\star).$$

Appendix A treats the mask as fixed and therefore applies to every realized mask. Here, the probabilistic assumptions introduced are used only to obtain statistical guarantees. For notational simplicity, we will refer to $\widehat{\mathbf{W}}^{\mathcal{F}}$ as $\widehat{\mathbf{W}}$ in this section.

C.1 Notation, Sampling Model, and Overlap Structure

For matrices of the same dimensions we use $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{A}^\top \mathbf{B})$, and for $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $c \in \mathcal{C}$ we write $\mathbf{A}_c := \mathbf{A}[I_c, :]$. The sampling operator is the binary mask $[\mathcal{P}_\Omega(\mathbf{A})]_{ij} := Z_{ij} A_{ij}$. Further, since $Z_{ij} \in \{0, 1\}$, $\mathcal{P}_\Omega^2 = \mathcal{P}_\Omega$.

Uniform Bernoulli sampling. Let $\overline{N} > 0$ be the desired expected number of observed entries and

$$q := \frac{\overline{N}}{nm} \leq 1. \tag{15}$$

Independently for every (i,j), $Z_{ij} \sim \text{Bernoulli}(q)$. Each entry is observed once or not at all, the realized sample size $N_{\text{obs}} := |\Omega|$ is random with $\mathbb{E}[N_{\text{obs}}] = \overline{N}$, and for every deterministic $\mathbf{A}$,

$$\mathbb{E}\left[\|\mathcal{P}_\Omega(\mathbf{A})\|_F^2\right] = q\,\|\mathbf{A}\|_F^2. \tag{16}$$

Thus, on average, the sampling operator preserves Frobenius norm up to the factor q. The RSC result of Section C.5 upgrades this one-matrix identity to a high-probability inequality holding simultaneously over a structured class, which is what lets us pass from the empirical loss $\|\mathcal{P}_\Omega(\boldsymbol{\Delta})\|_F^2$ to the error $\|\boldsymbol{\Delta}\|_F^2$.

Per-category counts and degrees. The realized number of observations in category c is $N_c := |\Omega \cap (I_c \times [m])|$, with $\mathbb{E}[N_c] = \overline{N} n_c / n$. The largest observed row or column degree within category c is

$$d_c^\Omega := \max\left\{ \max_{i \in I_c} \sum_{j=1}^m Z_{ij},\ \max_{j \in [m]} \sum_{i \in I_c} Z_{ij} \right\}, \tag{17}$$

whose expected scale is

$$\bar{d}_c := q\,(n_c \vee m) = \frac{\overline{N}(n_c \vee m)}{nm}, \tag{18}$$

since a row has expected degree qm and a column restricted to I_c has expected degree qn_c. The two quantities play different roles, where N_c measures whether a category is represented at all, while d_c^Ω measures how concentrated its observations are and governs the operator norm of its observed noise. We also fix the logarithmic factor

$$\ell_c := 2\log(2d_c), \qquad \text{so that} \qquad e^{-\ell_c} = \frac{1}{4d_c^2} \quad \text{and} \quad \ell_c \leq 4\log(n+m). \tag{19}$$

Row multiplicity. For each row $i \in [n]$ define

$$\kappa(i) := \big|\{c \in \mathcal{C} : i \in I_c\}\big|, \qquad \kappa_{\max} := \max_i \kappa(i), \qquad \kappa_{\min} := \min_i \kappa(i),$$

and let $\mathbf{D}_c \in \{0,1\}^{n_c \times n}$ be the row-selection matrix with $\mathbf{A}_c = \mathbf{D}_c \mathbf{A}$.

Model subspaces. The mechanism by which a nuclear-norm penalty confines the estimation error to a low-rank cone is decomposability over the tangent space of the truth. For each c, let $\mathbf{W}_c^\star = \mathbf{U}_c \mathbf{\Sigma}_c \mathbf{V}_c^\top$ be the thin SVD with $\mathbf{U}_c \in \mathbb{R}^{n_c \times r_c}$, $\mathbf{V}_c \in \mathbb{R}^{m \times r_c}$, and let $\mathcal{U}_c, \mathcal{V}_c$ be the column and row spaces of $\mathbf{W}_c^\star$. Define

$$\mathcal{P}_{\mathcal{M}_c^\perp}(\mathbf{B}) := \mathbf{P}_{\mathcal{U}_c^\perp} \mathbf{B}\, \mathbf{P}_{\mathcal{V}_c^\perp}, \qquad \mathcal{P}_{\mathcal{M}_c}(\mathbf{B}) := \mathbf{B} - \mathcal{P}_{\mathcal{M}_c^\perp}(\mathbf{B}).$$

These are the model-subspace projections of the M-estimation framework of [Negahban et al., 2012, Sec. 2], one per block; both are orthogonal projections in the Frobenius inner product, and

(D1) $\operatorname{rank}\big(\mathcal{P}_{\mathcal{M}_c}(\mathbf{B})\big) \le 2r_c$ for all $\mathbf{B}$, and the singular vectors of $\mathcal{P}_{\mathcal{M}_c^\perp}(\mathbf{B})$ are orthogonal to those of $\mathbf{W}_c^\star$;

(D2) (decomposability of the nuclear norm) by that orthogonality, $\|\mathbf{W}_c^\star + \mathcal{P}_{\mathcal{M}_c^\perp}(\mathbf{B})\|_* = \|\mathbf{W}_c^\star\|_* + \|\mathcal{P}_{\mathcal{M}_c^\perp}(\mathbf{B})\|_*$.

Categories with $r_c = 0$ are permitted, with the convention $\mathcal{P}_{\mathcal{M}_c}(\mathbf{B}) = \mathbf{0}$. Decomposability is what confines each block error $\mathbf{\Delta}_c$ to a low-rank cone in the proof of the main theorem.

C.2 Assumptions

Assumption C.1 (Uniform Bernoulli sampling)**.** The indicators $\{Z_{ij}\}$ are i.i.d. $\mathrm{Bernoulli}(q)$ with $q = \overline{N}/(nm)$, and the mask is independent of $\mathbf{E}$.

Assumption C.2 (Subexponential noise)**.** The entries E_{ij} are independent, mean zero, and subexponential with parameters (σ, R): for every integer $p \ge 2$, $\mathbb{E}|E_{ij}|^p \le \frac{p!}{2} R^{p-2}\sigma^2$. This is the moment form of [Tropp, 2012]; up to universal constants it coincides with Definition 2.7 of [Wainwright, 2019], with σ^2 a variance proxy (in particular $\mathbb{E}E_{ij}^2 \le \sigma^2$) and R the subexponential scale. Sub-Gaussian noise is the special case $R \asymp \sigma$.

Assumption C.3 (Spikiness)**.** $\|\mathbf{W}^\star\|_\infty \le \alpha^\star/\sqrt{nm}$, so $\mathbf{W}^\star$ is feasible for (14). Since $\widehat{\mathbf{W}}$ satisfies the same constraint, the error $\mathbf{\Delta} := \widehat{\mathbf{W}} - \mathbf{W}^\star$ obeys

$$\|\mathbf{\Delta}\|_\infty \le \frac{2\alpha^\star}{\sqrt{nm}}. \tag{20}$$

The constraint rules out matrices whose mass sits on a few entries, which cannot be identified from partial observations.

Assumption C.4 (Cover)**.** $\bigcup_{c \in \mathcal{C}} I_c = [n]$, equivalently $\kappa_{\min} \ge 1$. A row belonging to no category is unregularized and hence unidentifiable.

C.3 Consequences of the Sampling Design

Everything random about the design is captured by the mask, and we need two facts about it. The first is that sample sizes concentrate, so that rates stated in terms of $\overline{N}$ can be read as statements about realized counts. The second is that no row or column inside a category is observed much more often than expected, since this is what controls the noise operator norm in Lemma C.4, and it is the only property of the mask that enters the choice of λ_c.

Sample sizes. By the multiplicative Chernoff bound, since N_{obs} and each N_c are sums of independent Bernoulli variables,

$$\mathbb{P}\Big(\tfrac{1}{2}\overline{N} \le N_{\mathrm{obs}} \le \tfrac{3}{2}\overline{N}\Big) \ge 1 - 2e^{-\overline{N}/12}, \qquad \mathbb{P}\Big(N_c \ge \tfrac{1}{2}\,\tfrac{\overline{N} n_c}{n} \;\; \forall c\Big) \ge 1 - \sum_{c \in \mathcal{C}} e^{-\overline{N} n_c/(8n)}. \tag{21}$$

Thus every occurrence of $\overline{N}$ below may be replaced by N_{obs} up to a factor 2, and each category receives $N_c \asymp \overline{N} n_c/n$ observations as soon as $\overline{N} n_{\min}/n \gtrsim \log|\mathcal{C}|$, where $n_{\min} := \min_c n_c$. Neither fact is used in the proofs, both are used to interpret the rates.

Observed degrees. The operator norm of a noise matrix supported on the observed entries of a category scales as the square root of its largest row or column degree, whereby we need a high-probability ceiling on d_c^{Ω}.

Lemma C.1 (Observed-degree concentration)**.** *For every $c \in \mathcal{C}$,*

$$\mathbb{P}\Big(d_c^{\Omega} > 2(\bar{d}_c + \ell_c)\Big) \le \frac{1}{d_c}, \tag{22}$$

and hence $d_c^{\Omega} \le 2(\bar{d}_c + \ell_c)$ simultaneously for all c with probability at least $1 - \sum_c d_c^{-1}$.

Proof. For a row $i \in I_c$, $D_i := \sum_j Z_{ij}$ is a sum of independent Bernoulli variables with mean and variance at most $qm \le \bar{d}_c$, and likewise, a column degree $\sum_{i \in I_c} Z_{ij}$ has mean at most $qn_c \le \bar{d}_c$. Bernstein's inequality gives, for $u > 0$, $\mathbb{P}\big(D_i > \bar{d}_c + \sqrt{2\bar{d}_c u} + u/3\big) \le e^{-u}$, and $\sqrt{2\bar{d}_c u} \le \bar{d}_c + u/2$ shows the threshold is at most $2(\bar{d}_c + u)$. Take $u = \ell_c$ and union bound over the n_c rows and m columns, whence the failure probability is at most $d_c e^{-\ell_c} = 1/(4d_c)$. □

C.4 GAME Decomposability and Compatibility

The goal of this section is the standard conclusion of M-estimation theory, particularly, once each λ_c dominates the noise seen by block c, the error $\mathbf{\Delta}$ is forced into a cone in which nuclear-norm quantities are controlled by Frobenius quantities with factors depending only on the ranks r_c. The complication here is overlap, as the regularizer couples blocks that share rows, so the single-norm decomposability argument of [Negahban and Wainwright, 2011] does not apply. We proceed in four steps: decomposability one block at a time (Lemma C.2), a reweighting that distributes the noise inner product over blocks without double counting shared rows (Lemma C.3), a per-block bound on the reweighted noise (Lemma C.4), and the resulting cone constraint with its compatibility constants (Lemmas C.6–C.7).

