SMART: A Unified Framework for Structural Representation Attribution

Anonymous authors
Paper under double-blind review

Abstract

Recovering latent signals via representation learning is fundamental to analyzing high-dimensional complex systems. However, existing approaches predominantly focus on static single-environment representations, overlooking mechanism shifts across environments, which are crucial for causal discovery, change point detection, and transfer learning, among others. We propose SMART (Sparse Mechanism Attribution for RepresenTation), a unified framework for structural representation attribution. Our two-stage framework first performs signal recovery, then applies sparse regularization on structural differences to attribute distributional shifts to specific mechanisms, yielding an attribution that circumvents the notorious identifiability issue. Theoretically, we show that estimation errors for structure and representation inherit the convergence rates of the first stage. A consistent information criterion is also introduced to determine latent dimensionality. Furthermore, we develop a one-step estimation method for additive noise representations and extend SMART to multi-environment scenarios with theoretical guarantees. Extensive simulations and applications to two real data datasets demonstrate the effectiveness and practical utility of SMART.

1 Introduction

High-dimensional data are often driven by low-dimensional latent structures. Distilling valid signals from noise constitutes the foundational step in comprehending complex systems, with prominent methodologies ranging from matrix factorization and factor analysis to modern representation learning techniques like Variational Autoencoders (VAEs) (Mnih & Salakhutdinov, 2007; Kingma & Welling, 2013; Liu et al., 2023).

However, in scientific and engineering domains, the central challenge lies in change. For instance, causal discovery requires identifying mechanisms altered by interventions; change point detection seeks to localize specific components drifting during state transitions; and transfer learning distinguishes invariant knowledge from domain specific variations (Aminikhanghahi & Cook, 2017; Rothenhäusler et al., 2021; Zhu et al., 2023). Essentially, these distinct tasks focus on a unified objective: structural attribution. Consequently, the shared imperative extends beyond accurate estimation to precise structural attribution. The goal is to disentangle structural drifts from distributional shifts, explicitly pinpointing the specific components of the latent generative process that have undergone variation.

Despite this shared goal, existing approaches encounter a fundamental mismatch. First, methods optimizing reconstruction error or distributional distance detect whether a change occurred but fail to localize it. Second, the identifiability barrier arises from rotational ambiguity in latent variable models (Bai & Li, 2012; Cui & Xu, 2025). Comparing estimators across environments directly results in dense loading differences due to arbitrary rotations, rendering the attribution uninterpretable.

In this paper, we address these challenges by proposing **SMART** (**S**parse **M**echanism **A**ttribution for **R**represen**T**ation), a unified framework designed to reframe structural attribution as an identifiability problem after estimation. Adopting a two-stage strategy, SMART first performs signal recovery via representation learning, and subsequently shifts the focus from minimizing reconstruction error to explicitly optimizing the parsimony of structural differences. Rather than accepting the dense discrepancies resulting from separate estimations, our framework seeks a latent alignment that yields the sparsest differential structure. By enforcing

a row sparsity constraint on the loading differences, we recover the true, sparse locus of the mechanism shift from dense artifacts. Theoretically, we demonstrate that the estimation errors for both structure and representation inherit the convergence rates of the first stage. We also propose a consistent information criterion to determine the latent dimensionality. Furthermore, we develop a unified one-step algorithm for additive noise models and extend the framework to multi-environment scenarios via a fused group penalty, establishing rigorous theoretical guarantees for both extensions.

The contribution of SMART is therefore not another single-environment signal recovery model, but an inferential layer that turns recovered representations into identifiable structural attributions. In contrast to transfer-learning methods that usually treat differences across environments as nuisance terms to be corrected for, SMART treats these differences as the primary objects of inference and provides support-level identifiability under the rotational equivalence class of latent factor models.

2 Related Work

Representation Learning and Factor Analysis. Dimensionality reduction via representation learning is fundamental to analyzing high-dimensional data, distilling informative latent variables from observed features (Bengio et al., 2013). Prominent approaches include statistical factor analysis (Fan et al., 2013; Chen et al., 2020; Liu et al., 2023), positing a linear generative process, and AEs or VAEs, capturing complex non-linear dependencies (Kingma & Welling, 2013; Vahdat & Kautz, 2020). While these methods excel in static environments, they face challenges under distribution shifts. Standard multi-group latent factor models typically assume either identical structures across groups or use tensor techniques (Han et al., 2024; Bolivar et al., 2025). However, existing frameworks fail to explicitly model the sparse structural shifts within the latent generative mechanism, limiting their interpretability in dynamic environments.

Sparse Difference from Transfer Learning. Our work draws inspiration from transfer learning, where source and target domains are assumed to share a dominant common structure with sparse disparities. Pioneering statistical works leverage this sparsity assumption to transfer knowledge (Bastani, 2021; Tian & Feng, 2023). In the machine learning community, transferability is rigorously analyzed through representation learning, leveraging task diversity to guarantee generalization to target domains (Tripuraneni et al., 2020; Zhu et al., 2023). Crucially, most existing methods assume shifts in downstream task labels. By contrast, SMART addresses the fundamental challenge of mechanism shifts within the latent generative process itself. Our framework is better suited to scenarios where environmental changes affect specific modular mechanisms.

Causal Invariance and Mechanism Attribution. A growing synergy exists between representation learning and causal inference, particularly in identifying causal representations under distribution shifts (Schölkopf et al., 2021). Some methods seek invariant representations across environments to achieve robust generalization (Arjovsky et al., 2019; Rothenhäusler et al., 2021). While they aim to filter out environmental variations to isolate a stable core, SMART adopts a complementary perspective, utilizing representation invariance as an anchor to localize and attribute structural changes. Prioritizing "difference attribution" over "invariance discovery" renders SMART particularly valuable for scientific discovery tasks, such as identifying gene pathway alterations or detecting structural breaks in economic systems.

3 Methodology

In this section, we describe the two stages of SMART.

3.1 Stage I: Signal Recovery via Representation Learning

Let $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n \times p}$ denote the high-dimensional observation matrix. We assume each entry x_{ij} is drawn from a distribution $f(x_{ij}; m_{ij})$, where m_{ij} represents the parameter of interest. Let $\mathbf{M} = (m_{ij}) \in \mathbb{R}^{n \times p}$ denote the latent parameter matrix. We extract the low-dimensional representation $\mathbf{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)^\top \in \mathbb{R}^{n \times r}$ ($r \ll p$) from $\mathbf{X}$, where $\boldsymbol{f}_i \in \mathbb{R}^r$ denotes the latent vector for the i-th observation, and subsequently use $\mathbf{F}$ to recover the underlying parameter matrix $\mathbf{M}$. Formally, this procedure consists of two mappings:

Representation Learning: We first extract latent structures from the observations via an encoding function $h : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times r}$, yielding the estimated representation

$$\hat{\mathbf{F}} = h(\mathbf{X}).$$

Signal Reconstruction: We then reconstruct the parameter matrix via a decoding function $g : \mathbb{R}^{n\times r} \to \mathbb{R}^{n\times p}$, yielding the estimated signal

$$\tilde{\mathbf{M}} = g(\hat{\mathbf{F}}).$$

We will provide two examples to illustrate this procedure.

Example 1 (Additive Noise Models). Consider the additive noise model $\mathbf{X} = \mathbf{M} + \mathcal{E}$, where $\mathcal{E}$ is a noise matrix with $\mathbb{E}[\mathcal{E}] = \mathbf{0}$. Here, the parameter of interest $\mathbf{M}$ corresponds to the expectation $\mathbb{E}[\mathbf{X}]$. Under the low-rank assumption, we estimate $\mathbf{M}$ by solving the constrained optimization problem:

$$\tilde{\mathbf{M}} = \underset{\mathbf{M}:\mathrm{rank}(\mathbf{M})\le r}{\arg\min} \|\mathbf{X} - \mathbf{M}\|_F^2.$$

In this setting, the solution corresponds to Principal Component Analysis (PCA) or Singular Value Decomposition (SVD). Specifically, $\hat{\mathbf{F}}$ is derived from the top-r left singular vectors (scaled by singular values).

Example 2 (Generalized Linear Models). Assume x_{ij} follows an exponential family distribution with density

$$f(x_{ij}; m_{ij}) = \exp\left[\phi^{-1}\{x_{ij}m_{ij} - \psi(m_{ij})\} + c(x_{ij})\right],$$

where m_{ij} is the natural parameter, $\psi(\cdot)$ and $c(\cdot)$ are prespecified functions. The scale parameter ϕ is known. Assuming that $\mathbf{M}$ can be linearly represented by $\mathbf{F}$, i.e., $\mathbf{M} = \mathbf{F}\mathbf{B}^\top$, where $\mathbf{B} \in \mathbb{R}^{p\times r}$ is the loading matrix. Then this is known as Generalized Factor Models (GFM). The estimation of $\mathbf{F}$ and $\mathbf{B}$ is typically achieved by maximizing the log-likelihood via alternating optimization algorithms (Chen et al., 2020; Liu et al., 2023).

Beyond these, our framework naturally accommodates other dimensionality reduction methods. For instance, deep generative models such as VAEs can be employed, where $h(\cdot)$ and $g(\cdot)$ are parameterized by neural networks (encoder and decoder, respectively) to capture complex dependencies in $\mathbf{M}$ (see Girin et al., 2022, and references therein). Moreover, the approach remains robust to incomplete observations. When only $\mathbf{X} \circ \mathbf{\Omega}$ is observed, where $\mathbf{\Omega}$ denotes a binary mask matrix, variants of the aforementioned methods can be readily adapted to perform effective signal recovery (Gondara & Wang, 2018; Chen & Li, 2022).

3.2 Stage II: Sparse Mechanism Attribution

Consider data matrices $\mathbf{X}^{(1)} = (x_{ij}^{(1)})$ and $\mathbf{X}^{(2)} = (x_{ij}^{(2)})$ from two environments with parameter matrices $\mathbf{M}^{(1)} = (m_{ij}^{(1)})$ and $\mathbf{M}^{(2)} = (m_{ij}^{(2)})$, respectively. To facilitate attribution estimation and enhance interpretability, we postulate that the true parameters share a linear structure induced by a common low-dimensional representation that is invariant across environments. Specifically, we model

$$\mathbf{M}^{(1)} = \mathbf{F}(\mathbf{B}^{(1)})^\top, \quad \mathbf{M}^{(2)} = \mathbf{F}(\mathbf{B}^{(2)})^\top,$$

where $\mathbf{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)^\top \in \mathbb{R}^{n\times r}$ denotes the low-dimensional sample representation, and $\mathbf{B}^{(k)} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)^\top \in \mathbb{R}^{p\times r} (k \in \{1, 2\})$ captures the interaction between representations and features in environment k. We focus on the shift in interactions by attributing changes to the difference $\Delta\mathbf{B} = \mathbf{B}^{(2)} - \mathbf{B}^{(1)} = (\Delta\boldsymbol{b}_1, \ldots, \Delta\boldsymbol{b}_p)^\top \in \mathbb{R}^{p\times r}$. Assuming $\Delta\mathbf{B}$ is row sparse, identifying its non-zero rows reveals the features undergoing variation across two environments. Let $\mathcal{S} = \left\{j : \|\Delta\boldsymbol{b}_j\|_2 \neq 0\right\}$ denote the support of the structural differences, and let $s = |\mathcal{S}|$ be its cardinality.

The target $\mathcal{S}$ is a support of mechanism changes rather than a support of marginal distributional changes in the raw variables. In an additive factor model, the covariance difference contains terms such as $\mathbf{B}^{(1)}\Sigma_{\boldsymbol{f}}(\Delta\mathbf{B})^\top$ and $\Delta\mathbf{B}\Sigma_{\boldsymbol{f}}(\mathbf{B}^{(1)})^\top$; when $\mathbf{B}^{(1)}$ is dense, even a sparse $\Delta\mathbf{B}$ can induce dense changes in observed covariances. Conversely, as illustrated in Appendix C.7, a nonzero row of $\Delta\mathbf{B}$ can be constructed so that the corresponding observed marginal distribution remains unchanged. This is why SMART uses the shared latent representation as an anchor for structural attribution instead of relying on separate coordinate-wise distribution tests.

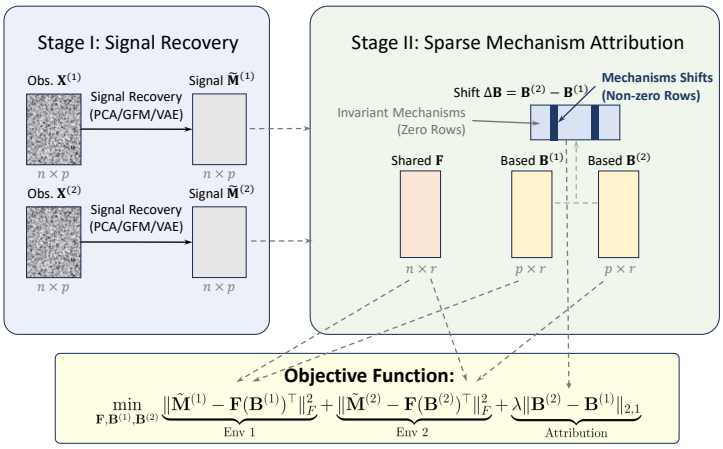


Figure 1: Overview of the SMART framework. Stage I recovers the denoised signals $\tilde{\mathbf{M}}^{(k)}$ from high dimensional observations $\mathbf{X}^{(k)}$ using standard representation learning methods, such as PCA, GFM, or VAE. Stage II performs structural disentanglement by minimizing the reconstruction error with a row sparse penalty on the shift matrix $\Delta\mathbf{B}$.

The proposed framework accommodates diverse scenarios characterized by mechanism shifts. Representative applications include, among others, causal inference for event shocks, where determining the sparse support of $\Delta\mathbf{B}$ reveals sensitive features; transfer learning, which exploits shared latent structures to isolate domain specific variations; and change point detection, which facilitates the simultaneous localization of temporal break points and structural coordinates.

To estimate the model parameters, we adopt a two-stage strategy. In the first stage, we obtain preliminary estimates $\tilde{\mathbf{M}}^{(k)}$ for each environment. This preliminary estimation proceeds without imposing a low-rank constraint, thereby enhancing flexibility in handling complex data structures. Subsequently, we construct the objective function for the second stage by aggregating the low-rank reconstruction losses of $\tilde{\mathbf{M}}^{(k)}$ and imposing a sparsity penalty on the environmental shift matrix

$$\mathcal{Q}(\mathbf{F}, \mathbf{B}^{(1)}, \mathbf{B}^{(2)}; \tilde{\mathbf{M}}^{(1)}, \tilde{\mathbf{M}}^{(2)}) = \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{F}(\mathbf{B}^{(1)})^\top\right\|_F^2 + \left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{F}(\mathbf{B}^{(2)})^\top\right\|_F^2 + \lambda \left\|\mathbf{B}^{(2)} - \mathbf{B}^{(1)}\right\|_{2,1},$$

where the $\|\cdot\|_{2,1}$ norm of a matrix denotes the sum of the ℓ_2-norms of its rows. Recalling the shift $\Delta\mathbf{B} = \mathbf{B}^{(2)} - \mathbf{B}^{(1)}$, we reformulate the optimization problem as

$$\mathcal{Q}(\mathbf{F}, \mathbf{B}^{(1)}, \Delta\mathbf{B}; \tilde{\mathbf{M}}^{(1)}, \tilde{\mathbf{M}}^{(2)}) = \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{F}(\mathbf{B}^{(1)})^\top\right\|_F^2 + \left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{F}(\mathbf{B}^{(1)} + \Delta\mathbf{B})^\top\right\|_F^2 + \lambda\|\Delta\mathbf{B}\|_{2,1} \tag{1}$$

with the estimator given by

$$\left(\hat{\mathbf{F}}, \hat{\mathbf{B}}^{(1)}, \hat{\Delta\mathbf{B}}\right) = \underset{\mathbf{F},\mathbf{B}^{(1)},\Delta\mathbf{B}}{\arg\min}\ \mathcal{Q}(\mathbf{F}, \mathbf{B}^{(1)}, \Delta\mathbf{B}; \tilde{\mathbf{M}}^{(1)}, \tilde{\mathbf{M}}^{(2)}).$$

Remark 3.1. The parameter set $\{\mathbf{F}, \mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \Delta\mathbf{B}\}$ is not uniquely identifiable. Specifically, for any invertible matrix $\mathbf{Q} \in \mathbb{R}^{r\times r}$, the transformed set $\{\mathbf{F}\mathbf{Q}, \mathbf{B}^{(1)}(\mathbf{Q}^{-1})^\top, \mathbf{B}^{(2)}(\mathbf{Q}^{-1})^\top, \Delta\mathbf{B}(\mathbf{Q}^{-1})^\top\}$ constitutes an equivalent solution. Imposing the constraint $n^{-1}\mathbf{F}^\top\mathbf{F} = \mathbf{I}_r$ can restrict $\mathbf{Q}$ to an orthogonal matrix. While the individual embeddings and loading matrices are not identifiable, their column spaces are estimable. Moreover, the row sparsity support of $\Delta\mathbf{B}$ is invariant under orthogonal rotation. Consequently, the identified key features remain unique and identifiable despite the rotational ambiguity.

To clarify the last statement, the ambiguity in Theorem 3.3 concerns the numerical values of $\Delta\mathbf{B}$, not its row support. For any orthogonal matrix $\mathbf{O}$ and any row vector $\Delta\boldsymbol{b}_j$, $\|\Delta\boldsymbol{b}_j\mathbf{O}\|_2 = \|\Delta\boldsymbol{b}_j\|_2$; hence $\Delta\boldsymbol{b}_j = \mathbf{0}$ if and only if $\Delta\boldsymbol{b}_j\mathbf{O} = \mathbf{0}$. The support $\mathcal{S}$ is therefore identifiable without choosing a particular coordinate system in the latent space. If a specific Stage-I factor model imposes stronger normalization and sign conventions, the remaining rotation may be further fixed (Bai & Li, 2012; Liu et al., 2023); SMART does not require this stronger identifiability of coefficient values because its inferential target is the support of the structural shift.

3.2.1 Estimation

Given the non-convex nature of the objective function, we employ an alternating minimization strategy to solve for the parameters. Following Chen et al. (2020), the estimation proceeds over a bounded region:

$$\mathcal{F}(C) = \left\{ \mathbf{F}, \mathbf{B}^{(k)} : \{\|\boldsymbol{f}_i\|_2, \|\boldsymbol{b}_j\|_2\} \leq C, n^{-1}\mathbf{F}^\top \mathbf{F} = \mathbf{I}_r \right\},$$

where C is a sufficiently large constant. The boundedness constraint ensures solution stability and facilitates theoretical analysis.

Step 0: Initialization. We initialize the parameters by performing SVD on the concatenated matrix $(\tilde{\mathbf{M}}^{(1)}, \tilde{\mathbf{M}}^{(2)}) \in \mathbb{R}^{n \times 2p}$. We set the initial representation $\mathbf{F}^{[0]}$ as the leading r left singular vectors scaled by $\sqrt{n}$, and initialize the structural shift matrix as $\Delta\mathbf{B}^{[0]} = \mathbf{0}$.

Step 1: Update $\mathbf{B}^{(1)}$. With $\mathbf{F}^{[t-1]}$ and $\Delta\mathbf{B}^{[t-1]}$ fixed, we minimize the Equation (1) with respect to $\mathbf{B}^{(1)}$. Omitting terms independent of $\mathbf{B}^{(1)}$, the problem becomes

$$\arg\min_{\mathbf{B}^{(1)}} \left\{ \left\| \tilde{\mathbf{M}}^{(1)} - \mathbf{F}^{[t-1]}(\mathbf{B}^{(1)})^\top \right\|_F^2 + \left\| \tilde{\mathbf{M}}^{(2)} - \mathbf{F}^{[t-1]}(\Delta\mathbf{B}^{[t-1]})^\top - \mathbf{F}^{[t-1]}(\mathbf{B}^{(1)})^\top \right\|_F^2 \right\},$$

yielding the closed form solution:

$$\breve{\mathbf{B}}^{(1)[t]} = \frac{1}{2}\left(\tilde{\mathbf{M}}^{(1)} + \tilde{\mathbf{M}}^{(2)}\right)^\top \mathbf{F}^{[t-1]} \left((\mathbf{F}^{[t-1]})^\top \mathbf{F}^{[t-1]}\right)^{-1} - \frac{1}{2}\Delta\mathbf{B}^{[t-1]}.$$

Step 2: Update $\Delta\mathbf{B}$. Given $\mathbf{F}^{[t-1]}$ and $\mathbf{B}^{(1)[t]}$, the optimization for $\Delta\mathbf{B}$ decomposes into p independent component-wise subproblems. Rewriting Equation (1), we have

$$\breve{\Delta\boldsymbol{b}}_j^{[t]} = \arg\min_{\Delta\boldsymbol{b}_j} \left(\sum_{i=1}^{n} \left[\left\{ \tilde{m}_{ij}^{(1)} - (\boldsymbol{b}_j^{(1)[t]})^\top \boldsymbol{f}_i^{[t-1]} \right\}^2 + \left\{ \tilde{m}_{ij}^{(2)} - (\boldsymbol{b}_j^{(1)[t]})^\top \boldsymbol{f}_i^{[t-1]} - \Delta\boldsymbol{b}_j^\top \boldsymbol{f}_i^{[t-1]} \right\}^2 \right] + \lambda \|\Delta\boldsymbol{b}_j\|_2 \right).$$

This formulation corresponds to a group Lasso problem; refer to Hastie et al. (2015) for details.

Step 3: Update $\mathbf{F}$. Fixing $\mathbf{B}^{(1)[t]}$ and $\Delta\mathbf{B}^{[t]}$, we solve for $\mathbf{F}$ by minimizing the low-rank reconstruction loss:

$$\breve{\mathbf{F}}^{[t]} = \arg\min_{\mathbf{F}} \left\{ \left\| \tilde{\mathbf{M}}^{(1)} - \mathbf{F}(\mathbf{B}^{(1)[t]})^\top \right\|_F^2 + \left\| \tilde{\mathbf{M}}^{(2)} - \mathbf{F}(\mathbf{B}^{(1)[t]} + \Delta\mathbf{B}^{[t]})^\top \right\|_F^2 \right\}.$$

The closed form update is given by

$$\breve{\mathbf{F}}^{[t]} = \left\{ \tilde{\mathbf{M}}^{(1)}\mathbf{B}^{(1)[t]} + \tilde{\mathbf{M}}^{(2)}\mathbf{B}^{(2)[t]} \right\} \times \left\{ (\mathbf{B}^{(1)[t]})^\top \mathbf{B}^{(1)[t]} + (\mathbf{B}^{(2)[t]})^\top \mathbf{B}^{(2)[t]} \right\}^{-1}.$$

Step 4: Orthogonalization and Projection. We first enforce the orthogonality condition via a normalization step. Let $\mathbf{S} = n^{-1}(\breve{\mathbf{F}}^{[t]})^\top \breve{\mathbf{F}}^{[t]}$ with eigendecomposition $\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$. Define $\mathbf{Q} = \mathbf{U}\boldsymbol{\Lambda}^{-\frac{1}{2}}$ and update the parameters as $\mathbf{F}^{[t]} = \breve{\mathbf{F}}^{[t]}\mathbf{Q}$, $\quad \mathbf{B}^{(1)[t]} = \breve{\mathbf{B}}^{(1)[t]}(\mathbf{Q}^{-1})^\top$, $\quad \Delta\mathbf{B}^{[t]} = \breve{\Delta\mathbf{B}}^{[t]}(\mathbf{Q}^{-1})^\top$. Subsequently, to ensure the solution resides within $\mathcal{F}(C)$, we project the row vectors of the updated matrices. Specifically, any row vector $\boldsymbol{v} \in \{\boldsymbol{f}_i^{[t]}, \boldsymbol{b}_j^{(1)[t]}, (\boldsymbol{b}_j^{(1)[t]} + \Delta\boldsymbol{b}_j^{[t]})\}$ is rescaled to $C\boldsymbol{v}/\|\boldsymbol{v}\|_2$ if $\|\boldsymbol{v}\|_2 > C$; otherwise, it remains unchanged.

Convergence. We iterate Steps 1–4 until convergence. The procedure is summarized in Appendix A.

3.3 Theoretical Property for SMART

For notational clarity, we use the superscript "$*$" to denote the true parameter values hereafter.

Condition (C1). $\|\boldsymbol{f}_i^*\|_2 \leq C$, $\left\|\boldsymbol{b}_j^{(k)*}\right\|_2 \leq C$, and $n^{-1}(\mathbf{F}^*)^\top \mathbf{F}^* = \mathbf{I}_r$.

Condition (C2). $\sigma_1\left(\mathbf{M}^{(1)*}\right) \asymp \sigma_{r^*}\left(\mathbf{M}^{(1)*}\right)$ and $\sigma_{r^*}\left(\mathbf{B}^{(k)*}\right) \asymp \sigma_{\mathbf{B}}$.

Condition (C1) requires the true parameters to be bounded, a standard assumption in theoretical analysis (e.g., Zhang et al., 2020; Chen & Li, 2022). Condition (C2) stipulates that the non-zero singular values of the true parameters share the same order of magnitude, consistent with Bai & Ng (2002) and Fan et al. (2022).

We first establish the estimation error bound for the parameter matrices in the following theorem.

Theorem 3.2. *Suppose that the estimation error in the first stage satisfies* $\left\|\tilde{\mathbf{M}}^{(k)} - \mathbf{M}^{(k)*}\right\|_F \lesssim \mathrm{err}$ *with probability tending to one for* $k \in \{1,2\}$*. Under Condition (C1) and assuming* $r \geq r^*$*, it holds that*

$$\left\|\mathbf{M}^{(k)} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^\top\right\|_F \lesssim \mathrm{err} + \sqrt{s\lambda}$$

for $k \in \{1,2\}$ *with probability tending to one.*

Theorem 3.2 quantifies the overall estimation error, which is dominated by two components: the error inherited from Stage I and the error induced by the regularization λ and sparsity level s.

Building on this result, we further characterize the estimation accuracy of the latent representation $\mathbf{F}$, the loading matrices $\mathbf{B}^{(k)}$, and the attribution matrix $\Delta\mathbf{B}$.

Theorem 3.3. *Suppose that the estimation error in the first stage satisfies* $\left\|\tilde{\mathbf{M}}^{(k)} - \mathbf{M}^{(k)*}\right\|_F \lesssim \mathrm{err}$ *with probability tending to one for* $k \in \{1,2\}$*. Under Conditions (C1) and (C2), and assuming* $r = r^*$*, there exists an orthogonal rotation matrix* $\mathbf{O}$ *such that the following convergence rates hold with probability tending to one:*

$$\frac{1}{\sqrt{n}}\left\|\hat{\mathbf{F}}\mathbf{O} - \mathbf{F}^*\right\|_F \lesssim \frac{\mathrm{err} + \sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}}, \quad \frac{1}{\sqrt{p}}\left\|\hat{\mathbf{B}}^{(k)}\mathbf{O} - \mathbf{B}^{(k)*}\right\|_F \lesssim \frac{\mathrm{err} + \sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}}, \quad \textit{for } k \in \{1,2\},$$

and

$$\frac{1}{\sqrt{s}}\left\|\hat{\Delta\mathbf{B}}\mathbf{O} - \Delta\mathbf{B}^*\right\|_F \lesssim \sqrt{\frac{p}{\sigma_{\mathbf{B}}^2}} \cdot \frac{\mathrm{err} + \sqrt{s\lambda}}{\sqrt{ns}}.$$

Theorem 3.3 indicates that a larger $\sigma_{\mathbf{B}}$ leads to improved estimation precision. While standard factor analysis typically assumes a strong signal regime with $\sigma_{\mathbf{B}} \asymp \sqrt{p}$ (Bai & Ng, 2002), our theory accommodates weaker signal regimes (Bai & Ng, 2023). Provided $\sqrt{n}\sigma_{\mathbf{B}} \gg \mathrm{err} + \sqrt{s\lambda}$, the consistency is guaranteed.

To demonstrate the versatility of our framework, we analyze data from generalized linear distributions, which encompass diverse distributions including Gaussian, Poisson, and Binary. Adopting the generative mechanism described in Example 2 and postulating a low-rank structure for the parameter matrix $\mathbf{M}$, we obtain the stage I estimate $\tilde{\mathbf{M}}$ via the maximum likelihood estimator proposed by Chen et al. (2020). The resulting theoretical guarantees are established in the following corollaries.

Corollary 3.4. *Under the conditions of Theorem 3.2, assume that the observations* x_{ij} *are independent given the common representation* $\boldsymbol{f}_i$*, and that the GFM estimator is utilized in the first stage. If* $s\lambda \ll n \vee p$*, then*

$$\frac{1}{\sqrt{np}}\left\|\mathbf{M}^{(k)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^\top\right\|_F \lesssim \frac{1}{\sqrt{n} \wedge \sqrt{p}}$$

with probability tending to one for $k \in \{1,2\}$*.*

Corollary 3.5. *Suppose that the conditions of Theorem 3.3 and Theorem 3.4 are satisfied. If we further assume* $\sigma_{\mathbf{B}} \asymp \sqrt{p}$*, then the following convergence rates hold with probability tending to one:*

$$\frac{1}{\sqrt{n}}\left\|\hat{\mathbf{F}}\mathbf{O} - \mathbf{F}^*\right\|_F \lesssim \frac{1}{\sqrt{n} \wedge \sqrt{p}}, \quad \frac{1}{\sqrt{p}}\left\|\hat{\mathbf{B}}^{(k)}\mathbf{O} - \mathbf{B}^{(k)*}\right\|_F \lesssim \frac{1}{\sqrt{n} \wedge \sqrt{p}}, \quad \textit{for } k \in \{1,2\},$$

and

$$\frac{1}{\sqrt{s}}\left\|\hat{\Delta\mathbf{B}}\mathbf{O} - \Delta\mathbf{B}^*\right\|_F \lesssim \frac{\sqrt{n} \vee \sqrt{p}}{\sqrt{ns}}.$$

3.4 Determination of Latent Dimensionality

While the preceding analysis presumes a given latent dimensionality r, in practice, it is typically unknown. Correctly specifying the latent dimensionality is critical (Bai & Ng, 2002; Chen & Li, 2022); underestimating r incurs signal loss, whereas overestimating it introduces extraneous noise. Following Chen & Li (2022) and Liu et al. (2023), we adopt an information criterion approach to determine r. Define

$$\mathrm{IC}(r) = \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{F}(\mathbf{B}^{(1)})^\top\right\|_F^2 + \left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{F}(\mathbf{B}^{(2)})^\top\right\|_F^2 + r \times \eta(n,p),$$

and estimate the dimensionality via

$$\hat{r} = \operatorname*{arg\,min}_{1\le r\le r_{\max}} \mathrm{IC}(r),$$

where $r_{\max}$ is a predetermined upper bound satisfying $r^* \le r_{\max} < n \wedge p$, and $\eta(n,p)$ is a penalty function. The following theorem establishes the selection consistency of our criterion.

Theorem 3.6. *Suppose that the estimation error in the first stage satisfies* $\left\|\tilde{\mathbf{M}}^{(k)} - \mathbf{M}^{(k)*}\right\|_F \lesssim \mathrm{err}$ *with probability tending to one for* $k \in \{1,2\}$*. Under Condition (C1), and provided that the penalty function satisfies* $(\mathrm{err})^2 + s\lambda \ll \eta(n,p) \ll n\sigma_{\mathbf{B}}^2$*, then* $\mathbb{P}(\hat{r} = r^*) \to 1$ *as* $n, p \to +\infty$*.*

We recommend the penalty functions $\eta_1(n,p) = 2(n+p+\hat{s})\log\{np/(n+p)\}$ and $\eta_2(n,p) = 2(n+p+\hat{s})\log(n\wedge p)$, where $\hat{s}$ is an estimate of s, owing to their empirical performance (see Section 6.1 for details), and establish their theoretical validity in the following corollary.

