Journal of Machine Learning Research 27 (2026) 1-41 Submitted 6/25; Revised 4/26; Published 9/26

Reluctant Interaction Inference after Additive Modeling

Yiling Huang YILINGH@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1107, USA

Snigdha Panigrahi PSNIGDHA@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1107, USA

Guo Yu GUOYU@UCSB.EDU
Department of Statistics and Applied Probability
University of California, Santa Barbara
Santa Barbara, CA 93106-3110, USA

Jacob Bien JBIEN@USC.EDU
Department of Data Sciences and Operations
University of Southern California
Los Angeles, CA 90089-0809, USA

Editor: Rahul Mazumder

Abstract

Additive models enjoy the flexibility of nonlinear models while still being readily understandable to humans. By contrast, other nonlinear models, which involve interactions between features, are not only harder to fit but also substantially more complicated to explain. Guided by the principle of parsimony, a data analyst therefore may naturally be reluctant to move beyond an additive model unless it is truly warranted.

To put this principle of interaction reluctance into practice, we formulate the problem as a hypothesis test with a fitted sparse additive model (SPAM) serving as the null. Because our hypotheses on interaction effects are formed after fitting a SPAM to the data, we adopt a selective inference approach to construct p-values that properly account for this data adaptivity. Our approach makes use of external randomization to obtain the distribution of test statistics conditional on the SPAM fit, allowing us to derive valid p-values, corrected for the over-optimism introduced by the data-adaptive process prior to the test. Through experiments on simulated and real data, we illustrate that—even with small amounts of external randomization—this rigorous modeling approach enjoys considerable advantages over naive methods and data splitting.

Keywords: selective inference, additive models, non-linearity, interpretable machine learning, sparse additive models

1. Introduction

The past decade has seen the widespread adoption of highly nonlinear models such as neural networks. This trend has been accompanied by concerns about the reliability of

©2026 Yiling Huang, Snigdha Panigrahi, Guo Yu, and Jacob Bien.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-1344.html.

such highly flexible models, when deployed in real-world settings (D'Amour et al., 2022). In particular, Fisher et al. (2019) emphasize Leo Breiman's *Rashomon effect*, the observation that there can often be many equivalently predictive models (Breiman, 2001). In light of this, Rudin et al. (2022) urge machine learners to adopt models that directly lend themselves to interpretation, such as additive models.

Additive models enjoy the flexibility of nonlinear models while still being readily understandable to humans. By contrast, other nonlinear models that involve interactions between features are more difficult to fit and far harder to interpret, as they require explaining how one or more features modulate the effects of other features on the response. Given the challenges of fitting and interpreting models with interactions, analysts should be cautious about moving beyond an additive model unless it is truly warranted. Motivated by this reasoning, Yu et al. (2019) introduce a *reluctance* principle that prioritizes the main effects in their modeling approach by including only those interactions that are highly correlated with the residuals after fitting a model with main effects and their squares.

Consider an illustrative example where $Y = X_1X_2 + \varepsilon$. Suppose X_2 is highly correlated with X_1, in the sense that $\mathrm{Var}(X_2 - E(X_2|X_1))$ is close to zero. Then $Y \approx X_1E(X_2|X_1) + \varepsilon$, so the interaction can be well-approximated by a nonlinear function of X_1 alone. In this case, the reluctance principle favors modeling the response as a simple additive model in the main effects, rather than introducing interaction terms. The reluctance principle motivates a sequential modeling strategy: we first fit an additive model that allows flexible (potentially nonlinear) effects of individual predictors, and only introduce interactions if there remains substantial signal that cannot be explained by these main effects.

In this paper, we seek a high standard of evidence for the existence of an interaction between two features, and in line with the reluctance principle described in Yu et al. (2019), we ask the following question: is there statistical evidence that a linear interaction between features is at play after having fit a SPAM to the data? We formulate this task as a hypothesis testing problem in which the SPAM fitted to our data acts as the null model. The p-values from our hypothesis testing approach guide us to a decision on whether to add specific interaction effects to the SPAM.

Taking such a testing approach, as natural as it may seem, falls outside the scope of classical statistical theory. This is because the hypotheses being tested are data-adaptive in nature, involving a SPAM fit that is obtained only *after* looking at the data. Typically, a SPAM is fitted by solving a group lasso problem, where each group consists of the basis expansions for a feature, and the group lasso penalty promotes the selection of relevant main effects in the model. A naive approach that ignores the data-adaptive nature of the tested hypotheses—using simple z-tests or t-tests for regression to compute p-values—may result in overly optimistic p-values. As a consequence, spurious interactions are more likely to be included in the model, even though they contribute no additional explanatory power beyond what is already captured by the SPAM fit, ultimately leading to a less accurate and a less interpretable model.

Our paper develops a selective inference method for constructing valid p-values for data-adaptive hypotheses on interaction effects, grounded in the principle of reluctance. Unlike the naive p-values, our approach properly accounts for the SPAM fit and corrects for the over-optimism introduced by the data-adaptive process that comes prior to the test. More precisely, we derive valid p-values from the conditional distribution of test statistics,

given the groups of main effects included in the SPAM fit. This conditional distribution is typically difficult to characterize and lacks a closed-form expression. To overcome this challenge, we make use of recently developed machinery in Panigrahi et al. (2023); Huang et al. (2025) that adds external Gaussian randomization to the group lasso objective. The main highlights of our randomized approach are as follows:

(1) The external randomization in our method facilitates tractable inference. By applying a change of variables to the Gaussian density of the randomization, we obtain a closed-form expression for the conditional distribution that accounts for the SPAM fit. We present an efficient algorithm for computing approximate p-values and confidence intervals based on the conditional distribution.

(2) The added randomization generates a noisy version of the data, enabling a flexible lever to control how much information in our data is used in the first-stage SPAM fit. This is similar to data splitting, another randomized method where the split proportion controls the amount of data used for selection versus inference. However, unlike data splitting, our p-values are based on statistics from the full dataset, rather than relying on just the holdout data used for inference in a splitting-based approach.

As a result, our p-values are more powerful than those from data splitting. Moreover, our method remains stable even as non-linear expansions of predictors and interactions create high correlations among features; by contrast, data splitting can fail to provide p-values in such scenarios, due to the smaller sample size of the holdout data which often leads to rank-deficient design matrices.

(3) In our numerical experiments, conducted across a wide range of settings, we employ Gaussian randomization variables with a small variance parameter. This ensures that the SPAM fit on the randomized (noisy) data closely matches the fit obtained using the full dataset. As a result, we are able to use nearly all the available data during both the main effects modeling stage and the interaction inference stage, thereby making better use of the available data. We show that by taking the more rigorous selective inference approach, we consistently obtain a more accurate non-linear model than both the naive approach and the data splitting method.

1.1 Related Work

Before proceeding, we briefly review some work in three areas related to our contributions: nonlinear additive modeling, interaction modeling, and selective inference.

Additive modeling. Additive models offer the best of both worlds: they are powerful tools for nonlinear modeling while still producing fairly interpretable models. The work by Ravikumar et al. (2009) introduced the use of a group lasso-type of penalty or ℓ_2-penalization to fit SPAMs to high-dimensional data (relatedly, Lin and Zhang (2006)). Lou et al. (2016) extended this idea to high-dimensional settings by proposing a nested group lasso penalty, which allows for an adaptive choice of features to be treated as either linear or nonlinear. Other extensions include Chouldechova and Hastie (2015); Tay and Tibshirani (2020); Haris et al. (2022). When a small number of interactions are incorporated, additive models have been shown to achieve predictive accuracy on par with state-of-the-art

nonlinear models, such as random forests, as demonstrated in the work of Caruana et al. (2015).

Interaction modeling. A standard approach for having main effects inform which interactions are included is through the hierarchical interaction assumption (also known as marginality and heredity) Nelder (1977); Peixoto (1987). Over the past few decades, several methods have formulated optimization-based approaches based on hierarchy, including Bien et al. (2013); Yuan et al. (2009); Haris et al. (2016); Hazimeh and Mazumder (2020). By contrast, Yu et al. (2019) move away from hierarchy and instead adopt the reluctance principle. In their approach, interactions are considered for inclusion if they contribute to explaining the variability of the response in addition to an additive model. These existing methods focus primarily on building a predictive model with interactions, with relatively few providing inferential guarantees about individual interaction effects. In our work, we aim to rigorously implement the reluctant approach for including interactions to additive models by adopting a confirmatory framework based on hypothesis testing.

Selective inference. To test data-adaptive hypotheses, which are formed after examining the data, selective inference tools are essential. A selective inference approach to our problem, where the null hypothesis is based on the SPAM fit to the given data, involves obtaining p-values by conditioning on this initial SPAM fit before testing for interactions. Early work along these lines, such as Lee et al. (2016), relied on the polyhedral nature of the conditioning event of the lasso to construct inferential procedures. As a result, this approach does not extend to many other selection problems, including the selection of a SPAM in our case, where the conditioning event is no longer polyhedral in shape.

Panigrahi et al. (2023) introduced a conditional selective inference method after solving a group lasso problem with Gaussian data, which was later developed for more general models using a maximum likelihood approach in Huang et al. (2025). This line of work proposed adding external randomization, in the form of Gaussian noise, to the group lasso optimization objective to obtain a tractable conditional distribution, without depending on the specific geometry of the selection event. Furthermore, the randomized approach has the added benefit of producing more powerful inference compared to non-randomized methods, as shown for high-dimensional regression problems in Panigrahi et al. (2023); Kivaranovic and Leeb (2024); Panigrahi et al. (2024).