Decomposability (D2) concerns a single nuclear norm at $\mathbf{W}_c^{\star}$ plus a perturbation orthogonal to it. The lemma below is the consequence we use, stating that adding $\mathbf{B}_c$ to $\mathbf{W}_c^{\star}$ can decrease the nuclear norm by at most the part of $\mathbf{B}_c$ which lives in the model subspace.

Lemma C.2 (Per-group decomposability)**.** *For every $\mathbf{B} \in \mathbb{R}^{n \times m}$ and every $c \in \mathcal{C}$,*

$$\|\mathbf{W}_c^{\star} + \mathbf{B}_c\|_* \;\ge\; \|\mathbf{W}_c^{\star}\|_* + \|\mathcal{P}_{\mathcal{M}_c^{\perp}}(\mathbf{B}_c)\|_* - \|\mathcal{P}_{\mathcal{M}_c}(\mathbf{B}_c)\|_*.$$

Proof. Write $\mathbf{W}_c^{\star} + \mathbf{B}_c = \big[\mathbf{W}_c^{\star} + \mathcal{P}_{\mathcal{M}_c^{\perp}}(\mathbf{B}_c)\big] + \mathcal{P}_{\mathcal{M}_c}(\mathbf{B}_c)$. By the triangle inequality and (D2),

$$\begin{aligned}\|\mathbf{W}_c^{\star} + \mathbf{B}_c\|_* &\ge \|\mathbf{W}_c^{\star} + \mathcal{P}_{\mathcal{M}_c^{\perp}}(\mathbf{B}_c)\|_* - \|\mathcal{P}_{\mathcal{M}_c}(\mathbf{B}_c)\|_* \\ &= \|\mathbf{W}_c^{\star}\|_* + \|\mathcal{P}_{\mathcal{M}_c^{\perp}}(\mathbf{B}_c)\|_* - \|\mathcal{P}_{\mathcal{M}_c}(\mathbf{B}_c)\|_*.\end{aligned}$$

□

The noise enters through the inner product $\langle \mathcal{P}_{\Omega}(\mathbf{E}), \mathbf{\Delta} \rangle$. To compare it with the regularizer we must write it as a sum over blocks of something paired with $\mathbf{\Delta}_c$. If a row belongs to $\kappa(i)$ blocks, summing $\langle \mathbf{A}_c, \mathbf{B}_c \rangle$ over c counts that row $\kappa(i)$ times, so dividing each row of $\mathbf{A}$ by its multiplicity compensates exactly. This is the only place overlap is handled, and it is what produces the $1/\kappa_{\min}$ factors in the regularization scale.

Lemma C.3 (Row-multiplicity split)**.** *Under Assumption C.4, for any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times m}$, define for each c the matrix $\widetilde{\mathbf{A}}^{(c)} \in \mathbb{R}^{n_c \times m}$ by $[\widetilde{\mathbf{A}}^{(c)}]_{ij} = A_{ij}/\kappa(i)$ for $i \in I_c$. Then*

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{c \in \mathcal{C}} \langle \widetilde{\mathbf{A}}^{(c)}, \mathbf{B}_c \rangle \le \sum_{c \in \mathcal{C}} \big\|\widetilde{\mathbf{A}}^{(c)}\big\|_{\mathrm{op}} \|\mathbf{B}_c\|_*.$$

Proof. By the cover assumption $\sum_{c : i \in I_c} 1 = \kappa(i) \ge 1$ for every i, so

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i,j} \frac{A_{ij}}{\kappa(i)} \sum_{c : i \in I_c} B_{ij} = \sum_c \sum_{i \in I_c,\, j} \frac{A_{ij}}{\kappa(i)} B_{ij} = \sum_c \langle \widetilde{\mathbf{A}}^{(c)}, \mathbf{B}_c \rangle.$$

The inequality is trace duality between $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_*$. □

Applying Lemma C.3 with $\mathbf{A} = \mathcal{P}_\Omega(\mathbf{E})$ defines the per-category reweighted noise matrix $\widetilde{\mathbf{E}}^{(c)} \in \mathbb{R}^{n_c \times m}$, with entries $Z_{ij}E_{ij}/\kappa(i)$ for $i \in I_c$, and $\langle \mathcal{P}_\Omega(\mathbf{E}), \mathbf{\Delta}\rangle = \sum_c \langle \widetilde{\mathbf{E}}^{(c)}, \mathbf{\Delta}_c\rangle$. The next lemma bounds its operator norm conditional on the mask. The bound has the familiar matrix-Bernstein shape, with a Gaussian-like term driven by σ and the largest observed degree, plus a subexponential correction driven by R. It is the sole input to the choice of λ_c.

Lemma C.4 (Per-category noise spectral bound)**.** *Condition on Ω. There is a universal constant $C_1 > 0$ such that*

$$\|\widetilde{\mathbf{E}}^{(c)}\|_{\mathrm{op}} \le \frac{C_1}{\kappa_{\min}} \left[\sigma\sqrt{d_c^\Omega \ell_c} + R\ell_c\right] \tag{23}$$

with conditional probability at least $1 - d_c^{-1}$, and simultaneously over all c with conditional probability at least $1 - \sum_c d_c^{-1}$.

Proof. Conditional on Ω, $\widetilde{\mathbf{E}}^{(c)} = \sum_{(i,j)\in\Omega,\, i\in I_c} \frac{E_{ij}}{\kappa(i)} \mathbf{e}_i^{(c)} \mathbf{e}_j^\top$ is a sum of independent mean-zero rank-one random matrices. Its rectangular variance proxy is at most $\sigma^2 d_c^\Omega/\kappa_{\min}^2$ (the row-side sum $\sum \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]$ is diagonal with entries $\sum_j Z_{ij}\mathbb{E}E_{ij}^2/\kappa(i)^2 \le \sigma^2 d_c^\Omega/\kappa_{\min}^2$, and similarly for columns), and each summand satisfies the subexponential moment condition with scale $R/\kappa_{\min}$. The subexponential matrix Bernstein inequality applied to the self-adjoint dilation of dimension d_c [Tropp, 2012, Theorem 6.2 and Remark 6.3] gives $\mathbb{P}(\|\widetilde{\mathbf{E}}^{(c)}\|_{\mathrm{op}} \ge \tau) \le d_c \exp\big(-\frac{\tau^2/2}{\sigma^2 d_c^\Omega/\kappa_{\min}^2 + R\tau/\kappa_{\min}}\big)$, and the right-hand side of (23) with C_1 large enough makes this at most $d_c e^{-\ell_c} = 1/(4d_c)$. A union bound gives the simultaneous statement. □

We now turn to the estimator. Comparing the objective at $\widehat{\mathbf{W}}$ with the objective at the feasible point $\mathbf{W}^\star$ bounds the empirical loss of the error by a noise term plus the decrease in regularizer.

Lemma C.5 (Basic inequality)**.** *Let $\widehat{\mathbf{W}}$ be any minimizer of* (14) *and $\mathbf{\Delta} := \widehat{\mathbf{W}} - \mathbf{W}^\star$. Then*

$$\tfrac{1}{2}\|\mathcal{P}_\Omega(\mathbf{\Delta})\|_F^2 \le \big\langle \mathcal{P}_\Omega(\mathbf{E}), \mathbf{\Delta}\big\rangle + \sum_{c\in\mathcal{C}} \lambda_c\big(\|\mathbf{W}_c^\star\|_* - \|\mathbf{W}_c^\star + \mathbf{\Delta}_c\|_*\big). \tag{24}$$

Proof. By optimality of $\widehat{\mathbf{W}}$ and feasibility of $\mathbf{W}^\star$,

$$\tfrac{1}{2}\|\mathcal{P}_\Omega(\widehat{\mathbf{W}} - \mathbf{X})\|_F^2 + \sum_c \lambda_c \|\widehat{\mathbf{W}}_c\|_* \le \tfrac{1}{2}\|\mathcal{P}_\Omega(\mathbf{W}^\star - \mathbf{X})\|_F^2 + \sum_c \lambda_c \|\mathbf{W}_c^\star\|_*.$$

Substitute $\mathbf{X} = \mathbf{W}^\star + \mathbf{E}$ and $\widehat{\mathbf{W}} = \mathbf{W}^\star + \mathbf{\Delta}$, expand $\|\mathcal{P}_\Omega(\mathbf{\Delta} - \mathbf{E})\|_F^2 = \|\mathcal{P}_\Omega(\mathbf{\Delta})\|_F^2 + \|\mathcal{P}_\Omega(\mathbf{E})\|_F^2 - 2\langle \mathcal{P}_\Omega(\mathbf{\Delta}), \mathcal{P}_\Omega(\mathbf{E})\rangle$, cancel $\frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{E})\|_F^2$ on both sides, and use $\langle \mathcal{P}_\Omega(\mathbf{\Delta}), \mathcal{P}_\Omega(\mathbf{E})\rangle = \langle \mathbf{\Delta}, \mathcal{P}_\Omega(\mathbf{E})\rangle$, since $\mathcal{P}_\Omega$ is self-adjoint and idempotent. □

We now combine the three preceding lemmas to show that the estimation error is itself nearly low rank. Nothing in (14) forces this. It is a consequence of the penalty being set above the noise level. Decompose each block error as $\mathbf{\Delta}_c = \mathcal{P}_{\mathcal{M}_c}(\mathbf{\Delta}_c) + \mathcal{P}_{\mathcal{M}_c^\perp}(\mathbf{\Delta}_c)$, a low-rank component aligned with $\mathbf{W}_c^\star$ and a component orthogonal to it. The two components affect the penalty differently. Adding the orthogonal component always increases the nuclear norm by its full size (D2), whereas adding the aligned component can decrease it, though by no more than its own size (Lemma C.2). The noise term, by contrast, can reward either component, but only at rate $\|\widetilde{\mathbf{E}}^{(c)}\|_{\mathrm{op}}$ per unit of nuclear norm (Lemma C.3). Hence if λ_c is at least twice that rate, every unit of orthogonal error costs the estimator more in penalty than the noise can repay, and a minimizer will not pay that cost unless it is offset by aligned error. The error is thereby confined to a cone around the low-rank model. Its orthogonal part is at most three times its aligned part, and the empirical loss is controlled by the aligned part alone. This is the step that lets the rest of the proof measure the error by the local ranks r_c.