Corollary 3.7. *Under the conditions of Theorem 3.6, assume that the observations* x_{ij} *are independent given the common representation* $\boldsymbol{f}_i$*, and that the GFM estimator is utilized in the first stage. Furthermore, assume* $s\lambda \ll n \vee p$ *and* $\sigma_{\mathbf{B}} \asymp \sqrt{p}$*. If the penalty function is specified as either* $\eta_1(n,p)$ *or* $\eta_2(n,p)$*, then* $\mathbb{P}(\hat{r} = r^*) \to 1$ *as* $n, p \to +\infty$*.*

4 Extension: Multi-Environment Scenario

While the preceding analysis focused on pairwise environmental shifts, extending SMART to multiple environments facilitates comprehensive structural shift detection. This setting is natural for time series analysis, change point detection, and transfer learning involving several source domains.

Consider data matrices $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)}$ from K environments. Retaining the two-stage strategy, we perform the signal recovery stage followed by a joint optimization across all K environments. The estimators are obtained by minimizing the cumulative reconstruction error penalized by adjacent structural shifts:

$$\left(\hat{\mathbf{F}}, \{\hat{\mathbf{B}}^{(k)}\}_{k=1}^K\right) = \operatorname*{arg\,min}_{\mathbf{F},\mathbf{B}^{(k)}} \left\{ \sum_{k=1}^K \left\|\tilde{\mathbf{M}}^{(k)} - \mathbf{F}(\mathbf{B}^{(k)})^\top\right\|_F^2 + \lambda \sum_{k=2}^K \sum_{j=1}^p \left\|\boldsymbol{b}_j^{(k)} - \boldsymbol{b}_j^{(k-1)}\right\|_2 \right\}.$$

We suppose that the sample representations remain invariant across environments, attributing observed variations to shifts in feature influences. The optimization problem is solved via an alternating iterative algorithm analogous to the pairwise case.

We establish the theoretical guarantees for the multi-environment setting as follows.

Theorem 4.1 (Multi-environment error)**.** *Suppose that the estimation error in the first stage satisfies* $\left\|\tilde{\mathbf{M}}^{(k)} - \mathbf{M}^{(k)*}\right\|_F \lesssim \mathrm{err}$ *with probability tending to one for all* $k \in \{1,\ldots,K\}$*. Under Condition (C1) and assuming* $r \ge r^*$*, the average reconstruction error satisfies*

$$\frac{1}{K}\sum_{k=1}^K \left\|\mathbf{M}^{(k)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^\top\right\|_F \lesssim \mathrm{err} + \sqrt{s\lambda}.$$

Furthermore, under Condition (C2) and assuming $r = r^*$*, there exists an orthogonal rotation matrix* $\mathbf{O}$ *such that the following convergence rates hold with probability tending to one:*

$$\frac{1}{\sqrt{n}}\left\|\hat{\mathbf{F}}\mathbf{O} - \mathbf{F}^*\right\|_F \lesssim \frac{\mathrm{err} + \sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}}, \quad \frac{1}{K}\sum_{k=1}^K \left(\frac{1}{\sqrt{p}}\left\|\hat{\mathbf{B}}^{(k)}\mathbf{O} - \mathbf{B}^{(k)*}\right\|_F\right) \lesssim \frac{\mathrm{err} + \sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}},$$

and

$$\frac{1}{K-1}\sum_{k=2}^{K}\left(\frac{1}{\sqrt{s}}\left\|\Delta\hat{\mathbf{B}}^{(k)}\mathbf{O}-\Delta\mathbf{B}^{(k)*}\right\|_F\right)\lesssim\frac{\sqrt{p}}{\sigma_{\mathbf{B}}}\cdot\frac{\mathrm{err}+\sqrt{s\lambda}}{\sqrt{ns}},$$

where $\Delta\mathbf{B}^{(k)*}=\mathbf{B}^{(k)*}-\mathbf{B}^{(k-1)*}$.

The result mirrors the pairwise theory: the second stage inherits the preliminary recovery error while the fused group penalty localizes which rows change between adjacent environments. Hence the inferential target becomes the union, or sequence, of row supports across the $K-1$ adjacent shifts.

For determining the latent dimensionality, we generalize the information criterion approach. Defining the penalized objective for K environments, we estimate the dimensionality via:

$$\hat{r}=\underset{1\le r\le r_{\max}}{\arg\min}\left\{\sum_{k=1}^{K}\left\|\tilde{\mathbf{M}}^{(k)}-\hat{\mathbf{F}}_r(\hat{\mathbf{B}}_r^{(k)})^\top\right\|_F^2+Kr\eta(n,p)\right\},$$

where $\hat{\mathbf{F}}_r$ and $\hat{\mathbf{B}}_r^{(k)}$ denote the estimators obtained under rank r. The theoretical consistency of this selection criterion is established below.

Theorem 4.2. *Suppose that the conditions established in Theorem 3.6 are satisfied for all* $k\in\{1,\dots,K\}$. *Then* $\mathbb{P}(\hat{r}=r^*)\to 1$ *as* $n,p\to+\infty$.

5 Extensions: One-Step Estimation

For additive noise models (as described in Example 1), we can integrate the two-stage procedure into a one-step estimation algorithm. In this one-step formulation, we substitute the raw observations $\mathbf{X}^{(k)}$ for the stage-one estimates $\tilde{\mathbf{M}}^{(k)}$. Specifically, we directly estimate by

$$\left(\hat{\mathbf{F}},\hat{\mathbf{B}}^{(1)},\hat{\Delta\mathbf{B}}\right)=\underset{\mathbf{F},\mathbf{B}^{(1)},\Delta\mathbf{B}}{\arg\min}\left\{\left\|\mathbf{X}^{(1)}-\mathbf{F}(\mathbf{B}^{(1)})^\top\right\|_F^2+\left\|\mathbf{X}^{(2)}-\mathbf{F}(\mathbf{B}^{(1)}+\Delta\mathbf{B})^\top\right\|_F^2+\lambda\left\|\Delta\mathbf{B}\right\|_{2,1}\right\}.$$

Minimizing this objective yields estimates for the representation $\mathbf{F}$, the loadings $\mathbf{B}^{(k)}$, and the structural shift $\Delta\mathbf{B}$. We employ an analogous alternating algorithm to solve this problem. Let ε_{ij} denote the (i,j)-th entry of the noise matrix $\mathcal{E}$.

To establish the theoretical guarantees for this one-step estimator, we introduce an additional regularity condition.

Condition (C3). (C3.1) $\mathbb{E}\left|\varepsilon_{ij}\right|^{\xi}\le C_1$ for $\xi>14$.

(C3.2) $\frac{1}{n}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\left\{p^{-1}\sum_{j=1}^{p}\mathbb{E}\left(\varepsilon_{i_1j}\varepsilon_{i_2j}\right)\right\}^2\le C_2$.

(C3.3) $\mathbb{E}\left(p^{-\frac{1}{2}}\sum_{j=1}^{p}\left[\varepsilon_{i_1j}\varepsilon_{i_2j}-\mathbb{E}\left(\varepsilon_{i_1j}\varepsilon_{i_2j}\right)\right]\right)^2\le C_3$ for any (i_1,i_2).

C_1,C_2, and C_3 are constants. Condition (C3) relaxes the independence assumption on the error terms ε_{ij}, thereby accommodating dependence structures among the observations x_{ij}. This flexibility is particularly useful when analyzing time series data.

Then we establish the convergence rates of the estimation error as follows.

Theorem 5.1. *Under Conditions (C1) and (C3), and assume* $r\ge r^*$, *then*

$$\left\|\hat{\mathbf{M}}^{(k)}-\mathbf{M}^{(k)*}\right\|_F\lesssim n^{\frac{1}{2}}p^{\frac{1}{4}}\vee n^{\frac{1}{4}}p^{\frac{1}{2}}+\sqrt{s\lambda}$$

with probability tending to one for $k\in\{1,2\}$. *Further assume Condition (C2) and* $r=r^*$, *then there exists an orthogonal rotation matrix* $\mathbf{O}$ *such that the following convergence rates hold with probability tending to one:*

$$\frac{1}{\sqrt{n}}\left\|\hat{\mathbf{F}}\mathbf{O}-\mathbf{F}^*\right\|_F\lesssim\frac{n^{\frac{1}{2}}p^{\frac{1}{4}}\vee n^{\frac{1}{4}}p^{\frac{1}{2}}+\sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}},\quad\frac{1}{\sqrt{p}}\left\|\hat{\mathbf{B}}^{(k)}\mathbf{O}-\mathbf{B}^{(k)*}\right\|_F\lesssim\frac{n^{\frac{1}{2}}p^{\frac{1}{4}}\vee n^{\frac{1}{4}}p^{\frac{1}{2}}+\sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}},$$

and

$$\frac{1}{\sqrt{s}}\left\|\hat{\Delta\mathbf{B}}\mathbf{O}-\Delta\mathbf{B}^*\right\|_F \lesssim \sqrt{\frac{p}{\sigma_{\mathbf{B}}^2}}\cdot\frac{n^{\frac{1}{2}}p^{\frac{1}{4}}\vee n^{\frac{1}{4}}p^{\frac{1}{2}}+\sqrt{s\lambda}}{\sqrt{ns}}.$$

Regarding the determination of the latent dimensionality, the information criterion approach proposed in the preceding section remains applicable.

The one-step estimator and the two-stage version of SMART serve different regimes. The one-step objective is attractive for additive noise data, because it avoids a separate preliminary denoising step and is computationally faster in our experiments. The two-stage procedure is more broadly applicable: it naturally accommodates non-Gaussian and mixed data through an appropriate Stage I signal recovery method, and its modularity allows PCA, GFM, VAE, or domain specific denoisers to be substituted without modifying the attribution stage. Theorem 3.2 formalizes this modularity by showing that Stage II inherits the convergence rate of the chosen Stage I estimator. Additional comparisons are reported in Appendix C.10.

6 Experiments on Synthetic and Real Data

6.1 Experiments on Synthetic Data

6.1.1 Experiment Settings

In this section, we evaluate the finite-sample performance of the SMART method through 200 Monte Carlo simulations. The data generation process follows the generalized linear models described in Example 2.

Regarding parameter settings, the elements of $\mathbf{F}$ and $\mathbf{B}^{(1)}$ are independently sampled from $U[-\alpha,\alpha]$. The structural difference $\Delta\mathbf{B}$ follows a row-sparse assumption, where a support set $\mathcal{S}^*$ of size s is randomly selected, with non-zero elements sampled from $U[-c,c]$. The simulations cover three typical distributions: Gaussian, Poisson, and Binary.

The experimental grid includes sample sizes and dimensions $(n,p)\in\{50,100,300\}^2$, true ranks $r^*\in\{2,4\}$, sparsity levels $s\in\{5,10,20\}$, signal strength $c=2$, and distribution parameter $\alpha=2$. To ensure numerical stability, data values are truncated at 2000 for the Poisson distribution. We benchmark the SMART method against Naïve Lasso (N-Lasso) and unpenalized SVD estimator. For all penalized methods, the regularization parameter λ is determined via Cross-Validation (CV). Detailed algorithmic implementations for these benchmarks are deferred to Appendix A. Due to space constraints, we focus the discussion on the Poisson distribution as a representative case. Results for Gaussian and Binary distributions, which exhibit similar trends, are provided in Appendix C. We employ two metrics to assess the accuracy of variable selection and subspace recovery:

The comparison is designed as a controlled ablation. N-Lasso uses the same Stage I recovery, initialization, alternating updates, and CV procedure as SMART, but replaces the rowwise group $\ell_{2,1}$ penalty with an entrywise ℓ_1 penalty on $\Delta\mathbf{B}$. This isolates the effect of enforcing sparsity at the row level. The unpenalized SVD baseline removes the sparsity penalty altogether and serves as a low-rank reconstruction reference for subspace recovery.

F_1 Score: Evaluates the attribution results. Let $\hat{\mathcal{S}}=\{j:\|\hat{\Delta\boldsymbol{b}}_j\|_2>10^{-5}\}$; the score is defined as: $F_1=2|\mathcal{S}^*\cap\hat{\mathcal{S}}|/(|\mathcal{S}^*|+|\hat{\mathcal{S}}|)$. Minimum Canonical Correlation ($\rho_{\min}$): Measures the alignment between the true subspace $\mathbf{U}$ and its estimate $\hat{\mathbf{U}}$: $\rho_{\min}(\mathbf{U},\hat{\mathbf{U}})=\sigma_{\min}(\mathbf{U}^\top\hat{\mathbf{U}})$. This metric corresponds to the cosine of the largest principal angle, where values closer to 1 indicate superior recovery. In the main text, we focus on the subspace recovery of $\mathbf{B}^{(k)}$. Analysis regarding the common representation matrix $\mathbf{F}$, as well as additional metrics, is provided in Appendix C.

We also performed a rank selection experiment to validate the effectiveness of our IC. The true rank was fixed at $r^*=2$, with $s=10$, $c=2$, and $r_{\max}=5$. We use $\eta_1(n,p)$ and $\eta_2(n,p)$ as penalty terms. We further validated the SMART model and the one-step estimation method on incomplete data in Appendix C.

Table 1: Variable selection performance (F_1 score) for Poisson data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$				$s = 10$				$s = 20$			
			N-Lasso		SMART		N-Lasso		SMART		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.750	0.214	0.822	0.167	0.815	0.140	0.873	0.124	0.891	0.088	0.931	0.056
		100	0.762	0.184	0.823	0.154	0.776	0.130	0.839	0.113	0.856	0.089	0.909	0.056
		300	0.766	0.167	0.830	0.146	0.767	0.134	0.818	0.108	0.769	0.104	0.826	0.087
	100	50	0.710	0.223	0.803	0.179	0.834	0.145	0.882	0.118	0.895	0.075	0.931	0.052
		100	0.739	0.209	0.823	0.159	0.775	0.125	0.845	0.106	0.846	0.088	0.898	0.064
		300	0.774	0.167	0.830	0.146	0.776	0.124	0.836	0.101	0.773	0.085	0.833	0.073
	300	50	0.732	0.213	0.815	0.156	0.790	0.151	0.858	0.120	0.889	0.074	0.927	0.050
		100	0.720	0.201	0.816	0.151	0.755	0.147	0.824	0.128	0.844	0.080	0.897	0.065
		300	0.753	0.191	0.840	0.145	0.763	0.121	0.826	0.098	0.785	0.087	0.847	0.074
4	50	50	0.915	0.108	0.977	0.048	0.933	0.070	0.984	0.027	0.967	0.031	0.990	0.015
		100	0.894	0.123	0.970	0.061	0.916	0.073	0.974	0.038	0.947	0.040	0.987	0.019
		300	0.889	0.115	0.974	0.051	0.901	0.076	0.969	0.039	0.904	0.056	0.974	0.027
	100	50	0.883	0.133	0.971	0.060	0.936	0.069	0.988	0.026	0.973	0.029	0.991	0.015
		100	0.880	0.131	0.972	0.058	0.923	0.072	0.981	0.033	0.950	0.044	0.987	0.019
		300	0.910	0.105	0.976	0.051	0.900	0.082	0.973	0.042	0.916	0.050	0.977	0.025
	300	50	0.876	0.143	0.971	0.051	0.915	0.089	0.985	0.030	0.971	0.030	0.991	0.016
		100	0.876	0.126	0.963	0.067	0.921	0.078	0.982	0.034	0.951	0.041	0.992	0.015
		300	0.912	0.106	0.978	0.049	0.909	0.070	0.973	0.036	0.917	0.049	0.978	0.024

Table 2: Subspace recovery accuracy ($\rho_{\min}$) of $\mathbf{B}^{(1)}$ for Poisson data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.995	0.002	0.987	0.004	0.990	0.004	0.996	0.002	0.982	0.006	0.987	0.005	0.996	0.002	0.974	0.009	0.983	0.006
		100	0.994	0.003	0.991	0.003	0.992	0.003	0.994	0.003	0.987	0.004	0.989	0.003	0.994	0.003	0.982	0.005	0.987	0.004
		300	0.993	0.004	0.993	0.004	0.993	0.004	0.993	0.004	0.991	0.004	0.992	0.004	0.993	0.004	0.988	0.004	0.990	0.004
	100	50	0.997	0.001	0.988	0.004	0.990	0.003	0.997	0.001	0.983	0.005	0.988	0.003	0.997	0.001	0.976	0.007	0.986	0.004
		100	0.998	0.001	0.993	0.002	0.994	0.001	0.998	0.001	0.990	0.002	0.992	0.002	0.998	0.001	0.985	0.003	0.989	0.003
		300	0.997	0.002	0.996	0.002	0.996	0.002	0.997	0.002	0.994	0.002	0.995	0.002	0.996	0.003	0.991	0.002	0.993	0.002
	300	50	0.995	0.005	0.986	0.006	0.988	0.005	0.995	0.005	0.981	0.006	0.986	0.005	0.995	0.004	0.975	0.008	0.984	0.005
		100	0.998	0.002	0.992	0.003	0.993	0.003	0.998	0.002	0.989	0.003	0.991	0.003	0.998	0.002	0.984	0.004	0.989	0.003
		300	0.999	0.000	0.997	0.001	0.998	0.001	0.999	0.000	0.996	0.001	0.996	0.001	0.999	0.000	0.993	0.001	0.994	0.001
4	50	50	0.990	0.004	0.979	0.006	0.987	0.005	0.990	0.005	0.970	0.009	0.984	0.006	0.990	0.004	0.954	0.013	0.980	0.007
		100	0.991	0.004	0.986	0.004	0.989	0.004	0.991	0.004	0.981	0.005	0.988	0.004	0.991	0.004	0.973	0.006	0.985	0.004
		300	0.986	0.007	0.985	0.007	0.986	0.007	0.985	0.007	0.983	0.007	0.984	0.007	0.985	0.007	0.979	0.007	0.983	0.007
	100	50	0.994	0.003	0.982	0.006	0.990	0.003	0.994	0.003	0.974	0.008	0.988	0.004	0.994	0.003	0.963	0.010	0.986	0.004
		100	0.995	0.002	0.988	0.003	0.992	0.002	0.994	0.003	0.985	0.003	0.991	0.003	0.994	0.003	0.976	0.006	0.988	0.003
		300	0.993	0.003	0.991	0.003	0.992	0.003	0.992	0.003	0.989	0.003	0.991	0.003	0.992	0.003	0.987	0.003	0.990	0.003
	300	50	0.994	0.003	0.982	0.006	0.990	0.003	0.993	0.003	0.975	0.007	0.988	0.003	0.994	0.003	0.963	0.010	0.986	0.004
		100	0.996	0.001	0.990	0.002	0.994	0.002	0.996	0.001	0.986	0.003	0.993	0.002	0.997	0.001	0.980	0.003	0.991	0.002
		300	0.997	0.001	0.995	0.001	0.996	0.001	0.997	0.001	0.993	0.001	0.995	0.001	0.997	0.001	0.990	0.001	0.994	0.001

6.1.2 Experiments Results

For variable selection of the structural difference $\mathbf{\Delta B}$, we focus our comparison on SMART and the Naïve Lasso, as the unpenalized SVD method is incapable of variable selection. The results for Poisson data are summarized in Table 1. Compared to Naïve Lasso, SMART achieves superior performance with high stability regarding sample size n. It remains robust against high dimensionality p, and notably, its selection capability strengthens as the true rank r^* and s increase, benefiting from richer signal information.

As shown in Table 2 and 3, SMART accurately recovers the latent structure, surpassing Naïve Lasso and aligning closely with unpenalized SVD. This confirms that introducing sparsity does not hinder the estimation of the latent representation space. The method exhibits consistent asymptotic properties, with accuracy improving as n and p increase, thereby successfully balancing structural selection and estimation fidelity.

We examine the reliability of the proposed rank selection criteria, with results presented in Table 4. This experiment evaluates the frequency with which the IC (IC_1 and IC_2 corresponding to penalty terms η_1 and η_2) correctly identify the true rank. For both Gaussian and Poisson data, IC_1 and IC_2 achieve almost 100%

Table 3: Subspace recovery accuracy ($\rho_{\min}$) of $\mathbf{B}^{(2)}$ for Poisson data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.994	0.004	0.987	0.005	0.989	0.005	0.993	0.004	0.982	0.006	0.986	0.005	0.992	0.006	0.978	0.008	0.985	0.005
		100	0.994	0.004	0.990	0.004	0.992	0.003	0.993	0.003	0.987	0.004	0.989	0.004	0.992	0.005	0.982	0.005	0.986	0.005
		300	0.993	0.004	0.993	0.004	0.993	0.004	0.993	0.004	0.991	0.004	0.992	0.004	0.993	0.004	0.988	0.004	0.990	0.004
	100	50	0.991	0.007	0.985	0.006	0.987	0.006	0.990	0.008	0.981	0.006	0.986	0.006	0.990	0.006	0.977	0.008	0.984	0.006
		100	0.996	0.003	0.992	0.003	0.993	0.003	0.995	0.003	0.989	0.003	0.991	0.003	0.994	0.004	0.985	0.004	0.988	0.003
		300	0.996	0.003	0.995	0.002	0.996	0.002	0.996	0.003	0.994	0.002	0.994	0.002	0.996	0.003	0.991	0.003	0.992	0.002
	300	50	0.986	0.010	0.982	0.008	0.985	0.008	0.984	0.010	0.977	0.009	0.981	0.008	0.984	0.010	0.974	0.009	0.980	0.008
		100	0.991	0.008	0.990	0.005	0.991	0.005	0.990	0.007	0.986	0.005	0.988	0.004	0.989	0.008	0.982	0.006	0.986	0.005
		300	0.998	0.002	0.997	0.001	0.997	0.001	0.997	0.002	0.995	0.001	0.996	0.001	0.996	0.003	0.992	0.002	0.993	0.002
4	50	50	0.986	0.008	0.977	0.008	0.984	0.007	0.984	0.008	0.969	0.008	0.981	0.006	0.977	0.012	0.958	0.012	0.975	0.010
		100	0.988	0.005	0.983	0.005	0.987	0.005	0.986	0.006	0.979	0.005	0.985	0.005	0.981	0.008	0.970	0.007	0.979	0.007
		300	0.985	0.007	0.984	0.006	0.985	0.006	0.983	0.007	0.982	0.007	0.983	0.007	0.981	0.008	0.978	0.007	0.981	0.007
	100	50	0.991	0.005	0.980	0.007	0.987	0.006	0.986	0.007	0.971	0.008	0.983	0.006	0.982	0.008	0.963	0.010	0.979	0.007
		100	0.992	0.004	0.986	0.004	0.990	0.003	0.990	0.005	0.982	0.004	0.988	0.004	0.987	0.007	0.975	0.005	0.984	0.005
		300	0.992	0.003	0.990	0.003	0.991	0.003	0.991	0.003	0.988	0.003	0.990	0.003	0.989	0.004	0.984	0.003	0.988	0.003
	300	50	0.985	0.009	0.974	0.010	0.982	0.009	0.982	0.008	0.968	0.009	0.979	0.008	0.981	0.007	0.961	0.010	0.978	0.007
		100	0.992	0.005	0.986	0.005	0.990	0.004	0.990	0.004	0.982	0.004	0.988	0.004	0.988	0.005	0.975	0.005	0.985	0.004
		300	0.996	0.001	0.994	0.002	0.995	0.001	0.995	0.002	0.992	0.002	0.994	0.002	0.994	0.002	0.988	0.002	0.992	0.002

Table 4: Frequency of correctly identifying the true rank with IC_1 and IC_2, with varying n and p for Gaussian, Poisson, and binary data over 200 repetitions.

		Binary		Gaussian		Poisson	
n	p	IC_1	IC_2	IC_1	IC_2	IC_1	IC_2
50	50	99.5%	100.0%	100.0%	100.0%	100.0%	100.0%
	100	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	300	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
100	50	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	100	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	300	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
300	50	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	100	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	300	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

success rate in recovering the true rank, regardless of variations in sample size n or dimension p. Even under the more challenging Binary distribution, the criteria maintain high precision.

We further validate two aspects raised by the reviewer feedback. First, Appendix C.6 provides log-log plots showing that the empirical estimation errors track the theoretical scaling with n and p. Second, Appendix C.7 demonstrates a setting in which direct tests on observed coordinates fail because the observed marginal distribution does not change, whereas SMART still recovers the changed mechanism support.

We added three sets of experiments to broaden the empirical evidence in the Appendix. First, for the multi-environment extension in Section 4, we varied the number of environments K from 2 to 5 and compared one-step and two-stage fitting. The union F_1 scores remain high across all values of K, and the fitting time grows smoothly with K; the results are reported in Appendix C.8. Second, we evaluated mixed data by combining Gaussian and Poisson coordinates. SMART remains competitive or superior to N-Lasso in F_1 score, supporting the modular Stage I design for heterogeneous data; see Appendix C.9. Third, we compared one-step and two-stage estimation under Gaussian data with varying noise levels. The one-step estimator is faster and can be slightly more accurate at low noise, while the two-stage estimator covers non-Gaussian and mixed data; timing and accuracy results are given in Appendix C.10.

6.2 Experiments on Real Data

In this section, we demonstrate the practical utility of the SMART framework. We primarily focus on identifying structural shifts in high-dimensional gene expression data to discover potential biomarkers for breast cancer. As an additional real data analysis, we also apply SMART to player performance metrics from the NBA 2023-24 season; the full NBA analysis is reported in Appendix D.3.

Breast cancer is a highly heterogeneous disease driven by complex molecular alterations. Distinguishing between "passenger" mutations and true "driver" structural shifts remains a critical challenge in oncology. By applying SMART to paired Tumor and Para-cancerous tissue samples, we aim to isolate significant mechanistic deviations ($\Delta\mathbf{B}$) that underpin tumorigenesis, thereby identifying robust candidate biomarkers for diagnosis and prognosis.

We utilize paired gene expression data from the TCGA-BRCA dataset. To mitigate sequencing noise, we filter genes by mean expression (10–5000), then compute the coefficient of variation (σ/μ) in tumor samples and select the top 300 genes as model input. The model was configured with $r = 6$ and $\lambda = 3$ based on the selection heuristics described in Appendix D.

The model identifies 13 significant genes in $\Delta\mathbf{B}$ (Figure 2). We highlight key findings here. Firstly, SMART captures a cohesive cluster of genes, including `CSN1S1`, `CSN3`, and `LALBA`, that are associated with aberrant alveogenesis. Langille et al. (2022) posits that the expression of these markers reflects lineage infidelity, where cells (potentially including basal cells) inappropriately activate the lactation program, a process linked to tumorigenesis. Additionally, the model identifies `CA6`, which is consistent with findings from the Oncomine database showing its significant differential expression between breast cancer and normal tissues Wu et al. (2020). Additional analysis can be found in Appendix D.2. These results suggest that SMART effectively pinpoints markers of specific pathogenic developmental programs.

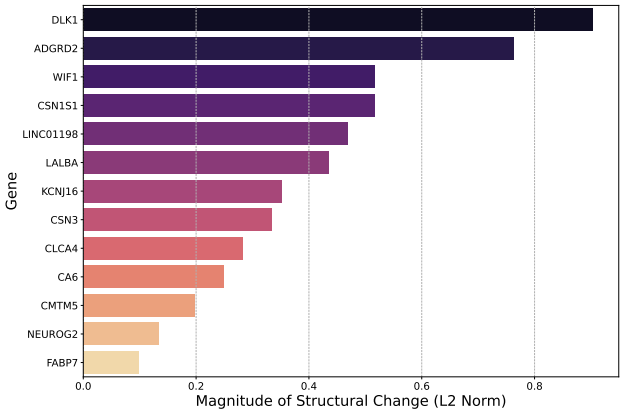


Figure 2: Significant nonzero difference loadings identified in the TCGA dataset and their magnitudes.

7 Conclusion

In this paper, we introduced SMART, a unified framework that bridges the gap between signal recovery and structural attribution. By integrating low-rank reconstruction error minimization with a group-structured sparse penalty, SMART effectively achieves the precise localization of mechanism shifts. This formulation ensures that the identified sparse structural drifts are invariant to rotational ambiguity, thereby explicitly disentangling invariant components from specific mechanism changes and providing interpretable insights into dynamic systems.

Theoretically, we established that the estimators via SMART inherit the convergence rates of the preliminary signal recovery and provided a consistent information criterion for latent dimensionality selection. The framework's versatility is further demonstrated through its extensions to a unified one-step estimator for additive noise models and multi-environment scenarios. Empirically, the efficacy of SMART is validated on both count data (BRCA Gene Expression) and continuous data (NBA), showcasing its broad applicability across diverse scientific domains.

Several avenues remain for future exploration. First, the current second-stage attribution assumes a linear interaction between latent representations and features in the attribution stage; extending this framework to accommodate non-linear structural disentanglement represents a valuable direction. Second, the performance of the attribution stage relies on the quality of the initial signal recovery. Future research could investigate more robust signal recovery strategies to mitigate potential error propagation.

References

L Ai, Q Tao, S Zhong, et al. Inactivation of Wnt inhibitory factor-1 (WIF1) expression by epigenetic silencing is a common event in breast cancer. *Carcinogenesis*, 27(7):1341–1348, 2006.

Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Jushan Bai and Kunpeng Li. Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436–465, 2012.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Jushan Bai and Serena Ng. Approximate factor models with weaker loadings. *Journal of Econometrics*, 235 (2):1893–1916, 2023.

Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5): 2964–2984, 2021.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Stevenson Bolivar, Shuo-Chieh Huang, and Rong Chen. Analysis of tensor time series. *Annual Review of Statistics and Its Application*, 13, 2025.

Z Budkova, A K Sigurdardottir, E Briem, et al. Expression of ncRNAs on the DLK1-DIO3 locus is associated with basal and mesenchymal phenotype in breast epithelial progenitor cells. *Frontiers in Cell and Developmental Biology*, 8:461, 2020.

Yunxiao Chen and Xiaoou Li. Determining the number of factors in high-dimensional generalized latent factor models. *Biometrika*, 109(3):769–782, 2022.

Yunxiao Chen, Xiaoou Li, and Siliang Zhang. Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, 115(532): 1756–1770, 2020.

A Cordero, D Kanojia, J Miska, et al. FABP7 is a key metabolic regulator in HER2+ breast cancer brain metastasis. *Oncogene*, 38(37):6445–6460, 2019.

Chengyu Cui and Gongjun Xu. Identifiability and inference for generalized latent factor models. *arXiv preprint arXiv:2508.05866*, 2025.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(4): 603–680, 2013.

Jianqing Fan, Jianhua Guo, and Shurong Zheng. Estimating number of factors by adjusted eigenvalues thresholding. *Journal of the American Statistical Association*, 117(538):852–861, 2022.

M. A. Fonseca-Montaño, K. I. Vazquez-Santillan, and A. Hidalgo-Miranda. The current advances of lncRNAs in breast cancer immunobiology research. *Frontiers in Immunology*, 14:1194300, 2023.

Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *Foundations and Trends in Machine Learning*, 15(1-2):1–175, 2022.

Lovedeep Gondara and Ke Wang. Mida: Multiple imputation using denoising autoencoders. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 260–272. Springer, 2018.

Yuefeng Han, Dan Yang, Cun-Hui Zhang, and Rong Chen. CP factor model for dynamic tensors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1383–1413, 2024.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL: CRC Press, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

M Laisné, S Benlamara, A Nicolas, et al. Cancer/Testis genes are predictive of breast tumor subtypes. *bioRxiv preprint bioRxiv:2021.10.27.465656*, 2021.

E Langille, K N Al-Zahrani, Z Ma, et al. Loss of epigenetic regulation disrupts lineage integrity, induces aberrant alveogenesis, and promotes breast cancer. *Cancer Discovery*, 12(12):2930–2953, 2022.

Arnar Laxdal, Andreas Ivarsson, Olafur Sigurgeirsson, and Tommy Haugen. Are the playoffs different from the regular season? a comparison of in-game statistics in icelandic elite handball. *International Journal of Performance Analysis in Sport*, 22(5):649–655, 2022.

Wei Liu, Huazhen Lin, Shurong Zheng, and Jin Liu. Generalized factor model for ultra-high dimensional correlated variables with mixed types. *Journal of the American Statistical Association*, 118(542):1385–1401, 2023.

Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. *Advances in Neural Information Processing Systems*, 20:1257–1264, 2007.