Panigrahi and Taylor (2023) proposed a maximum likelihood approach incorporating randomization for inference, which we adopt in this paper to perform inference on the interaction effects, given our initial SPAM fit. New variants of data splitting have been proposed for Gaussian data in Rasines and Young (2023) and extended to other parametric distributions in Leiner et al. (2025); Neufeld et al. (2024); Dharamshi et al. (2025). However, like the widely used sample splitting approach, these methods condition on all the data used for selection, thereby conditioning on substantially more information than is necessary. In a parallel line of work, simultaneous methods for selective inference, as pursued in Berk et al. (2013); McCloskey (2024); Zrnic and Fithian (2024), provide guarantees for the set of plausible data-adaptive targets that could have been formed, rather than for the observed event, such as the observed SPAM fit in our problem. Moreover, there is no existing work on how this approach could be applied to inference after the group lasso in an easily tractable manner. Finally, we note that selective inference methods have been developed for non-linear regression models that do not involve additive modeling. For example, Neufeld

et al. (2022); Bakshi et al. (2024) develop selective inference for regression trees, while Suzumura et al. (2017) develop methods for higher-order interaction models, leveraging the tree structure of various interaction orders.

The rest of this paper is organized as follows. In Section 2, we present the proposed method and illustrate its main features with a motivating data example. In Section 3, we derive the conditional distribution of key statistics, which yields p-values for testing the data-adaptive hypotheses based on the SPAM fit to given data. In Section 4, we study the performance of our method under various settings, and in Section 5, we apply our method to NYC flight data to predict flight arrival delay. We conclude the work with Section 6. All proofs and additional data analysis results for Section 5 are collected in the Appendix.

2. The Outline of Our Method

Motivated by the reluctance principle in modeling interactions, we introduce a hypothesis testing method to determine if there is an interaction between features after fitting a SPAM. Our method, described in Algorithm 1, proceeds in two steps.

Step 1: In Step 1, we fit a SPAM by solving a randomized group lasso problem, which we display in (1) after first introducing some notation. Each feature $X_j \in \mathbb{R}^n$ is expanded using a B_j-dimensional spline basis to give $\Psi_j \in \mathbb{R}^{n \times B_j}$, leading to a design matrix $\Psi = [\Psi_1 \cdots \Psi_p] \in \mathbb{R}^{n \times q}$, whose $q = \sum_{j=1}^p B_j$ columns are naturally partitioned into p groups.

For a user-specified positive definite matrix Ω, let $\omega \sim N(0_q, \Omega)$ be a randomization variable drawn independently of y. To identify the main effects in our model, we solve:

$$\widehat{\beta} = \underset{(\beta_{g_j} \in \mathbb{R}^{B_j})_{j \in [p]}}{\arg\min} \frac{1}{2} \left\| y - \sum_{j=1}^p \Psi_j \beta_{g_j} \right\|_2^2 + \sum_{j=1}^p \left(\lambda_j \|\beta_{g_j}\|_2 + \frac{\epsilon}{2} \|\beta_{g_j}\|_2^2 \right) - \left(\beta_{g_1}^\top \ \ldots \ \beta_{g_p}^\top \right) \omega, \quad (1)$$

where $\{\lambda_j\}_{j \in [p]}$ are positive tuning parameters, and $\epsilon > 0$ is a small ridge penalty parameter that guarantees strict convexity of the problem and thus the uniqueness of its solution.

Based on the solution $\widehat{\beta} = \left(\widehat{\beta}_{g_1}^\top \ \ldots \ \widehat{\beta}_{g_p}^\top \right)^\top$, let

$$\widehat{M} = \left\{ j \in [p] : \widehat{\beta}_{g_j} \neq 0_{B_j} \right\} \quad (2)$$

denote the set of selected groups (features). Finally, suppose M is the observed value of $\widehat{M}$ for the given data and randomization (y, ω), which determines the additive main effects selected in the SPAM fit. Let $q_M = \sum_{j \in M} B_j$, and let $\Psi_M \in \mathbb{R}^{n \times q_M}$ and $\Psi_{-M} \in \mathbb{R}^{n \times (q - q_M)}$ be a column concatenation of the matrices $\{\Psi_j : j \in M\}$ and $\{\Psi_j : j \notin M\}$, respectively.

Step 2: In Step 2, we perform a hypothesis test to determine if we must include linear interactions between features while taking into account the SPAM fit in Step 1.

For a pair $(j, k) \in [p]^2$, we denote the interaction between the features X_j and X_k by $\mathcal{I}_{jk} = X_j \circ X_k \in \mathbb{R}^n$, where $\circ$ represents the component-wise product of two vectors. We use $\mathcal{T}^M \subseteq [p]^2$ to represent the indices of interactions that we want to test for inclusion in our model. Note that we allow this set to depend on data through the main effects model M from the previous step, which is indicated by superscript M. For example, under a weak hierarchy rule for including interactions, $\mathcal{T}^M = \left\{ (j, k) \in [p]^2 : j \in M \text{ or } k \in M \right\}$.

In this step, we conduct a marginal screening of interactions in the set $\mathcal{T}^M$. Specifically, for each $(j,k) \in \mathcal{T}^M$, we obtain a test for $H_0^{jk} : \theta_{jk}^M = 0$ in the model:

$$y \sim N(\Psi_M \beta_{jk}^M + \mathcal{I}_{jk}\theta_{jk}^M, \sigma^2 I_n), \tag{3}$$

where $\beta_{jk}^M \in \mathbb{R}^{q_M}$ and $\theta_{jk}^M \in \mathbb{R}$ represent the main effect parameter vector and the interaction effect for $\mathcal{I}_{jk}$, respectively. More precisely, we provide a two-sided p-value $\widehat{p}_{jk}^M$ that satisfies

$$\mathbb{P}_{H_0^{jk}}\left(\widehat{p}_{jk}^M \leq \alpha \mid \widehat{M} = M\right) \leq \alpha, \tag{4}$$

at a predetermined test level α. Conditioning on $\{\widehat{M} = M\}$ enables us to account for the adaptivity in the choice of the null hypothesis, including both the SPAM fit to the data obtained in Step 1 and the choice of interaction given this fit.

Algorithm 1 The proposed reluctant interaction inference framework

Step 1:

Require: y: the response vector, Ψ: the design matrix using a spline expansion of features, ω: the randomization vector

Fit a SPAM by solving (1).

Output: M

Step 2:

Specify: $\mathcal{T}^M \subseteq [p]^2$, a set of interaction indices that is allowed to depend on M

Using the model in (3), obtain a p-value $\widehat{p}_{jk}^M$ to test the null hypothesis $H_0^{jk} : \theta_{jk}^M = 0$ for each $(j,k) \in \mathcal{T}^M$ that satisfies (4).

Output: $\mathcal{S}^M \subseteq \mathcal{T}^M$: the set of interactions with significant p-values

Furthermore, if simultaneous screening of a set of interactions $\mathcal{C} \subset \mathcal{T}^M$ is of interest, our proposed framework admits the extension to simultaneous screening of $\mathcal{C}$. For example, a simultaneous screening $\mathcal{C} = \left\{(j_0,k) \in [p]^2 : k \in [p]\right\}$ for some $j_0 \in M$ answers the question of whether j_0 has any interaction with other features that significantly improves the model fit. For any arbitrary $\mathcal{C}$, we develop a test for the global null:

$$H_0^{\mathcal{C}} : \theta_{jk}^M = 0, \quad \text{for all } (j,k) \in \mathcal{C},$$

with corresponding p-value $\widehat{p}_{\mathcal{C}}^M$ satisfying

$$\mathbb{P}_{H_0^{\mathcal{C}}}\left(\widehat{p}_{\mathcal{C}}^M \leq \alpha \mid \widehat{M} = M\right) \leq \alpha.$$

2.1 A Motivating Data Example

We simulate a data example to illustrate the inferential performance of our two-step method described in Algorithm 1. In particular, this example highlights the role of external randomization and demonstrates a trade-off between the quality of the first-step SPAM fit and

the inferential power for recovering interactions in the second step. The data are generated from a nonlinear model, with details deferred to Section 4.

We compare inference for the interaction coefficient θ^M_{jk} using three different methods: (i) naive inference, which ignores the fact that a main-effects model was selected using the data, and proceeds with a z-test as if the SPAM were fixed in advance; (ii) data splitting, which partitions the data into two independent subsamples, one for selecting the SPAM and the other for inference on θ^M_{jk}; and (iii) our proposed method using conditional inference.

For our method, we implement randomization using $\Omega = \sigma^2 \frac{1-r}{r} \Psi^\top \Psi$, a natural choice for the randomization covariance matrix discussed further in Section 4.2. In the data splitting approach, $(r \times 100)\%$ of the data is used for selection. The parameter $r \in (0, 1)$ determines the trade-off between data used for selection and data reserved for inference. While r plays a similar role in our method, an important distinction is that we use the full data for inference rather than using only the holdout data as is done in data splitting.