Lemma C.6 (GAME cone constraint)**.** *Suppose*

$$\lambda_c \ge 2\|\widetilde{\mathbf{E}}^{(c)}\|_{\mathrm{op}} \qquad \textit{for every } c \in \mathcal{C}. \tag{25}$$

Then

$$\sum_c \lambda_c \|\mathcal{P}_{\mathcal{M}_c^\perp}(\mathbf{\Delta}_c)\|_* \le 3\sum_c \lambda_c \|\mathcal{P}_{\mathcal{M}_c}(\mathbf{\Delta}_c)\|_*, \qquad \tfrac{1}{2}\|\mathcal{P}_\Omega(\mathbf{\Delta})\|_F^2 \le \tfrac{3}{2}\sum_c \lambda_c \|\mathcal{P}_{\mathcal{M}_c}(\mathbf{\Delta}_c)\|_*. \tag{26}$$

Proof. Apply Lemma C.3 to the noise term of (24) and Lemma C.2 to the regularizer term, then use $\|\mathbf{\Delta}_c\|_* \le \|\mathcal{P}_{\mathcal{M}_c}(\mathbf{\Delta}_c)\|_* + \|\mathcal{P}_{\mathcal{M}_c^\perp}(\mathbf{\Delta}_c)\|_*$ and (25):

$$\tfrac{1}{2}\|\mathcal{P}_\Omega(\mathbf{\Delta})\|_F^2 \le \sum_c \Big[\tfrac{3\lambda_c}{2}\|\mathcal{P}_{\mathcal{M}_c}(\mathbf{\Delta}_c)\|_* - \tfrac{\lambda_c}{2}\|\mathcal{P}_{\mathcal{M}_c^\perp}(\mathbf{\Delta}_c)\|_*\Big].$$

Nonnegativity of the left-hand side gives the first claim; dropping the nonpositive term gives the second. □

The cone constraint controls the error in nuclear norm, but restricted strong convexity is stated in Frobenius norm, so we need to pass from one to the other. The bridge is (D1). Since $\mathcal{P}_{\mathcal{M}_c}(\mathbf{\Delta}_c)$ has rank at most $2r_c$, its nuclear norm is at most $\sqrt{2r_c}$ times its Frobenius norm, and this is the only place the ranks enter the rate. Carrying this through the two inequalities of Lemma C.6 and summing over blocks yields two constants. The first, A_λ, converts the loss bound into the form $\frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{\Delta})\|_F^2 \lesssim A_\lambda\|\mathbf{\Delta}\|_F$, which is the upper half of the sandwich used in the main theorem. The second, B_λ, bounds the ratio $\|\mathbf{\Delta}\|_*/\|\mathbf{\Delta}\|_F$ of the global error. A bounded ratio is what certifies that $\mathbf{\Delta}$ is effectively low rank, and it is the condition under which the sampling operator retains curvature, so B_λ is what places $\mathbf{\Delta}$ inside the class on which RSC holds. Assume $\lambda_c > 0$ for every c and define

$$\begin{aligned} \theta_\lambda &:= \frac{\max_c \lambda_c}{\min_c \lambda_c}, \qquad r_{\mathrm{tot}} := \sum_{c\in\mathcal{C}} r_c, \\ A_\lambda &:= \Big(\sum_{c\in\mathcal{C}} \lambda_c^2 r_c\Big)^{1/2}, \qquad B_\lambda := \frac{4\sqrt{2\kappa_{\max}}}{\kappa_{\min}} \cdot \frac{A_\lambda}{\min_c \lambda_c} \le \frac{4\sqrt{2\kappa_{\max}}}{\kappa_{\min}}\,\theta_\lambda\sqrt{r_{\mathrm{tot}}}. \end{aligned} \tag{27}$$

Lemma C.7 (GAME compatibility bounds)**.** *Under* (25)*,*

$$\tfrac{1}{2}\|\mathcal{P}_\Omega(\mathbf{\Delta})\|_F^2 \le \frac{3\sqrt{2\kappa_{\max}}}{2} A_\lambda \|\mathbf{\Delta}\|_F, \qquad \textit{and} \qquad \|\mathbf{\Delta}\|_* \le B_\lambda \|\mathbf{\Delta}\|_F. \tag{28}$$

Proof. By (D1) and the fact that $\mathcal{P}_{\mathcal{M}_c}$ is a Frobenius orthogonal projection, $\|\mathcal{P}_{\mathcal{M}_c}(\mathbf{\Delta}_c)\|_* \le \sqrt{2r_c}\|\mathbf{\Delta}_c\|_F$. Cauchy–Schwarz and the identity $\sum_c \|\mathbf{\Delta}_c\|_F^2 = \sum_i \kappa(i)\|\mathbf{\Delta}_{i,:}\|_2^2 \le \kappa_{\max}\|\mathbf{\Delta}\|_F^2$ give

$$\sum_c \lambda_c \|\mathcal{P}_{\mathcal{M}_c}(\mathbf{\Delta}_c)\|_* \le \sqrt{2} A_\lambda \Big(\sum_c \|\mathbf{\Delta}_c\|_F^2\Big)^{1/2} \le \sqrt{2\kappa_{\max}} A_\lambda \|\mathbf{\Delta}\|_F,$$

and substitution into (26) proves the first inequality. For the second, (26) gives $\sum_c \lambda_c\|\mathbf{\Delta}_c\|_* \le 4\sum_c \lambda_c \|\mathcal{P}_{\mathcal{M}_c}(\mathbf{\Delta}_c)\|_* \le 4\sqrt{2\kappa_{\max}} A_\lambda \|\mathbf{\Delta}\|_F$. Let $\mathbf{K} := \mathrm{diag}(\kappa(1),\dots,\kappa(n))$ and $\mathbf{K}_c := \mathbf{D}_c\mathbf{K}\mathbf{D}_c^\top$. The rowwise identity $\mathbf{\Delta} = \sum_c \mathbf{D}_c^\top \mathbf{K}_c^{-1}\mathbf{\Delta}_c$ and the triangle inequality give $\|\mathbf{\Delta}\|_* \le \kappa_{\min}^{-1}\sum_c \|\mathbf{\Delta}_c\|_* \le (\kappa_{\min}\min_c \lambda_c)^{-1}\sum_c \lambda_c\|\mathbf{\Delta}_c\|_*$, which is the claim. □

C.5 Restricted Strong Convexity

The loss $\frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{\Delta})\|_F^2$ sees only the observed entries, so it cannot be strongly convex in all directions. In particular, any $\mathbf{\Delta}$ supported on unobserved entries has zero loss. The workaround of [Negahban and Wainwright, 2011] is that the loss is comparable to $q\|\mathbf{\Delta}\|_F^2$ on matrices that are neither too spiky (so mass cannot hide in a few entries) nor too high rank in the nuclear-to-Frobenius sense (so the class is not too rich). The cone constraint supplies the second property and the ℓ_∞ constraint in (14) the first. We prove the inequality for the Bernoulli mask following their argument.

For nonzero $\mathbf{A}$, define the spikiness and rank ratios

$$\alpha_{\mathrm{sp}}(\mathbf{A}) := \frac{\sqrt{nm}\,\|\mathbf{A}\|_\infty}{\|\mathbf{A}\|_F}, \qquad \beta_{\mathrm{ra}}(\mathbf{A}) := \frac{\|\mathbf{A}\|_*}{\|\mathbf{A}\|_F}, \tag{29}$$

both of which are at least one, and the class

$$\mathfrak{C}(c_0) := \left\{ \mathbf{A} \ne 0 :\ \alpha_{\mathrm{sp}}(\mathbf{A})\,\beta_{\mathrm{ra}}(\mathbf{A}) \le \frac{1}{c_0}\sqrt{\frac{\overline{N}}{d_0 \log(n+m)}} \right\}. \tag{30}$$

Lemma C.8 (RSC for the Bernoulli mask)**.** *There exist universal constants $c_0, c_1, c_2, C_0 > 0$ such that, if $\overline{N} \geq C_0 d_0 \log(n+m)$, then with probability at least $1 - c_1 \exp(-c_2 d_0 \log(n+m))$,*

$$\|\mathcal{P}_\Omega(\mathbf{A})\|_F^2 \geq \frac{q}{2} \|\mathbf{A}\|_F^2 \qquad \textit{simultaneously for every } \mathbf{A} \in \mathfrak{C}(c_0). \tag{31}$$

Proof. Define $Q_Z(\mathbf{A}) := q^{-1}\|\mathcal{P}_\Omega(\mathbf{A})\|_F^2 = q^{-1}\sum_{i,j} Z_{ij} A_{ij}^2$, so that $\mathbb{E}Q_Z(\mathbf{A}) = \|\mathbf{A}\|_F^2$ by (16). For $D, a, \rho > 0$ let $\mathcal{S}(D, a, \rho) := \{\mathbf{A} : \|\mathbf{A}\|_F \leq D,\ \|\mathbf{A}\|_\infty \leq a,\ \|\mathbf{A}\|_* \leq \rho\}$. Throughout, suprema over $\mathcal{S}(D, a, \rho)$ may be taken over a countable dense subset, since the class is separable and every process below is continuous in $\mathbf{A}$.