Joseph Price, Brian P Soebbing, David J Berri, and Brad R Humphreys. Tournament incentives, league policy, and nba team performance revisited. *Journal of Sports Economics*, 11(2):117–135, 2010.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697, 2023.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33:7852–7862, 2020.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

Fa Wang. Maximum likelihood estimation and inference for high dimensional generalized factor models with application to factor-augmented regressions. *Journal of Econometrics*, 229(1):180–200, 2022.

Z H Wu, Y Tang, and Y Zhou. A metabolic gene signature to predict overall survival in head and neck squamous cell carcinoma. *Mediators of Inflammation*, 2020(1):6716908, 2020.

Y Yu, V Walia, and R C Elble. Loss of CLCA4 promotes epithelial-to-mesenchymal transition in breast cancer cells. *PLoS One*, 8(12):e83943, 2013.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Xuefei Zhang, Songkai Xue, and Ji Zhu. A flexible latent space model for multilayer networks. In *International Conference on Machine Learning*, pp. 11288–11297. PMLR, 2020.

J Zhou, J Lei, J Wang, et al. Bioinformatics-based discovery of CKLF-like MARVEL transmembrane member 5 as a novel biomarker for breast cancer. *Frontiers in Cell and Developmental Biology*, 7:361, 2020.

Z Zhu, Z Lei, J Qian, et al. The ion channel-related gene signatures correlated with diagnosis, prognosis, and individualized treatment in patients with clear cell renal cell carcinoma. *Frontiers in Pharmacology*, 13: 889142, 2022.

Zhuangdi Zhu, Kaixiang Lin, Anil K Jain, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13344–13362, 2023.

A Algorithm and Implementation Details

A.1 Algorithms

In this section, we provide pseudocode for four algorithms: our proposed method, SMART; two compared methods, sparse mechanism attribution with Naïve Lasso and Unpenalized SVD Projection; and the Constrained Joint Maximum Likelihood Estimation (CJMLE) to solve model for generalized linear data (Example 2) (Chen et al., 2020; Chen & Li, 2022), which serves as the shared algorithm for the Stage I signal recovery.

Algorithm 1 CJMLE: constrained joint maximum likelihood estimation algorithm

Input: Observed data $\mathbf{X} \circ \mathbf{\Omega} = (x_{ij}\omega_{ij})$, latent dimensionality r, constraint constant C'.
  Initialize $\boldsymbol{f}_i^{[0]}$ and $\boldsymbol{b}_j^{[0]}$ for all i, j.
Alternating Minimization:
for $l = 1, 2, \ldots$ **do**
  Update Row Parameters ($\boldsymbol{f}$):
  for $i = 1, \ldots, n$ **do**
    Compute $\boldsymbol{f}_i^{[l]}$ by minimizing the negative log-likelihood:

$$\boldsymbol{f}_i^{[l]} \in \underset{\|\boldsymbol{f}\| \leq C'}{\arg\min} - \sum_{j=1}^{p} \omega_{ij} \left[x_{ij}(\boldsymbol{f}^\top \boldsymbol{b}_j^{(l-1)}) - \psi(\boldsymbol{f}^\top \boldsymbol{b}_j^{(l-1)}) \right].$$

  end for
  Update Column Parameters ($\boldsymbol{b}$):
  for $j = 1, \ldots, p$ **do**
    Compute $\boldsymbol{b}_j^{[l]}$ by minimizing the negative log-likelihood:

$$\boldsymbol{b}_j^{[l]} \in \underset{\|\boldsymbol{b}\| \leq C'}{\arg\min} - \sum_{i=1}^{n} \omega_{ij} \left[x_{ij}((\boldsymbol{f}_i^{[l]})^\top \boldsymbol{b}) - \psi((\boldsymbol{f}_i^{[l]})^\top \boldsymbol{b}) \right].$$

  end for
  if converged **then**
    break
  end if
end for
Iteratively perform Alternating Minimization until convergence.
Return: $\hat{\boldsymbol{f}}_i = \boldsymbol{f}_i^{[L]}$ and $\hat{\boldsymbol{b}}_j = \boldsymbol{b}_j^{[L]}$ where L is the last iteration.

A.2 Implementation Details

To determine the optimal regularization parameter λ, we implement an entry-wise 3-fold CV strategy specifically designed for matrix-structured data. This process integrates both Stage I and Stage II to ensure that the latent signal recovery does not leak information about the validation set.

Before initiating the pipeline, the indices of the observed data matrices are randomly partitioned into three disjoint sets. In each fold, one set is designated as the validation mask $\mathbf{\Omega}_{\text{val}}$, effectively treating these entries as missing values during the training phase. The remaining entries $\mathbf{\Omega}_{\text{train}}$ are fed into the CJMLE framework (Algorithm 1) to recover the training signals $\tilde{\mathbf{M}}_{\text{train}}$. Stage II is then executed for each value in a candidate grid of λ using these training signals. The candidate grids are defined as follows:

- **Gaussian:** $\{0.60, 0.61, \ldots, 0.70\}$
- **Poisson:** $\{0.80, 0.81, \ldots, 0.90\}$

Algorithm 2 SMART: sparse mechanism attribution for representation algorithm

1: **Input:** Observed data $\mathbf{X}^{(1)} \circ \mathbf{\Omega}^{(1)} = (x_{ij}^{(1)}\omega_{ij}^{(1)}), \mathbf{X}^{(2)} \circ \mathbf{\Omega}^{(2)} = (x_{ij}^{(2)}\omega_{ij}^{(2)})$, regularization parameter λ, latent dimensionality r, constraint constant C.
Stage I: Signal Recovery
$\tilde{\mathbf{M}}^{(1)}, \tilde{\mathbf{M}}^{(2)} \leftarrow \mathbf{X}^{(1)} \circ \mathbf{\Omega}^{(1)}, \mathbf{X}^{(2)} \circ \mathbf{\Omega}^{(2)}$ via CJMLE (Algorithm 1) or other methods.
Stage II: Sparse Mechanism Attribution
2: **Initialization:**
3: Compute SVD: $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^\top = \mathrm{SVD}(\tilde{\mathbf{M}}^{(1)}, \tilde{\mathbf{M}}^{(2)})$.
4: Initialize $\mathbf{F}$ and $\Delta\mathbf{B}$: $\mathbf{F}^{[0]} \leftarrow \sqrt{n}\mathbf{U}_{:,1:r}, \quad \Delta\mathbf{B}^{[0]} \leftarrow \mathbf{0}$.
5: **Iterative Solver**
6: **for** $t = 1, 2, \ldots$ **do**
7: **Update $\mathbf{B}^{(1)}$:**

$$\mathbf{B}^{(1)[t]} = \frac{1}{2}\left(\tilde{\mathbf{M}}^{(1)} + \tilde{\mathbf{M}}^{(2)}\right)^\top \mathbf{F}^{[t-1]}\left((\mathbf{F}^{[t-1]})^\top \mathbf{F}^{[t-1]}\right)^{-1} - \frac{1}{2}\Delta\mathbf{B}^{[t-1]}.$$

8: **Update $\Delta\mathbf{B}$** via group Lasso:

$$\Delta\boldsymbol{b}_j^{[t]} = \arg\min_{\Delta\boldsymbol{b}_j} \sum_{i=1}^{n}\left[\left\{\tilde{m}_{ij}^{(1)} - (\boldsymbol{b}_j^{(1)[t]})^\top \boldsymbol{f}_i^{[t-1]}\right\}^2 + \left\{\tilde{m}_{ij}^{(2)} - (\boldsymbol{b}_j^{(1)[t]})^\top \boldsymbol{f}_i^{[t-1]} - \Delta\boldsymbol{b}_j^\top \boldsymbol{f}_i^{[t-1]}\right\}^2\right] + \lambda\|\Delta\boldsymbol{b}_j\|_2.$$

9: **Update $\mathbf{F}$:**

$$\mathbf{F}^{[t]} = \arg\min_{\mathbf{F}} \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{F}(\mathbf{B}^{(1)[t]})^\top\right\|_F^2 + \left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{F}(\mathbf{B}^{(1)[t]} + \Delta\mathbf{B}^{[t]})^\top\right\|_F^2.$$

10: **Orthogonalization and Projection:** Ensure that the estimator lies within the feasible set $\mathcal{F}(C)$.
11: **end for**
12: **Return:** $\hat{\mathbf{F}} = \mathbf{F}^{[t]}, \hat{\mathbf{B}}^{(1)} = \mathbf{B}^{(1)[t]}, \hat{\Delta\mathbf{B}} = \Delta\mathbf{B}^{[t]}$.

- **Binary:** $\{\exp(0), \exp(0.2), \ldots, \exp(2.0)\}$

The optimal parameter is chosen by minimizing the prediction error on the masked validation entries. Once the optimal λ^* is selected, the final model is retrained on the full dataset. The complete selection process is detailed in Algorithm 5.

B Proof of Theoretical Guarantee

Proof of Theorem 3.2: By definition of the Maximum Likelihood Estimator (MLE), we obtain the following inequality

$$\begin{aligned}
&\left\|\tilde{\mathbf{M}}^{(1)} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F^2 + \left\|\tilde{\mathbf{M}}^{(2)} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)} + \hat{\Delta\mathbf{B}}\right)^\top\right\|_F^2 + \lambda\sum_{j=1}^{p}\left\|\hat{\Delta\boldsymbol{b}}_j\right\|_2 \\
\leq &\left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{F}^*\left(\mathbf{B}^{(1)*}\right)^\top\right\|_F^2 + \left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{F}^*\left(\mathbf{B}^{(1)*} + \Delta\mathbf{B}^*\right)^\top\right\|_F^2 + \lambda\sum_{j=1}^{p}\left\|\Delta\boldsymbol{b}_j^*\right\|_2.
\end{aligned}$$

Algorithm 3 Structural disentanglement with the Naïve Lasso algorithm

Input: Observed data $\mathbf{X}^{(1)} \circ \mathbf{\Omega}^{(1)} = (x_{ij}^{(1)}\omega_{ij}^{(1)}), \mathbf{X}^{(2)} \circ \mathbf{\Omega}^{(2)} = (x_{ij}^{(2)}\omega_{ij}^{(2)})$, regularization parameter λ, latent dimensionality r, constraint constant C.

Stage I: Signal Recovery

$\tilde{\mathbf{M}}^{(1)}, \tilde{\mathbf{M}}^{(2)} \leftarrow \mathbf{X}^{(1)} \circ \mathbf{\Omega}^{(1)}, \mathbf{X}^{(2)} \circ \mathbf{\Omega}^{(2)}$ via CJMLE (Algorithm 1) or other methods.

Stage II: Sparse Estimation via Lasso

Initialization:

Compute SVD: $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^\top = \mathrm{SVD}(\tilde{\mathbf{M}}^{(1)}, \tilde{\mathbf{M}}^{(2)})$

Initialize $\mathbf{F}$ and $\Delta\mathbf{B}$: $\mathbf{F}^{[0]} \leftarrow \sqrt{n}\mathbf{U}_{:,1:r}, \quad \Delta\mathbf{B}^{[0]} \leftarrow \mathbf{0}$

Iterative Solver

for $t = 1, 2, \ldots$ **do**

Update $\mathbf{B}^{(1)}$ via Lasso:

$$\mathbf{B}^{(1)[t]} = \frac{1}{2}\left(\tilde{\mathbf{M}}^{(1)} + \tilde{\mathbf{M}}^{(2)}\right)^\top \mathbf{F}^{[t-1]}\left((\mathbf{F}^{[t-1]})^\top \mathbf{F}^{[t-1]}\right)^{-1} - \frac{1}{2}\Delta\mathbf{B}^{[t-1]}.$$

Update $\Delta\mathbf{B}$**:**

$$\Delta\boldsymbol{b}_j^{[t]} = \arg\min_{\Delta\boldsymbol{b}_j} \sum_{i=1}^{n}\left[\left\{\tilde{m}_{ij}^{(1)} - (\boldsymbol{b}_j^{(1)[t]})^\top \boldsymbol{f}_i^{[t-1]}\right\}^2 + \left\{\tilde{m}_{ij}^{(2)} - (\boldsymbol{b}_j^{(1)[t]})^\top \boldsymbol{f}_i^{[t-1]} - \Delta\boldsymbol{b}_j^\top \boldsymbol{f}_i^{[t-1]}\right\}^2\right] + \lambda\|\Delta\boldsymbol{b}_j\|_1.$$

Update $\mathbf{F}$**:**

$$\mathbf{F}^{[t]} = \arg\min_{\mathbf{F}} \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{F}(\mathbf{B}^{(1)[t]})^\top\right\|_F^2 + \left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{F}(\mathbf{B}^{(1)[t]} + \Delta\mathbf{B}^{[t]})^\top\right\|_F^2.$$

Orthogonalization and Projection: Ensure that the estimator lies within the feasible set $\mathcal{F}(C)$.

end for

Return: $\hat{\mathbf{F}} = \mathbf{F}^{[t]}, \hat{\mathbf{B}}^{(1)} = \mathbf{B}^{(1)[t]}, \hat{\Delta\mathbf{B}} = \Delta\mathbf{B}^{[t]}$.

Algorithm 4 Unpenalized SVD algorithm

Input: Observed data $\mathbf{X}^{(1)} \circ \mathbf{\Omega}^{(1)} = (x_{ij}^{(1)}\omega_{ij}^{(1)}), \mathbf{X}^{(2)} \circ \mathbf{\Omega}^{(2)} = (x_{ij}^{(2)}\omega_{ij}^{(2)})$, latent dimensionality r.

Signal Recovery: $\tilde{\mathbf{M}}^{(1)}, \tilde{\mathbf{M}}^{(2)} \leftarrow \mathbf{X}^{(1)} \circ \mathbf{\Omega}^{(1)}, \mathbf{X}^{(2)} \circ \mathbf{\Omega}^{(2)}$ via CJMLE (Algorithm 1) or other methods.

SVD and Factor Extraction:

Compute SVD: $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^\top = \mathrm{SVD}(\tilde{\mathbf{M}}^{(1)}, \tilde{\mathbf{M}}^{(2)})$

Extract Common Representations: $\hat{\mathbf{F}} \leftarrow \sqrt{n}\mathbf{U}_{:,1:r}$.

Projection:

Compute Pseudo-inverse: $\mathbf{F}^\dagger \leftarrow (\hat{\mathbf{F}}^\top\hat{\mathbf{F}})^{-1}\hat{\mathbf{F}}^\top$.

Estimate Loadings for each environment:

$$\hat{\mathbf{B}}^{(1)} = (\mathbf{F}^\dagger\tilde{\mathbf{M}}^{(1)})^\top, \quad \hat{\mathbf{B}}^{(2)} = (\mathbf{F}^\dagger\tilde{\mathbf{M}}^{(2)})^\top.$$

Compute Structural Difference:

$$\hat{\Delta\mathbf{B}} = \hat{\mathbf{B}}^{(2)} - \hat{\mathbf{B}}^{(1)}.$$

Return: $\hat{\mathbf{F}}, \hat{\mathbf{B}}^{(1)}, \hat{\Delta\mathbf{B}}$.

Rearranging the terms yields

$$\begin{aligned}
0 \geq & \left\|\tilde{\mathbf{M}}^{(1)} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F^2 - \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{F}^*\left(\mathbf{B}^{(1)*}\right)^\top\right\|_F^2 \\
& + \left\|\tilde{\mathbf{M}}^{(2)} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)} + \hat{\Delta\mathbf{B}}\right)^\top\right\|_F^2 - \left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{F}^*\left(\mathbf{B}^{(1)^*} + \Delta\mathbf{B}^*\right)^\top\right\|_F^2 \\
& + \lambda\sum_{j=1}^{p}\left(\left\|\hat{\Delta\boldsymbol{b}_j}\right\|_2 - \left\|\Delta\boldsymbol{b}_j^*\right\|_2\right) \quad 18 \\
:= & \mathbf{I} + \mathbf{II} + \mathbf{III}.
\end{aligned}$$

Algorithm 5 Entrywise CV algorithm for hyperparameter tuning

Input: Observed data $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$, candidate grid $\mathbf{\Lambda}_{\text{grid}}$, latent dimensionality r.
  Partition observed entries of $\mathbf{\Omega}$ into three disjoint sets: $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$.
for $\lambda \in \mathbf{\Lambda}_{\text{grid}}$ **do**
    Initialize validation error: $E_\lambda \leftarrow 0$.
  for $k = 1$ to 3 **do**
    Masking:
      Define Validation Mask: $\mathbf{\Omega}_{\text{val}} \leftarrow \mathcal{I}_k$.
      Define Training Mask: $\mathbf{\Omega}_{\text{train}} \leftarrow \mathbf{\Omega} \setminus \mathcal{I}_k$.
      $\hat{\mathbf{F}}, \hat{\mathbf{B}}^{(1)}, \hat{\Delta\mathbf{B}} \leftarrow \text{SMART}(\mathbf{X}^{(1)} \circ \mathbf{\Omega}^{(1)}_{\text{train}}, \mathbf{X}^{(2)} \circ \mathbf{\Omega}^{(2)}_{\text{train}}, \lambda, r)$ (or Structural Disentanglement with Naïve Lasso).
    Predict and evaluate on validation set:
      $\hat{\mathbf{X}}^{(1)}_{\text{val}}, \hat{\mathbf{X}}^{(2)}_{\text{val}} \leftarrow \hat{\mathbf{F}}(\hat{\mathbf{B}}^{(1)})^\top, \hat{\mathbf{F}}(\hat{\mathbf{B}}^{(1)} + \hat{\Delta\mathbf{B}})^\top$.
      $\epsilon_{k,\lambda} \leftarrow \text{Loss}(\mathbf{X}^{(1)}[\mathbf{\Omega}_{\text{val}}], \mathbf{X}^{(2)}[\mathbf{\Omega}_{\text{val}}], \hat{\mathbf{X}}^{(1)}_{\text{val}}[\mathbf{\Omega}_{\text{val}}], \hat{\mathbf{X}}^{(2)}_{\text{val}}[\mathbf{\Omega}_{\text{val}}])$.
      $E_\lambda \leftarrow E_\lambda + \epsilon_{k,\lambda}$.
  end for
end for
Selection and Final Fit:
  Select optimal parameter: $\hat{\lambda} \leftarrow \arg\min_{\lambda \in \mathbf{\Lambda}_{\text{grid}}} (E_\lambda/3)$.
  Refit on full data:
  $\hat{\mathbf{F}}, \hat{\mathbf{B}}^{(1)}, \hat{\Delta\mathbf{B}} \leftarrow \text{SMART}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \hat{\lambda}, r)$ (or Structural Disentanglement with Naïve Lasso).
Return: optimal $\hat{\lambda}$ and final estimators $\hat{\mathbf{F}}, \hat{\mathbf{B}}^{(1)}, \hat{\Delta\mathbf{B}}$.

To bound the first term, **I**, we observe that

$$\begin{aligned}\left\|\tilde{\mathbf{M}}^{(1)} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F^2 &= \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*} + \mathbf{M}^{(1)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F^2 \\ &\geq \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F^2 + \left\|\mathbf{M}^{(1)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F^2 \\ &\quad - 2\left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F \cdot \left\|\mathbf{M}^{(1)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F.\end{aligned}$$

Consequently,

$$\mathbf{I} \geq \left\|\mathbf{M}^{(1)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F^2 - 2\left\|\hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F \cdot \left\|\mathbf{M}^{(1)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F. \tag{2}$$

Similarly, for the second term, **II**, we decompose it as

$$\begin{aligned}\left\|\tilde{\mathbf{M}}^{(2)} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)} + \hat{\Delta\mathbf{B}}\right)^\top\right\|_F^2 &= \left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*} + \mathbf{M}^{(2)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)} + \hat{\Delta\mathbf{B}}\right)^\top\right\|_F^2 \\ &\geq \left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F^2 + \left\|\mathbf{M}^{(2)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)} + \hat{\Delta\mathbf{B}}\right)^\top\right\|_F^2 \\ &\quad - 2\left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F \cdot \left\|\mathbf{M}^{(2)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)} + \hat{\Delta\mathbf{B}}\right)^\top\right\|_F.\end{aligned}$$

This leads to the bound

$$\mathbf{II} \geq \left\|\mathbf{M}^{(2)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)} + \hat{\Delta\mathbf{B}}\right)^\top\right\|_F^2 - 2\left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F \cdot \left\|\mathbf{M}^{(2)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)} + \hat{\Delta\mathbf{B}}\right)^\top\right\|_F.$$

Recall that $\mathcal{S}^* = \left\{ j : \left\| \Delta \boldsymbol{b}_j^* \right\|_2 \neq 0 \right\}$, and $s = |\mathcal{S}^*|$. Then for term **III**, we derive the following bound

$$
\begin{aligned}
\mathbf{III} &= \lambda \sum_{j=1}^{p} \left(\left\| \Delta \hat{\boldsymbol{b}}_j \right\|_2 - \left\| \Delta \boldsymbol{b}_j^* \right\|_2 \right) \\
&= \lambda \left\{ \sum_{j \in \mathcal{S}^*} \left(\left\| \Delta \hat{\boldsymbol{b}}_j \right\|_2 - \left\| \Delta \boldsymbol{b}_j^* \right\|_2 \right) + \sum_{j \notin \mathcal{S}^*} \left\| \Delta \hat{\boldsymbol{b}}_j \right\|_2 \right\} \\
&\geq \lambda \sum_{j \in \mathcal{S}^*} \left(\left\| \Delta \hat{\boldsymbol{b}}_j \right\|_2 - \left\| \Delta \boldsymbol{b}_j^* \right\|_2 \right) \\
&\geq -\lambda \sum_{j \in \mathcal{S}^*} \left\| \Delta \hat{\boldsymbol{b}}_j - \Delta \boldsymbol{b}_j^* \right\|_2 \\
&\geq -\lambda \sqrt{s} \left\| \Delta \hat{\mathbf{B}}_{\mathcal{S}^*} - \Delta \mathbf{B}^* \right\|_F ,
\end{aligned}
\tag{3}
$$

where $\Delta \hat{\mathbf{B}}_{\mathcal{S}^*}$ represents a matrix in which only the rows from the set $\mathcal{S}^*$ in the original $\Delta \hat{\mathbf{B}}$ matrix are non-zero, with all other rows set to zero.

Thus we can bound this term by relying on the boundedness of the parameters. Specifically, by assuming that $\|\boldsymbol{b}_j^{(1)*}\|_2 \leq C$ and $\|\boldsymbol{b}_j^{(2)*}\|_2 \leq C$ in Condition (C1), it follows that $\|\Delta \boldsymbol{b}_j^*\|_2 = \|\boldsymbol{b}_j^{(2)*} - \boldsymbol{b}_j^{(1)*}\|_2 \leq 2C$. By the same logic, the boundedness of the estimates implies $\|\Delta \hat{\boldsymbol{b}}_j\| \leq \|\hat{\boldsymbol{b}}_j^{(2)} - \hat{\boldsymbol{b}}_j^{(1)}\| \leq \|\hat{\boldsymbol{b}}_j^{(2)}\| + \|\hat{\boldsymbol{b}}_j^{(1)}\| \leq 2C$.

Thus, we have

$$
\mathbf{III} \geq -2Cs\lambda.
$$

Finally, combining the results for **I**, **II**, and **III**, we conclude that

$$
\begin{aligned}
0 \geq \mathbf{I} + \mathbf{II} + \mathbf{III} \geq & \left\| \mathbf{M}^{(1)*} - \hat{\mathbf{F}} \left(\hat{B}^{(1)} \right)^\top \right\|_F^2 + \left\| \mathbf{M}^{(2)*} - \hat{\mathbf{F}} \left(\hat{\mathbf{B}}^{(1)} + \Delta \hat{\mathbf{B}} \right)^\top \right\|_F^2 \\
& - 2 \left\| \tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*} \right\|_F \cdot \left\| \mathbf{M}^{(1)*} - \hat{\mathbf{F}} \left(\hat{\mathbf{B}}^{(1)} \right)^\top \right\|_F \\
& - 2 \left\| \tilde{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*} \right\|_F \cdot \left\| \mathbf{M}^{(2)*} - \hat{\mathbf{F}} \left(\hat{\mathbf{B}}^{(1)} + \Delta \hat{\mathbf{B}} \right)^\top \right\|_F \\
& - 2Cs\lambda.
\end{aligned}
$$

Rearranging the terms from the above inequality and applying the Cauchy-Schwarz inequality yields the following upper bound:

$$
\begin{aligned}
& \left\| \mathbf{M}^{(1)*} - \hat{\mathbf{F}} \left(\hat{\mathbf{B}}^{(1)} \right)^\top \right\|_F^2 + \left\| \mathbf{M}^{(2)*} - \hat{\mathbf{F}} \left(\hat{\mathbf{B}}^{(1)} + \Delta \hat{\mathbf{B}} \right)^\top \right\|_F^2 \\
\leq & 2 \left\| \tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*} \right\|_F \cdot \left\| \mathbf{M}^{(1)*} - \hat{\mathbf{F}} \left(\hat{\mathbf{B}}^{(1)} \right)^\top \right\|_F \\
& + 2 \left\| \tilde{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*} \right\|_F \cdot \left\| \mathbf{M}^{(2)*} - \hat{\mathbf{F}} \left(\hat{\mathbf{B}}^{(1)} + \Delta \hat{\mathbf{B}} \right)^\top \right\|_F + 2Cs\lambda \\
\leq & 2 \left(\left\| \tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*} \right\|_F^2 + \left\| \tilde{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*} \right\|_F^2 \right)^{\frac{1}{2}} \\
& \cdot \left(\left\| \mathbf{M}^{(1)*} - \hat{\mathbf{F}} \left(\hat{\mathbf{B}}^{(1)} \right)^\top \right\|_F^2 + \left\| \mathbf{M}^{(2)*} - \hat{\mathbf{F}} \left(\hat{\mathbf{B}}^{(1)} + \Delta \hat{\mathbf{B}} \right)^\top \right\|_F^2 \right)^{\frac{1}{2}} + 2Cs\lambda.
\end{aligned}
$$

Based on this algebraic structure, it implies that either

$$\begin{aligned}&\left\|\mathbf{M}^{(1)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^{\top}\right\|_F^2+\left\|\mathbf{M}^{(2)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}+\hat{\Delta\mathbf{B}}\right)^{\top}\right\|_F^2\\ \leq&16\left(\left\|\tilde{\mathbf{M}}^{(1)}-\mathbf{M}^{(1)*}\right\|_F^2+\left\|\tilde{\mathbf{M}}^{(2)}-\mathbf{M}^{(2)*}\right\|_F^2\right)\end{aligned} \tag{4}$$

or

$$\left\|\mathbf{M}^{(1)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^{\top}\right\|_F^2+\left\|\mathbf{M}^{(2)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}+\hat{\Delta\mathbf{B}}\right)^{\top}\right\|_F^2\leq 4Cs\lambda. \tag{5}$$

Furthermore, utilizing the elementary inequality $(a+b)^2\leq 2(a^2+b^2)$, we observe that

$$\begin{aligned}&\frac{1}{2}\left(\left\|\mathbf{M}^{(1)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^{\top}\right\|_F+\left\|\mathbf{M}^{(2)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}+\hat{\Delta\mathbf{B}}\right)^{\top}\right\|_F\right)^2\\ \leq&\left\|\mathbf{M}^{(1)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^{\top}\right\|_F^2+\left\|\mathbf{M}^{(2)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}+\hat{\Delta\mathbf{B}}\right)^{\top}\right\|_F^2.\end{aligned} \tag{6}$$

Combining Equations (4) and (5) with Equation (6) leads to the following bounds on the sum of the Frobenius norms

$$\begin{aligned}&\left\|\mathbf{M}^{(1)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^{\top}\right\|_F+\left\|\mathbf{M}^{(2)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}+\hat{\Delta\mathbf{B}}\right)^{\top}\right\|_F\\ \leq&4\sqrt{2}\left(\left\|\tilde{\mathbf{M}}^{(1)}-\mathbf{M}^{(1)*}\right\|_F^2+\left\|\tilde{\mathbf{M}}^{(2)}-\mathbf{M}^{(2)*}\right\|_F^2\right)^{\frac{1}{2}}\end{aligned}$$

or

$$\left\|\mathbf{M}^{(1)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^{\top}\right\|_F+\left\|\mathbf{M}^{(2)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}+\hat{\Delta\mathbf{B}}\right)^{\top}\right\|_F\leq 2\sqrt{2Cs\lambda}.$$

We now incorporate the error from the first stage. Assuming that

$$\left\|\tilde{\mathbf{M}}^{(1)}-\mathbf{M}^{(1)^*}\right\|_F+\left\|\tilde{\mathbf{M}}^{(2)}-\mathbf{M}^{(2)*}\right\|_F\lesssim \text{err}$$

with probability tending to one, it follows that

$$\left\|\mathbf{M}^{(1)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^{\top}\right\|_F+\left\|\mathbf{M}^{(2)^*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}+\hat{\Delta\mathbf{B}}\right)^{\top}\right\|_F\lesssim \text{err}+\sqrt{s\lambda}$$

with probability tending to one. Consequently, we obtain the individual consistency results

$$\begin{aligned}\left\|\mathbf{M}^{(1)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^{\top}\right\|_F&\lesssim \text{err}+\sqrt{s\lambda}\\ \left\|\mathbf{M}^{(2)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(2)}\right)^{\top}\right\|_F&\lesssim \text{err}+\sqrt{s\lambda}.\end{aligned} \tag{7}$$

with probability tending to one. □

Proof of Theorem 3.3: Assuming the rank is correctly specified (i.e., $r=r^*$), we invoke Theorem 3 of Yu et al. (2015). This ensures the existence of an orthogonal rotation matrix $\mathbf{O}$ such that the estimation error of the factor matrix satisfies

$$\begin{aligned}\frac{1}{\sqrt{n}}\left\|\hat{\mathbf{F}}\mathbf{O}-\mathbf{F}^*\right\|_F&\lesssim\frac{\sigma_1\left(\mathbf{M}^{(1)*}\right)\left\|\mathbf{M}^{(1)*}-\hat{\mathbf{F}}\hat{\mathbf{B}}^{(1)}\right\|_F}{\sigma_{r^*}^2\left(\mathbf{M}^{(1)*}\right)}\\ &\lesssim\frac{\text{err}+\sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}}.\end{aligned}$$

The validity of this bound is contingent on the Condition (C2) regarding the singular values and signal strength.