Figure 1 shows the empirical CDFs of the pivots for inference on θ^M_{jk} constructed using the three methods. These pivots are obtained by applying a probability integral transform to the distribution of the test statistic used in each method; see Section 4 for their exact definitions. If the inference is valid, these pivots are expected to follow a uniform distribution. While the pivots from naive inference deviate from the uniform distribution, as expected, the pivots from data splitting and the proposed method closely follow the uniform distribution, indicating the validity of inference with both approaches.

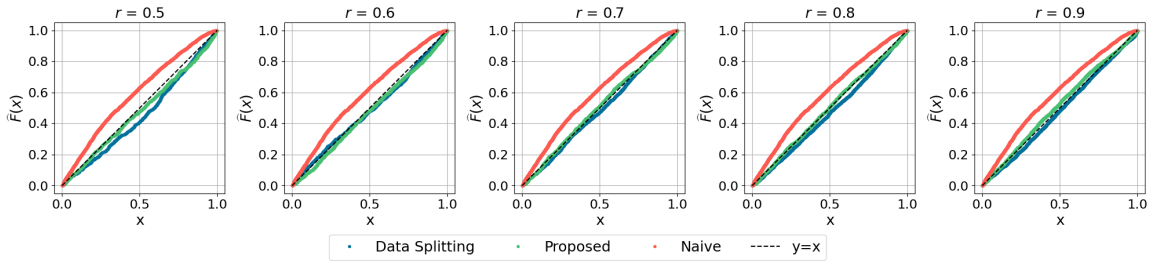


Figure 1: ECDF of the uniform pivots obtained by naive inference, data splitting, and the proposed method with varying amount of information used for selection (r)

One might be concerned that randomizing the SPAM will compromise the quality of main effects selection. However, when the amount of noise introduced during the randomization is sufficiently small, the randomized SPAM is able to recover selection performance comparable to that of the original, non-randomized approach. To highlight this, in Figure 2, we vary r along the x-axis of the plots. The left panel displays the F1 scores, defined as the harmonic mean of precision and recall, which measure the quality of main effects recovery, or equivalently, the SPAM fit in Step 1. As r approaches 1—meaning a larger fraction of the data is used for fitting the SPAM—we observe that the selection quality of both the proposed method and data splitting gradually converges to that of the naive method, which uses the full dataset for selecting main effects.

The right panel of Figure 2 shows that inference for θ^M_{jk} following the SPAM fit must account for the selection of the main effects. As expected, the average lengths of the

confidence intervals produced by data splitting and our method widen as r increases toward 1. However, the proposed method consistently yields shorter confidence intervals compared to data splitting, with the advantage in inferential power becoming even more pronounced for larger values of r.

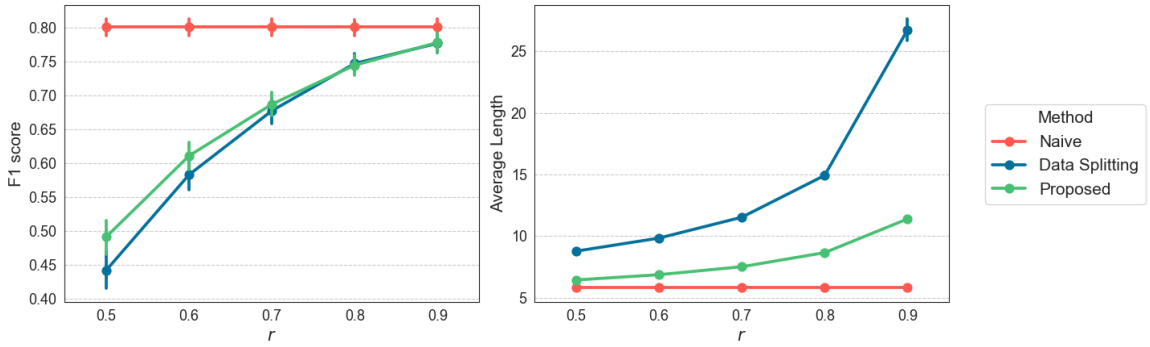


Figure 2: F1 scores for true main effects selection (left) and average confidence interval lengths for interaction coefficients (right) obtained by naive inference, data splitting, and the proposed method with varying amount of information used for selection (r)

Since we would like randomization to have minimal effect on the performance of Step 1, we fix $r = 0.9$ for all following numerical experiments, at which point our randomized method closely matches the selection quality of the naive approach. As such, our method enables the use of nearly the entire dataset for fitting the SPAM while still leveraging the full data for powerful inference on interactions. These observations empirically support the main highlights (1), (2), and (3) of our method, outlined in Section 1.

3. Inference Procedures

In this section, we derive the p-values that we will use in Step 2 of Algorithm 1. Although (y, ω) have a simple joint normal distribution due to their independence, the conditional distribution of (y, ω) given the SPAM fit lacks a straightforward characterization. To address this, we perform a change of variables on the known density of the randomization variables. This reduces the problem to characterizing the conditional distribution of y and the randomized group lasso estimators. The change of variables transformation serves a dual purpose: (i) it yields a simple representation of the selection event as a set of sign constraints on the randomized group lasso estimator, and (ii) it enables a closed-form characterization of the joint conditional distribution of y and the group lasso estimator.

We provide here a roadmap of what is to come. In particular, in Section 3.1, we show that the KKT conditions, when suitably parameterized, yield a simple characterization of the event that a certain set of features was selected by the SPAM fit in Step 1. In Section 3.2, we describe the key statistics that will be involved in the p-value and derive their distribution if the set M were fixed rather than selected based on the data. In Section 3.3, we work out the exact conditional distribution of these key statistics given the event that SPAM selected

a certain set of main effects. In Section 3.4, we develop a Laplace approximation to the likelihood derived from this conditional distribution, making inference easier to implement. In Section 3.5, we show how to perform inference in this approximate likelihood. Finally, in Section 3.6, we present an extension of our method to simultaneous screening of multiple interaction terms.

R.V. / realization	Space	Meaning
ω	$\mathbb{R}^q$	Randomization $\omega \sim N(0_q, \Omega)$
$\widehat{\beta}$	$\mathbb{R}^q$	Randomized group lasso estimator, defined in (1)
$\widehat{\beta}_{g_j}$	$\mathbb{R}^{B_j}$	The Ψ_j-segment of the randomized group lasso estimator
$\widehat{M}$ / M	$\mathcal{P}([p])$[a]	Index set of selected features: $\widehat{M} = \left\{ j \in [p] : \widehat{\beta}_{g_j} \neq 0_{B_j} \right\}$
$\begin{pmatrix} \widehat{\theta}^M_{jk} \\ \widehat{\beta}^M_{jk} \end{pmatrix}$	$\mathbb{R}^{q_M+1}$	Least squares estimator of parameters in model (3), defined in (8)
$\widehat{\mathcal{A}}^M_{jk}$ / $\mathcal{A}$	$\mathbb{R}^{q-q_M}$	The ancillary statistic, whose value we condition on, defined as: $\widehat{\mathcal{A}}^M_{jk} = -\Psi^\top_{-M} \left(y - \overline{Z}^M_{jk} \begin{pmatrix} \widehat{\theta}^M_{jk} \\ \widehat{\beta}^M_{jk} \end{pmatrix} \right)$
$\widehat{\gamma}_j$	$\mathbb{R}$	Norm of the Ψ_j-segment of the randomized group lasso estimator, defined as: $\widehat{\gamma}_j = \|\widehat{\beta}_{g_j}\|_2$
$\widehat{u}_j$	$\mathbb{R}^{B_j}$	Sub-gradient vector of the group lasso penalty for $j \in M$: $\widehat{u}_j = \frac{\widehat{\beta}_{g_j}}{\|\widehat{\beta}_{g_j}\|_2}$, where $\widehat{\beta}_{g_j} \neq 0_{B_j}$
$\widehat{z}_j$	$\mathbb{R}^{B_j}$	Sub-gradient vector of the group lasso penalty for $j \notin M$: $\|\widehat{z}_j\|_2 \leq 1$, where $\widehat{\beta}_{g_j} = 0_{B_j}$
$\widehat{\gamma}$ / γ	$\mathbb{R}^{\lvert M \rvert}$	The stacked vector of $\widehat{\gamma}_j$'s: $\widehat{\gamma} = (\widehat{\gamma}_j : j \in M)$
$\widehat{U}$ / U	$\mathbb{R}^{q_M}$	The stacked vector of $\widehat{u}_j$'s: $\widehat{U} = (\widehat{u}_j : j \in M)$
$\widehat{Z}$ / Z	$\mathbb{R}^{q-q_M}$	The stacked vector of $\widehat{z}_j$'s: $\widehat{Z} = (\widehat{z}_j : j \notin M)$

[a] $\mathcal{P}([p])$ denotes the power set of $[p] = \{1, \ldots, p\}$.

Table 1: Random variables for computing selective inference

3.1 The Conditioning Event

To account for the selection of the SPAM prior to conducting tests for screening interactions, we condition on the event $\{\widehat{M} = M\}$. Although this event does not have a simple description, there exists a subset of this event that can be characterized by a set of simple

sign constraints on the group lasso estimator $\widehat{\beta}$. Below, we review this characterization, which was identified in previous work by Panigrahi et al. (2023); Huang et al. (2025).