Step 1: Expected Deviation. Symmetrization, the Ledoux–Talagrand contraction inequality (since $Z_{ij} \in \{0,1\}$, $Z_{ij}A_{ij}^2 = (Z_{ij}A_{ij})^2$ with $|Z_{ij}A_{ij}| \leq a$, and the map $t \mapsto t^2$ is $2a$-Lipschitz on $[-a, a]$; we extend it $2a$-Lipschitz outside this interval, which does not affect any of the arguments), and trace duality give

$$\mathbb{E} \sup_{\mathbf{A} \in \mathcal{S}(D,a,\rho)} |Q_Z(\mathbf{A}) - \mathbb{E}Q_Z(\mathbf{A})| \leq \frac{C}{q}\, a\rho\, \mathbb{E}\Big\| \sum_{i,j} \varepsilon_{ij} Z_{ij} \mathbf{e}_i \mathbf{e}_j^\top \Big\|_{\mathrm{op}}, \tag{32}$$

with $\{\varepsilon_{ij}\}$ an independent Rademacher array. The random matrix on the right has summands of operator norm one and variance parameters $\max_i \sum_j q = qm$ and $\max_j \sum_i q = qn$, so the rectangular matrix Bernstein inequality [Tropp, 2012, Theorem 1.6], integrated, bounds its expected operator norm by $C\big[\sqrt{q\, d_0 \log(n+m)} + \log(n+m)\big] \leq 2C\sqrt{q\, d_0 \log(n+m)}$, the last step using $\log(n+m) \leq q\, d_0 = \overline{N}/\min(n,m)$, which $\overline{N} \geq C_0 d_0 \log(n+m)$ implies. Hence

$$\mathbb{E} \sup_{\mathbf{A} \in \mathcal{S}(D,a,\rho)} |Q_Z(\mathbf{A}) - \mathbb{E}Q_Z(\mathbf{A})| \leq C\, a\rho \sqrt{\frac{nm\, d_0 \log(n+m)}{\overline{N}}}. \tag{33}$$

Step 2: Concentration of the Supremum. Let $Z_{D,a,\rho}$ denote the supremum in (33); it is the supremum of a centered empirical process over the independent coordinates $\{Z_{ij}\}$, indexed by $\mathcal{S}(D, a, \rho)$ enlarged by its negative. Each summand $q^{-1}(Z_{ij} - q)A_{ij}^2$ is bounded by $b := a^2/q = nma^2/\overline{N}$, and the weak variance satisfies $v := \sup_{\mathbf{A}} \sum_{i,j} \mathrm{Var}\big(q^{-1}(Z_{ij} - q)A_{ij}^2\big) \leq q^{-1} \sup_{\mathbf{A}} \sum_{i,j} A_{ij}^4 \leq nmD^2a^2/\overline{N}$. Although the Z_{ij} are identically distributed, the summands, viewed as random vectors indexed by $\mathbf{A}$, are not, since the map $\mathbf{A} \mapsto A_{ij}^2$ depends on (i,j). We therefore use the form of Bousquet's inequality [Boucheron et al., 2013, Theorem 12.5] for independent, not necessarily identically distributed, coordinates due to Klein and Rio [2005], which gives, for a universal constant C, with probability at least $1 - 2e^{-t}$, $Z_{D,a,\rho} \leq \mathbb{E}Z_{D,a,\rho} + C\sqrt{(v + 2b\,\mathbb{E}Z_{D,a,\rho})t} + Cbt$. Using $2\sqrt{xy} \leq x + y$ on the mixed term and inserting (33), with probability at least $1 - 2e^{-t}$,

$$Z_{D,a,\rho} \leq C\left[a\rho\sqrt{\frac{nm\, d_0 \log(n+m)}{\overline{N}}} + Da\sqrt{\frac{nm\, t}{\overline{N}}} + \frac{nm\, a^2 t}{\overline{N}}\right]. \tag{34}$$

Step 3: Peeling Argument. Both sides of (31) are quadratic in $\mathbf{A}$, and $\mathfrak{C}(c_0)$ is invariant under positive scaling, so it suffices to treat $\|\mathbf{A}\|_F = 1$. Set $T := c_0^{-1}\sqrt{\overline{N}/(d_0 \log(n+m))}$ and take $C_0 \geq c_0^2$ so that $T \geq 1$. For a normalized $\mathbf{A} \in \mathfrak{C}(c_0)$, $\alpha_{\mathrm{sp}}(\mathbf{A}) = \sqrt{nm}\|\mathbf{A}\|_\infty$ and $\beta_{\mathrm{ra}}(\mathbf{A}) = \|\mathbf{A}\|_*$, with product at most T and each factor at least one; in particular $\alpha_{\mathrm{sp}} \leq T$. Partition the class into slices $2^k \leq \alpha_{\mathrm{sp}}(\mathbf{A}) < 2^{k+1}$, $0 \leq k \leq \lfloor \log_2 T \rfloor$, so that $2^k \leq T$ on every slice. On slice k, every matrix lies in $\mathcal{S}(D, a, \rho)$ with

$$D = 1, \qquad a = \frac{2^{k+1}}{\sqrt{nm}}, \qquad \rho = \frac{T}{2^k}, \qquad \text{so that} \qquad a\rho = \frac{2T}{\sqrt{nm}}:$$

spikier slices are allowed proportionally smaller nuclear norm, and the product driving the first term of (34) does not depend on k. Choose $t = c_t d_0 \log(n+m)$. Using $2^k \leq T$, the three terms in (34) are at most

$$\frac{2}{c_0}, \qquad 2^{k+1}\sqrt{\frac{t}{\overline{N}}} \leq 2T\sqrt{\frac{t}{\overline{N}}} = \frac{2\sqrt{c_t}}{c_0}, \qquad \frac{2^{2k+2}t}{\overline{N}} \leq \frac{4T^2 t}{\overline{N}} = \frac{4c_t}{c_0^2},$$

so choosing c_t small and then c_0 large (both universal, after which $C_0 \geq c_0^2$ is fixed) makes $Z_{D,a,\rho} \leq 1/2$ on every slice. A union bound over the $1 + \lfloor \log_2 T \rfloor \lesssim \log(nm)$ slices is absorbed into c_1, c_2. Hence, with the stated probability, $Q_Z(\mathbf{A}) \geq \|\mathbf{A}\|_F^2 - \frac{1}{2} = \frac{1}{2}$ for every normalized $\mathbf{A}$ in the class, which is (31). □

C.6 Frobenius Error Bound

We now assemble the pieces. Lemma C.7 shows that $\mathbf{\Delta}$ has bounded rank ratio, and the ℓ_∞ constraint shows that it has bounded spikiness relative to its own Frobenius norm. By (20),

$$\alpha_{\mathrm{sp}}(\mathbf{\Delta}) \leq \frac{2\alpha^\star}{\|\mathbf{\Delta}\|_F}, \qquad \beta_{\mathrm{ra}}(\mathbf{\Delta}) \leq B_\lambda, \qquad \text{so} \qquad \alpha_{\mathrm{sp}}(\mathbf{\Delta})\beta_{\mathrm{ra}}(\mathbf{\Delta}) \leq \frac{2\alpha^\star B_\lambda}{\|\mathbf{\Delta}\|_F}. \tag{35}$$

Thus $\mathbf{\Delta}$ belongs to $\mathfrak{C}(c_0)$ as soon as $\|\mathbf{\Delta}\|_F$ exceeds the tolerance

$$\tau := 2c_0\, \alpha^\star B_\lambda \sqrt{\frac{d_0 \log(n+m)}{\overline{N}}}, \tag{36}$$

while errors below τ are already small. Exactly one of the two cases occurs, and each yields a bound. This is how the spikiness term enters our matrix-completion bound.

It remains to choose λ_c. Lemma C.4 says λ_c must dominate $\sigma\sqrt{d_c^\Omega \ell_c} + R\ell_c$ up to $\kappa_{\min}$, and Lemma C.1 says the random degree d_c^Ω is at most $2(\bar{d}_c + \ell_c)$. We therefore set

$$\lambda_c := \frac{C_\lambda}{\kappa_{\min}} \left[\sigma\sqrt{(\bar{d}_c + \ell_c)\,\ell_c} + R\,\ell_c\right], \qquad C_\lambda := 2\sqrt{2}\,C_1, \tag{37}$$

a deterministic quantity depending only on $(\overline{N}, n, m, n_c)$, the noise parameters, and the overlap. Since $R > 0$ and $\ell_c \geq 1$, every $\lambda_c > 0$.

Theorem C.5 (GAME Frobenius error)**.** *Suppose Assumptions C.1–C.4 hold,*

$$\overline{N} \;\geq\; C_0\Big(1 + \frac{R^2}{\sigma^2}\Big)\, d_0 \log(n+m),$$

and the λ_c are chosen by (37)*. Then with probability at least*

$$1 - 2\sum_{c\in\mathcal{C}} d_c^{-1} - c_1 \exp\big(-c_2 d_0 \log(n+m)\big), \tag{38}$$

$$\frac{\|\widehat{\mathbf{W}} - \mathbf{W}^\star\|_F^2}{nm} \leq 72\,\frac{\kappa_{\max} nm}{\overline{N}^2} A_\lambda^2 + \frac{\tau^2}{nm}, \tag{39}$$

and consequently

$$\frac{\|\widehat{\mathbf{W}} - \mathbf{W}^\star\|_F^2}{nm} \lesssim \underbrace{\frac{\kappa_{\max}}{\kappa_{\min}^2} \cdot \frac{\sigma^2 \log(n+m)}{\overline{N}} \sum_{c\in\mathcal{C}} r_c\,(n_c \vee m)}_{\textit{noise term}} \;+\; \underbrace{(\alpha^\star)^2\,\frac{\kappa_{\max}}{\kappa_{\min}^2}\,\theta_\lambda^2\, r_{\mathrm{tot}} \cdot \frac{d_0 \log(n+m)}{nm\,\overline{N}}}_{\textit{spikiness term}}. \tag{40}$$

Proof. We work on the intersection of three events: (E1) the degree event of Lemma C.1, (E2) the noise event of Lemma C.4, and (E3) the RSC event of Lemma C.8. On (E1)∩(E2),

$$2\|\widetilde{\mathbf{E}}^{(c)}\|_{\mathrm{op}} \leq \frac{2C_1}{\kappa_{\min}}\Big[\sigma\sqrt{2(\bar{d}_c + \ell_c)\ell_c} + R\ell_c\Big] \leq \lambda_c \qquad \text{for every } c,$$

so (25) holds and Lemmas C.6 and C.7 apply. Lemma C.4 is conditional on Ω; integrating over the mask, (E1)∩(E2) has probability at least $1 - 2\sum_c d_c^{-1}$, and (E3) adds the remaining term of (38).