By decomposing the error term and applying the triangle inequality,

$$\begin{aligned}
&\left\|\mathbf{M}^{(1)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F = \left\|\mathbf{F}^*\left(\mathbf{B}^{(1)*}\right)^\top - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F \\
=&\left\|\left(\mathbf{F}^* - \hat{\mathbf{F}}\mathbf{O}\right)\left(\mathbf{B}^{(1)*}\right)^\top + \hat{\mathbf{F}}\mathbf{O}\left(\mathbf{B}^{(1)*}\right)^\top - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F \\
\geq&\left\|\hat{\mathbf{F}}\left(\mathbf{B}^{(1)*}\mathbf{O}^\top - \hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F - \left\|\left(\mathbf{F}^* - \hat{\mathbf{F}}\mathbf{O}\right)\left(\mathbf{B}^{(1)*}\right)^\top\right\|_F.
\end{aligned}$$

Rearranging these terms provides the following inequality

$$\left\|\hat{\mathbf{F}}\left(\mathbf{B}^{(1)*}\mathbf{O}^\top - \hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F \leq \left\|\mathbf{M}^{(1)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F + \left\|\left(\mathbf{F}^* - \hat{\mathbf{F}}\mathbf{O}\right)\left(\mathbf{B}^{(1)*}\right)^\top\right\|_F. \tag{8}$$

Furthermore, under the identification constraints $\hat{\mathbf{F}}^\top\hat{\mathbf{F}} = n\mathbf{I}_r$, the term on the left-hand side of Equation (8) can be simplified as

$$\begin{aligned}
\left\|\hat{\mathbf{F}}\left(\mathbf{B}^{(1)*}\mathbf{O}^\top - \hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F &= \operatorname{tr}\left\{\left(\mathbf{B}^{(1)*}\mathbf{O}^\top - \hat{\mathbf{B}}^{(1)}\right)^\top\hat{\mathbf{F}}^\top\hat{\mathbf{F}}\left(\mathbf{B}^{(1)*}\mathbf{O}^\top - \hat{\mathbf{B}}^{(1)}\right)^\top\right\} \\
&= n\left\|\hat{\mathbf{B}}^{(1)}\mathbf{O} - \mathbf{B}^{(1)*}\right\|_F^2.
\end{aligned}$$

Additionally, the second term on the right-hand side of Equation (8) is bounded by

$$\left\|\left(\mathbf{F}^* - \hat{\mathbf{F}}\mathbf{O}\right)\left(\mathbf{B}^{(1)*}\right)^\top\right\|_F \leq \left\|\mathbf{F}^* - \hat{\mathbf{F}}\mathbf{O}\right\|_F\left\|\mathbf{B}^{(1)*}\right\|_F \lesssim \sqrt{p}\left\|\mathbf{F}^* - \hat{\mathbf{F}}\mathbf{O}\right\|_F.$$

Substituting these results back into the inequality and normalizing by $\sqrt{p}$, we derive the convergence rate for the loading matrix $\hat{\mathbf{B}}^{(1)}$

$$\begin{aligned}
\frac{1}{\sqrt{p}}\left\|\hat{\mathbf{B}}^{(1)}\mathbf{O} - \mathbf{B}^{(1)*}\right\|_F &\lesssim \frac{1}{\sqrt{n}}\left\|\hat{\mathbf{F}}\mathbf{O} - \mathbf{F}^*\right\|_F + \frac{1}{\sqrt{np}}\left\|\mathbf{M}^{(1)*} - \hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(1)}\right)^\top\right\|_F \\
&\lesssim \frac{\mathrm{err} + \sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}} + \frac{\mathrm{err} + \sqrt{s\lambda}}{\sqrt{np}} \\
&\lesssim \frac{\mathrm{err} + \sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}}.
\end{aligned}$$

By an analogous argument, the same bound holds for $\hat{\mathbf{B}}^{(2)}$

$$\frac{1}{\sqrt{p}}\left\|\hat{\mathbf{B}}^{(2)}\mathbf{O} - \mathbf{B}^{(2)*}\right\|_F \lesssim \frac{\mathrm{err} + \sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}}.$$

Consequently, the error bound for the difference $\hat{\Delta\mathbf{B}} = \hat{\mathbf{B}}^{(2)} - \hat{\mathbf{B}}^{(1)}$ is given by

$$\frac{1}{\sqrt{p}}\left\|\hat{\Delta\mathbf{B}}\mathbf{O} - \Delta\mathbf{B}^*\right\|_F \lesssim \frac{\mathrm{err} + \sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}}.$$

Finally, adjusting the normalization for the sparsity level s, we conclude

$$\frac{1}{\sqrt{s}}\left\|\hat{\Delta\mathbf{B}}\mathbf{O} - \Delta\mathbf{B}^*\right\|_F \lesssim \sqrt{\frac{p}{\sigma_{\mathbf{B}}^2}} \cdot \frac{\mathrm{err} + \sqrt{s\lambda}}{\sqrt{ns}}.$$

Thus the proof is finished. □

Proof of Theorem 3.4: By invoking Theorem 1 in Chen & Li (2022), we have

$$\left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F + \left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F \lesssim \sqrt{n} \vee \sqrt{p}$$

with probability tending to one. This, along with $s\lambda \ll n \vee p$, confirms that the requisite error bounds for $\hat{\mathbf{M}}^{(1)}$ and $\hat{\mathbf{M}}^{(2)}$ holds. □

Proof of Theorem 3.5: Based on Theorem 3.3, under the conditions that $s\lambda \ll n \vee p$ and $\sigma_{\mathbf{B}} \asymp \sqrt{p}$, we observe that

$$\frac{s\lambda}{n\sigma_{\mathbf{B}}^2} \asymp \frac{s\lambda}{np} \ll \frac{1}{n \wedge p}.$$

Consequently, the final term in Equation (7) is dominated by the estimation error, and the proof is complete. □

Proof of Theorem 3.6: We begin by decomposing the difference in the information criterion values

$$\begin{aligned}
&\left\{\mathcal{L}\left(\mathbf{F}^*, \mathbf{B}^{(1)*}; \tilde{\mathbf{M}}^{(1)}\right) + \mathcal{L}\left(\mathbf{F}^*, \mathbf{B}^{(2)*}; \tilde{\mathbf{M}}^{(2)}\right) + r^*\eta(n,p)\right\} \\
&- \left\{\mathcal{L}\left(\hat{\mathbf{F}}_r, \hat{\mathbf{B}}_r^{(1)}; \tilde{\mathbf{M}}^{(1)}\right) + \mathcal{L}\left(\hat{\mathbf{F}}_r, \hat{\mathbf{B}}_r^{(2)}; \tilde{\mathbf{M}}^{(2)}\right) + r\eta(n,p)\right\} \\
=&\left\{\mathcal{L}\left(\mathbf{F}^*, \mathbf{B}^{(1)*}; \tilde{\mathbf{M}}^{(1)}\right) - \mathcal{L}\left(\hat{\mathbf{F}}_r, \hat{\mathbf{B}}_r^{(1)}; \tilde{\mathbf{M}}^{(1)}\right)\right\} + \left\{\mathcal{L}\left(\mathbf{F}^*, \mathbf{B}^{(2)*}; \tilde{\mathbf{M}}^{(2)}\right) - \mathcal{L}\left(\hat{\mathbf{F}}_r, \hat{\mathbf{B}}_r^{(2)}; \tilde{\mathbf{M}}^{(2)}\right)\right\} \\
&+ \left\{r^*\eta(n,p) - r\eta(n,p)\right\} \\
=&\mathbf{I} + \mathbf{II} + \mathbf{III}.
\end{aligned}$$

We analyze the proof in two scenarios.

Case 1: overestimation ($r \geq r^*$).

First, consider the term **I**. By applying Theorem 3.2, we bound the estimation error as follows

$$\begin{aligned}
\left\|\tilde{\mathbf{M}}^{(1)} - \hat{\mathbf{F}}_r\left(\hat{\mathbf{B}}_r^{(1)}\right)^\top\right\|_F &\leq \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F + \left\|\mathbf{M}^{(1)*} - \hat{\mathbf{F}}_r\left(\hat{\mathbf{B}}_r^{(1)}\right)^\top\right\|_F \\
&\lesssim \mathrm{err} + \sqrt{s\lambda}.
\end{aligned}$$

Consequently, the magnitude of **I** satisfies

$$\begin{aligned}
|\mathbf{I}| &\leq \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{F}^*\left(\mathbf{B}^{(1)*}\right)^\top\right\|_F^2 + \left\|\tilde{\mathbf{M}}^{(1)} - \hat{\mathbf{F}}_r\left(\hat{\mathbf{B}}_r^{(1)}\right)^\top\right\|_F^2 \\
&\lesssim (\mathrm{err})^2 + s\lambda
\end{aligned}$$

with probability tending to one.

Similarly, for the second term **II**, we have

$$\begin{aligned}
|\mathbf{II}| &\leq \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{F}^*\left(\mathbf{B}^{(1)*}\right)^\top\right\|_F^2 + \left\|\tilde{\mathbf{M}}^{(1)} - \hat{\mathbf{F}}_r\left(\hat{\mathbf{B}}_r^{(1)}\right)^\top\right\|_F^2 \\
&\lesssim (\mathrm{err})^2 + s\lambda.
\end{aligned}$$

For simplicity, we denote

$$\mathrm{IC}(*) = \mathcal{L}\left(\mathbf{F}^*, \mathbf{B}^{(1)*}; \tilde{\mathbf{M}}^{(1)}\right) + \mathcal{L}\left(\mathbf{F}^*, \mathbf{B}^{(2)*}; \tilde{\mathbf{M}}^{(2)}\right) + r^*\eta(n,p).$$

Thus, for any $r \geq r^*$, the difference in the criteria can be expressed as

$$\begin{aligned}
\mathrm{IC}(r^*) - \mathrm{IC}(r) &= \{\mathrm{IC}(r^*) - \mathrm{IC}(*)\} - \{\mathrm{IC}(r) - \mathrm{IC}(*)\} \\
&= O_p\left((\mathrm{err})^2 + s\lambda\right) + (r^* - r)\,\eta(n,p).
\end{aligned} \tag{9}$$

Imposing the condition $\eta(n,p) \gg (\text{err})^2 + s\lambda$, we observe that the penalty difference dominates the estimation error in Equation (9). Since $r > r^*$, the term $(r^* - r)\eta(n,p)$ is negative and dominant. Therefore,

$$\mathbb{P}\left\{\text{IC}\left(r^*\right) - \text{IC}(r) < 0\right\} \to 1$$

as $n, p \to +\infty$.

Case 2: Underestimation ($1 \leq r < r^*$).

We begin by decomposing the first term, **I**. Using the definition of the Frobenius norm and expanding the squared terms, we obtain

$$\begin{aligned}
\mathbf{I} &= \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{F}^*\left(\mathbf{B}^{(1)*}\right)^\top\right\|_F^2 - \left\|\tilde{\mathbf{M}}^{(1)} - \hat{\mathbf{F}}_r\left(\hat{\mathbf{B}}_r^{(1)}\right)^\top\right\|_F^2 \\
&= \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F^2 - \left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*} + \mathbf{M}^{(1)*} - \hat{\mathbf{M}}_r^{(1)}\right\|_F^2 \\
&\leq 2\left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F \cdot \left\|\hat{\mathbf{M}}_r^{(1)} - \mathbf{M}^{(1)*}\right\|_F - \left\|\hat{\mathbf{M}}_r^{(1)} - \mathbf{M}^{(1)*}\right\|_F^2.
\end{aligned}$$

Note that this expression represents a quadratic function with respect to $\left\|\hat{\mathbf{M}}_r^{(1)} - \mathbf{M}^{(1)*}\right\|_F$.

Recall that for any rank-r approximation where $r < r^*$, the error is lower-bounded by the singular values of the true matrix. Specifically,

$$\left\|\hat{\mathbf{M}}_r^{(1)} - \mathbf{M}^{(1)*}\right\|_F \geq \sigma_{r^*}\left(\mathbf{M}^{(1)*}\right).$$

To bound **I**, let us define the auxiliary function $\mu(a) = 2a\left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}_r^{(1)*}\right\|_F - a^2$. The inequality for **I** can then be expressed as

$$\mathbf{I} \leq \sup_{a \geq \sigma_{r^*}\left(\mathbf{M}^{(1)*}\right)} \mu(a).$$

If the singular value is sufficiently large such that

$$\sigma_{r^*}\left(\mathbf{M}^{(1)*}\right) \geq 4\left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F,$$

then the supremum of $\mu(a)$ over the domain $a \geq \sigma_{r^*}\left(\mathbf{M}^{(1)*}\right)$ is attained at the boundary $a = \sigma_{r^*}\left(\mathbf{M}^{(1)*}\right)$.

Substituting this value back into $\mu(a)$, we derive the upper bound

$$\begin{aligned}
\mathbf{I} \leq \mu\left(\sigma_{r^*}\left(\mathbf{M}^{(1)*}\right)\right) &= -\sigma_{r^*}^2\left(\mathbf{M}^{(1)*}\right) + 2\sigma_{r^*}\left(\mathbf{M}^{(1)*}\right)\left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F \\
&\leq -\frac{1}{2}\sigma_{r^*}^2\left(\mathbf{M}^{(1)*}\right).
\end{aligned}$$

By an analogous argument, the second term **II** satisfies

$$\mathbf{II} \leq -\frac{1}{2}\sigma_{r^*}^2\left(\mathbf{M}^{(2)*}\right),$$

provided that the condition

$$\sigma_{r^*}\left(\mathbf{M}^{(2)*}\right) \geq 4\left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F$$

is met.

By the conditions, $\sigma_{r^*}\left(\mathbf{M}^{(1)*}\right) \asymp \sqrt{n}\sigma_{\mathbf{B}}$, $\left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F \lesssim \text{err}$, and $n\sigma_{\mathbf{B}}^2 \gg (\text{err})^2 + s\lambda$, the required inequalities

$$\sigma_{r^*}\left(\mathbf{M}^{(1)*}\right) \geq 4\left\|\tilde{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F$$

and

$$\sigma_{r^*}\left(\mathbf{M}^{(2)*}\right) \geq 4\left\|\tilde{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F$$

hold with probability tending to one.

Combining these results, we obtain the following difference in the information criterion

$$\begin{aligned}
\mathrm{IC}\left(r^*\right) - \mathrm{IC}(r) &= \left\{\mathrm{IC}\left(r^*\right) - \mathrm{IC}(*)\right\} - \left\{\mathrm{IC}(r) - \mathrm{IC}(*)\right\} \\
&\leq -\frac{1}{2}\sigma_{r^*}^2\left(\mathbf{M}^{(1)*}\right) - \frac{1}{2}\sigma_{r^*}^2\left(\mathbf{M}^{(2)*}\right) + O_p\left((\mathrm{err})^2 + s\lambda\right) \\
&\quad + \left(r^* - r\right)\eta(n,p) \\
&\lesssim -n\sigma_{\mathbf{B}}^2 + O_p\left((\mathrm{err})^2 + s\lambda\right) + \left(r^* - r\right)\eta(n,p).
\end{aligned}$$

Given the condition $(\mathrm{err})^2 + s\lambda \ll \eta(n,p) \ll n\sigma_{\mathbf{B}}^2$, the term $-n\sigma_{\mathbf{B}}^2$ dominates the terms. Thus, we conclude that

$$\mathbb{P}\left\{\mathrm{IC}\left(r^*\right) - \mathrm{IC}(r) < 0\right\} \to 1$$

as $n, p \to +\infty$.

The above two cases lead to

$$\mathbb{P}\left(\hat{r} = r^*\right) \to 1$$

as $n, p \to +\infty$. □

Proof of Theorem 3.7: We first establish the asymptotic dominance of the penalty term $\eta(n,p)$. Specifically, it holds that

$$\eta_1(n,p) = 2(n+p+\hat{s})\log\left(\frac{np}{n+p}\right) \gg (n \vee p) \gtrsim \mathrm{err}.$$

Furthermore, we verify that the penalty vanishes relative to the signal strength,

$$\frac{\eta_1(n,p)}{n\sigma_{\mathbf{B}}^2} = \frac{2(n+p+\hat{s})}{np}\log\left(\frac{np}{n+p}\right) \to 0.$$

The proof for the alternative penalty specification, given by

$$\eta_2(n,p) = 2(n+p+\hat{s})\log\left(n \wedge p\right)$$

follows an analogous argument and is therefore omitted. □

Proof of Theorem 5.1: By the properties of the MLE, we obtain the following inequality

$$\begin{aligned}
0 \geq &\left\|\mathbf{X}^{(1)} - \hat{\mathbf{M}}^{(1)}\right\|_F^2 - \left\|\mathbf{X}^{(1)} - \mathbf{F}^*\mathbf{B}^{(1)*}\right\|_F^2 \\
&+ \left\|\mathbf{X}^{(2)} - \hat{\mathbf{M}}^{(2)}\right\|_F^2 - \left\|\mathbf{X}^{(2)} - \mathbf{F}^*\mathbf{B}^{(2)*}\right\|_F^2 \\
&+ \lambda\sum_{j=1}^{p}\left(\left\|\Delta\hat{\boldsymbol{b}}_j\right\|_2 - \left\|\Delta\boldsymbol{b}_j^*\right\|_2\right) \\
:=&\mathbf{I} + \mathbf{II} + \mathbf{III}.
\end{aligned}$$

Let $\langle\mathbf{A},\mathbf{B}\rangle$ denote the matrix inner product $\mathrm{tr}(\mathbf{A}^\top\mathbf{B})$. The first term $\mathbf{I}$ can be expanded as

$$\begin{aligned}
\mathbf{I} &= \sum_{i=1}^{n}\sum_{j=1}^{p}\left\{\left(x_{ij}^{(1)} - \hat{m}_{ij}^{(1)}\right)^2 - \left(x_{ij}^{(1)} - m_{ij}^{(1)*}\right)^2\right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{p}\left\{\left(\hat{m}_{ij}^{(1)} - m_{ij}^{(1)*}\right)^2 - 2\left(x_{ij}^{(1)} - m_{ij}^{(1)*}\right)\left(\hat{m}_{ij}^{(1)} - m_{ij}^{(1)*}\right)\right\} \\
&= \left\|\hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F^2 - 2\left\langle\mathbf{X}^{(1)} - \mathbf{M}^{(1)*}, \hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\rangle.
\end{aligned}$$

To bound the inner product, we utilize the matrix inequality $|\langle \mathbf{A}, \mathbf{B}\rangle| \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_* \leq \sqrt{\text{rank}(\mathbf{B})}\|\mathbf{A}\|_2\|\mathbf{B}\|_F$. Applying this to our context, where the rank of the difference matrix is at most $r + r^*$, we have

$$\begin{aligned}&\left|\left\langle \mathbf{X}^{(1)} - \mathbf{M}^{(1)*}, \hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\rangle\right| \\ &\leq \sqrt{r+r^*}\left\|\mathbf{X}^{(1)} - \mathbf{M}^{(1)*}\right\|_2 \left\|\hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F.\end{aligned}$$

Substituting this back into the expansion of $\mathbf{I}$ yields

$$\mathbf{I} \geq \left\|\hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F^2 - 2\sqrt{r+r^*}\left\|\mathbf{X}^{(1)} - \mathbf{M}^{(1)*}\right\|_2 \cdot \left\|\hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F.$$

Similarly, for the second term $\mathbf{II}$, we obtain

$$\mathbf{II} \geq \left\|\hat{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F^2 - 2\sqrt{r+r^*}\left\|\mathbf{X}^{(2)} - \mathbf{M}^{(2)*}\right\|_2 \cdot \left\|\hat{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F.$$

The analysis of the penalty term $\mathbf{III}$ follows the same logic as in Theorem 3.2, leading to the bound

$$\mathbf{III} \geq -2Cs\lambda.$$

Combining these results, the initial inequality becomes

$$\begin{aligned}0 \geq \mathbf{I} + \mathbf{II} + \mathbf{III} \geq &\left\|\hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)^*}\right\|_F^2 + \left\|\hat{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F^2 \\ &- 2\sqrt{r+r^*}\left\|\mathbf{X}^{(1)} - \mathbf{M}^{(1)*}\right\|_2 \cdot \left\|\hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F \\ &- 2\sqrt{r+r^*}\left\|\mathbf{X}^{(2)} - \mathbf{M}^{(2)*}\right\|_2 \cdot \left\|\hat{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F \\ &- 2Cs\lambda.\end{aligned}$$

Rearranging the terms and applying the Cauchy-Schwarz inequality, we derive

$$\begin{aligned}&\left\|\hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F^2 + \left\|\hat{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F^2 \\ \leq &2\sqrt{r+r^*}\left(\left\|\mathbf{X}^{(1)} - \mathbf{M}^{(1)*}\right\|_2 \cdot \left\|\hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F + \left\|\mathbf{X}^{(2)} - \mathbf{M}^{(2)*}\right\|_2 \cdot \left\|\hat{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F\right) \\ &+ 2Cs\lambda \\ \leq &2\sqrt{r+r^*}\left(\left\|\hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F^2 + \left\|\hat{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F^2\right)^{\frac{1}{2}} \cdot \left(\left\|\mathbf{X}^{(1)} - \mathbf{M}^{(1)*}\right\|_2^2 + \left\|\mathbf{X}^{(2)} - \mathbf{M}^{(2)*}\right\|_2^2\right)^{\frac{1}{2}} \\ &+ 2Cs\lambda.\end{aligned}$$

We next bound the spectral norms of the noise matrices, $\left\|\mathbf{X}^{(1)} - \mathbf{M}^{(1)*}\right\|_2$ and $\left\|\mathbf{X}^{(2)} - \mathbf{M}^{(2)*}\right\|_2$, by invoking Lemma 1 in Wang (2022). Assumption 2 in Wang (2022) is satisfied under our linear setting, and Assumption 3 is verified by our Condition (C3). Consequently, we have

$$\left\|\mathbf{X}^{(1)} - \mathbf{M}^{(1)*}\right\|_2 = O_p\left(n^{\frac{1}{2}}p^{\frac{1}{4}} \vee n^{\frac{1}{4}}p^{\frac{1}{2}}\right)$$

and

$$\left\|\mathbf{X}^{(2)} - \mathbf{M}^{(2)*}\right\|_2 = O_p\left(n^{\frac{1}{2}}p^{\frac{1}{4}} \vee n^{\frac{1}{4}}p^{\frac{1}{2}}\right).$$

Following the convergence analysis framework established in Theorem 3.2, these bounds imply that

$$\left\|\hat{\mathbf{M}}^{(1)} - \mathbf{M}^{(1)*}\right\|_F \lesssim n^{\frac{1}{2}}p^{\frac{1}{4}} \vee n^{\frac{1}{4}}p^{\frac{1}{2}} + \sqrt{s\lambda}$$

and

$$\left\|\hat{\mathbf{M}}^{(2)} - \mathbf{M}^{(2)*}\right\|_F \lesssim n^{\frac{1}{2}}p^{\frac{1}{4}} \vee n^{\frac{1}{4}}p^{\frac{1}{2}} + \sqrt{s\lambda}.$$

with probability tending to one.

Finally, analogously to Theorem 3.3, we deduce the rates for the factor and loading matrices,

$$\frac{1}{\sqrt{n}}\left\|\hat{\mathbf{F}}\mathbf{O}-\mathbf{F}^*\right\|_F \lesssim \frac{n^{\frac{1}{2}}p^{\frac{1}{4}} \vee n^{\frac{1}{4}}p^{\frac{1}{2}}+\sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}},$$

$$\frac{1}{\sqrt{p}}\left\|\hat{\mathbf{B}}^{(k)}\mathbf{O}-\mathbf{B}^*\right\|_F \lesssim \frac{n^{\frac{1}{2}}p^{\frac{1}{4}} \vee n^{\frac{1}{4}}p^{\frac{1}{2}}+\sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}},$$

and

$$\frac{1}{\sqrt{s}}\left\|\Delta\hat{\mathbf{B}}\mathbf{O}-\Delta\mathbf{B}^*\right\|_F \lesssim \sqrt{\frac{p}{\sigma_{\mathbf{B}}^2}}\cdot\frac{n^{\frac{1}{2}}p^{\frac{1}{4}} \vee n^{\frac{1}{4}}p^{\frac{1}{2}}+\sqrt{s\lambda}}{\sqrt{ns}}.$$

This concludes the proof. □

Proof of Theorem 4.1: The proof of this theorem parallels that of Theorem 3.2. Therefore, we focus primarily on the necessary modifications for the multi-environment setting.

By the definition of the MLE, we have the following inequality

$$\begin{aligned}
0 \geq & \sum_{k=1}^{K}\left\{\left\|\tilde{\mathbf{M}}^{(k)}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^\top\right\|_F^2-\left\|\tilde{\mathbf{M}}^{(k)}-\mathbf{F}^*\left(\mathbf{B}^{(k*)}\right)^\top\right\|_F^2\right\} \\
& +\lambda\sum_{k=2}^{K}\sum_{j=1}^{p}\left\{\left\|\hat{\boldsymbol{b}}_j^{(k)}-\hat{\boldsymbol{b}}_j^{(k-1)}\right\|_2-\left\|\boldsymbol{b}_j^{(k)*}-\boldsymbol{b}_j^{(k-1)*}\right\|_2\right\} \\
:= & \sum_{k=1}^{K}\mathbf{I}_k+\sum_{k=2}^{K}\mathbf{II}_k.
\end{aligned}$$

For the first term $\mathbf{I}_k$, applying an argument analogous to Equation (2) yields the lower bound

$$\mathbf{I}_k \geq \left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^\top\right\|_F^2-2\left\|\tilde{\mathbf{M}}^{(k)}-\mathbf{M}^{(k)*}\right\|_F\cdot\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^\top\right\|_F.$$

Similarly, summing the penalty terms leads to a bound analogous to Equation (3),

$$\sum_{k=2}^{K}\mathbf{II}_k \geq -2C(K-1)s\lambda.$$

Aggregating these bounds, we obtain

$$\begin{aligned}
& \sum_{k=1}^{K}\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^\top\right\|_F^2 \\
\leq & 2\sum_{k=1}^{K}\left\|\tilde{\mathbf{M}}^{(k)}-\mathbf{M}^{(k)*}\right\|_F\cdot\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^\top\right\|_F+2C(K-1)s\lambda \\
\leq & 2\left(\sum_{k=1}^{K}\left\|\tilde{\mathbf{M}}^{(k)}-\mathbf{M}^{(k)*}\right\|_F^2\right)^{\frac{1}{2}}\cdot\left(\sum_{k=1}^{K}\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^\top\right\|_F^2\right)^{\frac{1}{2}}+2C(K-1)s\lambda.
\end{aligned}$$

Thus, either

$$\sum_{k=1}^{K}\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^\top\right\|_F^2 \leq 16\sum_{k=1}^{K}\left\|\tilde{\mathbf{M}}^{(k)}-\mathbf{M}^{(k)*}\right\|_F^2$$

or

$$\sum_{k=1}^{K}\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^\top\right\|_F^2 \leq 4C(K-1)s\lambda$$

holds.

Applying the Cauchy-Schwarz inequality allows us to relate the sum of squares to the square of the sum

$$K\sum_{k=1}^{K}\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^{\top}\right\|_F^2 \geq \left\{\sum_{k=1}^{K}\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^{\top}\right\|_F\right\}^2.$$

Dividing by K and substituting the bounds derived above, we find

$$\frac{1}{K}\left\{\sum_{k=1}^{K}\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^{\top}\right\|_F\right\}^2 \leq 16\sum_{k=1}^{K}\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^{\top}\right\|_F^2 \lesssim K(\mathrm{err})^2$$

and consequently,

$$\frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^{\top}\right\|_F \lesssim \mathrm{err},$$

or, in the case where the $s\lambda$ dominates,

$$\frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^{\top}\right\|_F \lesssim \sqrt{\frac{K-1}{K}s\lambda} \lesssim \sqrt{s\lambda}.$$

Therefore, the average estimation error satisfies

$$\frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{M}^{(k)*}-\hat{\mathbf{F}}\left(\hat{\mathbf{B}}^{(k)}\right)^{\top}\right\|_F \lesssim \mathrm{err}+\sqrt{s\lambda}.$$

with probability tending to one.

Next, we derive the error bounds for $\hat{\mathbf{F}}$ and $\hat{\mathbf{B}}^{(k)}$. Let us define the following stacked matrices

$$\hat{\bar{\mathbf{M}}} = \left(\hat{\mathbf{M}}^{(1)}, \hat{\mathbf{M}}^{(2)}, \ldots, \hat{\mathbf{M}}^{(K)}\right),$$

$$\bar{\mathbf{B}}^{*} = \left(\left(\mathbf{B}^{(1)^{*}}\right)^{\top}, \ldots, \left(\mathbf{B}^{(K)*}\right)^{\top}\right)^{\top},$$

and

$$\bar{\mathbf{M}}^{*} = \left(\mathbf{M}^{(1)*}, \mathbf{M}^{(2)*}, \ldots, \mathbf{M}^{(K)*}\right) = \mathbf{F}^{*}\left(\bar{\mathbf{B}}^{*}\right)^{\top}.$$

Under the Condition (C2), the singular values of the stacked signal matrix scale as

$$\sigma^2\left(\bar{\mathbf{M}}^{*}\right) \asymp Kn\sigma_{\mathbf{B}}^2.$$

We know that the total reconstruction error for the stacked matrix is bounded by

$$\left\|\bar{\mathbf{M}}-\bar{\mathbf{M}}^{*}\right\|_F^2 = \sum_{k=1}^{K}\left\|\hat{\mathbf{M}}^{(k)}-\mathbf{M}^{(k)*}\right\|_F^2 \lesssim K\left\{(\mathrm{err})^2+s\lambda\right\}$$

Applying the perturbation bound for the factor matrix $\hat{\mathbf{F}}$, we obtain

$$\frac{1}{n}\left\|\hat{\mathbf{F}}\mathbf{O}-\mathbf{F}^{*}\right\|_F^2 \lesssim \frac{K\left\{(\mathrm{err})^2+s\lambda\right\}}{Kn\sigma_{\mathbf{B}}^2},$$

which simplifies to

$$\frac{1}{\sqrt{n}}\left\|\hat{\mathbf{F}}\mathbf{O}-\mathbf{F}^{*}\right\|_F \lesssim \frac{\mathrm{err}+\sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}}.$$

We now turn to the analysis of the $\hat{\mathbf{B}}^{(k)}$. This process follows the logic of Theorem 3.3, except that

$$\left\|\left(\mathbf{F}^* - \hat{\mathbf{F}}\mathbf{O}\right)\left(\bar{\mathbf{B}}^*\right)^\top\right\|_F \leq \left\|\mathbf{F}^* - \hat{\mathbf{F}}\mathbf{O}\right\|_F \left\|\bar{\mathbf{B}}^*\right\|_F \lesssim \sqrt{Kp}\left\|\mathbf{F}^* - \hat{\mathbf{F}}\mathbf{O}\right\|_F.$$

Substituting this into the error decomposition yields

$$\begin{aligned}
\frac{1}{Kp}\left\|\hat{\mathbf{B}}\mathbf{O} - \bar{\mathbf{B}}^*\right\|_F^2 &\leq \frac{1}{n}\left\|\hat{\mathbf{F}}\mathbf{O} - \mathbf{F}^*\right\|_F^2 + \frac{1}{Knp}\left\|\hat{\bar{\mathbf{M}}} - \bar{\mathbf{M}}^*\right\|_F^2 \\
&\lesssim \frac{(\text{err})^2 + s\lambda}{n\sigma_{\mathbf{B}}^2} + \frac{1}{Knp}\sum_{k=1}^{K}\left\|\mathbf{M}^{(k)*} - \hat{\mathbf{M}}^{(k)*}\right\|_F^2 \\
&\lesssim \frac{(\text{err})^2 + s\lambda}{n\sigma_{\mathbf{B}}^2}.
\end{aligned}$$

This implies a bound on the average squared error of $\hat{\mathbf{B}}^{(k)}$,

$$\frac{1}{p}\frac{1}{K}\sum_{k=1}^{K}\left\|\hat{\mathbf{B}}^{(k)}\mathbf{O} - \mathbf{B}^{(k)*}\right\|_F^2 \lesssim \frac{(\text{err})^2 + s\lambda}{n\sigma_{\mathbf{B}}^2}.$$

Using the inequality relating the sum of squares to the square of the sum

$$K\sum_{k=1}^{K}\left\|\hat{\mathbf{B}}^{(k)}\mathbf{O} - \mathbf{B}^{(k)*}\right\|_F^2 \geq \left(\sum_{k=1}^{K}\left\|\hat{\mathbf{B}}^{(k)}\mathbf{O} - \mathbf{B}^{(k)*}\right\|\right)^2,$$

we deduce that the average estimation error for $\hat{\mathbf{B}}^{(k)}$ satisfies,

$$\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{\sqrt{p}}\left\|\hat{\mathbf{B}}^{(k)}\mathbf{O} - \mathbf{B}^{(k)*}\right\|_F\right) \lesssim \frac{\text{err} + \sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}}.$$

Finally, we analyze the structural changes. By the triangle inequality, the error of the differences is bounded by the sum of the individual estimation errors

$$\begin{aligned}
&\sum_{k=2}^{K}\left\|\hat{\Delta\mathbf{B}}^{(k)} - \Delta\mathbf{B}^{(k)*}\right\|_F \\
&= \sum_{k=2}^{K}\left\|\left(\hat{\mathbf{B}}^{(k)} - \hat{\mathbf{B}}^{(k-1)}\right)\mathbf{O} - \left(\mathbf{B}^{(k)*} - \mathbf{B}^{(k-1)*}\right)\right\|_F \\
&\leq \sum_{k=2}^{K}\left\{\left\|\hat{\mathbf{B}}^{(k)}\mathbf{O} - \mathbf{B}^{(k)*}\right\|_F + \left\|\hat{\mathbf{B}}^{(k-1)}\mathbf{O} - \mathbf{B}^{(k-1)*}\right\|_F\right\} \\
&< 2\sum_{k=1}^{K}\left\|\hat{\mathbf{B}}^{(k)}\mathbf{O} - \mathbf{B}^{(k)*}\right\|_F.
\end{aligned}$$

Combining this with the average bound derived in Equation (3),

$$\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{\sqrt{p}}\left\|\hat{\mathbf{B}}^{(k)}\mathbf{O} - \mathbf{B}^{(k)*}\right\|_F\right) \lesssim \frac{\text{err} + \sqrt{s\lambda}}{\sqrt{n}\sigma_{\mathbf{B}}},$$

we conclude that

$$\frac{1}{K-1}\sum_{k=2}^{K}\left\|\hat{\Delta\mathbf{B}}^{(k)} - \Delta\mathbf{B}^{(k)*}\right\|_F \lesssim \sqrt{\frac{p}{n\sigma_{\mathbf{B}}^2}}\left(\text{err} + \sqrt{s\lambda}\right).$$

Adjusting for the sparsity level s, we arrive at the final rate

$$\frac{1}{K-1}\sum_{k=2}^{K}\left(\frac{1}{\sqrt{s}}\left\|\Delta\hat{\mathbf{B}}^{(k)}\mathbf{O}-\Delta\mathbf{B}^{(k)*}\right\|_F\right)\lesssim\frac{\sqrt{p}}{\sigma_{\mathbf{B}}}\cdot\frac{\mathrm{err}+\sqrt{s\lambda}}{\sqrt{ns}}.$$

Then the proof is finished. □

Proof of Theorem 4.2: The proof is similar to Theorem 3.6 and thus omitted. □

C Additional Results on Synthetic Data

C.1 Additional Results for Poisson Data

Table 5: Variable selection recall for Poisson data, with varying n, p, r^*, and s over 200 repetitions.