Hereafter, in the solution vector $\widehat{\beta}$, we assume without loss of generality that the nonzero groups where $\widehat{\beta}_{g_j} \neq 0$ are stacked above the zero groups where $\widehat{\beta}_{g_j} = 0$. Consider the stationarity conditions of (1), which imply

$$0_q \in -\Psi^\top y - \omega + \left(\Psi^\top \Psi + \epsilon I_q\right) \begin{pmatrix} \left(\widehat{\beta}_{g_j}\right)_{j\in M} \\ \left(0_{B_j}\right)_{j\notin M} \end{pmatrix} + \begin{pmatrix} \left(\lambda_j \dfrac{\widehat{\beta}_{g_j}}{\|\widehat{\beta}_{g_j}\|_2}\right)_{j\in M} \\ \left(\lambda_j \partial \|\widehat{\beta}_{g_j}\|_2\right)_{j\notin M} \end{pmatrix}. \tag{5}$$

Even though the stationarity conditions only need a subgradient to exist, the strictly convex quadratic loss function ensures that the subgradient is always unique. More precisely, for $j \notin M$, there exists a unique subgradient $\widehat{z}_j \in \partial \|\widehat{\beta}_{g_j}\|_2$ ($\iff$ $\|\widehat{z}_j\|_2 \leq 1$) such that equality in (5) holds, i.e.,

$$\omega = -\Psi^\top y + \left(\Psi^\top \Psi + \epsilon I_q\right) \begin{pmatrix} (\widehat{\gamma}_j \widehat{u}_j)_{j\in M} \\ \left(0_{B_j}\right)_{j\notin M} \end{pmatrix} + \begin{pmatrix} (\lambda_j \widehat{u}_j)_{j\in M} \\ (\lambda_j \widehat{z}_j)_{j\notin M} \end{pmatrix}, \tag{6}$$

where, for $j \in M$, $\widehat{\gamma}_j = \|\widehat{\beta}_{g_j}\|_2$, and $\widehat{\gamma}_j > 0$, $\widehat{u}_j = \frac{\widehat{\beta}_{g_j}}{\|\widehat{\beta}_{g_j}\|_2}$.

Additionally, we define

$$\widehat{\gamma} = (\widehat{\gamma}_j : j \in M) \in \mathbb{R}^{|M|}, \quad \widehat{U} = (\widehat{u}_j : j \in M) \in \mathbb{R}^{q_M}, \quad \widehat{Z} = (\widehat{z}_j : j \notin M) \in \mathbb{R}^{q-q_M},$$

as estimators that satisfy (6), where $(\widehat{u}_j : j \in M)$ denotes the stacking of $|M|$ vectors $\widehat{u}_j$'s, indexed by the set M, into one vector, and similarly for $(\widehat{z}_j : j \notin M)$. These notations, along with a few additional ones introduced next, are summarized in Table 1 for the reader's convenience.

We have written the KKT conditions in terms of $(\widehat{\gamma}, \widehat{U}, \widehat{Z})$ in place of $(\widehat{\beta}, \widehat{U}, \widehat{Z})$ because of the following lemma, which establishes a simplifying equivalence of events.

Lemma 1 (Equivalent characterization of model selection (Panigrahi et al., 2023; Huang et al., 2025))**.** *Consider solving* (1)*. Then, it holds that*

$$\left\{\widehat{M} = M,\ \widehat{U} = U,\ \widehat{Z} = Z\right\} = \left\{\widehat{\gamma} \succ 0_{|M|},\ \widehat{U} = U,\ \widehat{Z}\ = Z\right\}. \tag{7}$$

The event on the left-hand side of (7) is a proper subset of $\{\widehat{M} = M\}$, which is characterized by a set of positive constraints on the ℓ_2-norms of the nonzero groups in the solution vector $\widehat{\beta}$. Conditioning on this event takes into account the selection of main effects in the SPAM fit.

3.2 The Key Statistics

In this section, we introduce what we call the "key statistics" that are involved in constructing a p-value for the null hypothesis $H_0^{jk} : \theta_{jk}^M = 0$ in (3). In particular, our key statistics form a $(q+1)$-dimensional function of y, from which a p-value for testing H_0^{jk} can be constructed based on their distribution conditional on $\{\widehat{M} = M\}$.

Denote by $\overline{Z}_{jk}^M = \left(\mathcal{I}_{jk} : \Psi_M\right) \in \mathbb{R}^{n\times(q_M+1)}$, the interaction-augmented design matrix obtained by concatenating Ψ_M with $\mathcal{I}_{jk}$. Let

$$\begin{pmatrix}\widehat{\theta}_{jk}^M \\ \widehat{\beta}_{jk}^M\end{pmatrix} = \underset{\theta\in\mathbb{R},\ \beta\in\mathbb{R}^{q_M}}{\arg\min} \|y - \Psi_M\beta - \mathcal{I}_{jk}\theta\|_2^2 = \left(\overline{Z}_{jk}^M\right)^+ y \tag{8}$$

be the least squares estimator when we perform a regression of y against the derived features in $\overline{Z}_{jk}^M$, where $A^+ = (A^\top A)^{-1}A^\top$ denotes the Moore-Penrose inverse of matrix A. However, the selection of the main effects not only depends on the least squares estimator based on the model (3), but also on the residuals from this model fit and the features excluded from the SPAM fit. To see this, observe that y can be written as the sum of the predicted value and the residuals from the fit as

$$y = \overline{Z}_{jk}^M \begin{pmatrix}\widehat{\theta}_{jk}^M \\ \widehat{\beta}_{jk}^M\end{pmatrix} + \left(y - \overline{Z}_{jk}^M \begin{pmatrix}\widehat{\theta}_{jk}^M \\ \widehat{\beta}_{jk}^M\end{pmatrix}\right).$$

This in turn allows us to decompose $-\Psi^\top y$ involved in the KKT conditions in (6) as

$$-\Psi^\top y = -\Psi^\top \overline{Z}_{jk}^M \begin{pmatrix}\widehat{\theta}_{jk}^M \\ \widehat{\beta}_{jk}^M\end{pmatrix} + \begin{pmatrix}0_{q_M} \\ \widehat{\mathcal{A}}_{jk}^M\end{pmatrix}, \tag{9}$$

where $\widehat{\mathcal{A}}_{jk}^M = -\Psi_{-M}^\top \left(y - \overline{Z}_{jk}^M \begin{pmatrix}\widehat{\theta}_{jk}^M \\ \widehat{\beta}_{jk}^M\end{pmatrix}\right) \in \mathbb{R}^{q-q_M}$ involves the main effects excluded from the SPAM fit. The 0_{q_M} follows from the normal equations of (8).

We denote the collection of these key statistics by

$$\widehat{S}_{jk}^M = \left(\widehat{\theta}_{jk}^M,\ \left(\widehat{\beta}_{jk}^M\right)^\top,\ \left(\widehat{\mathcal{A}}_{jk}^M\right)^\top\right)^\top \in \mathbb{R}^{q+1}. \tag{10}$$

We start by stating their distribution for a fixed M (i.e., not conditioning on the event $\{\widehat{M} = M\}$) in Lemma 2. Directly doing inference based on Lemma 2 but with M replaced with $\widehat{M}$ would not properly account for the selection of main effects from SPAM and corresponds to what we will call throughout the naive approach.

Lemma 2 (Marginal distribution of the key statistics)**.** *For any fixed set of additive main effects M and any fixed pair of indices (j,k), consider the model*

$$y \sim N_n(\Psi_M\beta_{jk}^M + \mathcal{I}_{jk}\theta_{jk}^M, \sigma^2 I_n).$$

In this model, $\widehat{S}_{jk}^M \sim N_{q+1}\left(\nu_{jk}^M, \Sigma_{jk}^M\right)$ where

$$\nu_{jk}^M = \begin{pmatrix}\theta_{jk}^M \\ \beta_{jk}^M \\ 0_{q-q_M}\end{pmatrix}, \qquad \Sigma_{jk}^M = \begin{pmatrix}\overline{\Sigma}_{jk}^M & 0_{(q_M+1)\times(q-q_M)} \\ 0_{(q-q_M)\times(q_M+1)} & \widetilde{\Sigma}_{jk}^M\end{pmatrix},$$

$$\overline{\Sigma}_{jk}^M = \sigma^2\left(\left(\overline{Z}_{jk}^M\right)^\top \overline{Z}_{jk}^M\right)^{-1} \in \mathbb{R}^{(q_M+1)\times(q_M+1)},\ \textit{and}$$

$$\widetilde{\Sigma}_{jk}^M = \sigma^2\Psi_{-M}^\top\left(I_n - \overline{Z}_{jk}^M\left(\overline{Z}_{jk}^M\right)^+\right)\Psi_{-M} \in \mathbb{R}^{(q-q_M)\times(q-q_M)}.$$

Note that $\widehat{\mathcal{A}}_{jk}^M$ is an ancillary statistic for the parameters $\theta_{jk}^M, \beta_{jk}^M$ in our model. A p-value for θ_{jk}^M, when the set of main effects M and the indices (j,k) are fixed, can be

obtained from the Gaussian distribution of $\left(\widehat{\theta}^M_{jk}, (\widehat{\beta}^M_{jk})^\top\right)^\top \mid \widehat{\mathcal{A}}^M_{jk} \stackrel{d}{=} \left(\widehat{\theta}^M_{jk}, (\widehat{\beta}^M_{jk})^\top\right)^\top$, where the equality in distribution is due to independence between $\left(\widehat{\theta}^M_{jk}, (\widehat{\beta}^M_{jk})^\top\right)^\top$ and $\widehat{\mathcal{A}}^M_{jk}$.