Let $\mathbf{\Delta} := \widehat{\mathbf{W}} - \mathbf{W}^\star$. If $\|\mathbf{\Delta}\|_F \leq \tau$, the second term of (39) applies. Otherwise (35) and (36) give $\alpha_{\mathrm{sp}}(\mathbf{\Delta})\beta_{\mathrm{ra}}(\mathbf{\Delta}) < c_0^{-1}\sqrt{\overline{N}/(d_0\log(n+m))}$, so $\mathbf{\Delta} \in \mathfrak{C}(c_0)$ and Lemma C.8 yields $\|\mathcal{P}_\Omega(\mathbf{\Delta})\|_F^2 \geq \frac{q}{2}\|\mathbf{\Delta}\|_F^2$. Combined with (28),

$$\frac{q}{4}\|\mathbf{\Delta}\|_F^2 \leq \frac{3\sqrt{2\kappa_{\max}}}{2} A_\lambda \|\mathbf{\Delta}\|_F, \qquad \text{i.e.} \qquad \|\mathbf{\Delta}\|_F \leq \frac{6\sqrt{2\kappa_{\max}}}{q} A_\lambda = \frac{6\sqrt{2\kappa_{\max}}\, nm}{\overline{N}} A_\lambda.$$

Squaring and dividing by nm gives the first term of (39); summing the two cases proves it.

For (40), note that under the sample-size hypothesis $\bar{d}_c \geq \overline{N}/n \geq C_0 \log(n+m) \geq \frac{C_0}{4}\ell_c$ by (19), so $\bar{d}_c + \ell_c \leq 2\bar{d}_c$ for $C_0 \geq 4$, and $(x+y)^2 \leq 2(x^2+y^2)$ gives $\lambda_c^2 \lesssim \kappa_{\min}^{-2}[\sigma^2 \bar{d}_c \ell_c + R^2 \ell_c^2]$. Substituting $\bar{d}_c = \overline{N}(n_c \vee m)/(nm)$ and $\ell_c \asymp \log(n+m)$,

$$72\,\frac{\kappa_{\max} nm}{\overline{N}^2} A_\lambda^2 \lesssim \frac{\kappa_{\max}}{\kappa_{\min}^2}\left[\frac{\sigma^2 \log(n+m)}{\overline{N}} \sum_c r_c(n_c \vee m) + \frac{R^2 nm \log^2(n+m)}{\overline{N}^2}\, r_{\mathrm{tot}}\right].$$

Since $\sum_c r_c(n_c \vee m) \geq m\, r_{\mathrm{tot}}$, the second summand is at most the first whenever $\overline{N} \geq (R^2/\sigma^2)\, n \log(n+m)$, which the hypothesis guarantees because $d_0 \geq n$. For the spikiness term, insert $B_\lambda \leq \frac{4\sqrt{2\kappa_{\max}}}{\kappa_{\min}}\theta_\lambda \sqrt{r_{\mathrm{tot}}}$ from (27) into (36):

$$\frac{\tau^2}{nm} = 4c_0^2(\alpha^\star)^2 B_\lambda^2 \frac{d_0 \log(n+m)}{nm\,\overline{N}} \leq 128\, c_0^2(\alpha^\star)^2 \frac{\kappa_{\max}}{\kappa_{\min}^2}\theta_\lambda^2 r_{\mathrm{tot}} \frac{d_0 \log(n+m)}{nm\,\overline{N}}. \qquad \square$$

Remark C.6 (Reading the regularization rule). Under the hypothesis of the theorem the $R\ell_c$ term in (37) is dominated, and the rule reduces to

$$\lambda_c \asymp \frac{\sigma}{\kappa_{\min}} \sqrt{\frac{(\overline{N} n_c/n)\, \log d_c}{\min(n_c, m)}}. \tag{41}$$

Intuitively, the penalty on category c scales with the noise level, with the square root of the category's expected number of observations $\overline{N} n_c/n$ divided by its smaller block dimension, and with a log of the block size. The ratio θ_λ is then at most $\sqrt{\max_c(n_c \vee m)/\min_c(n_c \vee m)}$, which is $O(1)$ when the categories are comparably sized and equals one when every $n_c \leq m$.

Remark C.7 (Data-driven regularization). The expected degree $\bar{d}_c$ in (37) may be replaced by the observed degree d_c^Ω. The resulting λ_c depends on the data only through the mask, which is independent of the noise, so the proof goes through unchanged and (40) holds up to a constant factor. This is the rule used in the experiments.

C.7 Comparison with Global Nuclear-Norm Regularization

The purpose of GAME is to control the complexity with respect to categories rather than solely for the complexity of the entire matrix. To make this precise, compare (40) with the bound for the global nuclear-norm estimator

$$\widehat{\mathbf{W}}^{\mathrm{global}} \in \arg \min_{\|\mathbf{W}\|_\infty \leq \alpha^\star/\sqrt{nm}} \left\{ \tfrac{1}{2}\|\mathcal{P}_\Omega(\mathbf{X} - \mathbf{W})\|_F^2 + \lambda \|\mathbf{W}\|_* \right\}, \qquad r^\star := \mathrm{rank}(\mathbf{W}^\star). \tag{42}$$

The global estimator is the one-category specialization of GAME: take $\mathcal{C} = \{0\}$ with $I_0 = [n]$, so that $\kappa_{\min} = \kappa_{\max} = 1$, $r_0 = r_{\mathrm{tot}} = r^\star$, $\theta_\lambda = 1$, and $n_0 \vee m = n \vee m = d_0$. Theorem C.5 then gives

$$\frac{\|\widehat{\mathbf{W}}^{\mathrm{global}} - \mathbf{W}^\star\|_F^2}{nm} \lesssim \frac{\sigma^2\, r^\star d_0 \log(n+m)}{\overline{N}} + (\alpha^\star)^2\, r^\star\, \frac{d_0 \log(n+m)}{nm\,\overline{N}}, \tag{43}$$

whose noise term is the minimax-optimal rate for noisy nuclear-norm matrix completion [Negahban and Wainwright, 2011, Cor. 1]. The comparison below is therefore simultaneously a comparison with our own one-category bound and with the literature.

Term-by-term ratios. Dividing (40) by (43) term by term, the common factors cancel and

$$\frac{\text{GAME noise term}}{\text{global noise term}} \asymp \frac{\kappa_{\max}}{\kappa_{\min}^2} \cdot \frac{\sum_{c \in \mathcal{C}} r_c(n_c \vee m)}{r^\star(n \vee m)}, \qquad \frac{\text{GAME spikiness term}}{\text{global spikiness term}} \asymp \frac{\kappa_{\max}}{\kappa_{\min}^2}\, \theta_\lambda^2\, \frac{r_{\mathrm{tot}}}{r^\star}. \tag{44}$$

The global estimator pays rank times dimension once, for the whole matrix. GAME pays rank times dimension once per category, with that category's rank and that category's block size. Each term in the

sum has a smaller rank and a smaller block than the global product, so the sum can be far smaller than $r^\star(n \vee m)$. This requires tall categories: if $n \le m$, then $\sum_c r_c(n_c \vee m) = m\, r_{\mathrm{tot}} \ge \kappa_{\min}\, m\, r^\star$, since row-subset rank is submodular and hence fractionally subadditive over the cover, so the noise-term ratio is at least $\kappa_{\max}/\kappa_{\min} \ge 1$. The prefactor $\kappa_{\max}/\kappa_{\min}^2$ is the cost of overlap and equals one when the categories are disjoint.

The spikiness terms do not favor GAME. There the ratio is $\theta_\lambda^2\, \kappa_{\max} r_{\mathrm{tot}}/(\kappa_{\min}^2 r^\star)$, which is at least a constant because $r_{\mathrm{tot}} \ge r^\star$. The reason is that this term comes from the curvature of the sampling operator, which is a property of the whole matrix and does not see the block structure. In the regime where noise rather than sampling limits accuracy, which is the regime of Theorem C.5, the noise term dominates and (44) governs the comparison.

When the result is sharper. The ratio (44) is small when the matrix is only moderately low rank as a whole but much lower rank within each category. The clearest case is a fixed number of categories per row and each $c \in \mathcal{C}$ is comparably sized. Suppose each row belongs to at most ϱ categories (e.g. each user in a matrix has only an age and sex label, implying $\varrho = 2$), each category has $n_c \asymp \varrho n/|\mathcal{C}|$ rows and local rank r_{loc}, and the categories are tall matrices with $n_c \gg m$. If the categories' row spaces are largely distinct, so that $r^\star \gtrsim |\mathcal{C}|\, r_{\mathrm{loc}}/\varrho$, then

$$\frac{\text{GAME noise term}}{\text{global noise term}} \lesssim \frac{\varrho^2\, r_{\mathrm{loc}}\, n}{(|\mathcal{C}| r_{\mathrm{loc}}/\varrho)\, n} = \frac{\varrho^3}{|\mathcal{C}|}. \tag{45}$$

With a bounded number of categories per row, GAME improves the dominant term by a factor of order $|\mathcal{C}|$. The condition on ϱ restricts how many categories a single row can belong to, not how many categories there are. If rows are tagged by a fixed set of attributes, such as age bucket and sex, or brain region and recording session, then ϱ is the number of attributes and refining any attribute grows $|\mathcal{C}|$ without changing ϱ.

C.8 Subspace Recovery

Many downstream uses of GAME care less about entrywise reconstruction than about the per-category right singular subspaces, i.e. the latent directions along which category c varies. A Frobenius bound on the global error immediately bounds each block error, and block errors control singular subspaces through Wedin's $\sin\Theta$ theorem, provided the perturbation is smaller than the r_c-th singular value of the block. Let $\mathcal{B}^2$ denote nm times the right-hand side of (39). On the event of Theorem C.5, then we have $\|\widehat{\mathbf{W}} - \mathbf{W}^\star\|_F \le \mathcal{B}$.