			$s=5$				$s=10$				$s=20$			
			N-Lasso		SMART		N-Lasso		SMART		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.641	0.249	0.730	0.221	0.708	0.181	0.793	0.169	0.815	0.125	0.878	0.092
		100	0.647	0.219	0.726	0.208	0.652	0.168	0.738	0.160	0.758	0.129	0.838	0.092
		300	0.649	0.210	0.733	0.198	0.640	0.169	0.706	0.150	0.636	0.135	0.712	0.124
	100	50	0.593	0.252	0.706	0.223	0.739	0.190	0.807	0.168	0.820	0.117	0.878	0.090
		100	0.624	0.233	0.726	0.205	0.650	0.163	0.746	0.154	0.743	0.125	0.821	0.102
		300	0.660	0.210	0.733	0.198	0.650	0.160	0.730	0.140	0.637	0.112	0.721	0.104
	300	50	0.618	0.247	0.721	0.211	0.682	0.193	0.777	0.163	0.812	0.114	0.878	0.080
		100	0.596	0.221	0.717	0.201	0.628	0.181	0.720	0.171	0.739	0.116	0.819	0.102
		300	0.638	0.230	0.749	0.198	0.631	0.155	0.715	0.139	0.655	0.115	0.742	0.108
4	50	50	0.861	0.164	0.963	0.083	0.883	0.116	0.973	0.049	0.940	0.055	0.988	0.024
		100	0.828	0.179	0.947	0.103	0.854	0.118	0.953	0.067	0.903	0.071	0.980	0.034
		300	0.819	0.167	0.956	0.088	0.829	0.119	0.945	0.069	0.831	0.092	0.953	0.050
	100	50	0.812	0.189	0.951	0.101	0.888	0.112	0.980	0.047	0.949	0.054	0.990	0.024
		100	0.808	0.189	0.951	0.099	0.865	0.119	0.965	0.060	0.909	0.075	0.978	0.035
		300	0.851	0.163	0.957	0.089	0.827	0.127	0.953	0.075	0.848	0.083	0.958	0.046
	300	50	0.804	0.192	0.949	0.090	0.855	0.136	0.973	0.054	0.947	0.054	0.992	0.021
		100	0.800	0.182	0.935	0.112	0.862	0.123	0.967	0.062	0.909	0.071	0.984	0.029
		300	0.853	0.158	0.960	0.085	0.840	0.113	0.951	0.065	0.850	0.082	0.958	0.045

Table 6: Variable selection precision for Poisson data, with varying n, p, r^*, and s over 200 repetitions.

			$s=5$				$s=10$				$s=20$			
			N-Lasso		SMART		N-Lasso		SMART		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.975	0.157	0.999	0.014	1.000	0.000	1.000	0.000	0.999	0.009	0.997	0.013
		100	0.985	0.122	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	0.999	0.006
		300	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.006	1.000	0.000	1.000	0.000
	100	50	0.979	0.141	0.987	0.103	1.000	0.000	0.999	0.012	0.998	0.010	0.996	0.016
		100	0.970	0.171	0.995	0.071	1.000	0.007	0.999	0.013	1.000	0.000	1.000	0.000
		300	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
	300	50	0.977	0.144	0.991	0.056	0.990	0.053	0.988	0.059	0.993	0.025	0.988	0.037
		100	0.975	0.157	0.995	0.040	1.000	0.000	0.997	0.039	0.999	0.016	0.999	0.010
		300	0.990	0.100	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
4	50	50	0.998	0.021	0.996	0.026	0.999	0.010	0.998	0.013	0.997	0.014	0.992	0.021
		100	1.000	0.000	1.000	0.000	0.999	0.011	0.999	0.010	0.998	0.010	0.995	0.016
		300	0.999	0.020	0.998	0.023	0.999	0.009	0.998	0.016	0.999	0.009	0.998	0.015
	100	50	1.000	0.000	0.998	0.017	1.000	0.006	0.998	0.013	0.999	0.008	0.993	0.019
		100	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	0.999	0.005	0.998	0.011
		300	1.000	0.000	1.000	0.000	1.000	0.000	0.998	0.013	0.999	0.006	0.999	0.007
	300	50	0.994	0.072	0.999	0.012	1.000	0.000	0.999	0.011	0.999	0.012	0.992	0.022
		100	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
		300	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.004	1.000	0.003

To provide a comprehensive dissection of the SMART method's internal behavior regarding variable selection mechanisms, parameter estimation, and model fit, we examine four sets of auxiliary metrics in this supplementary section, complementing the primary indicators presented in the main text.

Specifically, we report Precision and Recall for the difference loading $\mathbf{\Delta B}$ to investigate the model's specific tendencies in sparse recovery. While the F_1 score used in the main text offers a harmonic mean of these

Table 7: Subspace recovery accuracy ($\rho_{\min}$) of $\mathbf{F}$ for Poisson data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.997	0.002	0.997	0.002	0.997	0.002	0.996	0.002	0.996	0.002	0.996	0.002	0.995	0.003	0.996	0.003	0.996	0.003
		100	0.996	0.003	0.996	0.003	0.996	0.003	0.995	0.003	0.995	0.003	0.995	0.003	0.995	0.004	0.995	0.003	0.995	0.003
		300	0.995	0.004	0.995	0.004	0.995	0.004	0.995	0.004	0.995	0.004	0.995	0.004	0.995	0.004	0.995	0.004	0.995	0.004
	100	50	0.995	0.003	0.996	0.003	0.995	0.003	0.995	0.004	0.995	0.003	0.995	0.003	0.995	0.003	0.995	0.003	0.995	0.003
		100	0.998	0.002	0.998	0.002	0.998	0.002	0.997	0.002	0.998	0.002	0.997	0.002	0.997	0.002	0.997	0.002	0.997	0.002
		300	0.997	0.002	0.997	0.002	0.997	0.002	0.997	0.003	0.997	0.003	0.997	0.003	0.997	0.002	0.997	0.002	0.997	0.002
	300	50	0.992	0.006	0.992	0.006	0.992	0.006	0.991	0.005	0.992	0.005	0.991	0.005	0.991	0.005	0.992	0.005	0.992	0.005
		100	0.995	0.004	0.996	0.004	0.996	0.004	0.995	0.004	0.995	0.003	0.995	0.003	0.994	0.004	0.995	0.004	0.995	0.004
		300	0.999	0.001	0.999	0.001	0.999	0.001	0.999	0.001	0.999	0.001	0.999	0.001	0.998	0.002	0.998	0.001	0.998	0.001
4	50	50	0.991	0.004	0.991	0.004	0.991	0.004	0.990	0.005	0.990	0.005	0.990	0.005	0.987	0.006	0.988	0.006	0.987	0.006
		100	0.994	0.003	0.994	0.003	0.994	0.003	0.993	0.003	0.993	0.003	0.993	0.003	0.992	0.003	0.993	0.003	0.993	0.003
		300	0.992	0.004	0.992	0.004	0.992	0.004	0.991	0.004	0.991	0.004	0.991	0.004	0.991	0.004	0.991	0.004	0.991	0.004
	100	50	0.991	0.004	0.991	0.004	0.991	0.004	0.989	0.005	0.989	0.004	0.989	0.004	0.986	0.006	0.987	0.005	0.987	0.005
		100	0.994	0.003	0.994	0.003	0.994	0.003	0.993	0.003	0.994	0.003	0.994	0.003	0.991	0.004	0.991	0.004	0.991	0.004
		300	0.995	0.002	0.995	0.002	0.995	0.002	0.995	0.002	0.995	0.002	0.995	0.002	0.994	0.002	0.994	0.002	0.994	0.002
	300	50	0.985	0.007	0.985	0.007	0.985	0.007	0.981	0.008	0.982	0.008	0.982	0.008	0.980	0.010	0.981	0.009	0.981	0.009
		100	0.992	0.004	0.992	0.004	0.992	0.004	0.991	0.003	0.991	0.003	0.991	0.003	0.989	0.004	0.990	0.004	0.989	0.004
		300	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.995	0.001	0.995	0.001	0.995	0.001

Table 8: Estimation accuracy of latent parameters (R^2) for the control group under Poisson data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.989	0.004	0.978	0.006	0.981	0.005	0.988	0.003	0.969	0.007	0.975	0.006	0.988	0.004	0.957	0.010	0.970	0.007
		100	0.986	0.007	0.981	0.007	0.983	0.007	0.986	0.007	0.974	0.007	0.978	0.007	0.985	0.007	0.966	0.008	0.973	0.008
		300	0.985	0.012	0.984	0.011	0.984	0.011	0.985	0.009	0.982	0.009	0.983	0.009	0.985	0.009	0.977	0.009	0.979	0.009
	100	50	0.990	0.005	0.977	0.007	0.980	0.007	0.989	0.005	0.967	0.008	0.974	0.007	0.989	0.004	0.957	0.009	0.970	0.006
		100	0.993	0.003	0.987	0.003	0.988	0.003	0.993	0.003	0.980	0.004	0.984	0.004	0.992	0.003	0.971	0.005	0.978	0.005
		300	0.992	0.005	0.990	0.005	0.991	0.005	0.992	0.005	0.987	0.006	0.988	0.006	0.991	0.006	0.982	0.006	0.984	0.006
	300	50	0.984	0.010	0.970	0.013	0.973	0.012	0.983	0.010	0.959	0.013	0.966	0.012	0.983	0.008	0.950	0.012	0.963	0.010
		100	0.991	0.005	0.983	0.008	0.984	0.008	0.991	0.005	0.976	0.008	0.980	0.007	0.990	0.005	0.967	0.008	0.974	0.007
		300	0.997	0.001	0.995	0.002	0.995	0.002	0.997	0.001	0.991	0.002	0.992	0.002	0.996	0.002	0.985	0.003	0.988	0.003
4	50	50	0.984	0.005	0.975	0.006	0.981	0.005	0.983	0.005	0.965	0.006	0.977	0.005	0.981	0.005	0.950	0.007	0.971	0.005
		100	0.986	0.005	0.981	0.005	0.984	0.004	0.985	0.004	0.975	0.005	0.982	0.004	0.984	0.005	0.965	0.005	0.977	0.005
		300	0.979	0.008	0.978	0.008	0.979	0.008	0.978	0.007	0.975	0.007	0.977	0.007	0.977	0.008	0.970	0.008	0.974	0.008
	100	50	0.986	0.005	0.976	0.005	0.982	0.004	0.984	0.005	0.965	0.006	0.977	0.005	0.982	0.005	0.951	0.006	0.972	0.005
		100	0.989	0.003	0.983	0.003	0.986	0.003	0.988	0.004	0.977	0.004	0.984	0.004	0.986	0.004	0.966	0.005	0.978	0.004
		300	0.987	0.004	0.985	0.004	0.986	0.004	0.987	0.004	0.983	0.004	0.985	0.004	0.986	0.004	0.977	0.004	0.982	0.004
	300	50	0.981	0.006	0.970	0.007	0.976	0.006	0.978	0.006	0.959	0.007	0.971	0.007	0.977	0.007	0.945	0.008	0.967	0.007
		100	0.988	0.003	0.982	0.004	0.985	0.004	0.987	0.003	0.976	0.004	0.982	0.003	0.986	0.003	0.965	0.004	0.978	0.004
		300	0.992	0.002	0.990	0.002	0.991	0.002	0.991	0.002	0.987	0.002	0.989	0.002	0.991	0.002	0.982	0.002	0.987	0.002

Table 9: Estimation accuracy of latent parameters (R^2) for the treatment group under Poisson data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.986	0.007	0.977	0.008	0.980	0.008	0.984	0.008	0.968	0.009	0.974	0.008	0.983	0.010	0.962	0.011	0.971	0.009
		100	0.986	0.008	0.981	0.008	0.982	0.008	0.984	0.008	0.974	0.008	0.977	0.008	0.983	0.009	0.967	0.009	0.973	0.009
		300	0.984	0.011	0.984	0.011	0.984	0.011	0.984	0.010	0.981	0.009	0.982	0.009	0.984	0.009	0.977	0.009	0.979	0.009
	100	50	0.981	0.013	0.973	0.011	0.976	0.011	0.979	0.014	0.965	0.012	0.971	0.012	0.979	0.011	0.959	0.010	0.969	0.009
		100	0.991	0.006	0.986	0.005	0.987	0.005	0.989	0.007	0.979	0.006	0.983	0.005	0.987	0.007	0.971	0.006	0.977	0.006
		300	0.991	0.006	0.990	0.006	0.991	0.006	0.990	0.007	0.987	0.006	0.988	0.006	0.989	0.006	0.982	0.006	0.984	0.006
	300	50	0.972	0.020	0.965	0.016	0.969	0.016	0.967	0.020	0.955	0.016	0.961	0.015	0.969	0.018	0.951	0.015	0.960	0.014
		100	0.983	0.014	0.980	0.010	0.982	0.010	0.980	0.014	0.974	0.009	0.977	0.009	0.978	0.015	0.965	0.011	0.972	0.011
		300	0.996	0.004	0.994	0.002	0.995	0.002	0.994	0.004	0.991	0.003	0.992	0.003	0.991	0.006	0.985	0.004	0.987	0.004
4	50	50	0.980	0.009	0.969	0.009	0.976	0.008	0.976	0.009	0.958	0.010	0.970	0.009	0.967	0.012	0.942	0.011	0.959	0.011
		100	0.983	0.006	0.978	0.006	0.981	0.006	0.980	0.007	0.971	0.007	0.977	0.006	0.975	0.009	0.958	0.008	0.970	0.008
		300	0.978	0.008	0.977	0.008	0.978	0.008	0.976	0.008	0.973	0.008	0.975	0.008	0.973	0.009	0.967	0.008	0.971	0.008
	100	50	0.982	0.007	0.971	0.008	0.977	0.007	0.976	0.010	0.957	0.011	0.969	0.010	0.969	0.011	0.943	0.011	0.960	0.010
		100	0.987	0.005	0.980	0.005	0.984	0.005	0.984	0.006	0.973	0.006	0.980	0.006	0.977	0.008	0.960	0.007	0.972	0.007
		300	0.986	0.004	0.984	0.004	0.985	0.004	0.984	0.005	0.981	0.004	0.983	0.004	0.981	0.005	0.974	0.005	0.979	0.005
	300	50	0.972	0.012	0.961	0.012	0.968	0.011	0.966	0.012	0.948	0.012	0.960	0.012	0.961	0.013	0.935	0.012	0.953	0.012
		100	0.984	0.007	0.978	0.006	0.982	0.006	0.981	0.006	0.970	0.006	0.977	0.006	0.975	0.008	0.957	0.007	0.969	0.007
		300	0.991	0.002	0.989	0.002	0.990	0.002	0.989	0.003	0.985	0.003	0.987	0.003	0.987	0.004	0.978	0.003	0.983	0.003

aspects, independent analysis allows for finer granular insight. Based on the estimated support set $\hat{\mathcal{S}} = \{j : \|\hat{\Delta}\boldsymbol{b}_j\|_2 > 10^{-5}\}$ and the true support set $\mathcal{S}^*$, Precision is defined as $|\mathcal{S}^* \cap \hat{\mathcal{S}}|/(|\hat{\mathcal{S}}| + \epsilon)$, measuring the proportion of identified difference variables that correspond to true non-zero signals. High precision indicates a low False Positive Rate. Conversely, Recall is defined as $|\mathcal{S}^* \cap \hat{\mathcal{S}}|/(|\mathcal{S}^*| + \epsilon)$, where ϵ is a small constant to ensure numerical stability, quantifying the proportion of actual difference variables successfully captured by the model. The specific results for these metrics are detailed in Table 5 (Recall) and Table 6 (Precision).

Additionally, the subspace recovery results for the shared representation matrix $\mathbf{F}$, which were omitted from the main text, are provided in Table 7.

Finally, we evaluate the model's parameter estimation capability via R^2. This metric directly assesses the model's ability to reconstruct the true parameter matrix $\mathbf{M}^*$. Since the true parameters $\mathbf{M}^{(k)*} = \mathbf{F}(\mathbf{B}^{(k)*})^\top$ are known in the simulation, we calculate the denoised coefficient of determination:

$$R^2(\mathbf{M}^*, \hat{\mathbf{M}}) = 1 - \frac{\sum_{i,j}(m^*_{ij} - \hat{m}_{ij})^2}{\sum_{i,j}(m^*_{ij} - \bar{m}^*)^2},$$

where $\bar{m}^* = (np)^{-1}\sum_{i=1}^n \sum_{j=1}^p m^*_{ij}$. For Poisson and Gaussian distributions, $\mathbf{M}$ corresponds to the linear predictor, while for the Binary distribution, it corresponds to the Logit probability. An R^2 value approaching 1 indicates that the model has effectively stripped away observational noise to reveal the underlying data generation mechanism. The performance is summarized in Table 8 (Environment 1) and Table 9 (Environment 2).

C.2 Results for Gaussian Data

Table 10: Variable selection performance (F_1 score) for Gaussian data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$				$s = 10$				$s = 20$			
			N-Lasso		SMART		N-Lasso		SMART		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.888	0.128	0.913	0.116	0.918	0.083	0.938	0.072	0.952	0.038	0.960	0.033
		100	0.895	0.112	0.919	0.092	0.884	0.100	0.910	0.081	0.930	0.049	0.944	0.043
		300	0.880	0.122	0.912	0.101	0.870	0.088	0.902	0.072	0.882	0.062	0.916	0.051
	100	50	0.892	0.110	0.914	0.097	0.914	0.082	0.936	0.070	0.954	0.037	0.965	0.032
		100	0.887	0.116	0.918	0.098	0.895	0.092	0.924	0.066	0.934	0.051	0.953	0.039
		300	0.898	0.112	0.929	0.090	0.876	0.077	0.909	0.071	0.880	0.061	0.908	0.050
	300	50	0.898	0.118	0.924	0.102	0.933	0.065	0.951	0.056	0.956	0.038	0.967	0.033
		100	0.870	0.141	0.900	0.129	0.886	0.092	0.913	0.071	0.930	0.046	0.950	0.038
		300	0.892	0.105	0.915	0.090	0.885	0.081	0.913	0.068	0.889	0.056	0.917	0.048
4	50	50	0.977	0.051	0.990	0.032	0.985	0.027	0.980	0.034	0.989	0.017	0.965	0.027
		100	0.977	0.053	0.993	0.027	0.982	0.032	0.993	0.020	0.988	0.018	0.982	0.021
		300	0.976	0.053	0.974	0.115	0.980	0.033	0.993	0.019	0.979	0.025	0.993	0.013
	100	50	0.978	0.050	0.990	0.035	0.990	0.025	0.997	0.013	0.996	0.011	0.999	0.006
		100	0.980	0.046	0.993	0.030	0.985	0.029	0.995	0.016	0.992	0.016	0.998	0.007
		300	0.978	0.050	0.993	0.027	0.978	0.032	0.993	0.019	0.981	0.021	0.995	0.011
	300	50	0.988	0.037	0.994	0.025	0.992	0.021	0.999	0.005	0.996	0.011	1.000	0.003
		100	0.979	0.048	0.995	0.023	0.984	0.030	0.994	0.017	0.993	0.013	0.998	0.008
		300	0.977	0.052	0.991	0.035	0.979	0.033	0.995	0.018	0.981	0.022	0.994	0.012

For the Gaussian distribution experiments, the evaluation metrics are consistent with those in the main text, with the addition of the Minimum Canonical Correlation for $\mathbf{F}$. The variable selection performance F_1 Score is reported in Table 10. Subspace recovery accuracies for $\mathbf{F}$, $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ are detailed in Tables 11, 12, and 13.

C.3 Results for Binary Data

For the Binary distribution experiments, the evaluation metrics are identical to those used in the Gaussian setting. The variable selection metrics (F_1 Score) are presented in Table 14. Subspace recovery performance is shown in Tables 15, 16 and 17.

Table 11: Subspace recovery accuracy ($\rho_{\min}$) of $\mathbf{F}$ for Gaussian data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.996	0.001	0.996	0.001	0.996	0.001	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001
		100	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000
		300	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000
	100	50	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001
		100	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000
		300	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000
	300	50	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.000	0.997	0.000	0.997	0.000	0.997	0.000	0.997	0.000	0.997	0.000
		100	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000
		300	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000
4	50	50	0.995	0.001	0.995	0.001	0.995	0.001	0.995	0.001	0.995	0.001	0.995	0.001	0.996	0.001	0.996	0.001	0.996	0.001
		100	0.998	0.001	0.998	0.001	0.998	0.001	0.998	0.001	0.998	0.001	0.998	0.001	0.998	0.001	0.998	0.001	0.998	0.001
		300	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000
	100	50	0.995	0.001	0.995	0.001	0.995	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001
		100	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000
		300	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000
	300	50	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001
		100	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000
		300	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000

Table 12: Subspace recovery accuracy ($\rho_{\min}$) of $\mathbf{B}^{(1)}$ for Gaussian data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.992	0.002	0.990	0.003	0.992	0.002	0.993	0.002	0.987	0.004	0.991	0.003	0.992	0.002	0.983	0.005	0.988	0.003
		100	0.993	0.001	0.993	0.002	0.994	0.001	0.993	0.001	0.991	0.002	0.992	0.002	0.993	0.001	0.988	0.003	0.991	0.002
		300	0.993	0.001	0.995	0.001	0.996	0.001	0.993	0.001	0.994	0.001	0.995	0.001	0.993	0.001	0.992	0.001	0.994	0.001
	100	50	0.996	0.001	0.992	0.002	0.994	0.002	0.996	0.001	0.989	0.003	0.993	0.002	0.996	0.001	0.986	0.004	0.991	0.002
		100	0.997	0.001	0.995	0.001	0.996	0.001	0.997	0.001	0.993	0.002	0.994	0.001	0.997	0.001	0.990	0.002	0.993	0.001
		300	0.997	0.000	0.997	0.000	0.997	0.000	0.997	0.000	0.996	0.001	0.997	0.001	0.997	0.000	0.994	0.001	0.996	0.001
	300	50	0.999	0.000	0.993	0.002	0.995	0.001	0.999	0.000	0.991	0.002	0.994	0.002	0.999	0.000	0.988	0.003	0.993	0.002
		100	0.999	0.000	0.996	0.001	0.997	0.001	0.999	0.000	0.994	0.001	0.996	0.001	0.999	0.000	0.991	0.002	0.994	0.001
		300	0.999	0.000	0.998	0.000	0.998	0.000	0.999	0.000	0.997	0.000	0.998	0.000	0.999	0.000	0.995	0.001	0.997	0.000
4	50	50	0.989	0.003	0.986	0.004	0.991	0.003	0.989	0.002	0.981	0.005	0.990	0.003	0.990	0.003	0.976	0.007	0.988	0.003
		100	0.990	0.002	0.990	0.002	0.993	0.002	0.990	0.002	0.988	0.003	0.992	0.002	0.990	0.002	0.982	0.004	0.991	0.002
		300	0.991	0.001	0.994	0.001	0.995	0.001	0.991	0.001	0.993	0.001	0.995	0.001	0.991	0.001	0.990	0.002	0.994	0.001
	100	50	0.995	0.001	0.990	0.003	0.994	0.002	0.995	0.001	0.986	0.004	0.994	0.002	0.995	0.001	0.980	0.006	0.992	0.002
		100	0.996	0.001	0.993	0.002	0.996	0.001	0.996	0.001	0.991	0.002	0.995	0.001	0.996	0.001	0.987	0.002	0.994	0.001
		300	0.996	0.000	0.996	0.001	0.997	0.000	0.996	0.001	0.995	0.001	0.997	0.001	0.996	0.000	0.993	0.001	0.996	0.001
	300	50	0.998	0.000	0.992	0.002	0.996	0.001	0.998	0.000	0.988	0.003	0.996	0.001	0.998	0.000	0.983	0.004	0.994	0.001
		100	0.999	0.000	0.994	0.001	0.997	0.001	0.999	0.000	0.993	0.001	0.997	0.001	0.999	0.000	0.989	0.002	0.996	0.001
		300	0.999	0.000	0.998	0.000	0.999	0.000	0.999	0.000	0.996	0.001	0.998	0.000	0.999	0.000	0.995	0.001	0.998	0.000

Table 13: Subspace recovery accuracy ($\rho_{\min}$) of $\mathbf{B}^{(2)}$ for Gaussian data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.993	0.002	0.990	0.003	0.992	0.002	0.993	0.002	0.988	0.004	0.991	0.003	0.994	0.002	0.987	0.004	0.991	0.003
		100	0.993	0.001	0.993	0.002	0.994	0.002	0.993	0.001	0.990	0.002	0.992	0.002	0.993	0.002	0.988	0.003	0.991	0.002
		300	0.993	0.001	0.995	0.001	0.995	0.001	0.993	0.001	0.994	0.001	0.995	0.001	0.993	0.001	0.992	0.001	0.993	0.001
	100	50	0.996	0.001	0.992	0.003	0.994	0.002	0.996	0.001	0.990	0.003	0.993	0.002	0.997	0.001	0.989	0.003	0.993	0.002
		100	0.996	0.001	0.994	0.002	0.995	0.002	0.996	0.001	0.992	0.002	0.994	0.002	0.996	0.001	0.990	0.002	0.993	0.002
		300	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001	0.996	0.001	0.996	0.001	0.997	0.001	0.994	0.001	0.995	0.001
	300	50	0.999	0.001	0.993	0.002	0.995	0.002	0.999	0.001	0.992	0.002	0.994	0.002	0.998	0.001	0.990	0.003	0.994	0.002
		100	0.999	0.001	0.995	0.002	0.996	0.001	0.998	0.001	0.993	0.002	0.995	0.001	0.998	0.001	0.992	0.002	0.994	0.001
		300	0.999	0.000	0.998	0.001	0.998	0.001	0.999	0.000	0.997	0.001	0.997	0.001	0.998	0.001	0.995	0.001	0.996	0.001
4	50	50	0.991	0.003	0.988	0.003	0.992	0.002	0.991	0.002	0.984	0.005	0.991	0.003	0.992	0.002	0.982	0.006	0.991	0.003
		100	0.991	0.002	0.991	0.002	0.994	0.001	0.991	0.002	0.989	0.003	0.993	0.002	0.992	0.002	0.986	0.004	0.992	0.002
		300	0.992	0.001	0.994	0.001	0.995	0.001	0.992	0.001	0.993	0.001	0.995	0.001	0.992	0.001	0.991	0.001	0.994	0.001
	100	50	0.995	0.001	0.991	0.003	0.995	0.001	0.996	0.001	0.988	0.003	0.995	0.001	0.996	0.001	0.985	0.005	0.994	0.002
		100	0.996	0.001	0.993	0.001	0.996	0.001	0.996	0.001	0.992	0.002	0.995	0.001	0.996	0.001	0.990	0.002	0.995	0.001
		300	0.996	0.000	0.996	0.001	0.997	0.000	0.996	0.000	0.995	0.001	0.997	0.000	0.996	0.000	0.994	0.001	0.996	0.000
	300	50	0.999	0.000	0.993	0.002	0.997	0.001	0.999	0.000	0.990	0.003	0.996	0.001	0.999	0.000	0.988	0.004	0.996	0.001
		100	0.999	0.000	0.995	0.001	0.998	0.001	0.999	0.000	0.993	0.001	0.997	0.001	0.999	0.000	0.991	0.002	0.997	0.001
		300	0.999	0.000	0.998	0.000	0.999	0.000	0.999	0.000	0.996	0.001	0.998	0.000	0.999	0.000	0.995	0.001	0.998	0.000

Table 14: Variable selection performance (F_1 score) for binary data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$				$s = 10$				$s = 20$			
			N-Lasso		SMART		N-Lasso		SMART		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.224	0.249	0.266	0.240	0.345	0.248	0.397	0.248	0.543	0.261	0.597	0.238
		100	0.173	0.218	0.223	0.233	0.207	0.206	0.287	0.242	0.295	0.253	0.378	0.277
		300	0.142	0.195	0.190	0.200	0.156	0.167	0.180	0.172	0.155	0.149	0.200	0.178
	100	50	0.251	0.258	0.369	0.270	0.355	0.262	0.474	0.267	0.495	0.246	0.657	0.200
		100	0.174	0.211	0.268	0.230	0.204	0.185	0.310	0.216	0.217	0.160	0.329	0.207
		300	0.148	0.199	0.212	0.224	0.148	0.155	0.223	0.183	0.151	0.113	0.219	0.132
	300	50	0.129	0.189	0.213	0.225	0.182	0.177	0.277	0.214	0.275	0.203	0.443	0.219
		100	0.061	0.132	0.115	0.179	0.084	0.122	0.145	0.150	0.107	0.106	0.175	0.138
		300	0.029	0.101	0.055	0.132	0.041	0.090	0.086	0.123	0.053	0.078	0.089	0.098
4	50	50	0.168	0.186	0.171	0.190	0.312	0.217	0.302	0.212	0.560	0.177	0.580	0.141
		100	0.107	0.158	0.139	0.195	0.221	0.204	0.218	0.207	0.398	0.202	0.409	0.155
		300	0.060	0.132	0.099	0.171	0.103	0.156	0.131	0.185	0.235	0.213	0.262	0.201
	100	50	0.403	0.208	0.370	0.145	0.590	0.132	0.495	0.073	0.753	0.050	0.672	0.037
		100	0.501	0.270	0.574	0.175	0.590	0.194	0.671	0.109	0.699	0.124	0.736	0.070
		300	0.344	0.266	0.611	0.221	0.345	0.216	0.664	0.143	0.404	0.150	0.718	0.098
	300	50	0.655	0.244	0.884	0.124	0.754	0.165	0.915	0.069	0.888	0.068	0.933	0.037
		100	0.311	0.260	0.693	0.205	0.388	0.202	0.778	0.139	0.465	0.176	0.861	0.085
		300	0.154	0.202	0.513	0.279	0.164	0.162	0.502	0.194	0.195	0.129	0.541	0.147