To account for the selection of M using the same data, we adjust the naive distribution of the key statistics by using the conditioning event in Lemma 1. In what follows, we derive a valid p-value using the conditional distribution

$$\left(\widehat{\theta}^M_{jk}, (\widehat{\beta}^M_{jk})^\top\right)^\top \Big| \left\{\widehat{\gamma} \succ 0_{|M|},\ \widehat{U} = U,\ \widehat{Z}\ = Z,\ \widehat{\mathcal{A}}^M_{jk} = \mathcal{A}\right\}. \tag{11}$$

After conditioning, $\widehat{\mathcal{A}}^M_{jk}$ is no longer an ancillary statistic, and we have the option to either integrate it out or condition on it. We do the latter to avoid calculating an integral over $q - q_M$ dimensions. In the next section, we delve into the derivation of this distribution, which leads to a likelihood in $\left(\theta^M_{jk}, (\beta^M_{jk})^\top\right)^\top$.

3.3 Conditional Distribution of the Key Statistics

First, we fix some more notation and identify a mapping that allows us to compute the desired conditional distribution in (11).

Let $\mathrm{bd}(\widehat{U}) = \mathrm{diag}(\widehat{u}_j : j \in M) \in \mathbb{R}^{q_M \times |M|}$ denote the block-diagonal matrix with $\widehat{u}_j$ on the diagonal blocks for $j \in M$. Define

$$\mathcal{U}_M = \left\{u = (u_j : j \in M) \in \mathbb{R}^{q_M} : u_j \in \mathbb{S}^{|g_j|-1}\right\} \subseteq \mathbb{R}^{q_M},$$
$$\mathcal{Z}_M = \left\{z = (z_j : j \notin M) \in \mathbb{R}^{q-q_M} : \|z_j\|_2 \le 1\right\} \subseteq \mathbb{R}^{q-q_M},$$

where $(v_j : j \in M)$ denotes the stacking of $|M|$ vectors v_j, indexed by the set M, into one vector. Furthermore, denote by $\mathbb{S}^{k-1} = \{v \in \mathbb{R}^k : ||v||_2 = 1\}$ the unit sphere in $\mathbb{R}^k$. Let $Q_M = \Psi_M^\top \Psi_M + \epsilon I_{q_M}$ and let $\mathcal{U}_\perp^{(j)}(\widehat{u}_j) \in \mathbb{R}^{B_j \times (B_j - 1)}$ be an orthogonal basis completion of $\widehat{u}_j \in \mathbb{S}^{B_j - 1}$ for $j \in M$, and let

$$\mathcal{U}_\perp(\widehat{U}) = \mathrm{diag}\left(\left(\mathcal{U}_\perp^{(j)}(\widehat{u}_j)\right)_{j\in M}\right),\ \Gamma(\widehat{\gamma}) = \mathrm{diag}\left(\left(\widehat{\gamma}_j I_{B_j - 1}\right)_{j\in M}\right),\ \Lambda = \mathrm{diag}\left(\left(\lambda_j I_{B_j}\right)_{j\in M}\right).$$

Lemma 3. *Define the mapping* $\Pi_{\widehat{S}^M_{jk}}(\cdot,\cdot,\cdot) : \mathbb{R}^{|M|} \times \mathcal{U}_M \times \mathbb{R}^{q-q_M} \to \mathbb{R}^q$ *by*

$$\Pi_{\widehat{S}^M_{jk}}(\gamma, U, Z) = A^M_{jk} \begin{pmatrix} \widehat{\theta}^M_{jk} \\ \widehat{\beta}^M_{jk} \end{pmatrix} + B(U)\gamma + c(U, Z),$$

where

$$A^M_{jk} = -\Psi^\top \overline{Z}^M_{jk},\ B(U) = \left(\Psi^\top \Psi_M + \begin{pmatrix} \epsilon I_{q_M} \\ 0_{(q-q_M)\times q_M} \end{pmatrix}\right) \mathrm{bd}(U),$$
$$c(U, Z) = \begin{pmatrix} 0_{q_M} \\ \widehat{\mathcal{A}}^M_{jk} \end{pmatrix} + \begin{pmatrix} (\lambda_j u_j)_{j\in M} \\ (\lambda_j z_j)_{j\notin M} \end{pmatrix}.$$

Consider the estimators $(\widehat{\gamma}, \widehat{U}, \widehat{Z})$ *that satisfy the relationship in* (6)*. Then, it holds that*

$$(\widehat{\gamma}, \widehat{U}, \widehat{Z}) = \Pi^{-1}_{\widehat{S}^M_{jk}}(\omega).$$

As will be shown in the proof of Proposition 1, Lemma 3 provides us a tool for characterizing the distribution of the randomized group lasso statistics $(\widehat{\gamma},\ \widehat{U},\ \widehat{Z})$ through a change-of-variables applied to the known distribution of ω.

Proposition 1. *Denote $\phi(x; a, B)$ as the density function of $N(a, B)$ evaluated at x. The density function of the conditional distribution of*

$$\left(\widehat{S}^M_{jk},\ \widehat{\gamma},\ \widehat{U},\ \widehat{Z}\right)\ \Big|\ \left\{\widehat{\gamma} \succ 0_{|M|},\ \widehat{U} \in \mathcal{U}_M,\ \widehat{Z} \in \mathcal{Z}_M\right\}$$

when evaluated at (S, γ, U, Z) is proportional to

$$\phi(S; \nu^M_{jk}, \Sigma^M_{jk}) \cdot \phi\left(\Pi_S(\gamma, U, Z); 0_q, \Omega\right) \cdot \det D_{\Pi_S}(\gamma, U, Z) \cdot 1_{\mathbb{R}^{|M|}_+}(\gamma) \cdot 1_{\mathcal{U}_M}(U) \cdot 1_{\mathcal{Z}_M}(Z)$$

where $D_{\Pi_S}(\gamma, U, Z) = \Gamma(\gamma) + (\mathcal{U}_\perp(U))^\top Q_M^{-1} \Lambda\left(\mathcal{U}_\perp(U)\right)$.

We are now ready to derive the likelihood, referred to as the selective likelihood, that accounts for the selection of the main effects during the SPAM fit. We obtain an expression for this likelihood from the conditional density in Proposition 1, by conditioning further on the observed values of $\widehat{U}$, $\widehat{Z}$ and $\widehat{\mathcal{A}}^M_{jk}$. Before we do so, define

$$\begin{gathered}\overline{\Omega} = \left(B(\widehat{U})^\top \Omega^{-1} B(\widehat{U})\right)^{-1}, \quad \overline{A} = -\overline{\Omega} B(\widehat{U})^\top \Omega^{-1} A^M_{jk}, \quad \overline{b} = -\overline{\Omega} B(\widehat{U})^\top \Omega^{-1} c(\widehat{U}, \widehat{Z}), \\ \overline{\Theta} = \left((\overline{\Sigma}^M_{jk})^{-1} - \overline{A}^\top \overline{\Omega}^{-1} \overline{A} + (A^M_{jk})^\top \Omega^{-1} A^M_{jk}\right)^{-1}, \quad \overline{R} = \overline{\Theta}(\overline{\Sigma}^M_{jk})^{-1}, \\ \overline{s} = \overline{\Theta}\left(\overline{A}^\top \overline{\Omega}^{-1} \overline{b} - (A^M_{jk})^\top \Omega^{-1} c(\widehat{U}, \widehat{Z})\right),\end{gathered} \tag{12}$$

where A^M_{jk}, $B(\widehat{U})$, and $c(\widehat{U}, \widehat{Z})$ are as defined in Lemma 3.

Additionally, let

$$c(\theta^M_{jk}, \beta^M_{jk}) = \int\int_{g' \succ 0_{|M|}} \phi\left(b'; \overline{R}\begin{pmatrix}\theta^M_{jk} \\ \beta^M_{jk}\end{pmatrix} + \overline{s}, \overline{\Theta}\right) \cdot \phi\left(g'; \overline{A} b' + \overline{b}, \overline{\Omega}\right) \cdot \det D_{\Pi_S}(g', U, Z)\ dg'\ db'.$$

Theorem 1 (Selective log-likelihood)**.** *The log-likelihood of $\left(\widehat{\theta}^M_{jk}, (\widehat{\beta}^M_{jk})^\top\right)^\top$ conditional on*

$$\left\{\widehat{\gamma} \succ 0_{|M|},\ \widehat{U} = U,\ \widehat{Z}\ = Z,\ \widehat{\mathcal{A}}^M_{jk} = \mathcal{A}\right\},$$

is equal to

$$\ell_M(\theta^M_{jk},\ \beta^M_{jk}; U, Z, \mathcal{A}) = -\log c(\theta^M_{jk}, \beta^M_{jk}) + \log \phi\left(\begin{pmatrix}\widehat{\theta}^M_{jk} \\ \widehat{\beta}^M_{jk}\end{pmatrix}; \overline{R}\begin{pmatrix}\theta^M_{jk} \\ \beta^M_{jk}\end{pmatrix} + \overline{s}, \overline{\Theta}\right). \tag{13}$$

Theorem 1 provides us with an exact likelihood that accounts for the SPAM fit on the given data. This likelihood yields valid p-values for the hypothesis $H_0 : \theta^M_{jk} = 0$ for the interaction between X_j and X_k, where $(j, k) \in \mathcal{T}^M$.