Corollary C.8 (Per-category right-subspace recovery)**.** *Let $r_c \ge 1$, let $\mathbf{Q}_c^\star \in \mathbb{R}^{m \times r_c}$ contain the top r_c right singular vectors of $\mathbf{W}_c^\star$, and let $\widehat{\mathbf{Q}}_c$ contain the top r_c right singular vectors of $\widehat{\mathbf{W}}_c$. Let*

$$\mathrm{O}(r_c) = \{\mathbf{O} \in \mathbb{R}^{r_c \times r_c} : \mathbf{O}^\top \mathbf{O} = \mathbf{I}_{r_c}\}.$$

define the orthogonal group. On the event of Theorem C.5, if $\mathcal{B} < \sigma_{r_c}(\mathbf{W}_c^\star)$, then

$$\min_{\mathbf{O} \in \mathrm{O}(r_c)} \|\widehat{\mathbf{Q}}_c \mathbf{O} - \mathbf{Q}_c^\star\|_F^2 \le \frac{2\mathcal{B}^2}{\left[\sigma_{r_c}(\mathbf{W}_c^\star) - \mathcal{B}\right]^2}, \tag{46}$$

and in particular, if $\mathcal{B} \le \frac{1}{2}\sigma_{r_c}(\mathbf{W}_c^\star)$,

$$\min_{\mathbf{O} \in \mathrm{O}(r_c)} \|\widehat{\mathbf{Q}}_c \mathbf{O} - \mathbf{Q}_c^\star\|_F^2 \le 8\, \frac{\mathcal{B}^2}{\sigma_{r_c}^2(\mathbf{W}_c^\star)}. \tag{47}$$

Proof. Let $\mathbf{\Delta}_c := \widehat{\mathbf{W}}_c - \mathbf{W}_c^\star$, so $\|\mathbf{\Delta}_c\|_{\mathrm{op}} \le \|\mathbf{\Delta}_c\|_F \le \mathcal{B}$. Weyl's inequality gives $\sigma_{r_c+1}(\widehat{\mathbf{W}}_c) \le \sigma_{r_c+1}(\mathbf{W}_c^\star) + \|\mathbf{\Delta}_c\|_{\mathrm{op}} = \|\mathbf{\Delta}_c\|_{\mathrm{op}} < \sigma_{r_c}(\mathbf{W}_c^\star)$, so the gap condition of Wedin's theorem [YU et al., 2015] holds and

$$\|\sin\Theta(\widehat{\mathbf{Q}}_c, \mathbf{Q}_c^\star)\|_F \le \frac{\|\mathbf{\Delta}_c\|_F}{\sigma_{r_c}(\mathbf{W}_c^\star) - \|\mathbf{\Delta}_c\|_{\mathrm{op}}} \le \frac{\mathcal{B}}{\sigma_{r_c}(\mathbf{W}_c^\star) - \mathcal{B}}.$$

The Procrustes–principal-angle inequality $\min_{\mathbf{O}} \|\widehat{\mathbf{Q}}_c \mathbf{O} - \mathbf{Q}_c^\star\|_F^2 \le 2\|\sin\Theta(\widehat{\mathbf{Q}}_c, \mathbf{Q}_c^\star)\|_F^2$ [YU et al., 2015] gives (46). Under the stronger hypothesis the denominator is at least $\sigma_{r_c}(\mathbf{W}_c^\star)/2$. □

Since the conditioning constant $\sigma_{r_c}(\mathbf{W}_c^\star)$ depends only on $\mathbf{W}_c^\star$, the same statement holds for the global estimator with $\mathcal{B}$ replaced by its own error bound, and the ratio of per-category subspace errors reduces to the ratio of Frobenius errors analyzed in Section C.7.

C.9 Extension to Nonuniform Sampling

Structured missingness, where some subgroups are observed far less often than others, is one of the motivating settings for GAME. We extend the preceding guarantees to category-dependent sampling probabilities, with explicit factors quantifying sampling imbalance.

Sampling model. Let each row i have sampling mass $p_i > 0$, with $\sum_i p_i = 1$, and suppose

$$Z_{ij} \sim \text{Bernoulli}\left(\frac{Np_i}{m}\right)$$

independently across entries, with $N \max_i p_i \le m$. The sampling mask is independent of the noise $\mathbf{E}$.

A natural construction uses meta-category intersections: the intersections partition the rows into disjoint groups g, each assigned mass $\rho_g > 0$, with $\sum_g \rho_g = 1$. For $i \in g$, set $p_i = \rho_g/|g|$. This permits different observation probabilities across subgroups, including block-wise missingness.

Define the relative sampling weights

$$w_i = np_i, \qquad w_{\min} = \min_i w_i, \qquad w_{\max} = \max_i w_i, \qquad \chi = \frac{w_{\max}}{w_{\min}}.$$

Under uniform sampling, $w_i = 1$ and $\chi = 1$. Assume $w_{\min} \ge 1/L_s$ for a fixed constant $L_s \ge 1$.

For each category c, define

$$\bar{d}_c := N\left(\max_{i \in I_c} p_i \ \vee \ \frac{1}{m}\sum_{i \in I_c} p_i\right),$$

the largest expected row or column degree within category c. Choose λ_c according to Equation (37), replacing the uniform expected degree by $\bar{d}_c$.

Theorem C.9 (GAME under nonuniform sampling)**.** *Suppose the sampling model above holds, together with Assumptions C.2–C.4. Assume*

$$N \ge C_0(L_s) d_0 \log(n+m), \qquad \bar{d}_c \ge C_{\deg}\ell_c \quad \textit{for all } c \in \mathcal{C},$$

where $d_0 = \max\{n, m\}$, and $C_0(L_s)$ may depend on the fixed lower-weight bound L_s. With the category-specific regularization parameters defined above, the restricted GAME estimator satisfies, with high probability,

$$\begin{aligned}\frac{\|\widehat{\mathbf{W}} - \mathbf{W}^\star\|_F^2}{nm} \lesssim \frac{\kappa_{\max}}{\kappa_{\min}^2 w_{\min}^2}\frac{nm\log(n+m)}{N}\left[\sigma^2 \sum_{c \in \mathcal{C}} r_c \frac{\bar{d}_c}{N} + \frac{R^2\log(n+m)}{N} r_{\text{tot}}\right] \\ + \chi^2(\alpha^\star)^2 \frac{\kappa_{\max}}{\kappa_{\min}^2}\theta_\lambda^2 r_{\text{tot}} \frac{d_0 \log(n+m)}{nmN}.\end{aligned} \tag{48}$$

Here the implicit constants may depend on L_s, but not on n, m, N, or the number of categories. The high-probability statement has the same form as Theorem C.5, with constants in the exponential tail allowed to depend on L_s.

Proof. The proof follows that of Theorem C.5, with modifications to the population curvature, category-level degrees, and restricted strong convexity argument.

1. Population curvature. Let $R_w = \text{diag}(w_1, \dots, w_n)$ and define

$$\|A\|_{\omega(F)} := \|R_w^{1/2} A\|_F.$$

Since $\mathbb{E}Z_{ij} = Np_i/m$,

$$\begin{aligned}\mathbb{E}\|\mathcal{P}_\Omega(A)\|_F^2 &= \sum_{i,j} \frac{Np_i}{m} A_{ij}^2 \\ &= \frac{N}{nm}\sum_{i,j} w_i A_{ij}^2 \\ &= \frac{N}{nm}\|A\|_{\omega(F)}^2.\end{aligned}$$

Thus, the population loss retains curvature in the weighted Frobenius norm.

2. Degrees and noise control. For category c, the expected observed degree of row $i \in I_c$ is Np_i, while that of column j, restricted to I_c, is

$$\frac{N}{m} \sum_{i \in I_c} p_i.$$

Both are bounded by $\bar{d}_c$. Since the observation indicators remain independent, Bernstein's inequality gives the same degree concentration as Lemma C.1, with $\bar{d}_c$ replacing the uniform expected degree.

The deterministic decomposability, row-multiplicity, and compatibility arguments of Lemmas C.2, C.3, and C.5–C.7 remain unchanged. Lemma C.4 is a conditional probabilistic statement rather than a deterministic one. It continues to apply because, conditional on the realized mask Ω, the noise entries remain independent with the same subexponential moment bounds. Consequently, the per-category reweighted noise satisfies

$$\|\widetilde{\mathbf{E}}^{(c)}\|_{\mathrm{op}} \le \frac{C_1}{\kappa_{\min}} \left[\sigma \sqrt{d_c^{\Omega} \ell_c} + R \ell_c \right]$$

with the probability stated in Lemma C.4. Combining this event with the degree concentration event ensures

$$\lambda_c \ge 2 \|\widetilde{\mathbf{E}}^{(c)}\|_{\mathrm{op}} \qquad \text{for every } c.$$

Hence the cone constraint and compatibility bounds of Lemmas C.6–C.7 continue to hold.

3. Weighted restricted strong convexity. We extend Lemma C.8 to the nonuniform sampling operator. Set

$$q = \frac{N}{nm}, \qquad \Gamma = R_w^{1/2} A,$$

so that $Z_{ij} \sim \mathrm{Bernoulli}(q w_i)$. Define

$$Q_Z(\Gamma) := \frac{1}{q} \sum_{i,j} \frac{Z_{ij}}{w_i} \Gamma_{ij}^2.$$

Then

$$\mathbb{E} Q_Z(\Gamma) = \|\Gamma\|_F^2.$$

For $D, a, \rho > 0$, consider

$$\mathcal{S}(D, a, \rho) = \{\Gamma : \|\Gamma\|_F \le D,\ \|\Gamma\|_\infty \le a,\ \|\Gamma\|_* \le \rho\}.$$

By symmetrization, contraction, and trace duality,

$$\begin{aligned} &\mathbb{E} \sup_{\Gamma \in \mathcal{S}(D,a,\rho)} \left| Q_Z(\Gamma) - \|\Gamma\|_F^2 \right| \\ &\qquad \lesssim \frac{a\rho}{q} \mathbb{E} \left\| \sum_{i,j} \frac{\varepsilon_{ij} Z_{ij}}{w_i} e_i e_j^\top \right\|_{\mathrm{op}}, \end{aligned}$$

where ε_{ij} are independent Rademacher variables.