Table 15: Subspace recovery accuracy ($\rho_{\min}$) of $\mathbf{F}$ for binary data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.918	0.024	0.919	0.023	0.919	0.023	0.924	0.022	0.925	0.021	0.925	0.022	0.932	0.015	0.933	0.015	0.933	0.015
		100	0.959	0.010	0.959	0.010	0.959	0.010	0.960	0.011	0.960	0.011	0.960	0.011	0.962	0.009	0.963	0.008	0.963	0.008
		300	0.987	0.003	0.987	0.003	0.987	0.003	0.987	0.003	0.987	0.003	0.987	0.003	0.987	0.003	0.987	0.003	0.987	0.003
	100	50	0.933	0.012	0.933	0.012	0.933	0.012	0.937	0.012	0.937	0.012	0.937	0.012	0.942	0.011	0.942	0.011	0.942	0.011
		100	0.967	0.006	0.967	0.006	0.967	0.006	0.966	0.006	0.966	0.006	0.966	0.006	0.969	0.005	0.969	0.005	0.969	0.005
		300	0.989	0.002	0.989	0.002	0.989	0.002	0.989	0.002	0.989	0.002	0.989	0.002	0.989	0.002	0.989	0.002	0.989	0.002
	300	50	0.942	0.008	0.942	0.008	0.942	0.008	0.943	0.008	0.943	0.008	0.943	0.008	0.947	0.007	0.947	0.007	0.947	0.007
		100	0.970	0.003	0.970	0.003	0.970	0.003	0.971	0.003	0.971	0.003	0.971	0.003	0.973	0.003	0.973	0.003	0.973	0.003
		300	0.990	0.001	0.990	0.001	0.990	0.001	0.990	0.001	0.990	0.001	0.990	0.001	0.991	0.001	0.991	0.001	0.991	0.001
4	50	50	0.827	0.087	0.833	0.082	0.832	0.083	0.838	0.061	0.845	0.048	0.844	0.049	0.860	0.042	0.864	0.040	0.864	0.040
		100	0.921	0.018	0.923	0.017	0.922	0.017	0.923	0.018	0.925	0.017	0.925	0.017	0.927	0.018	0.928	0.018	0.928	0.018
		300	0.976	0.005	0.976	0.005	0.976	0.005	0.976	0.005	0.976	0.005	0.976	0.005	0.977	0.005	0.977	0.005	0.977	0.005
	100	50	0.880	0.024	0.883	0.024	0.882	0.024	0.886	0.020	0.888	0.019	0.887	0.020	0.898	0.018	0.900	0.017	0.899	0.017
		100	0.945	0.008	0.946	0.008	0.946	0.008	0.948	0.008	0.948	0.008	0.948	0.008	0.949	0.007	0.949	0.007	0.949	0.007
		300	0.984	0.002	0.984	0.002	0.984	0.002	0.984	0.002	0.984	0.002	0.984	0.002	0.984	0.002	0.984	0.002	0.984	0.002
	300	50	0.907	0.012	0.908	0.012	0.908	0.012	0.912	0.010	0.913	0.010	0.913	0.010	0.917	0.011	0.918	0.011	0.918	0.011
		100	0.959	0.004	0.959	0.004	0.959	0.004	0.960	0.004	0.960	0.004	0.960	0.004	0.961	0.004	0.961	0.004	0.961	0.004
		300	0.987	0.001	0.987	0.001	0.987	0.001	0.987	0.001	0.987	0.001	0.987	0.001	0.988	0.001	0.988	0.001	0.988	0.001

Table 16: Subspace recovery accuracy ($\rho_{\min}$) of $\mathbf{B}^{(1)}$ for binary data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.836	0.040	0.894	0.027	0.895	0.026	0.841	0.040	0.879	0.031	0.884	0.029	0.846	0.035	0.860	0.034	0.871	0.033
		100	0.869	0.028	0.916	0.016	0.917	0.016	0.866	0.028	0.908	0.016	0.910	0.016	0.869	0.025	0.892	0.020	0.897	0.020
		300	0.887	0.014	0.935	0.009	0.935	0.009	0.886	0.016	0.931	0.010	0.931	0.010	0.887	0.015	0.925	0.009	0.926	0.009
	100	50	0.915	0.019	0.934	0.016	0.936	0.016	0.916	0.018	0.916	0.020	0.921	0.020	0.915	0.020	0.888	0.025	0.902	0.025
		100	0.933	0.011	0.953	0.008	0.954	0.008	0.931	0.011	0.941	0.011	0.943	0.011	0.933	0.010	0.922	0.013	0.925	0.012
		300	0.943	0.006	0.967	0.004	0.967	0.004	0.942	0.006	0.962	0.004	0.962	0.004	0.942	0.007	0.956	0.005	0.956	0.005
	300	50	0.973	0.006	0.962	0.011	0.963	0.010	0.973	0.006	0.940	0.014	0.942	0.014	0.972	0.006	0.901	0.022	0.910	0.021
		100	0.977	0.003	0.976	0.006	0.976	0.006	0.977	0.003	0.964	0.008	0.964	0.008	0.978	0.003	0.942	0.011	0.943	0.011
		300	0.981	0.002	0.986	0.002	0.986	0.002	0.981	0.002	0.981	0.002	0.982	0.002	0.981	0.002	0.974	0.004	0.974	0.004
4	50	50	0.704	0.106	0.797	0.080	0.798	0.081	0.696	0.092	0.778	0.067	0.779	0.069	0.718	0.078	0.772	0.061	0.773	0.059
		100	0.783	0.038	0.863	0.022	0.863	0.022	0.787	0.038	0.855	0.023	0.857	0.023	0.781	0.040	0.839	0.025	0.840	0.028
		300	0.831	0.020	0.900	0.013	0.900	0.013	0.833	0.020	0.899	0.013	0.899	0.013	0.832	0.022	0.892	0.014	0.893	0.014
	100	50	0.843	0.037	0.895	0.026	0.897	0.024	0.847	0.035	0.882	0.026	0.890	0.025	0.846	0.035	0.865	0.030	0.878	0.027
		100	0.893	0.016	0.931	0.011	0.935	0.010	0.897	0.017	0.921	0.013	0.931	0.012	0.896	0.016	0.903	0.015	0.922	0.012
		300	0.920	0.009	0.954	0.005	0.955	0.005	0.920	0.009	0.949	0.006	0.951	0.005	0.922	0.008	0.942	0.006	0.947	0.006
	300	50	0.951	0.009	0.943	0.014	0.957	0.010	0.951	0.010	0.925	0.017	0.952	0.011	0.951	0.011	0.893	0.027	0.944	0.013
		100	0.967	0.004	0.965	0.007	0.969	0.005	0.967	0.005	0.952	0.009	0.961	0.007	0.968	0.005	0.926	0.013	0.948	0.010
		300	0.975	0.002	0.982	0.002	0.982	0.002	0.974	0.002	0.977	0.003	0.978	0.003	0.975	0.002	0.968	0.004	0.970	0.003

Table 17: Subspace recovery accuracy ($\rho_{\min}$) of $\mathbf{B}^{(2)}$ for binary data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.856	0.037	0.894	0.025	0.896	0.025	0.874	0.036	0.886	0.032	0.891	0.032	0.889	0.029	0.882	0.032	0.892	0.030
		100	0.875	0.024	0.917	0.016	0.918	0.016	0.881	0.023	0.909	0.017	0.910	0.017	0.891	0.020	0.901	0.017	0.904	0.018
		300	0.888	0.015	0.934	0.009	0.934	0.009	0.889	0.015	0.931	0.009	0.931	0.009	0.894	0.013	0.926	0.009	0.926	0.009
	100	50	0.924	0.021	0.935	0.017	0.937	0.017	0.931	0.016	0.923	0.019	0.927	0.018	0.945	0.013	0.915	0.023	0.925	0.021
		100	0.938	0.010	0.953	0.009	0.954	0.009	0.938	0.011	0.943	0.009	0.944	0.009	0.945	0.010	0.931	0.012	0.933	0.012
		300	0.943	0.006	0.966	0.004	0.966	0.004	0.944	0.006	0.962	0.005	0.962	0.005	0.946	0.007	0.956	0.005	0.956	0.005
	300	50	0.972	0.007	0.959	0.012	0.959	0.012	0.974	0.008	0.944	0.014	0.945	0.015	0.978	0.006	0.931	0.016	0.937	0.016
		100	0.978	0.004	0.973	0.007	0.974	0.007	0.978	0.004	0.963	0.007	0.963	0.007	0.980	0.003	0.950	0.008	0.951	0.008
		300	0.981	0.002	0.985	0.002	0.985	0.002	0.981	0.002	0.981	0.003	0.981	0.003	0.982	0.002	0.974	0.003	0.974	0.003
4	50	50	0.723	0.094	0.796	0.084	0.798	0.082	0.749	0.075	0.794	0.060	0.794	0.063	0.788	0.060	0.806	0.051	0.812	0.053
		100	0.795	0.037	0.864	0.023	0.864	0.023	0.799	0.034	0.857	0.025	0.857	0.025	0.812	0.034	0.849	0.028	0.852	0.028
		300	0.833	0.021	0.900	0.014	0.900	0.014	0.839	0.020	0.899	0.013	0.899	0.013	0.839	0.022	0.892	0.014	0.893	0.014
	100	50	0.862	0.031	0.897	0.025	0.901	0.024	0.874	0.026	0.894	0.021	0.903	0.020	0.895	0.022	0.894	0.023	0.910	0.018
		100	0.900	0.016	0.930	0.012	0.934	0.011	0.905	0.014	0.924	0.012	0.932	0.011	0.912	0.013	0.914	0.013	0.929	0.011
		300	0.922	0.007	0.953	0.005	0.953	0.005	0.923	0.007	0.949	0.005	0.950	0.005	0.925	0.007	0.942	0.006	0.946	0.006
	300	50	0.954	0.010	0.946	0.014	0.957	0.012	0.957	0.010	0.935	0.015	0.955	0.012	0.962	0.009	0.925	0.020	0.958	0.012
		100	0.968	0.005	0.965	0.007	0.968	0.007	0.969	0.005	0.955	0.008	0.963	0.007	0.971	0.004	0.941	0.010	0.956	0.008
		300	0.975	0.002	0.980	0.003	0.981	0.003	0.975	0.002	0.975	0.003	0.976	0.003	0.976	0.002	0.968	0.004	0.969	0.004

Table 18: Variable selection performance (F_1 score) for Poisson data under missingness, with varying n, p, r^*, and ρ over 200 repetitions.

			$\rho = 0.1$				$\rho = 0.2$			
			N-Lasso		SMART		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std
2	50	50	0.925	0.086	0.951	0.063	0.885	0.113	0.922	0.086
		100	0.899	0.082	0.929	0.068	0.835	0.116	0.888	0.087
		300	0.842	0.115	0.884	0.095	0.761	0.147	0.827	0.123
	100	50	0.936	0.064	0.957	0.052	0.897	0.091	0.934	0.072
		100	0.886	0.100	0.919	0.082	0.852	0.118	0.894	0.105
		300	0.845	0.101	0.891	0.083	0.777	0.123	0.842	0.103
	300	50	0.928	0.074	0.957	0.049	0.891	0.097	0.931	0.069
		100	0.909	0.089	0.938	0.070	0.847	0.120	0.905	0.097
		300	0.850	0.098	0.896	0.078	0.782	0.110	0.851	0.087
4	50	50	0.959	0.052	0.995	0.016	0.931	0.062	0.987	0.027
		100	0.933	0.083	0.982	0.038	0.882	0.086	0.966	0.045
		300	0.896	0.078	0.972	0.040	0.820	0.109	0.945	0.054
	100	50	0.968	0.051	0.995	0.015	0.925	0.072	0.987	0.033
		100	0.931	0.072	0.986	0.028	0.876	0.093	0.974	0.036
		300	0.895	0.081	0.972	0.038	0.822	0.111	0.950	0.053
	300	50	0.964	0.049	0.996	0.014	0.908	0.089	0.989	0.023
		100	0.932	0.073	0.986	0.026	0.876	0.095	0.978	0.035
		300	0.884	0.088	0.974	0.042	0.815	0.120	0.947	0.054

Table 19: Subspace recovery ($\rho_{\min}$) of $\mathbf{F}$ for Poisson data under missingness, with varying n, p, r^*, and ρ over 200 repetitions.

			$\rho = 0.1$						$\rho = 0.2$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.987	0.004	0.987	0.004	0.987	0.004	0.987	0.004	0.987	0.004	0.987	0.004
		100	0.992	0.002	0.992	0.002	0.992	0.002	0.991	0.002	0.991	0.002	0.991	0.002
		300	0.995	0.001	0.995	0.001	0.995	0.001	0.994	0.001	0.994	0.001	0.994	0.001
	100	50	0.988	0.003	0.988	0.003	0.988	0.003	0.987	0.003	0.987	0.003	0.987	0.003
		100	0.992	0.002	0.992	0.002	0.992	0.002	0.991	0.002	0.991	0.002	0.991	0.002
		300	0.995	0.001	0.995	0.001	0.995	0.001	0.995	0.001	0.995	0.001	0.995	0.001
	300	50	0.988	0.003	0.988	0.003	0.988	0.003	0.987	0.003	0.987	0.003	0.987	0.003
		100	0.992	0.002	0.992	0.002	0.992	0.002	0.992	0.002	0.992	0.002	0.992	0.002
		300	0.995	0.001	0.995	0.001	0.995	0.001	0.995	0.001	0.995	0.001	0.995	0.001
4	50	50	0.973	0.010	0.973	0.010	0.973	0.010	0.970	0.012	0.970	0.011	0.970	0.012
		100	0.987	0.004	0.987	0.003	0.987	0.003	0.985	0.004	0.985	0.004	0.985	0.004
		300	0.994	0.001	0.994	0.001	0.994	0.001	0.993	0.001	0.993	0.001	0.993	0.001
	100	50	0.973	0.008	0.974	0.008	0.974	0.008	0.972	0.008	0.973	0.007	0.973	0.008
		100	0.988	0.003	0.988	0.003	0.988	0.003	0.987	0.003	0.987	0.003	0.987	0.003
		300	0.994	0.001	0.994	0.001	0.994	0.001	0.994	0.001	0.994	0.001	0.994	0.001
	300	50	0.975	0.007	0.976	0.007	0.976	0.007	0.974	0.007	0.974	0.007	0.974	0.007
		100	0.988	0.003	0.988	0.003	0.988	0.003	0.987	0.003	0.987	0.003	0.987	0.003
		300	0.994	0.001	0.994	0.001	0.994	0.001	0.994	0.001	0.994	0.001	0.994	0.001

Table 20: Subspace recovery ($\rho_{\min}$) of $\mathbf{B}^{(1)}$ for Poisson data under missingness, with varying n, p, r^*, and ρ over 200 repetitions.

			$\rho = 0.1$						$\rho = 0.2$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.984	0.004	0.967	0.010	0.976	0.007	0.982	0.005	0.960	0.012	0.971	0.008
		100	0.985	0.003	0.976	0.005	0.981	0.004	0.983	0.003	0.971	0.006	0.977	0.005
		300	0.985	0.003	0.983	0.003	0.985	0.003	0.984	0.003	0.981	0.003	0.983	0.003
	100	50	0.990	0.002	0.972	0.008	0.981	0.005	0.989	0.003	0.965	0.010	0.976	0.006
		100	0.991	0.002	0.980	0.004	0.985	0.003	0.990	0.002	0.976	0.005	0.982	0.004
		300	0.991	0.001	0.987	0.002	0.989	0.002	0.991	0.001	0.985	0.002	0.987	0.002
	300	50	0.994	0.001	0.976	0.006	0.984	0.004	0.994	0.001	0.968	0.009	0.979	0.005
		100	0.994	0.001	0.983	0.003	0.987	0.002	0.994	0.001	0.978	0.005	0.984	0.003
		300	0.995	0.001	0.989	0.001	0.991	0.001	0.994	0.001	0.988	0.002	0.990	0.001
4	50	50	0.973	0.007	0.940	0.018	0.965	0.010	0.968	0.009	0.922	0.022	0.955	0.013
		100	0.974	0.006	0.957	0.010	0.971	0.007	0.972	0.006	0.951	0.011	0.967	0.008
		300	0.976	0.004	0.974	0.004	0.978	0.004	0.973	0.005	0.970	0.005	0.974	0.005
	100	50	0.985	0.003	0.950	0.014	0.975	0.006	0.984	0.003	0.938	0.017	0.969	0.008
		100	0.986	0.003	0.966	0.007	0.981	0.003	0.985	0.003	0.960	0.008	0.977	0.005
		300	0.987	0.002	0.982	0.003	0.986	0.003	0.986	0.002	0.979	0.003	0.984	0.003
	300	50	0.993	0.001	0.956	0.011	0.982	0.004	0.993	0.001	0.943	0.014	0.976	0.006
		100	0.993	0.001	0.972	0.006	0.986	0.002	0.993	0.001	0.965	0.008	0.982	0.003
		300	0.994	0.001	0.986	0.002	0.991	0.001	0.994	0.001	0.984	0.002	0.990	0.001

Table 21: Subspace recovery ($\rho_{\min}$) of $\mathbf{B}^{(2)}$ for Poisson data under missingness, with varying n, p, r^*, and ρ over 200 repetitions.

			$\rho = 0.1$						$\rho = 0.2$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.971	0.011	0.947	0.016	0.955	0.014	0.969	0.011	0.939	0.017	0.949	0.015
		100	0.974	0.008	0.954	0.011	0.960	0.011	0.971	0.010	0.948	0.014	0.954	0.013
		300	0.979	0.005	0.970	0.007	0.972	0.007	0.978	0.005	0.967	0.007	0.969	0.007
	100	50	0.976	0.008	0.953	0.013	0.961	0.011	0.975	0.009	0.946	0.014	0.956	0.012
		100	0.978	0.007	0.958	0.010	0.963	0.009	0.976	0.008	0.950	0.012	0.957	0.011
		300	0.985	0.004	0.973	0.006	0.976	0.006	0.984	0.004	0.970	0.007	0.973	0.007
	300	50	0.980	0.008	0.956	0.012	0.964	0.010	0.979	0.008	0.948	0.012	0.958	0.010
		100	0.982	0.007	0.961	0.010	0.966	0.009	0.982	0.006	0.955	0.010	0.962	0.009
		300	0.988	0.004	0.975	0.006	0.978	0.006	0.987	0.004	0.973	0.006	0.975	0.006
4	50	50	0.943	0.019	0.903	0.023	0.926	0.020	0.938	0.021	0.889	0.024	0.917	0.021
		100	0.952	0.013	0.918	0.016	0.936	0.015	0.949	0.013	0.912	0.016	0.929	0.016
		300	0.965	0.007	0.950	0.008	0.957	0.008	0.962	0.008	0.947	0.009	0.954	0.009
	100	50	0.958	0.013	0.917	0.020	0.940	0.015	0.957	0.014	0.905	0.020	0.933	0.017
		100	0.963	0.010	0.928	0.013	0.945	0.012	0.962	0.010	0.923	0.015	0.941	0.014
		300	0.974	0.007	0.956	0.009	0.963	0.008	0.973	0.007	0.953	0.009	0.961	0.009
	300	50	0.967	0.009	0.923	0.015	0.947	0.011	0.968	0.009	0.911	0.018	0.942	0.013
		100	0.969	0.009	0.934	0.013	0.951	0.012	0.969	0.008	0.924	0.013	0.945	0.011
		300	0.980	0.006	0.960	0.008	0.968	0.008	0.980	0.006	0.958	0.008	0.966	0.008

C.4 Results for Poisson Data under Missingness

In this section, we evaluate the applicability and robustness of the SMART framework under incomplete count data scenarios. We assume a Missing Completely At Random (MCAR) mechanism. For the latent matrices $\mathbf{X}_{\text{full}}^{(k)}$ (where $k \in \{1, 2\}$), we introduce a binary mask matrix $\mathbf{\Omega}^{(k)} \in \{0, 1\}^{n \times p}$, where elements are sampled independently from a Bernoulli distribution with a missing rate ρ. The observed data are defined as the sparse matrix $\mathbf{X}_{\text{obs}}^{(k)} = \mathbf{X}_{\text{full}}^{(k)} \circ \mathbf{\Omega}^{(k)}$.

We investigate the capability of the two-stage SMART method to handle missingness with sample sizes and dimensions $(n, p) \in \{50, 100, 300\}^2$, latent ranks $r^* \in \{2, 4\}$, sparsity $s = 10$, and missing rates $\rho \in \{0.1, 0.2\}$. Unobserved entries are excluded from the optimization during training.

The methodology remains identical to the full-data case, with the sole exception of the CV procedure. To accommodate missingness, we employ an entry-wise CV where only the observed indices are partitioned into folds. In each fold, the designated validation entries are masked and treated as additional missing values during the Stage I CJMLE signal recovery. This allows the model to select λ by minimizing the prediction error specifically on observed but held-out entries. Once Stage I produces the completed matrix $\tilde{\mathbf{M}}_{\text{complete}}$, the Stage II attribution proceeds exactly as described in the complete-data framework. The results, summarizing variable selection performance and subspace recovery precision, are presented in Tables 18–21.

C.5 Results for One-step estimation

Table 22: Variable selection performance (F_1 score) for one-step estimation under missingness, with varying n, p, r^*, and ρ over 200 repetitions.

			$\rho = 0.1$				$\rho = 0.2$			
			N-Lasso		SMART		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std
2	50	50	0.902	0.090	0.930	0.076	0.844	0.121	0.892	0.099
		100	0.862	0.105	0.897	0.091	0.801	0.121	0.858	0.097
		300	0.850	0.089	0.882	0.087	0.816	0.103	0.861	0.089
	100	50	0.898	0.089	0.919	0.077	0.866	0.096	0.904	0.075
		100	0.878	0.089	0.910	0.068	0.817	0.118	0.866	0.098
		300	0.853	0.098	0.889	0.087	0.818	0.106	0.856	0.094
	300	50	0.898	0.082	0.929	0.072	0.849	0.118	0.890	0.096
		100	0.866	0.089	0.897	0.078	0.816	0.102	0.862	0.091
		300	0.858	0.088	0.893	0.072	0.820	0.092	0.867	0.077
4	50	50	0.981	0.028	0.990	0.022	0.954	0.057	0.985	0.027
		100	0.969	0.043	0.992	0.020	0.946	0.059	0.986	0.026
		300	0.962	0.047	0.989	0.025	0.936	0.056	0.979	0.035
	100	50	0.984	0.030	0.996	0.014	0.969	0.040	0.993	0.021
		100	0.977	0.036	0.994	0.017	0.956	0.053	0.987	0.028
		300	0.968	0.042	0.991	0.021	0.939	0.057	0.983	0.032
	300	50	0.982	0.034	0.996	0.014	0.968	0.048	0.992	0.025
		100	0.975	0.041	0.990	0.024	0.956	0.059	0.990	0.024
		300	0.966	0.042	0.990	0.021	0.952	0.049	0.986	0.028

Beyond the two-stage framework, we explore an efficient One-step Estimation strategy (Theorem 5.1) specifically designed for continuous (Gaussian) data under the MCAR assumption. This extension performs end-to-end structured decomposition and missing value imputation simultaneously by minimizing a regularized objective function directly on the observed samples:

$$\min_{\mathbf{F}, \mathbf{B}^{(1)}, \Delta\mathbf{B}} \sum_{k=1}^{2} \sum_{(i,j) \in \mathbf{\Omega}^{(k)}} \left(x_{ij}^{(k)} - m_{ij}^{(k)} \right)^2 + \lambda \sum_{j=1}^{p} \|\Delta \boldsymbol{b}_j\|_2$$

where $\mathbf{M}^{(1)} = \mathbf{F}(\mathbf{B}^{(1)})^\top$ and $\mathbf{M}^{(2)} = \mathbf{F}(\mathbf{B}^{(1)} + \Delta\mathbf{B})^\top$. This one-step approach circumvents the need for separate pre-imputation, providing a computationally efficient alternative for continuous data. Simulation results across different missing rates are presented in Tables 22–25.

Table 23: Subspace recovery ($\rho_{\min}$) of $\mathbf{F}$ for one-step estimation under missingness, with varying n, p, r^*, and ρ over 200 repetitions.

			$\rho = 0.1$						$\rho = 0.2$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001
		100	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.001	0.998	0.001	0.998	0.001
		300	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000
	100	50	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001
		100	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000
		300	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000
	300	50	0.997	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001	0.996	0.001
		100	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000
		300	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000
4	50	50	0.994	0.001	0.994	0.001	0.994	0.001	0.993	0.002	0.993	0.002	0.993	0.002
		100	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001	0.997	0.001
		300	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000
	100	50	0.995	0.001	0.995	0.001	0.995	0.001	0.994	0.001	0.994	0.001	0.994	0.001
		100	0.998	0.000	0.998	0.000	0.998	0.000	0.997	0.000	0.997	0.000	0.997	0.000
		300	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000
	300	50	0.995	0.001	0.995	0.001	0.995	0.001	0.995	0.001	0.995	0.001	0.995	0.001
		100	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000	0.998	0.000
		300	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000	0.999	0.000

Table 24: Subspace recovery ($\rho_{\min}$) of $\mathbf{B}^{(1)}$ for one-step estimation under missingness, with varying n, p, r^*, and ρ over 200 repetitions.

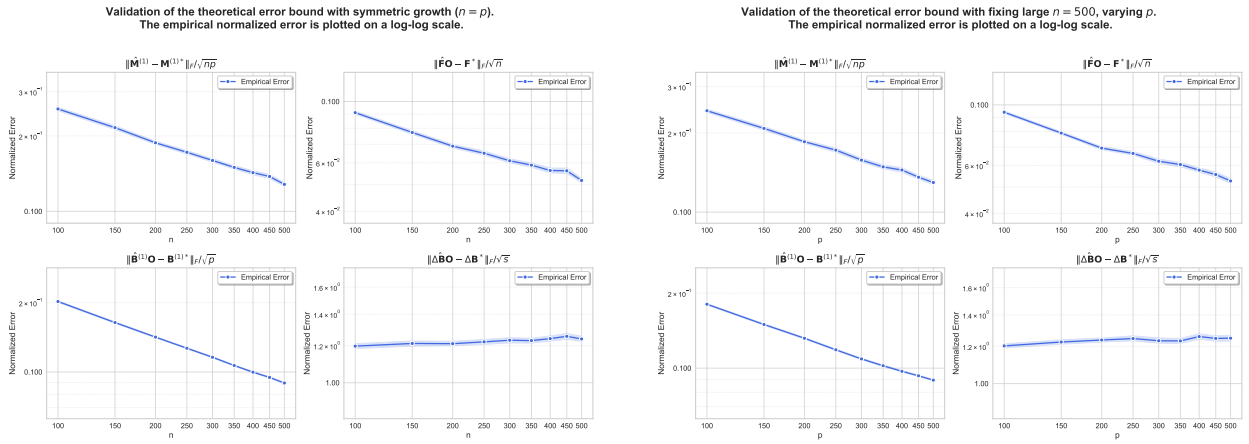
			$\rho = 0.1$						$\rho = 0.2$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.992	0.002	0.985	0.005	0.989	0.003	0.991	0.002	0.981	0.006	0.986	0.004
		100	0.992	0.002	0.989	0.003	0.991	0.002	0.991	0.002	0.987	0.003	0.989	0.002
		300	0.993	0.001	0.993	0.001	0.994	0.001	0.992	0.001	0.993	0.001	0.993	0.001
	100	50	0.996	0.001	0.987	0.003	0.991	0.002	0.996	0.001	0.984	0.005	0.988	0.004
		100	0.996	0.001	0.991	0.002	0.993	0.001	0.996	0.001	0.990	0.002	0.992	0.002
		300	0.997	0.000	0.996	0.001	0.996	0.001	0.996	0.000	0.995	0.001	0.996	0.001
	300	50	0.999	0.000	0.989	0.003	0.993	0.002	0.999	0.000	0.985	0.004	0.990	0.003
		100	0.999	0.000	0.993	0.001	0.995	0.001	0.999	0.000	0.991	0.002	0.993	0.001
		300	0.999	0.000	0.997	0.001	0.997	0.000	0.999	0.000	0.996	0.001	0.997	0.001
4	50	50	0.989	0.003	0.978	0.007	0.988	0.004	0.987	0.003	0.971	0.010	0.984	0.005
		100	0.990	0.002	0.985	0.003	0.991	0.002	0.988	0.002	0.983	0.004	0.989	0.003
		300	0.991	0.002	0.991	0.002	0.993	0.001	0.989	0.002	0.990	0.002	0.992	0.001
	100	50	0.995	0.001	0.982	0.005	0.992	0.002	0.994	0.001	0.977	0.007	0.990	0.003
		100	0.995	0.001	0.989	0.002	0.994	0.001	0.995	0.001	0.986	0.003	0.993	0.001
		300	0.996	0.000	0.994	0.001	0.996	0.001	0.995	0.001	0.994	0.001	0.996	0.001
	300	50	0.998	0.000	0.984	0.004	0.994	0.001	0.998	0.000	0.980	0.005	0.992	0.002
		100	0.999	0.000	0.991	0.002	0.996	0.001	0.998	0.000	0.989	0.002	0.995	0.001
		300	0.999	0.000	0.996	0.001	0.998	0.000	0.999	0.000	0.995	0.001	0.997	0.000

Table 25: Subspace recovery ($\rho_{\min}$) of $\mathbf{B}^{(2)}$ for one-step estimation under missingness, with varying n, p, r^*, and ρ over 200 repetitions.

			$\rho = 0.1$						$\rho = 0.2$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.993	0.002	0.988	0.004	0.991	0.003	0.992	0.002	0.984	0.005	0.988	0.004
		100	0.993	0.002	0.990	0.003	0.992	0.002	0.992	0.002	0.988	0.003	0.990	0.003
		300	0.993	0.001	0.994	0.001	0.994	0.001	0.992	0.001	0.993	0.001	0.994	0.001
	100	50	0.997	0.001	0.990	0.003	0.993	0.002	0.996	0.001	0.986	0.004	0.990	0.003
		100	0.997	0.001	0.992	0.002	0.994	0.001	0.996	0.001	0.991	0.002	0.993	0.002
		300	0.997	0.000	0.996	0.001	0.996	0.001	0.996	0.000	0.995	0.001	0.996	0.001
	300	50	0.999	0.000	0.991	0.003	0.994	0.002	0.999	0.000	0.988	0.004	0.992	0.003
		100	0.999	0.000	0.994	0.001	0.995	0.001	0.999	0.000	0.992	0.002	0.994	0.001
		300	0.999	0.000	0.997	0.000	0.998	0.000	0.999	0.000	0.996	0.001	0.997	0.000
4	50	50	0.990	0.003	0.981	0.007	0.990	0.003	0.988	0.003	0.977	0.008	0.987	0.005
		100	0.991	0.002	0.987	0.003	0.992	0.002	0.989	0.002	0.984	0.004	0.990	0.002
		300	0.991	0.001	0.992	0.001	0.994	0.001	0.989	0.002	0.991	0.002	0.993	0.001
	100	50	0.995	0.001	0.985	0.004	0.993	0.002	0.995	0.001	0.981	0.006	0.991	0.002
		100	0.996	0.001	0.990	0.002	0.995	0.001	0.995	0.001	0.988	0.003	0.994	0.001
		300	0.996	0.000	0.995	0.001	0.997	0.000	0.995	0.001	0.994	0.001	0.996	0.001
	300	50	0.999	0.000	0.987	0.004	0.995	0.001	0.998	0.000	0.984	0.005	0.994	0.002
		100	0.999	0.000	0.992	0.001	0.997	0.001	0.998	0.000	0.990	0.002	0.996	0.001
		300	0.999	0.000	0.996	0.001	0.998	0.000	0.999	0.000	0.995	0.001	0.998	0.000

C.6 Numerical Validation of Error Bounds

We include an additional numerical check of the theoretical rates. For each configuration, we record the normalized estimation errors of $\hat{\mathbf{M}}$, $\hat{\mathbf{F}}$, $\hat{\mathbf{B}}$, and $\hat{\Delta\mathbf{B}}$ and plot them against the problem size on a log-log scale. Figure 3 reports two representative regimes: symmetric growth with $n = p$ and a regime with fixed $n = 500$ and varying p. The empirical curves show clear monotone scaling trends, consistent with the rate statements in Theorems 3.2 and 3.3. The fixed-n regime also illustrates the blessing-of-dimensionality behavior in factor models: increasing p can improve latent-space recovery because more coordinates contribute information about the common representation.