Additionally, we note that this likelihood readily yields valid inference on the main effects β^M_{jk}, although this is not our focus in this paper. In the next section, we present a test statistic based on the maximum likelihood estimator for θ^M_{jk} in this likelihood and offer a computationally efficient algorithm for calculating a p-value for our hypothesis.

3.4 A Laplace Approximation for the Selective Likelihood

Following the line of recent research of approximate selective inference (Panigrahi et al., 2017; Huang et al., 2025; Panigrahi et al., 2023), we adopt a Laplace approximation to the normalizing constant $c(\theta^M_{jk}, \beta^M_{jk})$, where

$$c(\theta^M_{jk}, \beta^M_{jk}) \approx \sup_{g' \succ 0_{|M|}, b'} \phi\left(b'; \overline{R}\begin{pmatrix}\theta^M_{jk} \\ \beta^M_{jk}\end{pmatrix} + \overline{s}, \overline{\Theta}\right) \cdot \phi\left(g'; \overline{A} b' + \overline{b}, \overline{\Omega}\right) \cdot \det D_{\Pi_S}(g', U, Z).$$

We refer the readers to Huang et al. (2025) for formal probabilistic justifications for this approximation.

In practice, we handle the constrained optimization through the addition to the objective of a *barrier function*,

$$\mathrm{Barr}(g) = \sum_{k=1}^{|M|} \log(1 + 1/(g_k)),$$

which imposes an increasingly severe penalty as the variable g approaches the boundary of the positive orthant constraint. In particular, we approximate $c(\theta^M_{jk}, \beta^M_{jk})$ with $\widehat{c}(\theta^M_{jk}, \beta^M_{jk})$ where

$$\begin{aligned}\log \widehat{c}(\theta^M_{jk}, \beta^M_{jk}) = -\inf_{b',g'} \Bigg\{ & \frac{1}{2}\left(b' - \overline{R}\begin{pmatrix}\theta^M_{jk}\\ \beta^M_{jk}\end{pmatrix} - \overline{s}\right)^\top \overline{\Theta}^{-1}\left(b' - \overline{R}\begin{pmatrix}\theta^M_{jk}\\ \beta^M_{jk}\end{pmatrix} - \overline{s}\right) \\ & + \frac{1}{2}\left(g' - \overline{A}b' - \overline{b}\right)^\top \overline{\Omega}^{-1}\left(g' - \overline{A}b' - \overline{b}\right) \\ & - \log\det D_{\Pi_S}(g', U, Z) + \mathrm{Barr}(g')\Bigg\}.\end{aligned}$$

Dropping constants that do not depend on the parameters, we obtain our approximate selective log-likelihood:

$$\log \widehat{\mathcal{L}}_M(\theta^M_{jk},\ \beta^M_{jk}; U, Z, \mathcal{A}) = \log \phi\left(\begin{pmatrix}\widehat{\theta}^M_{jk}\\ \widehat{\beta}^M_{jk}\end{pmatrix}; \overline{R}\begin{pmatrix}\theta^M_{jk}\\ \beta^M_{jk}\end{pmatrix} + \overline{s}, \overline{\Theta}\right) - \log \widehat{c}(\theta^M_{jk}, \beta^M_{jk}). \tag{14}$$

3.5 Maximum Likelihood Inference Using the Selective Likelihood

We now present a result that gives an expression for the MLE $(\widehat{\theta}^{jk}_{mle}, \widehat{\beta}^{jk}_{mle})$ of $(\theta^M_{jk}, \beta^M_{jk})$, using the approximate selective log-likelihood given in (14). The expressions of the MLE and the corresponding observed Fisher information matrix are obtained as the solution to a low $|M|$-dimensional convex optimization problem. Using these expressions, we conveniently construct p-values based on a Wald-type test statistic (and Wald-type confidence intervals) for inference on the interaction effects.

Theorem 2. *Consider solving the $|M|$-dimensional optimization problem* $g^*(\widehat{\theta}^M_{jk}, \widehat{\beta}^M_{jk}) =$

$$\arg\min_g\ \frac{1}{2}\left(g - \overline{A}\begin{pmatrix}\widehat{\theta}^M_{jk}\\ \widehat{\beta}^M_{jk}\end{pmatrix} - \overline{b}\right)^\top \overline{\Omega}^{-1}\left(g - \overline{A}\begin{pmatrix}\widehat{\theta}^M_{jk}\\ \widehat{\beta}^M_{jk}\end{pmatrix} - \overline{b}\right) - \log\det D_{\Pi_S}(g, U, Z) + \mathrm{Barr}(g). \tag{15}$$

Then, the joint maximum likelihood estimator for $\begin{pmatrix}\theta^M_{jk}\\ \beta^M_{jk}\end{pmatrix}$ *is obtained as*

$$\begin{pmatrix}\widehat{\theta}^{jk}_{mle}\\ \widehat{\beta}^{jk}_{mle}\end{pmatrix} = \overline{R}^{-1}\begin{pmatrix}\widehat{\theta}^M_{jk}\\ \widehat{\beta}^M_{jk}\end{pmatrix} - \overline{R}^{-1}\overline{s} + \overline{\Sigma}^M_{jk}\overline{A}^\top\overline{\Omega}^{-1}\left(\overline{A}\begin{pmatrix}\widehat{\theta}^M_{jk}\\ \widehat{\beta}^M_{jk}\end{pmatrix} + \overline{b} - g^*(\widehat{\theta}^M_{jk}, \widehat{\beta}^M_{jk})\right).$$

Define the $(q_M + 1) \times (q_M + 1)$ *matrix*

$$\begin{aligned}K = & \overline{\Theta}^{-1} + \overline{A}^\top\overline{\Omega}^{-1}\overline{A} \\ & - \overline{A}^\top\overline{\Omega}^{-1}\left[\overline{\Omega}^{-1} - \nabla^2 \log\det D_{\Pi_S}(g^*(\widehat{\theta}^M_{jk}, \widehat{\beta}^M_{jk}), U, Z) + \nabla^2 \mathrm{Barr}(g^*(\widehat{\theta}^M_{jk}, \widehat{\beta}^M_{jk}))\right]^{-1}\overline{\Omega}^{-1}\overline{A}.\end{aligned}$$

The observed Fisher information matrix based on the approximate selective log-likelihood in (14) *is equal to* $I_{mle} = (\overline{\Sigma}^M_{jk})^{-1}K^{-1}(\overline{\Sigma}^M_{jk})^{-1}$.

Proof The proof of this result is similar to that of Theorem 4.1 in Huang et al. (2025) and is omitted for brevity. ■

Using the expression for the MLE and the observed Fisher information matrix from Theorem 2, we approximate the distribution of $\widehat{\theta}^{jk}_{mle}$ using $N\left(\theta^M_{jk}, I^{-1}_{mle}[1,1]\right)$, where $I^{-1}_{mle}[1,1]$ denotes the $(1,1)^{\text{th}}$ entry in I^{-1}_{mle}. This leads to an approximate Uniform$(0,1)$ pivot $\Phi(z_{jk})$, with $\Phi(\cdot)$ being the standard normal CDF, where

$$z_{jk} = (\widehat{\theta}^{jk}_{mle} - \theta^M_{jk})/\sqrt{I^{-1}_{mle}[1,1]},$$

and an approximate $(1-\alpha)$-level confidence interval for the parameter θ^M_{jk}, given by

$$\widehat{\theta}^{jk}_{mle} \pm z_{1-\alpha/2}\sqrt{I^{-1}_{mle}[1,1]}.$$

In Algorithm 2, we conclude by summarizing our inferential method for each parameter θ^M_{jk} of interest. In particular, we note that inference for an individual θ^M_{jk} can be computed efficiently by (i) solving a (q_M+1)-dimensional linear regression problem, (ii) computing (q_M+1)-dimensional matrices/vectors, and (iii) solving a $|M|$-dimensional convex optimization problem, corresponding to lines 6–8 of Algorithm 2, respectively. Moreover, the test for each interaction parameter θ^M_{jk} for all candidates in $\mathcal{T}^M$ can be parallelized.

In Appendix C, we provide an additional simulation study examining the computational complexity of the algorithm for varying values of $|M|$, where $q_M = 2|M|$, which empirically demonstrates that the computational time for $(\widehat{\theta}^{jk}_{mle}, \widehat{\beta}^{jk}_{mle})$ for a single interaction (j,k) scales as $O(|M|)$.

3.6 Simultaneous Screening of Multiple Interaction Terms

In this section, we extend the marginal screening of individual interaction terms $\mathcal{I}_{jk} \in \mathcal{T}^M$ to a more general framework that enables the simultaneous screening of a set of interaction terms.

Formally, let $\mathcal{C} \subset \mathcal{T}^M$ denote a collection of candidate interaction terms of interest. For example, under the weak hierarchy rule where

$$\mathcal{T}^M = \left\{(j,k) \in [p]^2 : j \in M \text{ or } k \in M\right\},$$

one could define $\mathcal{C} = \left\{(j_0,k) \in [p]^2 : k \in [p]\right\}$ for $j_0 \in M$, thereby addressing the question of whether any interaction term involving X_{j_0} should be considered.