Each summand in the random matrix has operator norm at most L_s. The row and column variance parameters obey

$$\begin{aligned} \max_i \sum_j \frac{\mathbb{E} Z_{ij}}{w_i^2} &= \max_i \frac{N}{n w_i} \le \frac{L_s N}{n}, \\ \max_j \sum_i \frac{\mathbb{E} Z_{ij}}{w_i^2} &= \frac{N}{nm} \sum_i \frac{1}{w_i} \le \frac{L_s N}{m}. \end{aligned}$$

Matrix Bernstein therefore gives

$$\mathbb{E} \left\| \sum_{i,j} \frac{\varepsilon_{ij} Z_{ij}}{w_i} e_i e_j^\top \right\|_{\mathrm{op}} \lesssim \sqrt{L_s q\, d_0 \log(n+m)} + L_s \log(n+m).$$

Under $N \geq C_0(L_s) d_0 \log(n+m)$, the second term is absorbed into the first. It follows that

$$\mathbb{E} \sup_{\Gamma \in \mathcal{S}(D,a,\rho)} \left| Q_Z(\Gamma) - \|\Gamma\|_F^2 \right| \lesssim a\rho \sqrt{\frac{L_s nm\, d_0 \log(n+m)}{N}}.$$

For concentration, each centered summand

$$\frac{Z_{ij} - qw_i}{qw_i} \Gamma_{ij}^2$$

has absolute value at most $L_s a^2/q$, and the sum of its variances is bounded by $L_s D^2 a^2/q$. Applying the bounded empirical-process concentration inequality of Klein and Rio Klein and Rio [2005], which permits independent, non-identically distributed coordinates, and absorbing the mixed expectation term yields, with probability at least $1 - 2e^{-t}$,

$$\sup_{\Gamma \in \mathcal{S}(D,a,\rho)} \left| Q_Z(\Gamma) - \|\Gamma\|_F^2 \right| \lesssim a\rho \sqrt{\frac{L_s nm\, d_0 \log(n+m)}{N}} + Da \sqrt{\frac{L_s nm\, t}{N}} + \frac{L_s nm\, a^2 t}{N}.$$

Define the weighted spikiness and rank ratios

$$\alpha_{\mathrm{sp},\omega}(A) := \frac{\sqrt{nm}\|\Gamma\|_\infty}{\|\Gamma\|_F}, \qquad \beta_{\mathrm{ra},\omega}(A) := \frac{\|\Gamma\|_*}{\|\Gamma\|_F}.$$

By homogeneity, normalize $\|\Gamma\|_F = 1$ and consider the restricted class

$$\alpha_{\mathrm{sp},\omega}(A)\beta_{\mathrm{ra},\omega}(A) \leq T, \qquad T = \frac{1}{c_0(L_s)} \sqrt{\frac{N}{d_0 \log(n+m)}}.$$

Partition this class into slices satisfying

$$2^k \leq \alpha_{\mathrm{sp},\omega}(A) < 2^{k+1}.$$

On slice k, take

$$D = 1, \qquad a = \frac{2^{k+1}}{\sqrt{nm}}, \qquad \rho = \frac{T}{2^k}.$$

Choosing $t = c_t d_0 \log(n+m)$, the three terms in the deviation bound are bounded, up to universal constants, by

$$\frac{\sqrt{L_s}}{c_0(L_s)}, \qquad \frac{\sqrt{L_s c_t}}{c_0(L_s)}, \qquad \frac{L_s c_t}{c_0(L_s)^2}.$$

Taking $c_0(L_s)$ sufficiently large and $C_0(L_s)$ large enough that $T \geq 1$, these terms sum to at most $1/2$. A union bound over the logarithmically many slices therefore yields

$$Q_Z(\Gamma) \geq \frac{1}{2}\|\Gamma\|_F^2$$

uniformly over the restricted class, with failure probability bounded by

$$c_1(L_s) \exp\{-c_2(L_s) d_0 \log(n+m)\}.$$

Returning to $A = R_w^{-1/2}\Gamma$,

$$\begin{aligned} \|\mathcal{P}_\Omega(A)\|_F^2 &= qQ_Z(\Gamma) \\ &\geq \frac{N}{2nm}\|\Gamma\|_F^2 \\ &= \frac{N}{2nm}\|A\|_{\omega(F)}^2 \\ &\geq \frac{N w_{\min}}{2nm}\|A\|_F^2. \end{aligned}$$

This establishes weighted restricted strong convexity, with constants depending on L_s.

4. Error-bound assembly. Let $\Delta = \widehat{\mathbf{W}} - \mathbf{W}^\star$. The compatibility bounds of Lemma C.7 give

$$\frac{1}{2}\|\mathcal{P}_\Omega(\Delta)\|_F^2 \le \frac{3\sqrt{2\kappa_{\max}}}{2} A_\lambda \|\Delta\|_F, \qquad \|\Delta\|_* \le B_\lambda \|\Delta\|_F,$$

where

$$A_\lambda^2 = \sum_c \lambda_c^2 r_c, \qquad B_\lambda \le \frac{4\sqrt{2\kappa_{\max}}}{\kappa_{\min}} \theta_\lambda \sqrt{r_{\text{tot}}}.$$

The weighted norm comparisons

$$\begin{aligned} \sqrt{w_{\min}}\|A\|_F \le \|A\|_{\omega(F)} \le \sqrt{w_{\max}}\|A\|_F, \\ \|A\|_{\omega(*)} \le \sqrt{w_{\max}}\|A\|_*, \\ \|A\|_{\omega(\infty)} \le \sqrt{w_{\max}}\|A\|_\infty \end{aligned}$$

and the spikiness constraint imply

$$\alpha_{\text{sp},\omega}(\Delta)\beta_{\text{ra},\omega}(\Delta) \le \frac{2\chi\alpha^\star B_\lambda}{\|\Delta\|_F}.$$

Consequently, the weighted RSC inequality applies whenever

$$\|\Delta\|_F > \tau_\omega, \qquad \tau_\omega := 2c_0(L_s)\chi\alpha^\star B_\lambda \sqrt{\frac{d_0 \log(n+m)}{N}}.$$

If $\|\Delta\|_F \le \tau_\omega$, the error is already bounded by the sampling-tolerance term. Otherwise, combining weighted RSC with the compatibility inequality gives

$$\frac{N w_{\min}}{4nm}\|\Delta\|_F^2 \le \frac{3\sqrt{2\kappa_{\max}}}{2} A_\lambda \|\Delta\|_F,$$

hence

$$\|\Delta\|_F \le \frac{6\sqrt{2\kappa_{\max}}\, nm}{N w_{\min}} A_\lambda.$$

Combining the two cases,

$$\frac{\|\Delta\|_F^2}{nm} \le \frac{72\kappa_{\max} nm}{N^2 w_{\min}^2} A_\lambda^2 + \frac{\tau_\omega^2}{nm}.$$

Finally, the regularization rule and the degree condition $\bar{d}_c \gtrsim \ell_c$ give

$$\lambda_c^2 \lesssim \frac{1}{\kappa_{\min}^2}\left[\sigma^2 \bar{d}_c \ell_c + R^2 \ell_c^2\right].$$

Substituting this into A_λ^2, using $\ell_c \lesssim \log(n+m)$, and bounding B_λ^2 as above yields (48).

The degree, conditional noise, and weighted RSC events hold simultaneously by a union bound. All constants introduced by the weighted argument depend at most on the fixed parameter L_s. □

Interpretation. The factor $w_{\min}^{-2}$ quantifies the cost of rows observed less often than average, while χ^2 measures sampling imbalance in the spikiness term. Both equal one under uniform sampling, recovering the corresponding uniform guarantee. The theorem therefore accommodates category-dependent block missingness, provided every subgroup retains sufficient observations and sampling imbalance remains controlled.

C.10 Reading the Theoretical Results

The theoretical results above allow for nice interpretation, giving insight for hyperparameter tuning and requirements for the number of observed entries for each category. Here, we list important features of our theory in helping understand the mechanisms behind GAME.

Category-specific regularization. The rule (37), in its simplified form (41), says that λ_c should dominate the noise visible inside category c, which scales with three quantities: the noise level σ, the expected (or observed) largest row/column degree of the category, and a logarithmic factor in the block dimension $d_c = n_c + m$. Smaller or more sparsely covered categories therefore need stronger relative regularization, while larger and better-sampled categories receive weaker relative shrinkage. All three inputs are observable or estimable, leading to a practical rule.

Category-specific sample budgets. Under uniform sampling, category c receives approximately $\overline{N} n_c/n$ observations, with (21) ensuring concentration when $\overline{N} n_{\min}/n \gtrsim \log|\mathcal{C}|$. GAME's leading error term scales with $\sum_c r_c(n_c \vee m)$, so smaller, lower-rank categories contribute less to estimation error. The overall sample-size requirement remains $\overline{N} \gtrsim (n \vee m)\log(n+m)$, as in standard matrix completion. Under nonuniform sampling, Theorem C.9 additionally requires $\bar{d}_c \gtrsim \log(n+m)$ and accounts for sampling imbalance. Thus, reliable recovery depends not only on the total number of observations but also on how they are distributed across categories. In practice, categories with inadequate observation coverage should be sampled more heavily or merged with structurally similar categories, balancing local specificity against statistical reliability.

The price of overlap. The factor $\kappa_{\max}/\kappa_{\min}^2$ equals one for a disjoint cover and quantifies the cost of letting rows belong to several categories, where the same row then contributes to several nuclear-norm penalties and to several blocks of the loss decomposition. It stays bounded as long as each row belongs to a bounded number of categories, regardless of how many categories there are in total. In practice, it is reasonable to assume that this overlap is bounded, e.g. age and gender overlap (since they both describe a user), therefore $\kappa = 2$.

When GAME helps. Section C.7 shows that GAME has the sharper upper bound when the local complexity $\sum_c r_c(n_c \vee m)$ is smaller than the global complexity $r^\star(n \vee m)$ by more than the overlap cost, and that the gain is confined to the noise term; in the spikiness term GAME pays the global rank proxy r_{tot}. The favorable regime is therefore: a matrix that is only moderately low rank globally but substantially lower rank within categories, a bounded number of categories per row, and moderately balanced sampling.

Subspace recovery. Corollary C.8 connects GAME's reconstruction guarantee to the recovery of category-specific latent directions. By Wedin's perturbation theorem, the subspace error satisfies

$$\text{subspace distance}_c^2 \;\lesssim\; \frac{\mathcal{B}^2}{\sigma_{r_c}^2(\mathbf{W}_c^\star)},$$

provided $\mathcal{B} \le \frac{1}{2}\sigma_{r_c}(\mathbf{W}_c^\star)$, where $\mathcal{B}$ is the global Frobenius error bound and $\sigma_{r_c}(\mathbf{W}_c^\star)$ is the weakest nonzero singular value of category c. Thus, accurate reconstruction translates into accurate subspace recovery when the category's latent directions are sufficiently strong relative to the estimation error. Weak directions, by contrast, are more sensitive to estimation noise and require greater accuracy to recover reliably.

Although this result follows from a standard perturbation argument, it gives an important interpretation of GAME's reconstruction guarantees. When GAME achieves a sharper error bound than global nuclear-norm completion, this improvement also yields a tighter bound on category-specific subspace error. This connects the local low-rank advantages established in Section C.7 to the preservation of latent geometry, which is central to our clustering and neural-dynamics experiments.