(a) Symmetric growth ($n = p$).

(b) Fixed $n = 500$ with varying p.

Figure 3: Numerical validation of theoretical error-bound scaling. Empirical normalized errors are plotted on a log-log scale.

C.7 Why Direct Tests on Observed Coordinates Are Insufficient

This subsection clarifies why SMART targets row support in $\Delta\mathbf{B}$ rather than marginal changes in $\mathbf{X}$. Consider the additive model $X_j = \boldsymbol{f}^\top \boldsymbol{b}_j + \varepsilon_j$ with $\boldsymbol{f}$ having a distribution symmetric under coordinate sign changes. Let a changed feature satisfy $\boldsymbol{b}_j^{(2)} = \boldsymbol{b}_j^{(1)} + \Delta\boldsymbol{b}_j$ with $\Delta\boldsymbol{b}_j \neq \mathbf{0}$, but choose $\boldsymbol{b}_j^{(2)}$ so that $\boldsymbol{f}^\top \boldsymbol{b}_j^{(1)}$ and $\boldsymbol{f}^\top \boldsymbol{b}_j^{(2)}$ have the same distribution. If the noise distribution is unchanged, then the entire marginal distribution of X_j is identical across the two environments even though the structural mechanism has changed. Thus, mean tests, variance tests, and more general marginal distributional comparisons can miss mechanism shifts in principle.

We verified this phenomenon numerically using 100 repetitions. We compared two-sample t-tests for mean shifts, F-tests for variance shifts, their union, and SMART. Direct-test p-values were adjusted across variables by the Benjamini–Hochberg procedure. Table 26 shows that the direct tests have nearly zero precision and recall, whereas SMART recovers the mechanism support perfectly. Halving the noise level does not change this conclusion, indicating that the failure is structural rather than a lack of sample size or signal-to-noise ratio.

Table 26: Direct observed-coordinate tests can fail when the mechanism changes but the observed marginal distribution does not. Results are reported as mean (standard deviation) over 100 repetitions.

Noise level	Method	Precision	Recall
$\sigma = 1$	t-test	0.0075 (0.0750)	0.0060 (0.0600)
$\sigma = 1$	F-test	0.0018 (0.0182)	0.0010 (0.0100)
$\sigma = 1$	Union	0.0093 (0.0770)	0.0070 (0.0607)
$\sigma = 1$	SMART	1.0000 (0.0000)	1.0000 (0.0000)
$\sigma = 0.5$	t-test	0.0100 (0.1000)	0.0100 (0.1000)
$\sigma = 0.5$	F-test	0.0000 (0.0000)	0.0000 (0.0000)
$\sigma = 0.5$	Union	0.0100 (0.1000)	0.0100 (0.1000)
$\sigma = 0.5$	SMART	1.0000 (0.0000)	1.0000 (0.0000)

C.8 Multi-Environment Simulations

This subsection reports the full results for the multi-environment extension in Section 4. We vary the number of environments $K \in \{2, 3, 4, 5\}$ while keeping the same latent shift design. Table 30 shows that both one-step and two-stage estimators retain high union F_1 scores across all values of K. Tables 28 and 29 show that the one-step estimator often attains slightly smaller subspace error, but the difference becomes modest when n and p increase. Tables 31 and 32 confirm that model fitting time grows smoothly with K, and that the additional cost of the two-stage procedure mainly comes from Stage I signal recovery.

Table 27: Summary of additional revision experiments. Full tables are provided in the appendix.

Experiment	Purpose	Main observation
Multi-environment simulation	Vary K and assess the extension in Section 4.	Union F_1 remains high for $K = 2, \ldots, 5$; fitting time increases smoothly with K.
Mixed data simulation	Combine Gaussian and Poisson coordinates in the same observation matrix.	SMART remains competitive or better than N-Lasso, especially for row support recovery.
One-step versus two-stage	Vary Gaussian noise and compare accuracy with computational cost.	One-step is faster and useful at low noise; two-stage is broader because it handles non-Gaussian and mixed data.

Table 28: Subspace error of $\mathbf{F}$ comparing one-step and two-stage methods, with varying n, p, and K over 200 repetitions.

		$K = 2$		$K = 3$		$K = 4$		$K = 5$	
n	p	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage
50	50	0.083 (0.010)	0.083 (0.010)	0.066 (0.009)	0.07 (0.009)	0.054 (0.007)	0.071 (0.014)	0.047 (0.006)	0.074 (0.017)
	100	0.058 (0.007)	0.059 (0.007)	0.047 (0.005)	0.052 (0.007)	0.04 (0.004)	0.052 (0.010)	0.035 (0.004)	0.056 (0.014)
	300	0.034 (0.003)	0.036 (0.004)	0.028 (0.003)	0.032 (0.005)	0.023 (0.003)	0.032 (0.006)	0.021 (0.002)	0.036 (0.008)
100	50	0.081 (0.008)	0.082 (0.008)	0.064 (0.007)	0.069 (0.007)	0.053 (0.006)	0.068 (0.012)	0.046 (0.005)	0.074 (0.016)
	100	0.057 (0.005)	0.058 (0.005)	0.046 (0.004)	0.05 (0.006)	0.039 (0.004)	0.052 (0.010)	0.034 (0.003)	0.055 (0.014)
	300	0.033 (0.003)	0.036 (0.003)	0.027 (0.002)	0.034 (0.005)	0.023 (0.002)	0.035 (0.006)	0.02 (0.002)	0.038 (0.009)
300	50	0.078 (0.006)	0.079 (0.006)	0.063 (0.006)	0.068 (0.008)	0.053 (0.005)	0.068 (0.012)	0.046 (0.004)	0.075 (0.015)
	100	0.056 (0.004)	0.057 (0.004)	0.045 (0.003)	0.05 (0.005)	0.038 (0.002)	0.051 (0.010)	0.033 (0.002)	0.058 (0.014)
	300	0.032 (0.002)	0.033 (0.002)	0.026 (0.001)	0.028 (0.003)	0.022 (0.001)	0.03 (0.006)	0.02 (0.001)	0.034 (0.008)

Table 29: Average subspace error of $\mathbf{B}^{(k)}$ comparing one-step and two-stage methods, with varying n, p, and K over 200 repetitions.

		$K = 2$		$K = 3$		$K = 4$		$K = 5$	
n	p	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage
50	50	0.144 (0.018)	0.147 (0.017)	0.181 (0.023)	0.193 (0.021)	0.194 (0.020)	0.221 (0.020)	0.204 (0.021)	0.241 (0.021)
	100	0.125 (0.014)	0.128 (0.013)	0.144 (0.015)	0.158 (0.016)	0.158 (0.014)	0.181 (0.017)	0.165 (0.015)	0.201 (0.019)
	300	0.099 (0.008)	0.101 (0.008)	0.103 (0.008)	0.113 (0.009)	0.107 (0.007)	0.126 (0.012)	0.111 (0.007)	0.14 (0.016)
100	50	0.129 (0.015)	0.134 (0.015)	0.172 (0.020)	0.188 (0.019)	0.19 (0.022)	0.217 (0.021)	0.198 (0.021)	0.238 (0.021)
	100	0.108 (0.010)	0.113 (0.010)	0.134 (0.012)	0.149 (0.014)	0.152 (0.013)	0.179 (0.016)	0.159 (0.012)	0.196 (0.019)
	300	0.08 (0.006)	0.083 (0.007)	0.091 (0.007)	0.102 (0.010)	0.098 (0.007)	0.12 (0.013)	0.103 (0.006)	0.136 (0.015)
300	50	0.119 (0.014)	0.125 (0.014)	0.168 (0.022)	0.184 (0.020)	0.184 (0.020)	0.212 (0.020)	0.198 (0.019)	0.237 (0.020)
	100	0.098 (0.009)	0.104 (0.010)	0.13 (0.012)	0.147 (0.015)	0.145 (0.012)	0.173 (0.017)	0.154 (0.013)	0.195 (0.019)
	300	0.066 (0.005)	0.071 (0.007)	0.083 (0.006)	0.097 (0.009)	0.096 (0.006)	0.118 (0.013)	0.104 (0.006)	0.137 (0.015)

Table 30: Variable selection performance (union F_1 score, identifying a variable if selected in at least one of the $K - 1$ shifts) comparing one-step and two-stage methods, with varying n, p, and K over 200 repetitions.

		$K = 2$		$K = 3$		$K = 4$		$K = 5$	
n	p	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage
50	50	0.923 (0.070)	0.923 (0.071)	0.978 (0.036)	0.977 (0.037)	0.986 (0.029)	0.987 (0.027)	0.979 (0.033)	0.983 (0.030)
	100	0.914 (0.075)	0.912 (0.078)	0.975 (0.039)	0.974 (0.039)	0.989 (0.025)	0.989 (0.023)	0.991 (0.019)	0.991 (0.019)
	300	0.909 (0.067)	0.909 (0.068)	0.974 (0.037)	0.973 (0.038)	0.988 (0.026)	0.988 (0.025)	0.989 (0.022)	0.99 (0.022)
100	50	0.924 (0.064)	0.923 (0.064)	0.984 (0.030)	0.981 (0.033)	0.993 (0.017)	0.992 (0.020)	0.999 (0.008)	0.998 (0.009)
	100	0.917 (0.070)	0.916 (0.071)	0.984 (0.031)	0.982 (0.032)	0.992 (0.021)	0.992 (0.021)	0.997 (0.012)	0.997 (0.013)
	300	0.919 (0.070)	0.921 (0.067)	0.981 (0.032)	0.98 (0.032)	0.992 (0.020)	0.992 (0.020)	0.996 (0.014)	0.995 (0.015)
300	50	0.932 (0.067)	0.932 (0.068)	0.985 (0.026)	0.984 (0.027)	0.994 (0.017)	0.993 (0.018)	0.999 (0.007)	0.998 (0.009)
	100	0.92 (0.067)	0.919 (0.068)	0.981 (0.035)	0.98 (0.034)	0.992 (0.020)	0.992 (0.021)	0.998 (0.010)	0.998 (0.011)
	300	0.922 (0.066)	0.922 (0.066)	0.975 (0.035)	0.975 (0.035)	0.99 (0.025)	0.991 (0.022)	0.994 (0.019)	0.994 (0.019)

Table 31: Computational time comparing one-step and two-stage methods, with varying n, p, and K over 200 repetitions.

		$K = 2$		$K = 3$		$K = 4$		$K = 5$	
n	p	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage
50	50	0.068 (0.019)	0.084 (0.020)	0.093 (0.038)	0.115 (0.040)	0.115 (0.055)	0.147 (0.053)	0.153 (0.072)	0.2 (0.074)
	100	0.038 (0.010)	0.055 (0.012)	0.067 (0.040)	0.09 (0.039)	0.09 (0.055)	0.127 (0.055)	0.125 (0.074)	0.173 (0.075)
	300	0.031 (0.007)	0.062 (0.009)	0.046 (0.010)	0.095 (0.016)	0.063 (0.014)	0.129 (0.020)	0.087 (0.026)	0.176 (0.030)
100	50	0.078 (0.018)	0.096 (0.021)	0.109 (0.049)	0.141 (0.049)	0.136 (0.060)	0.185 (0.059)	0.171 (0.075)	0.228 (0.080)
	100	0.05 (0.013)	0.073 (0.013)	0.095 (0.051)	0.133 (0.053)	0.115 (0.071)	0.167 (0.070)	0.173 (0.090)	0.242 (0.090)
	300	0.049 (0.009)	0.106 (0.012)	0.075 (0.017)	0.161 (0.024)	0.111 (0.028)	0.223 (0.033)	0.153 (0.041)	0.292 (0.051)
300	50	0.099 (0.023)	0.129 (0.026)	0.128 (0.050)	0.178 (0.052)	0.176 (0.070)	0.239 (0.071)	0.217 (0.089)	0.295 (0.095)
	100	0.075 (0.019)	0.127 (0.021)	0.136 (0.068)	0.212 (0.070)	0.209 (0.108)	0.32 (0.112)	0.277 (0.153)	0.421 (0.159)
	300	0.205 (0.037)	0.384 (0.045)	0.309 (0.061)	0.588 (0.070)	0.433 (0.101)	0.859 (0.097)	0.659 (0.191)	1.262 (0.328)

Table 32: Individual stage-wise model fitting time comparing one-step and two-stage methods, with varying n, p, and K over 200 repetitions.

		$K=2$			$K=3$			$K=4$			$K=5$		
n	p	one-step	two-stage stage I	two-stage stage II	one-step	two-stage stage I	two-stage stage II	one-step	two-stage stage I	two-stage stage II	one-step	two-stage stage I	two-stage stage II
50	50	0.068	0.016	0.068	0.093	0.024	0.09	0.115	0.034	0.113	0.153	0.046	0.154
		(0.019)	(0.005)	(0.019)	(0.038)	(0.005)	(0.039)	(0.055)	(0.006)	(0.053)	(0.072)	(0.008)	(0.072)
	100	0.038	0.018	0.037	0.067	0.027	0.063	0.09	0.039	0.088	0.125	0.048	0.125
		(0.010)	(0.005)	(0.010)	(0.040)	(0.006)	(0.037)	(0.055)	(0.008)	(0.053)	(0.074)	(0.009)	(0.074)
	300	0.031	0.031	0.031	0.046	0.049	0.046	0.063	0.066	0.063	0.087	0.085	0.09
		(0.007)	(0.006)	(0.006)	(0.010)	(0.011)	(0.011)	(0.014)	(0.013)	(0.015)	(0.026)	(0.015)	(0.027)
100	50	0.078	0.02	0.076	0.109	0.037	0.104	0.136	0.047	0.137	0.171	0.062	0.166
		(0.018)	(0.005)	(0.020)	(0.049)	(0.008)	(0.048)	(0.060)	(0.008)	(0.059)	(0.075)	(0.009)	(0.078)
	100	0.05	0.026	0.048	0.095	0.042	0.091	0.115	0.058	0.11	0.173	0.073	0.169
		(0.013)	(0.004)	(0.012)	(0.051)	(0.008)	(0.052)	(0.071)	(0.008)	(0.069)	(0.090)	(0.009)	(0.089)
	300	0.049	0.057	0.048	0.075	0.085	0.076	0.111	0.113	0.111	0.153	0.14	0.152
		(0.009)	(0.008)	(0.008)	(0.017)	(0.012)	(0.018)	(0.028)	(0.016)	(0.027)	(0.041)	(0.019)	(0.044)
300	50	0.099	0.031	0.097	0.128	0.049	0.129	0.176	0.071	0.169	0.217	0.088	0.207
		(0.023)	(0.007)	(0.024)	(0.050)	(0.010)	(0.051)	(0.070)	(0.013)	(0.069)	(0.089)	(0.015)	(0.094)
	100	0.075	0.056	0.072	0.136	0.085	0.127	0.209	0.116	0.204	0.277	0.147	0.273
		(0.019)	(0.008)	(0.017)	(0.068)	(0.012)	(0.068)	(0.108)	(0.014)	(0.110)	(0.153)	(0.020)	(0.154)
	300	0.205	0.18	0.204	0.309	0.285	0.303	0.433	0.417	0.442	0.659	0.581	0.681
		(0.037)	(0.025)	(0.037)	(0.061)	(0.032)	(0.063)	(0.101)	(0.048)	(0.084)	(0.191)	(0.115)	(0.257)

C.9 Mixed Data Simulations

We also consider heterogeneous observations formed by combining Gaussian and Poisson coordinates under the same latent loading structure. This experiment examines whether the modular Stage I design remains effective when the observation matrix is not generated from a single likelihood family. Table 33 shows that SMART usually improves the support recovery F_1 score over N-Lasso, especially when sparsity is moderate or large. The additional tables show that recovery of the shared representation is essentially identical across methods, and that the loading subspaces recovered by SMART remain close to the unpenalized SVD benchmark while improving on or matching N-Lasso in most settings.

Table 33: Variable selection performance (F_1 score) for mixed data, with varying n, p, r^*, and s over 200 repetitions.

			$s=5$				$s=10$				$s=20$			
			N-Lasso		SMART		N-Lasso		SMART		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.897	0.147	0.915	0.122	0.931	0.087	0.947	0.073	0.955	0.049	0.965	0.038
		100	0.912	0.125	0.930	0.105	0.910	0.112	0.930	0.092	0.940	0.066	0.955	0.052
		300	0.886	0.137	0.917	0.114	0.901	0.100	0.927	0.080	0.899	0.098	0.923	0.075
	100	50	0.910	0.131	0.922	0.119	0.941	0.078	0.958	0.059	0.962	0.045	0.973	0.034
		100	0.900	0.128	0.923	0.107	0.912	0.105	0.933	0.087	0.943	0.066	0.960	0.047
		300	0.906	0.126	0.930	0.104	0.910	0.093	0.933	0.078	0.909	0.089	0.932	0.067
	300	50	0.912	0.137	0.927	0.119	0.941	0.080	0.961	0.058	0.962	0.051	0.972	0.040
		100	0.898	0.141	0.923	0.114	0.910	0.109	0.931	0.089	0.944	0.064	0.960	0.049
		300	0.896	0.132	0.925	0.104	0.902	0.104	0.924	0.084	0.907	0.088	0.929	0.067
3	50	50	0.964	0.078	0.981	0.050	0.971	0.048	0.980	0.034	0.962	0.072	0.975	0.029
		100	0.954	0.088	0.977	0.052	0.965	0.057	0.981	0.036	0.972	0.034	0.982	0.024
		300	0.949	0.090	0.961	0.094	0.953	0.066	0.972	0.052	0.958	0.050	0.977	0.032
	100	50	0.965	0.068	0.979	0.047	0.975	0.044	0.989	0.025	0.972	0.069	0.992	0.016
		100	0.957	0.077	0.980	0.045	0.965	0.059	0.982	0.035	0.978	0.033	0.989	0.018
		300	0.957	0.076	0.977	0.054	0.952	0.059	0.974	0.042	0.958	0.053	0.978	0.029
	300	50	0.961	0.080	0.978	0.055	0.978	0.043	0.991	0.024	0.984	0.044	0.995	0.013
		100	0.957	0.082	0.976	0.053	0.967	0.055	0.982	0.035	0.981	0.029	0.991	0.018
		300	0.955	0.079	0.975	0.056	0.954	0.064	0.978	0.039	0.961	0.049	0.981	0.029
4	50	50	0.976	0.057	0.977	0.050	0.971	0.041	0.962	0.048	0.932	0.114	0.953	0.047
		100	0.962	0.070	0.969	0.057	0.968	0.045	0.970	0.044	0.971	0.030	0.961	0.036
		300	0.938	0.086	0.881	0.135	0.957	0.051	0.946	0.057	0.961	0.034	0.961	0.036
	100	50	0.977	0.051	0.987	0.039	0.984	0.031	0.986	0.027	0.971	0.079	0.986	0.021
		100	0.970	0.062	0.982	0.042	0.981	0.034	0.986	0.029	0.984	0.021	0.984	0.021
		300	0.963	0.072	0.945	0.088	0.968	0.044	0.968	0.040	0.970	0.033	0.976	0.026
	300	50	0.982	0.051	0.993	0.028	0.989	0.026	0.994	0.019	0.983	0.062	0.995	0.011
		100	0.974	0.063	0.992	0.031	0.982	0.035	0.995	0.018	0.991	0.016	0.995	0.011
		300	0.974	0.070	0.985	0.046	0.974	0.046	0.991	0.023	0.982	0.026	0.991	0.015

Table 34: Subspace error of $\mathbf{F}$ for mixed data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.073	0.010	0.072	0.010	0.072	0.010	0.071	0.011	0.071	0.011	0.071	0.011	0.067	0.011	0.066	0.011	0.066	0.011
		100	0.056	0.011	0.056	0.011	0.056	0.011	0.056	0.011	0.056	0.011	0.056	0.011	0.054	0.012	0.054	0.012	0.054	0.012
		300	0.044	0.018	0.044	0.018	0.044	0.018	0.042	0.017	0.042	0.017	0.042	0.017	0.044	0.018	0.044	0.018	0.044	0.018
	100	50	0.071	0.008	0.071	0.008	0.071	0.008	0.069	0.008	0.068	0.008	0.068	0.008	0.065	0.010	0.065	0.010	0.065	0.010
		100	0.051	0.005	0.051	0.005	0.051	0.005	0.050	0.005	0.050	0.005	0.050	0.005	0.049	0.007	0.049	0.007	0.049	0.007
		300	0.037	0.011	0.037	0.011	0.037	0.011	0.037	0.012	0.037	0.012	0.037	0.012	0.038	0.013	0.038	0.013	0.038	0.013
	300	50	0.075	0.012	0.075	0.012	0.075	0.012	0.072	0.012	0.072	0.012	0.072	0.012	0.066	0.012	0.066	0.012	0.066	0.012
		100	0.052	0.007	0.052	0.007	0.052	0.007	0.051	0.006	0.050	0.006	0.050	0.006	0.049	0.007	0.049	0.007	0.049	0.007
		300	0.029	0.003	0.029	0.003	0.029	0.003	0.030	0.003	0.030	0.003	0.030	0.003	0.029	0.003	0.029	0.003	0.029	0.003
3	50	50	0.081	0.011	0.081	0.011	0.081	0.011	0.078	0.011	0.078	0.011	0.078	0.011	0.073	0.012	0.073	0.011	0.073	0.011
		100	0.061	0.010	0.061	0.011	0.061	0.010	0.060	0.009	0.060	0.009	0.060	0.009	0.058	0.010	0.058	0.010	0.058	0.010
		300	0.052	0.018	0.053	0.018	0.053	0.018	0.052	0.019	0.053	0.019	0.053	0.019	0.051	0.020	0.052	0.020	0.052	0.020
	100	50	0.079	0.010	0.079	0.010	0.079	0.010	0.075	0.010	0.075	0.010	0.075	0.010	0.070	0.010	0.070	0.010	0.070	0.010
		100	0.056	0.006	0.056	0.006	0.056	0.006	0.054	0.006	0.054	0.006	0.054	0.006	0.052	0.006	0.052	0.006	0.052	0.006
		300	0.044	0.013	0.044	0.013	0.044	0.013	0.044	0.012	0.044	0.013	0.044	0.013	0.043	0.013	0.043	0.013	0.043	0.013
	300	50	0.082	0.011	0.082	0.012	0.082	0.012	0.079	0.013	0.079	0.013	0.079	0.013	0.073	0.012	0.073	0.013	0.073	0.013
		100	0.056	0.007	0.056	0.007	0.056	0.007	0.055	0.007	0.055	0.007	0.055	0.007	0.053	0.007	0.053	0.007	0.053	0.007
		300	0.033	0.003	0.033	0.003	0.033	0.003	0.032	0.003	0.032	0.003	0.032	0.003	0.032	0.003	0.032	0.003	0.032	0.003
4	50	50	0.092	0.013	0.092	0.013	0.092	0.013	0.089	0.014	0.089	0.014	0.089	0.014	0.081	0.013	0.081	0.014	0.081	0.013
		100	0.069	0.010	0.069	0.010	0.069	0.010	0.067	0.010	0.067	0.011	0.067	0.011	0.065	0.011	0.065	0.011	0.065	0.011
		300	0.056	0.013	0.057	0.014	0.057	0.014	0.057	0.015	0.058	0.016	0.058	0.016	0.055	0.014	0.056	0.015	0.055	0.015
	100	50	0.089	0.012	0.089	0.013	0.089	0.013	0.085	0.012	0.085	0.013	0.085	0.012	0.079	0.012	0.079	0.012	0.079	0.012
		100	0.067	0.009	0.067	0.009	0.067	0.009	0.064	0.008	0.065	0.008	0.065	0.008	0.062	0.009	0.062	0.009	0.062	0.009
		300	0.051	0.009	0.051	0.009	0.051	0.009	0.050	0.009	0.050	0.009	0.050	0.009	0.048	0.009	0.049	0.009	0.049	0.009
	300	50	0.093	0.013	0.093	0.013	0.093	0.013	0.088	0.012	0.089	0.013	0.089	0.013	0.081	0.014	0.081	0.014	0.081	0.014
		100	0.068	0.008	0.068	0.008	0.068	0.008	0.065	0.008	0.065	0.009	0.065	0.009	0.062	0.008	0.062	0.008	0.062	0.008
		300	0.046	0.005	0.046	0.005	0.046	0.005	0.045	0.005	0.045	0.005	0.045	0.005	0.044	0.005	0.045	0.005	0.044	0.005

Table 35: Subspace error of $\mathbf{B}^{(1)}$ for mixed data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.092	0.018	0.135	0.023	0.119	0.019	0.093	0.018	0.156	0.028	0.130	0.024	0.093	0.018	0.175	0.039	0.142	0.029
		100	0.096	0.021	0.118	0.014	0.106	0.013	0.097	0.021	0.133	0.019	0.117	0.017	0.096	0.021	0.151	0.024	0.127	0.021
		300	0.103	0.028	0.097	0.011	0.092	0.011	0.101	0.026	0.109	0.012	0.099	0.011	0.104	0.029	0.123	0.014	0.112	0.013
	100	50	0.069	0.014	0.123	0.022	0.106	0.018	0.070	0.015	0.142	0.028	0.115	0.023	0.070	0.015	0.161	0.035	0.128	0.026
		100	0.065	0.010	0.104	0.013	0.091	0.011	0.065	0.011	0.119	0.016	0.102	0.014	0.066	0.011	0.138	0.025	0.112	0.021
		300	0.071	0.018	0.080	0.008	0.072	0.008	0.071	0.020	0.093	0.009	0.082	0.008	0.072	0.020	0.107	0.012	0.095	0.011
	300	50	0.075	0.038	0.119	0.024	0.102	0.023	0.074	0.038	0.138	0.030	0.111	0.027	0.071	0.033	0.155	0.038	0.123	0.031
		100	0.055	0.021	0.096	0.013	0.084	0.012	0.054	0.020	0.112	0.017	0.094	0.017	0.055	0.021	0.130	0.023	0.103	0.021
		300	0.041	0.008	0.067	0.007	0.058	0.005	0.042	0.007	0.081	0.006	0.070	0.005	0.042	0.008	0.097	0.010	0.083	0.009
3	50	50	0.106	0.021	0.149	0.023	0.119	0.018	0.107	0.022	0.173	0.032	0.129	0.023	0.107	0.022	0.197	0.042	0.140	0.026
		100	0.110	0.023	0.128	0.017	0.108	0.014	0.110	0.022	0.146	0.020	0.117	0.016	0.108	0.021	0.170	0.027	0.127	0.020
		300	0.124	0.033	0.109	0.012	0.099	0.013	0.125	0.033	0.120	0.013	0.105	0.014	0.125	0.035	0.136	0.016	0.115	0.015
	100	50	0.083	0.017	0.135	0.021	0.104	0.016	0.083	0.018	0.158	0.030	0.112	0.019	0.082	0.016	0.179	0.037	0.122	0.022
		100	0.075	0.013	0.113	0.014	0.090	0.012	0.074	0.012	0.129	0.016	0.097	0.013	0.074	0.014	0.150	0.023	0.106	0.016
		300	0.085	0.022	0.087	0.008	0.074	0.008	0.085	0.021	0.100	0.008	0.082	0.008	0.085	0.021	0.116	0.013	0.092	0.011
	300	50	0.089	0.031	0.129	0.023	0.097	0.018	0.093	0.034	0.154	0.032	0.108	0.025	0.092	0.033	0.172	0.039	0.118	0.027
		100	0.063	0.018	0.104	0.015	0.080	0.012	0.063	0.017	0.120	0.016	0.087	0.014	0.064	0.018	0.141	0.023	0.096	0.017
		300	0.049	0.007	0.073	0.007	0.056	0.005	0.049	0.007	0.087	0.006	0.066	0.005	0.049	0.007	0.102	0.011	0.077	0.008
4	50	50	0.135	0.031	0.167	0.030	0.129	0.021	0.138	0.035	0.199	0.035	0.142	0.026	0.137	0.032	0.231	0.047	0.154	0.029
		100	0.135	0.028	0.141	0.018	0.116	0.016	0.138	0.029	0.163	0.021	0.126	0.018	0.138	0.030	0.190	0.028	0.136	0.021
		300	0.153	0.034	0.120	0.014	0.109	0.015	0.153	0.034	0.132	0.015	0.114	0.015	0.151	0.032	0.147	0.015	0.121	0.015
	100	50	0.104	0.025	0.150	0.025	0.107	0.016	0.104	0.025	0.177	0.031	0.117	0.020	0.105	0.024	0.202	0.040	0.128	0.021
		100	0.105	0.025	0.123	0.014	0.095	0.012	0.104	0.023	0.143	0.019	0.103	0.014	0.104	0.023	0.166	0.025	0.112	0.017
		300	0.115	0.023	0.099	0.010	0.083	0.010	0.115	0.022	0.110	0.010	0.089	0.009	0.115	0.022	0.127	0.013	0.098	0.012
	300	50	0.100	0.025	0.139	0.023	0.097	0.017	0.100	0.024	0.167	0.031	0.106	0.021	0.098	0.024	0.190	0.040	0.118	0.022
		100	0.083	0.017	0.110	0.015	0.080	0.011	0.081	0.015	0.131	0.016	0.088	0.011	0.081	0.014	0.152	0.024	0.095	0.014
		300	0.079	0.011	0.081	0.007	0.061	0.005	0.079	0.013	0.094	0.008	0.069	0.006	0.080	0.011	0.111	0.010	0.079	0.008

Table 36: Subspace error of $\mathbf{B}^{(2)}$ for mixed data, with varying n, p, r^*, and s over 200 repetitions.