For the general theory, we consider an arbitrary subset $\mathcal{C}$ of $\mathcal{T}^M$. Specifically, let

$$\mathcal{I}_{\mathcal{C}} = \left(\mathcal{I}_{jk} : (j,k) \in \mathcal{C}\right) \in \mathbb{R}^{n \times |\mathcal{C}|}$$

denote the matrix of interaction terms. We aim to obtain a p-value for the global null hypothesis

$$H_0^{\mathcal{C}} : \theta^M_{jk} = 0, \quad \text{for all } (j,k) \in \mathcal{C} \tag{16}$$

in the model

$$y \sim N\left(\Psi_M \beta^M_{\mathcal{C}} + \mathcal{I}_{\mathcal{C}} \theta^M_{\mathcal{C}},\ \sigma^2 I_n\right), \tag{17}$$

where $\beta^M_{\mathcal{C}} \in \mathbb{R}^{q_M}$ and $\theta^M_{\mathcal{C}} = \left(\theta^M_{jk} : (j,k) \in \mathcal{C}\right) \in \mathbb{R}^{|\mathcal{C}|}$ represent the main effect parameter vector and the interaction effects corresponding to $\{\mathcal{I}_{jk} : (j,k) \in \mathcal{C}\}$, respectively.

Analogously, we derive a p-value $\widehat{p}_{\mathcal{C}}^{M}$ for testing this global null $H_0^{\mathcal{C}}$, using a $\chi^2_{|\mathcal{C}|}$-distributed test statistic under the null. The theoretical development of the extension to simultaneous screening is provided in Appendix B.

Algorithm 2 p-values for interactions in $\mathcal{T}^M$ conditional on the SPAM fit

Require: $y \in \mathbb{R}^n$, $\Psi \in \mathbb{R}^{n\times q}$, $\lambda_j \geq 0,\ \forall j \in [p]$, $\epsilon > 0$, covariance matrix Ω

1: Sample $\omega \sim N(0_q, \Omega)$;

2: Solve

$$\widehat{\beta} = \underset{\left(\beta_{g_j}\in\mathbb{R}^{B_j}\right)_{j\in[p]}}{\arg\min} \frac{1}{2}\left\| y - \sum_{j=1}^{p} \Psi_j \beta_{g_j} \right\|_2^2 + \sum_{j=1}^{p} \left(\lambda_j \|\beta_{g_j}\|_2 + \frac{\epsilon}{2}\|\beta_{g_j}\|_2^2 \right) - \left(\beta_{g_1}^\top \ \ldots \ \beta_{g_p}^\top \right) \omega,$$

and record U, Z from the solver that meets the stationarity condition (6);

3: Obtain selection $M = \left\{ j \in [p] : \widehat{\beta}_{g_j} \neq 0_{B_j} \right\}$ and the set of candidate interactions $\mathcal{T}^M$;

4: Compute $\overline{\Omega}, \overline{b}$ which are shared for all $(j,k) \in \mathcal{T}^M$, according to (12);

5: **for** $(j,k) \in \mathcal{T}^M$ **do**

6: Solve $\begin{pmatrix} \widehat{\theta}_{jk}^M \\ \widehat{\beta}_{jk}^M \end{pmatrix} = \left(\overline{Z}_{jk}^M\right)^+ y$ according to (8);

7: Compute $\overline{A}, \overline{\Theta}, \overline{R}, \overline{s}$ according to (12);

8: Solve

$$g^*(\widehat{\theta}_{jk}^M, \widehat{\beta}_{jk}^M) \leftarrow \underset{g}{\arg\min}\ \frac{1}{2}\left(g - \overline{A}\begin{pmatrix} \widehat{\theta}_{jk}^M \\ \widehat{\beta}_{jk}^M \end{pmatrix} - \overline{b} \right)^\top \overline{\Omega}^{-1} \left(g - \overline{A}\begin{pmatrix} \widehat{\theta}_{jk}^M \\ \widehat{\beta}_{jk}^M \end{pmatrix} - \overline{b} \right) - \log\det D_{\Pi_S}(g, U, Z) + \mathrm{Barr}(g);$$

9: Compute MLE

$$\begin{pmatrix} \widehat{\theta}_{mle}^{jk} \\ \widehat{\beta}_{mle}^{jk} \end{pmatrix} = \overline{R}^{-1}\begin{pmatrix} \widehat{\theta}_{jk}^M \\ \widehat{\beta}_{jk}^M \end{pmatrix} - \overline{R}^{-1}\overline{s} + \overline{\Sigma}_{jk}^M \overline{A}^\top \overline{\Omega}^{-1}\left(\overline{A}\begin{pmatrix} \widehat{\theta}_{jk}^M \\ \widehat{\beta}_{jk}^M \end{pmatrix} + \overline{b} - g^*(\widehat{\theta}_{jk}^M, \widehat{\beta}_{jk}^M) \right);$$

10: Compute Fisher information $I_{mle} = (\overline{\Sigma}_{jk}^M)^{-1} K^{-1} (\overline{\Sigma}_{jk}^M)^{-1}$ as in Theorem 2;

11: Compute p-value under the null $H_0^{jk} : \theta_{jk}^M = 0$:

$$\widehat{p}_{jk}^M \leftarrow 2 * \min\left\{ \Phi\left(\widehat{\theta}_{mle}^{jk} / \sqrt{I_{mle}^{-1}[1,1]} \right),\ 1 - \Phi\left(\widehat{\theta}_{mle}^{jk} / \sqrt{I_{mle}^{-1}[1,1]} \right) \right\};$$

12: **end for**

13: **return** $\left\{ \widehat{p}_{jk}^M \right\}_{(j,k)\in\mathcal{T}^M}$.

4. Simulation

First, we describe the simulation settings explored in this section, along with the metrics used to evaluate our method. We then present the key findings from these simulations. The implementation for this section as well as Section 5 can be found at `https://github.com/yiling-h/SI-Interaction`.

4.1 Simulation Settings

We simulate $n = 200$ independent observations of $p = 20$ features. For each observation i, the features are sampled as follows: we first draw $(Z_{i,1}, ..., Z_{i,20}) \sim N_{20}\left(0_{20}, \Sigma(\rho_1, \rho_2, \rho_{cross})\right)$, where

$$\Sigma(\rho_1, \rho_2, \rho_{cross}) = \begin{pmatrix} (1-\rho_1)I_3 + \rho_1 1_3 1_3^\top & \rho_{cross} 1_3 1_{17}^\top \\ \rho_{cross} 1_{17} 1_3^\top & (1-\rho_2)I_{17} + \rho_2 1_{17} 1_{17}^\top \end{pmatrix}.$$

We then transform the samples as $X_{i,j} = 2.5 \cdot \Phi(Z_{i,j})$ for $j = 1, 2, 3$, and $X_{i,j} = \Phi(Z_{i,j})$ for $j = 4, ..., 20$, thereby generating features with varying effect strengths, where Φ is the standard Gaussian CDF. As a result, marginally, $X_{i,1}, X_{i,2}, X_{i,3} \sim \text{Uniform}(0, 2.5)$, and $X_{i,4}, ..., X_{i,20} \sim \text{Uniform}(0, 1)$, while their joint distribution depends on the matrix $\Sigma(\rho_1, \rho_2, \rho_{cross})$ through the parameters ρ_1, ρ_2 and ρ_{cross}. In particular, ρ_1 controls the dependence among signal features, ρ_2 controls the dependence among noise features, and ρ_{cross} controls the dependence between these two groups of features.

Next, we randomly sample a subset $\mathcal{G}$ of interaction pairs from the set $\{(j, k) : j \neq k;\ j, k \in [20]\}$, with $|\mathcal{G}| = s_{inter}$. We then construct the mean response vector $\mu \in \mathbb{R}^n$, where the i-th entry is given by:

$$\mu_i = \gamma_{main}\left[2\sin(2X_{i,1}) + X_{i,2}^2 + \exp(-X_{i,3})\right] + \gamma_{inter} \sum_{(j,k)\in\mathcal{G}} X_{i,j}X_{i,k}, \tag{18}$$

and $\gamma_{main}, \gamma_{inter} \in \mathbb{R}$ are constants that determine the relative strength of the main and interaction signals. Finally, the response vector $Y \in \mathbb{R}^n$ is generated as: $Y = \mu + \epsilon$, where $\epsilon \sim N_n(0_n, \sigma^2 I_n)$.

We construct four main simulation settings, which we refer to as Settings 1–4, each simulated in 1000 replications.

1. Setting 1: Varying noise level. We fix $s_{inter} = 5$, $\gamma_{main} = 2$, $\gamma_{inter} = 2$, $\rho_1 = \rho_2 = 0.6$, and $\rho_{cross} = 0.48$, and vary the standard deviation of the noise ϵ by setting $\sigma \in \{0.5, 1, 2, 4\}$.

2. Setting 2: Varying cross-correlation. We fix $s_{inter} = 5$, $\gamma_{main} = 2$, $\gamma_{inter} = 2$, $\rho_1 = \rho_2 = 0.6$, and $\sigma = 2$, and vary the cross-correlation parameter in $\Sigma(\rho_1, \rho_2, \rho_{cross})$ by setting $\rho_{cross} \in \{0, 0.2, 0.4, 0.6\}$.