D Additional Experiments and Details

D.1 Missing Metadata

If we are in the scenario of no group metadata, one could consider a post-hoc procedure by cleverly clustering a naively imputed (e.g. column mean imputation) version of the data, and use the produced cluster labels as group labels.

To exemplify this empirically, we extend the synthetic clustering experiment with subgroup overlap from Section 4.1 to evaluate whether group metadata can be estimated automatically. We retain the original data-generating process but do not provide GAME with the 10 oracle groups in $\mathcal{G}$. Instead, we first impute missing entries using column means and then apply K-means with $K = 10$, the true number of groups, to estimate the meta-categories from the imputed matrix.

After constructing these approximate meta-categories, we then run completion on the original (non-imputed) matrix with missing entries as described in Section 4.1 for GAME and the remaining benchmark models (all side-information baselines receive the same estimated labels). We present the ARI results in Figure 8, analogous to the results in Figure 3(a).

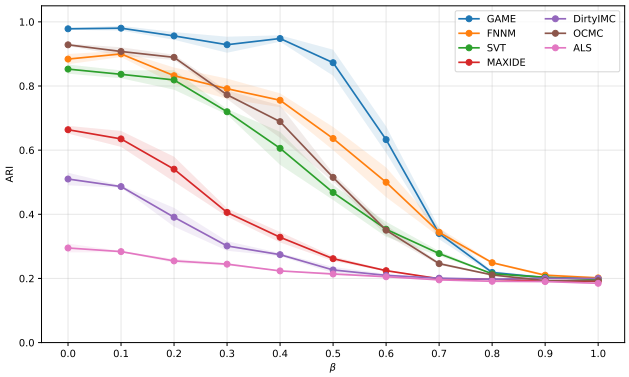


Figure 8: ARI (↑) of hidden-subcluster recovery as a function of β for the synthetic clustering experiment (Section 4.1), with meta-categories estimated by K-means on the column-mean-imputed matrix. Shaded regions denote ± 1 standard error across 3 masking replications.

GAME attains the highest ARI for $\beta \leq 0.6$, and by a wide margin for $\beta \leq 0.5$, even though the estimated groups **do not perfectly recover** the oracle meta-categories used to generate the data. For $\beta \geq 0.7$, all methods approach the same floor ARI value. Compared with oracle groups (Figure 3(a)), GAME's breakdown shifts earlier from $\beta \approx 0.7$ to $\beta \approx 0.6$. These results demonstrate GAME's robustness to imperfect group assignments and suggest that data-driven clustering can provide useful group information when metadata is unavailable.

A caveat is that the clustering results may be a function of nuisance variation in the observed data matrix, which can result in poorly assigned group labels. Figure 2(c) shows that GAME degrades gracefully under random mislabeling, but errors from clustering may be systematic (i.e. non-random), so we recommend validating estimated groups (e.g. by held-out completion error) before use.

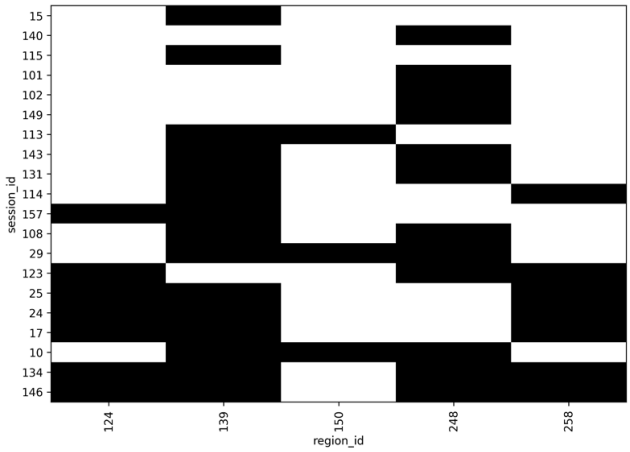


Figure 9: **Structured missingness of regions across Neuropixels recording sessions from [Chen et al., 2024].** White shows observed region-session pairs and black shows structured missingness as a result of experimental design and technological limitations.

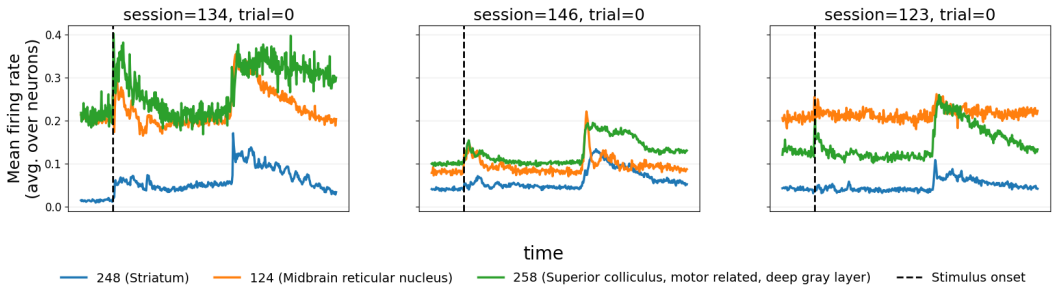


Figure 11: **Mean firing rates of three brain regions over three recording sessions from [Chen et al., 2024]**. Neuron spike trains are averaged to get mean firing rates for each brain region over time. Though single regions exhibit similar temporal dynamics across the sessions, firing rate comparisons between regions are visibly inconsistent.

D.2 Experiments Compute Resources

Experiments on synthetic, MovieLens, and BirdSet datasets were executed locally on a 2024 Apple Silicon MacBook environment. Due to the larger scale of the Svoboda Lab Neuropixels dataset, those experiments were executed on a university high-performance computing cluster using GPU-enabled compute nodes. Typical Neuropixels runs required approximately one hour for 500 GAME iterations.

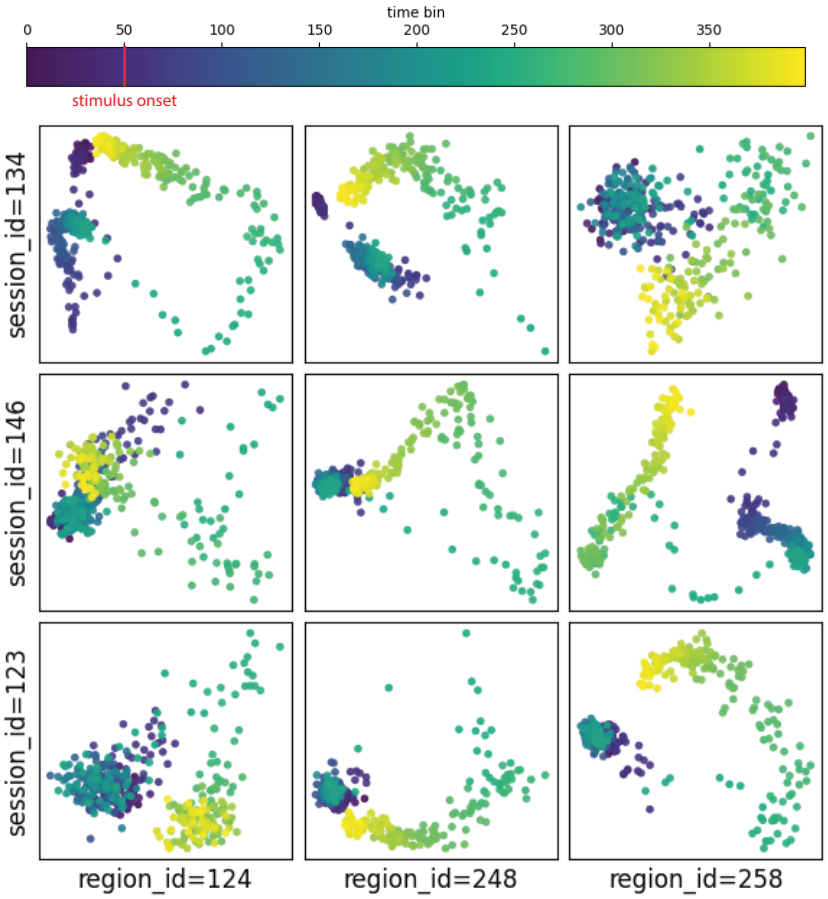


Figure 10: **Principal dynamics by subsetted session-region combinations from [Chen et al., 2024]**. We bin the data into 10 ms windows over fixed 4-second trials, for $T = 400$, and align the trials to the stimulus ($t = 50$). We further restrict analyses to successful trial types from either “hit left" or “hit right", keeping neurons with more than 10 spikes along the 400 timepoints, and specifically select the most populous single trial type (trial ID 0). We randomly select 20 of the most measured recording sessions and 5 of the most heavily measured brain regions across those recordings. PCA is computed on each subsetted pair of brain region spike trains by recording session to obtain temporal dynamics across a task. Region 124 corresponds to midbrain reticular nucleus, 248 corresponds to striatum, and 258 corresponds to the superior colliculus.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contribution of the paper and of our method is accurately represented in the abstract and introduction, both in method, theory, experiments, and case study.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations both in the theoretical assumptions necessary to obtain guarantees, and in implementation and tuning of our algorithm.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide extensive details of the assumptions and complete proofs for our theoretical guarantees, with a fully developed section in the main text and substantial supporting material in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All datasets are cited and publicly available, processing steps defined, and comparison methods referenced in order for the experimental results to be reproduced.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have linked a GitHub repository with sufficient instructions to faithfully reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Full data descriptions are given, and we have sections on hyperparameter tuning and selection for our method (Section 4, Theorem 3.5, Remark C.7).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Confidence bands are reported in figures when appropriate.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our method is not computationally expensive, and we have a brief section in the appendix describing what resources we used for our experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research conforms in every respect to the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There are no wide-ranging societal impacts of our results. GAME is a general-purpose matrix estimation method with no direct connection to specific applications that raise societal concerns; our contributions are methodological and theoretical, and the datasets used are either synthetic or publicly available scientific benchmarks.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our method poses no such risks, and has little to no potential for misuse since it is not a deployed tool, and hence there are no relevant safeguards to discuss.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We are the creators of the method, all contributors are appropriately credited and authored, and all the code used to generate experimental results is our own and available at `https://github.com/mattyshen/game`. We appropriately credit the datasets and comparison methods we use as well.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code is published at `https://github.com/mattyshen/game`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Our core methods do not involve LLM development or deployment in any manner.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.