			$s = 5$						$s = 10$						$s = 20$					
			SVD		N-Lasso		SMART		SVD		N-Lasso		SMART		SVD		N-Lasso		SMART	
r^*	n	p	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
2	50	50	0.111	0.017	0.125	0.029	0.111	0.026	0.105	0.019	0.137	0.040	0.116	0.034	0.094	0.022	0.135	0.053	0.111	0.041
		100	0.113	0.012	0.114	0.019	0.103	0.017	0.109	0.014	0.125	0.025	0.111	0.024	0.103	0.016	0.131	0.035	0.112	0.030
		300	0.114	0.009	0.096	0.012	0.091	0.013	0.113	0.009	0.107	0.013	0.098	0.014	0.111	0.011	0.119	0.020	0.108	0.019
	100	50	0.078	0.012	0.114	0.028	0.098	0.026	0.074	0.015	0.122	0.037	0.100	0.032	0.068	0.018	0.121	0.052	0.097	0.040
		100	0.080	0.008	0.101	0.017	0.089	0.016	0.078	0.011	0.112	0.025	0.096	0.024	0.074	0.015	0.120	0.038	0.098	0.033
		300	0.080	0.006	0.079	0.009	0.072	0.010	0.079	0.006	0.093	0.011	0.082	0.012	0.078	0.009	0.103	0.019	0.092	0.018
	300	50	0.047	0.011	0.107	0.030	0.091	0.028	0.045	0.013	0.117	0.038	0.093	0.034	0.044	0.018	0.114	0.051	0.089	0.041
		100	0.048	0.009	0.093	0.019	0.081	0.019	0.048	0.012	0.105	0.026	0.089	0.026	0.047	0.016	0.112	0.036	0.090	0.033
		300	0.047	0.006	0.068	0.011	0.060	0.011	0.048	0.008	0.083	0.011	0.071	0.013	0.048	0.011	0.095	0.020	0.083	0.019
3	50	50	0.123	0.018	0.135	0.027	0.108	0.021	0.114	0.018	0.148	0.035	0.111	0.025	0.099	0.019	0.148	0.052	0.108	0.033
		100	0.122	0.012	0.119	0.018	0.101	0.015	0.117	0.013	0.130	0.022	0.105	0.018	0.110	0.016	0.142	0.034	0.108	0.025
		300	0.123	0.010	0.105	0.012	0.095	0.013	0.120	0.010	0.115	0.014	0.100	0.014	0.117	0.010	0.125	0.018	0.105	0.017
	100	50	0.084	0.011	0.120	0.025	0.091	0.019	0.078	0.012	0.131	0.035	0.092	0.024	0.069	0.013	0.134	0.049	0.091	0.031
		100	0.084	0.008	0.105	0.017	0.084	0.013	0.081	0.008	0.116	0.021	0.088	0.017	0.075	0.010	0.125	0.031	0.088	0.022
		300	0.083	0.005	0.084	0.008	0.071	0.009	0.082	0.006	0.095	0.010	0.077	0.010	0.080	0.006	0.107	0.015	0.085	0.013
	300	50	0.048	0.006	0.114	0.026	0.082	0.021	0.046	0.007	0.126	0.036	0.084	0.026	0.041	0.009	0.125	0.048	0.079	0.031
		100	0.048	0.004	0.097	0.017	0.074	0.014	0.046	0.005	0.108	0.021	0.077	0.018	0.044	0.007	0.117	0.031	0.077	0.023
		300	0.048	0.003	0.072	0.007	0.055	0.007	0.047	0.003	0.084	0.008	0.063	0.009	0.046	0.005	0.096	0.016	0.072	0.014
4	50	50	0.133	0.019	0.146	0.028	0.112	0.019	0.125	0.020	0.164	0.036	0.115	0.024	0.110	0.021	0.167	0.056	0.111	0.030
		100	0.132	0.014	0.130	0.019	0.106	0.015	0.127	0.014	0.145	0.023	0.109	0.016	0.118	0.016	0.155	0.035	0.109	0.023
		300	0.130	0.011	0.116	0.013	0.103	0.014	0.128	0.011	0.124	0.015	0.106	0.015	0.123	0.011	0.133	0.016	0.108	0.015
	100	50	0.091	0.011	0.130	0.025	0.090	0.016	0.085	0.013	0.147	0.035	0.094	0.022	0.074	0.013	0.148	0.050	0.090	0.028
		100	0.089	0.007	0.112	0.017	0.086	0.012	0.085	0.008	0.124	0.021	0.087	0.014	0.079	0.010	0.135	0.032	0.087	0.020
		300	0.087	0.006	0.093	0.009	0.078	0.009	0.086	0.006	0.103	0.011	0.082	0.009	0.083	0.006	0.114	0.015	0.086	0.012
	300	50	0.051	0.006	0.119	0.025	0.078	0.018	0.048	0.007	0.136	0.034	0.080	0.022	0.043	0.008	0.136	0.048	0.076	0.027
		100	0.050	0.004	0.100	0.017	0.071	0.012	0.049	0.004	0.115	0.020	0.074	0.014	0.045	0.005	0.124	0.030	0.073	0.019
		300	0.048	0.002	0.078	0.007	0.058	0.006	0.048	0.003	0.088	0.009	0.063	0.007	0.047	0.003	0.100	0.013	0.069	0.010

C.10 Additional Comparison Between One-Step and Two-Stage Estimation

We also conducted additional comparisons between the one-step estimator and the two-stage SMART estimator for Gaussian data. The two methods use the same rowwise group penalty and the same CV protocol for λ; the difference is whether signal recovery is performed before attribution. The results support a simple practical guideline. When the data are Gaussian with low noise, one-step estimation can slightly reduce subspace error because it avoids the extra Stage I estimation step. As the noise level increases, the performance gap becomes negligible. For example, under $n = 100$ and $p = 100$, the subspace error of $\mathbf{B}^{(2)}$ is 0.082 for one-step and 0.096 for two-stage at $\sigma = 0.1$, whereas it becomes 0.142 and 0.151 at $\sigma = 2.0$. The F_1 scores remain similar across noise levels.

The computational comparison is also consistent with the methodological trade-off. The one-step method is faster in the Gaussian setting because it avoids Stage I; for instance, under $n = 300$, $p = 300$, and $\sigma = 1.0$, the model fitting time is 0.394 seconds for one-step and 0.831 seconds for two-stage in our implementation. However, the two-stage approach is more flexible: it covers Poisson, Binary, mixed data, and settings with missing values through the choice of Stage I estimator, while the one-step formulation in Section 4 is specialized to additive Gaussian observations. Additional simulations varying r^*, s, signal strength c, and the number of environments K show qualitatively similar conclusions.

Tables 37–41 report the full Gaussian comparison under varying noise levels. Together they support the role of the two-stage estimator as the default, broadly applicable version of SMART, while the one-step estimator is a useful efficient specialization for low noise additive data.

Table 37: Subspace error of $\mathbf{F}$ comparing one-step and two-stage methods, with varying n, p, and σ over 200 repetitions.

		$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.5$		$\sigma = 1.0$		$\sigma = 1.5$		$\sigma = 2.0$	
n	p	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage
50	50	0.008 (0.001)	0.017 (0.010)	0.016 (0.002)	0.022 (0.007)	0.042 (0.006)	0.044 (0.007)	0.084 (0.011)	0.085 (0.011)	0.123 (0.015)	0.123 (0.015)	0.168 (0.023)	0.168 (0.022)
	100	0.006 (0.001)	0.015 (0.008)	0.012 (0.001)	0.018 (0.005)	0.029 (0.003)	0.033 (0.005)	0.059 (0.007)	0.061 (0.007)	0.088 (0.011)	0.089 (0.011)	0.120 (0.013)	0.121 (0.013)
	300	0.003 (0.000)	0.013 (0.004)	0.007 (0.001)	0.014 (0.004)	0.017 (0.002)	0.021 (0.004)	0.034 (0.004)	0.036 (0.004)	0.051 (0.006)	0.053 (0.006)	0.068 (0.007)	0.069 (0.007)
100	50	0.008 (0.001)	0.017 (0.011)	0.016 (0.002)	0.023 (0.008)	0.040 (0.004)	0.043 (0.005)	0.080 (0.009)	0.081 (0.008)	0.120 (0.012)	0.120 (0.012)	0.162 (0.015)	0.162 (0.015)
	100	0.006 (0.000)	0.013 (0.007)	0.011 (0.001)	0.016 (0.006)	0.028 (0.002)	0.030 (0.003)	0.057 (0.005)	0.058 (0.005)	0.085 (0.007)	0.085 (0.007)	0.115 (0.010)	0.116 (0.010)
	300	0.003 (0.000)	0.017 (0.005)	0.007 (0.001)	0.018 (0.005)	0.016 (0.001)	0.023 (0.004)	0.033 (0.003)	0.036 (0.004)	0.050 (0.004)	0.052 (0.005)	0.066 (0.005)	0.067 (0.006)
300	50	0.008 (0.001)	0.018 (0.010)	0.016 (0.001)	0.023 (0.008)	0.040 (0.003)	0.042 (0.005)	0.079 (0.006)	0.080 (0.006)	0.118 (0.009)	0.118 (0.009)	0.157 (0.014)	0.156 (0.014)
	100	0.006 (0.000)	0.014 (0.008)	0.011 (0.001)	0.018 (0.007)	0.028 (0.002)	0.030 (0.004)	0.056 (0.003)	0.057 (0.003)	0.084 (0.005)	0.084 (0.005)	0.111 (0.007)	0.111 (0.007)
	300	0.003 (0.000)	0.006 (0.004)	0.006 (0.000)	0.009 (0.003)	0.016 (0.001)	0.017 (0.002)	0.032 (0.002)	0.033 (0.002)	0.048 (0.002)	0.048 (0.002)	0.064 (0.003)	0.064 (0.003)

Table 38: Subspace error of $\mathbf{B}^{(1)}$ comparing one-step and two-stage methods, with varying n, p, and σ over 200 repetitions.

		$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.5$		$\sigma = 1.0$		$\sigma = 1.5$		$\sigma = 2.0$	
n	p	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage
50	50	0.104 (0.016)	0.105 (0.016)	0.105 (0.016)	0.105 (0.016)	0.114 (0.019)	0.115 (0.019)	0.138 (0.021)	0.138 (0.021)	0.168 (0.024)	0.169 (0.024)	0.209 (0.033)	0.210 (0.033)
	100	0.091 (0.010)	0.092 (0.010)	0.093 (0.012)	0.093 (0.012)	0.101 (0.014)	0.102 (0.013)	0.124 (0.014)	0.125 (0.014)	0.155 (0.019)	0.156 (0.019)	0.191 (0.021)	0.192 (0.021)
	300	0.059 (0.006)	0.059 (0.006)	0.061 (0.007)	0.061 (0.007)	0.071 (0.007)	0.071 (0.007)	0.101 (0.009)	0.101 (0.009)	0.136 (0.011)	0.136 (0.011)	0.174 (0.013)	0.174 (0.013)
100	50	0.101 (0.014)	0.102 (0.014)	0.100 (0.014)	0.101 (0.014)	0.104 (0.014)	0.105 (0.014)	0.118 (0.016)	0.118 (0.016)	0.138 (0.018)	0.138 (0.017)	0.161 (0.023)	0.161 (0.023)
	100	0.089 (0.009)	0.090 (0.009)	0.090 (0.009)	0.090 (0.009)	0.094 (0.010)	0.094 (0.010)	0.106 (0.011)	0.107 (0.011)	0.124 (0.013)	0.124 (0.013)	0.148 (0.016)	0.149 (0.016)
	300	0.059 (0.006)	0.059 (0.006)	0.059 (0.006)	0.059 (0.006)	0.064 (0.006)	0.064 (0.006)	0.080 (0.006)	0.080 (0.006)	0.102 (0.007)	0.102 (0.007)	0.127 (0.009)	0.128 (0.009)
300	50	0.100 (0.012)	0.101 (0.012)	0.100 (0.012)	0.101 (0.012)	0.103 (0.013)	0.104 (0.013)	0.106 (0.013)	0.107 (0.013)	0.113 (0.014)	0.113 (0.014)	0.124 (0.017)	0.124 (0.017)
	100	0.088 (0.008)	0.089 (0.008)	0.089 (0.009)	0.091 (0.009)	0.089 (0.008)	0.091 (0.008)	0.094 (0.008)	0.095 (0.008)	0.102 (0.011)	0.103 (0.011)	0.111 (0.011)	0.112 (0.011)
	300	0.057 (0.004)	0.057 (0.004)	0.057 (0.004)	0.057 (0.004)	0.059 (0.005)	0.060 (0.005)	0.065 (0.005)	0.065 (0.005)	0.074 (0.005)	0.074 (0.005)	0.086 (0.006)	0.086 (0.006)

Table 39: Subspace error of $\mathbf{B}^{(2)}$ comparing one-step and two-stage methods, with varying n, p, and σ over 200 repetitions.

		$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.5$		$\sigma = 1.0$		$\sigma = 1.5$		$\sigma = 2.0$	
n	p	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage
50	50	0.092 (0.016)	0.103 (0.017)	0.092 (0.017)	0.102 (0.017)	0.102 (0.018)	0.111 (0.019)	0.126 (0.021)	0.133 (0.021)	0.157 (0.024)	0.163 (0.024)	0.196 (0.029)	0.201 (0.028)
	100	0.085 (0.010)	0.095 (0.016)	0.087 (0.013)	0.095 (0.013)	0.093 (0.013)	0.104 (0.014)	0.118 (0.014)	0.126 (0.016)	0.150 (0.017)	0.157 (0.017)	0.183 (0.020)	0.189 (0.019)
	300	0.056 (0.006)	0.066 (0.012)	0.058 (0.007)	0.066 (0.012)	0.069 (0.007)	0.077 (0.012)	0.099 (0.008)	0.104 (0.009)	0.135 (0.011)	0.138 (0.011)	0.172 (0.013)	0.175 (0.013)
100	50	0.090 (0.015)	0.100 (0.016)	0.089 (0.014)	0.099 (0.016)	0.093 (0.013)	0.104 (0.015)	0.106 (0.017)	0.115 (0.018)	0.124 (0.019)	0.134 (0.020)	0.148 (0.020)	0.157 (0.020)
	100	0.082 (0.009)	0.096 (0.014)	0.083 (0.010)	0.095 (0.014)	0.087 (0.010)	0.098 (0.013)	0.100 (0.011)	0.110 (0.014)	0.118 (0.013)	0.127 (0.014)	0.142 (0.016)	0.151 (0.015)
	300	0.056 (0.006)	0.064 (0.011)	0.056 (0.005)	0.067 (0.012)	0.061 (0.006)	0.070 (0.012)	0.077 (0.006)	0.085 (0.010)	0.101 (0.007)	0.106 (0.009)	0.125 (0.008)	0.130 (0.010)
300	50	0.089 (0.013)	0.100 (0.014)	0.089 (0.013)	0.101 (0.015)	0.091 (0.014)	0.103 (0.016)	0.095 (0.014)	0.105 (0.015)	0.102 (0.015)	0.112 (0.016)	0.113 (0.018)	0.125 (0.019)
	100	0.081 (0.008)	0.094 (0.013)	0.082 (0.008)	0.097 (0.015)	0.082 (0.008)	0.095 (0.013)	0.087 (0.008)	0.100 (0.014)	0.095 (0.010)	0.109 (0.015)	0.104 (0.011)	0.115 (0.014)
	300	0.054 (0.004)	0.067 (0.011)	0.055 (0.004)	0.069 (0.012)	0.057 (0.005)	0.070 (0.011)	0.062 (0.005)	0.076 (0.011)	0.072 (0.004)	0.083 (0.010)	0.084 (0.005)	0.093 (0.009)

Table 40: Variable selection performance (F_1 score) comparing one-step and two-stage methods, with varying n, p, and σ over 200 repetitions.

		$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.5$		$\sigma = 1.0$		$\sigma = 1.5$		$\sigma = 2.0$	
n	p	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage
50	50	0.946 (0.059)	0.945 (0.059)	0.944 (0.059)	0.943 (0.060)	0.949 (0.063)	0.948 (0.063)	0.941 (0.063)	0.941 (0.064)	0.866 (0.074)	0.864 (0.075)	0.702 (0.077)	0.690 (0.078)
	100	0.918 (0.073)	0.917 (0.073)	0.911 (0.077)	0.910 (0.078)	0.908 (0.077)	0.906 (0.077)	0.920 (0.071)	0.920 (0.071)	0.874 (0.086)	0.872 (0.086)	0.649 (0.092)	0.642 (0.097)
	300	0.912 (0.074)	0.913 (0.071)	0.923 (0.071)	0.921 (0.071)	0.917 (0.065)	0.917 (0.065)	0.916 (0.080)	0.916 (0.077)	0.836 (0.140)	0.824 (0.149)	0.457 (0.139)	0.446 (0.137)
100	50	0.946 (0.055)	0.944 (0.055)	0.948 (0.063)	0.947 (0.063)	0.952 (0.056)	0.950 (0.057)	0.948 (0.056)	0.948 (0.057)	0.944 (0.060)	0.942 (0.059)	0.886 (0.082)	0.880 (0.080)
	100	0.918 (0.070)	0.917 (0.071)	0.918 (0.068)	0.915 (0.072)	0.915 (0.073)	0.913 (0.073)	0.912 (0.069)	0.912 (0.070)	0.913 (0.072)	0.912 (0.073)	0.906 (0.070)	0.902 (0.072)
	300	0.915 (0.067)	0.914 (0.066)	0.914 (0.071)	0.914 (0.071)	0.918 (0.068)	0.917 (0.068)	0.913 (0.065)	0.914 (0.065)	0.910 (0.072)	0.910 (0.071)	0.874 (0.109)	0.862 (0.123)
300	50	0.943 (0.063)	0.941 (0.066)	0.944 (0.062)	0.942 (0.064)	0.945 (0.062)	0.942 (0.063)	0.943 (0.056)	0.942 (0.059)	0.946 (0.058)	0.945 (0.059)	0.944 (0.055)	0.945 (0.054)
	100	0.921 (0.072)	0.919 (0.073)	0.926 (0.070)	0.922 (0.071)	0.915 (0.072)	0.913 (0.073)	0.925 (0.064)	0.923 (0.063)	0.923 (0.070)	0.922 (0.070)	0.920 (0.074)	0.919 (0.075)
	300	0.926 (0.059)	0.923 (0.061)	0.914 (0.066)	0.911 (0.068)	0.923 (0.066)	0.920 (0.066)	0.916 (0.069)	0.914 (0.070)	0.915 (0.074)	0.916 (0.071)	0.914 (0.066)	0.913 (0.067)

Table 41: Computational time comparing one-step and two-stage methods, with varying n, p, and σ over 200 repetitions.

		$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.5$		$\sigma = 1.0$		$\sigma = 1.5$		$\sigma = 2.0$	
n	p	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage	one-step	two-stage
50	50	0.150 (0.009)	0.182 (0.015)	0.164 (0.017)	0.197 (0.021)	0.163 (0.015)	0.196 (0.017)	0.177 (0.019)	0.211 (0.022)	0.184 (0.025)	0.219 (0.028)	0.196 (0.045)	0.235 (0.052)
	100	0.104 (0.033)	0.146 (0.040)	0.108 (0.036)	0.154 (0.040)	0.112 (0.044)	0.160 (0.055)	0.125 (0.053)	0.171 (0.057)	0.130 (0.054)	0.178 (0.068)	0.216 (0.093)	0.282 (0.109)
	300	0.096 (0.037)	0.205 (0.070)	0.098 (0.040)	0.205 (0.072)	0.099 (0.038)	0.208 (0.071)	0.101 (0.040)	0.206 (0.073)	0.101 (0.048)	0.203 (0.076)	0.088 (0.054)	0.200 (0.080)
100	50	0.262 (0.095)	0.327 (0.119)	0.259 (0.092)	0.316 (0.111)	0.257 (0.085)	0.313 (0.098)	0.263 (0.091)	0.297 (0.089)	0.258 (0.081)	0.310 (0.090)	0.206 (0.027)	0.254 (0.033)
	100	0.120 (0.026)	0.182 (0.037)	0.122 (0.029)	0.178 (0.034)	0.125 (0.030)	0.179 (0.034)	0.130 (0.038)	0.186 (0.042)	0.127 (0.036)	0.178 (0.034)	0.135 (0.036)	0.189 (0.043)
	300	0.125 (0.025)	0.294 (0.052)	0.128 (0.027)	0.293 (0.047)	0.125 (0.024)	0.291 (0.050)	0.125 (0.029)	0.292 (0.046)	0.126 (0.029)	0.298 (0.046)	0.127 (0.032)	0.289 (0.046)
300	50	0.310 (0.046)	0.403 (0.060)	0.314 (0.052)	0.411 (0.070)	0.328 (0.064)	0.414 (0.070)	0.307 (0.038)	0.391 (0.046)	0.302 (0.042)	0.391 (0.042)	0.303 (0.034)	0.396 (0.044)
	100	0.220 (0.049)	0.371 (0.054)	0.227 (0.054)	0.377 (0.059)	0.225 (0.055)	0.379 (0.058)	0.222 (0.049)	0.382 (0.065)	0.228 (0.052)	0.377 (0.061)	0.225 (0.048)	0.370 (0.055)
	300	0.389 (0.053)	0.810 (0.080)	0.395 (0.063)	0.840 (0.097)	0.398 (0.053)	0.844 (0.092)	0.394 (0.060)	0.831 (0.092)	0.384 (0.051)	0.790 (0.087)	0.372 (0.049)	0.740 (0.097)

D Additional Results on Real Data

D.1 Regularization Parameter Selection

In real-world applications, identifying the most interpretable model structure is often prioritized over minimizing pure prediction error. Consequently, we adopted a heuristic approach rather than CV. Specifically, we determined the search range for λ via preliminary experiments to cover the full solution path. We then selected 100 points uniformly within this range as the candidate grid of λ values. For the latent rank r, we performed a preliminary assessment based on IC_2 criterion (range 1–8). After fixing r, we plotted the number of non-zero features against λ and adopted the "elbow method" to select the λ where the curve's slope shifts from a sharp decline to a gradual plateau. As shown in Figure 4 and Figure 5, both datasets exhibit clear elbow structures.

D.2 Additional TCGA Data Analysis

The remaining selected genes provide a broader biological validation of the structural shifts detected by SMART. Several markers have tumor suppression or epigenetic roles. `WIF1` is a frequent target of epigenetic silencing in breast cancer; its inactivation removes inhibition of the Wnt signaling pathway and thereby promotes oncogenesis Ai et al. (2006). `CLCA4` has been characterized as a tumor suppressor whose loss facilitates cancer progression and metastasis Yu et al. (2013), while `CMTM5` is reported as a putative suppressor in breast tissue Zhou et al. (2020). `LINC01198` has also been documented to play a dual role as both a tumor suppressor and a promoter of anti-tumor immune responses Fonseca-Montaño et al. (2023).

SMART also selects genes linked to malignancy and metabolic adaptation. `FABP7` supports breast cancer cell survival and proliferation within the brain microenvironment Cordero et al. (2019). The aberrant activation of the `DLK1-DIO3` locus, particularly its noncoding RNAs, is closely linked to epithelial mesenchymal transition, metastasis, and poor prognosis Budkova et al. (2020). Broadly altered expression profiles were also detected for `NEUROG2` Laisné et al. (2021) and `KCNJ16`, the latter of which shows significant alterations in pan-cancer analyses Zhu et al. (2022). Collectively, these findings confirm that the rows selected in $\Delta\mathbf{B}$ correspond to biologically meaningful drivers and indicators of breast cancer pathology.

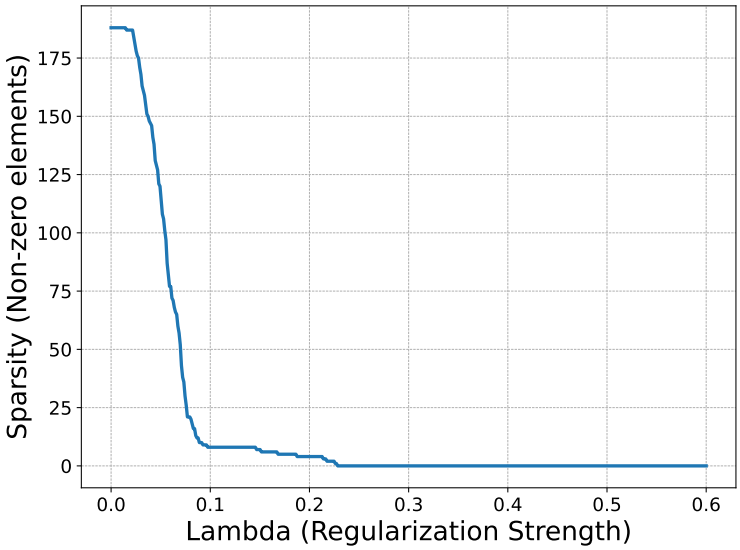


Figure 4: Sparsity path for the NBA dataset: number of nonzero elements as a function of λ.

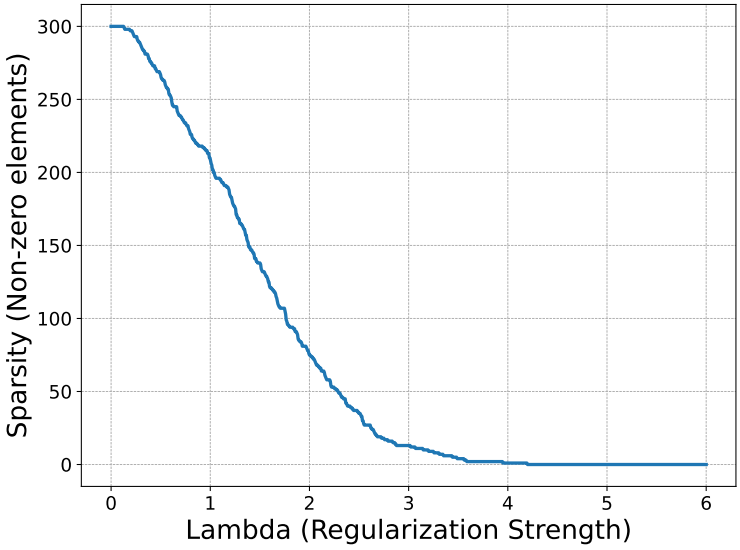


Figure 5: Sparsity path for the TCGA dataset: number of nonzero elements as a function of λ.

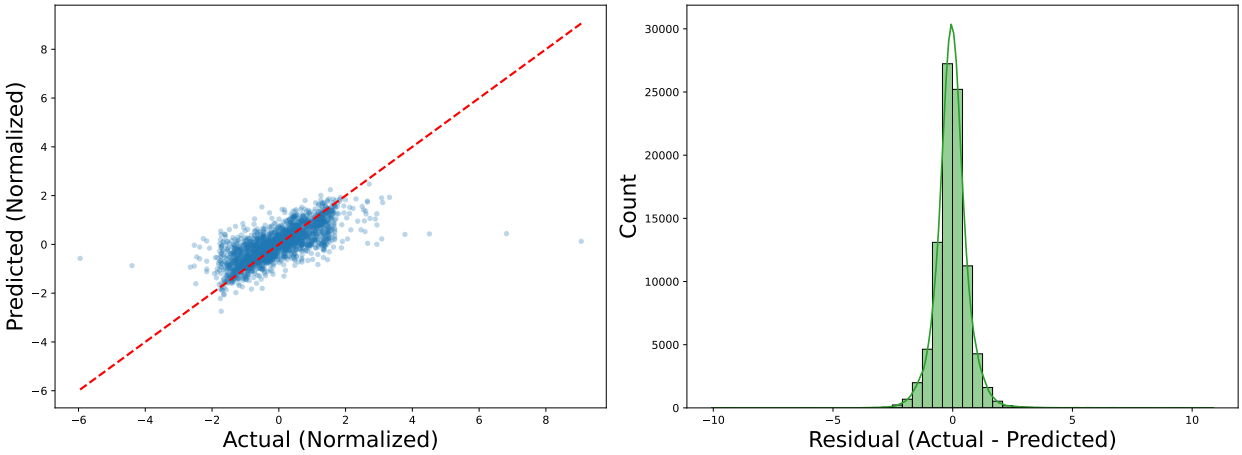


Figure 6: Model diagnostics for the period after the All-Star break. Left: goodness of fit, comparing predicted and actual values; right: residual distribution.

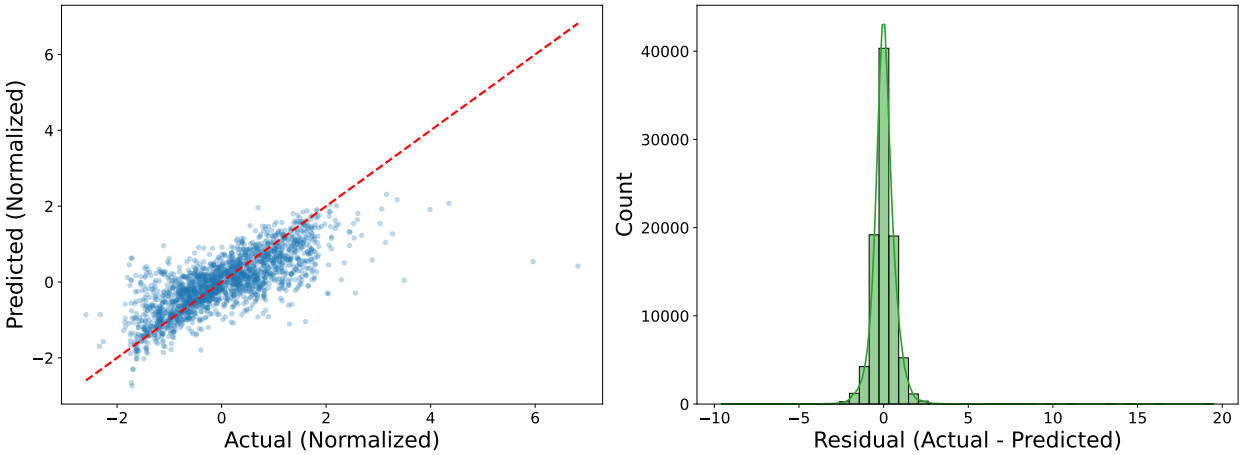


Figure 7: Model diagnostics for the period before the All-Star break. Left: goodness of fit, comparing predicted and actual values; right: residual distribution.

D.3 NBA 2023-24 Season Analysis

Professional basketball (NBA) represents a complex, high stakes dynamic system where team strategies and player incentives evolve significantly over the course of a season. The All-Star break traditionally serves as a structural inflection point, marking the transition from the regular season to a more intense playoff environment. We apply SMART to continuous player performance metrics to validate the model's ability to detect these latent strategic shifts ("gameplay evolution").

We analyzed paired data comparing the environments before and after the All-Star break. Following the selection strategy, we set the latent rank $r = 3$ and the regularization parameter $\lambda = 0.1$. We evaluated the model's goodness of fit for both periods. Figure 6 presents the diagnostic plots for the period after the All-Star break. The tight clustering of points along the $y = x$ line indicates high predictive accuracy, while the symmetric residual distribution centered at zero confirms that the model assumptions hold without significant systematic bias. Detailed diagnostic plots for the period before the All-Star break (Figure 7) exhibit similarly robust fit characteristics. The consistency of these diagnostic patterns across both temporal environments confirms that the SMART framework effectively captures the underlying data structure without overfitting to a specific period. Through sparse regularization on $\mathbf{\Delta B}$, the model identifies nine key nonzero rows

(features) in the difference loading matrix (Figure 8), which we map to three distinct mechanisms driving late season gameplay transformations:

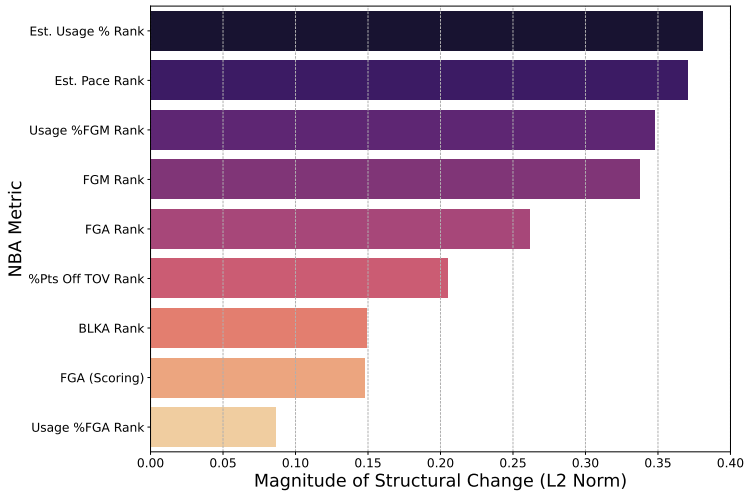


Figure 8: Nonzero features identified in the NBA dataset and their magnitudes of structural change.

Possession Distribution (Mechanism I): The structural shift is most heavily weighted towards offensive usage and volume metrics. Specifically, the model identifies *Est. Usage % Rank*, *Usage %FGM Rank*, and *Usage %FGA Rank*, alongside aggregate volume indicators like *FGM Rank*, *FGA Rank*, and *FGA (Scoring)*. The concentration of these six features confirms a consolidation of offensive responsibilities, consistent with usage patterns centered on star players predicted by Tournament Theory Price et al. (2010): as teams vie for playoff positions, the marginal value of each possession increases, intensifying individual incentives.

Game Pace (Mechanism II): The second mechanism is characterized by *Est. Pace Rank* and *%Pts Off TOV Rank*. The prominence of *Est. Pace Rank* (the second-highest magnitude feature) quantifies the slowdown in gameplay. Concurrently, the shift in *%Pts Off TOV Rank* reflects the downstream effect of this slowdown: fewer possessions and tighter defensive transition discipline reduce the frequency of opportunistic transition scoring.

Physical Intensity (Mechanism III): The identification of *BLKA Rank* (Blocked Attempts) captures the escalation in defensive physicality. Unlike simple foul counts, which are subject to whistle variability, a structural shift in blocked shots indicates a tangible increase in shot contention and rim protection.

These latter two mechanisms, Pace and Physicality, exhibit strong concordance with established shifts in officiating standards. Empirical observations indicate a distinct variation in whistle frequency between the periods before and after the All-Star break. As noted by Laxdal et al. (2022), officiating logic functions as a primary exogenous variable that directly regulates both the speed of play (Mechanism II) and the permissible level of physical contact (Mechanism III). The ability of SMART to isolate these specific evolutions, which map directly to these institutional adjustments, strongly corroborates the empirical validity of our structural shift detection.