3. Setting 3: Varying strength of interactions. We fix $s_{inter} = 5$, $\gamma_{main} = 2$, $\rho_1 = \rho_2 = 0.6$, $\rho_{cross} = 0.48$, and $\sigma = 2$, and vary $\gamma_{inter} \in \{0.5, 1, 2, 5\}$.

4. Setting 4: Varying the number of interactions. We fix $\gamma_{main} = 2$, $\gamma_{inter} = 2$, $\rho_1 = \rho_2 = 0.6$, $\rho_{cross} = 0.48$, and $\sigma = 2$, and vary $s_{inter} \in \{5, 10, 15, 20\}$.

Furthermore, the motivating example in Section 2.1 is simulated using $s_{inter} = 5$, $\gamma_{main} = 2$, $\gamma_{inter} = 2$, $\rho_1 = \rho_2 = 0.6$, $\rho_{cross} = 0.48$, and $\sigma = 2$.

4.2 Methods and Metrics

We compare our proposed method with two alternative approaches: the naive approach and the data-splitting approach. All three methods for obtaining p-values for the interaction effects are described below. To model the non-linear main effects in all three methods, we construct a B-spline basis expansion Ψ_j for each feature X_j, using degree 2 and 6 knots. This produces a design matrix Ψ with dimensions 200×40. To allow for greater flexibility, one can increase the degree of the B-spline basis expansion. As demonstrated in our simulation results, we observe substantial advantages over the naive and the data-splitting approaches, even when using degree 2.

Naive approach. The first method we consider is a "naive" approach, where the entire observed dataset is used to obtain $\widehat{M} = M$ by solving (1) with $\omega = 0_q$, i.e., without randomization. Inference for θ^M_{jk} under the model (3) is conducted via the standard z-test. Specifically, we ignore the randomness in $\widehat{M}$ and the fact that it is estimated from the same data, proceeding to make inferences for θ^M_{jk} for each $(j,k) \in \mathcal{T}^M$ as if the realization M were a fixed, predetermined set. The corresponding Gaussian pivot for each $(j,k) \in \mathcal{T}^M$ is given by $z^{naive}_{jk} = (\widehat{\theta}^M_{jk} - \theta^M_{jk})/(\sigma_{jk})$, where $\sigma_{jk} = \sigma\sqrt{\left(\left(\overline{Z}^M_{jk}\right)^\top \overline{Z}^M_{jk}\right)^{-1}_{1,1}}$. Note that if M were a fixed set, then $z^{naive}_{jk} \sim N(0,1)$, and $\Phi(z^{naive}_{jk}) \sim \text{Uniform}(0,1)$, where $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of a standard normal variable.

Data splitting approach. The second method we consider is the data splitting method, where we randomly partition the data into two subsamples: $[n] = \mathcal{S}_{sel} \cup \mathcal{S}_{inf}$, where $\mathcal{S}_{sel} \cap \mathcal{S}_{inf} = \emptyset$, with $\mathcal{S}_{sel}$ reserved for the selection of $\widehat{M}$ and $|\mathcal{S}_{sel}| = n_1 = \lfloor r * n \rfloor$, and $\mathcal{S}_{inf}$, with $|\mathcal{S}_{inf}| = n_2 = n - n_1$, reserved for inference of θ^M_{jk} for each $(j,k) \in \mathcal{T}^M$, after observing $\widehat{M} = M$. Specifically, denote y^{sel} and $\mathcal{I}^{sel}_{jk}$ as the entries of y and $\mathcal{I}_{jk}$ with indices in $\mathcal{S}_{sel}$, and Ψ^{sel} as the submatrix of Ψ with rows indexed by $\mathcal{S}_{sel}$. We define y^{inf}, Ψ^{inf}, and $\mathcal{I}^{inf}_{jk}$ analogously. We then estimate $\widehat{M}$ by solving (1) with $\omega = 0_q$, replacing y and Ψ_j with y^{sel} and Ψ^{sel}_j, respectively. Inference for θ^M_{jk} is performed using:

$$y^{inf} \sim N(\Psi^{inf}_M \beta^M_{jk} + \mathcal{I}^{inf}_{jk} \theta^M_{jk}, \sigma^2 I_{n_2}).$$

The same procedure as in the naive approach is used to compute the pivots for the interaction effects, except that they are computed solely on y^{inf}, the data reserved for inference.

The proposed approach. We compute M by solving (1), and then calculate $\widehat{\theta}^M_{jk}$ according to (8). The corresponding approximate selective Gaussian pivot for each $(j,k) \in \mathcal{T}^M$ is given by $z^{mle}_{jk} = (\widehat{\theta}^{jk}_{mle} - \theta^M_{jk})/(\sigma_{jk})$, where $\sigma_{jk} = \sqrt{I^{-1}_{mle}[1,1]}$. Here, $\widehat{\theta}^{jk}_{mle}$ is the selective MLE for θ^M_{jk} from Section 3.5, and $I^{-1}_{mle}[1,1]$ is its estimated variance.

In our method, we set the randomization covariance $\Omega = \sigma^2 \frac{1-r}{r} \Psi^\top \Psi$, for some $r \in (0,1)$. As established in previous work Panigrahi et al. (2021); Huang et al. (2025), this choice of Ω leads to an asymptotically equivalent solution to data splitting, where $(100 \times r)\%$ of the data is used for selection, provided the main effects regression model is correctly specified. This equivalence provides a natural choice for setting Ω, even when the main effects model is misspecified, i.e., when additional omitted interaction terms are present in the true data generation process.

To achieve main effects selection similar to naive inference, the proposed method is implemented using $\Omega = \sigma^2 \frac{1-r}{r} \Psi^\top \Psi$ with $r = 0.9$, and data splitting is also performed using 90% of the data for selection. For all methods, we fix the tuning parameter $\lambda_j = 0.5\sigma\sqrt{n}\sqrt{B_j}\sqrt{2\log(q)}$ in (1), where $B_j = 2$ is the group size for the j-th variable. Furthermore, whenever the dispersion parameter σ is not known, we replace it with the plug-in estimate $\widehat{\sigma} = \sqrt{||y - \Psi(\Psi^\top\Psi)^{-1}\Psi^\top y||_2^2/(n-q)}$. Finally, for all methods, we construct $\mathcal{T}^M$ using the weak hierarchy rule, where $\mathcal{T}^M = \big\{(j,k) \in [p]^2 : j \in M \text{ or } k \in M\big\}$.

To assess the performance of our method and compare it with the other two approaches, we compute the following metrics:

(i) Empirical cumulative distribution function (ECDF) of pivots. As noted before, if the inferential method is valid, with corresponding Type-I error rate control, then the ECDF of the pivots should approximately follow a Uniform$(0,1)$ distribution. Therefore, for each inferential method, we compute pivots for the interaction effects, for each $(j,k) \in \mathcal{T}^M$, and plot the ECDF of these pivots, aggregated across all the interaction effects in $\mathcal{T}^M$ and all simulation runs.

(ii) Lengths of confidence intervals. A common way to assess inferential power is by examining the lengths of the confidence intervals. For confidence intervals $\text{CI}(\theta^M_{jk}) = (L^M_{jk},\ U^M_{jk})$ corresponding to each θ^M_{jk}, for $(j,k) \in \mathcal{T}^M$, we compute the average confidence interval length within each simulation, defined as:

$$\text{Average Length} = \frac{\sum_{(j,k)\in\mathcal{T}^M}\left(U^M_{jk} - L^M_{jk}\right)}{|\mathcal{T}^M|},$$

where the average is taken over all $(j,k) \in \mathcal{T}^M$.

(iii) F1 score evaluation of interaction effect recovery. Suppose we obtain a collection of p-values from each method, denoted by $\mathcal{P}^M = \big\{p_{jk} : (j,k) \in \mathcal{T}^M\big\}$. We define the test F1 score as

$$\text{F1 Score}(\mathcal{P}^M) = 2 \times \frac{\text{Precision}(\mathcal{P}^M) \times \text{Recall}(\mathcal{P}^M)}{\text{Precision}(\mathcal{P}^M) + \text{Recall}(\mathcal{P}^M)},$$

where the set of true discoveries consists of interaction effects whose magnitudes are greater than or equal to a predefined threshold $t_0 > 0$. More specifically, to compute the F1 score, we calculate:

$$\text{Precision}(\mathcal{P}^M) = \frac{\left|\left\{(j,k) \in \mathcal{T}^M : |\theta^M_{jk}| \geq t_0\right\} \cap \left\{(j,k) \in \mathcal{T}^M : p_{jk} < \alpha\right\}\right|}{|\{(j,k) \in \mathcal{T}^M : p_{jk} < \alpha\}|};$$

$$\text{Recall}(\mathcal{P}^M) = \frac{\left|\left\{(j,k) \in \mathcal{T}^M : |\theta^M_{jk}| \geq t_0\right\} \cap \left\{(j,k) \in \mathcal{T}^M : p_{jk} < \alpha\right\}\right|}{\left|\left\{(j,k) \in \mathcal{T}^M : |\theta^M_{jk}| \geq t_0\right\}\right|}.$$

In our simulations, we set $t_0 = 0.1$.

4.3 Results

From Figure 3, we observe the following trends: (1) The proposed method produces valid pivots across all settings. Note that their empirical CDF closely aligns with the Uniform$(0,1)$ distribution. (2) The data splitting approach also yields valid pivots under most scenarios,

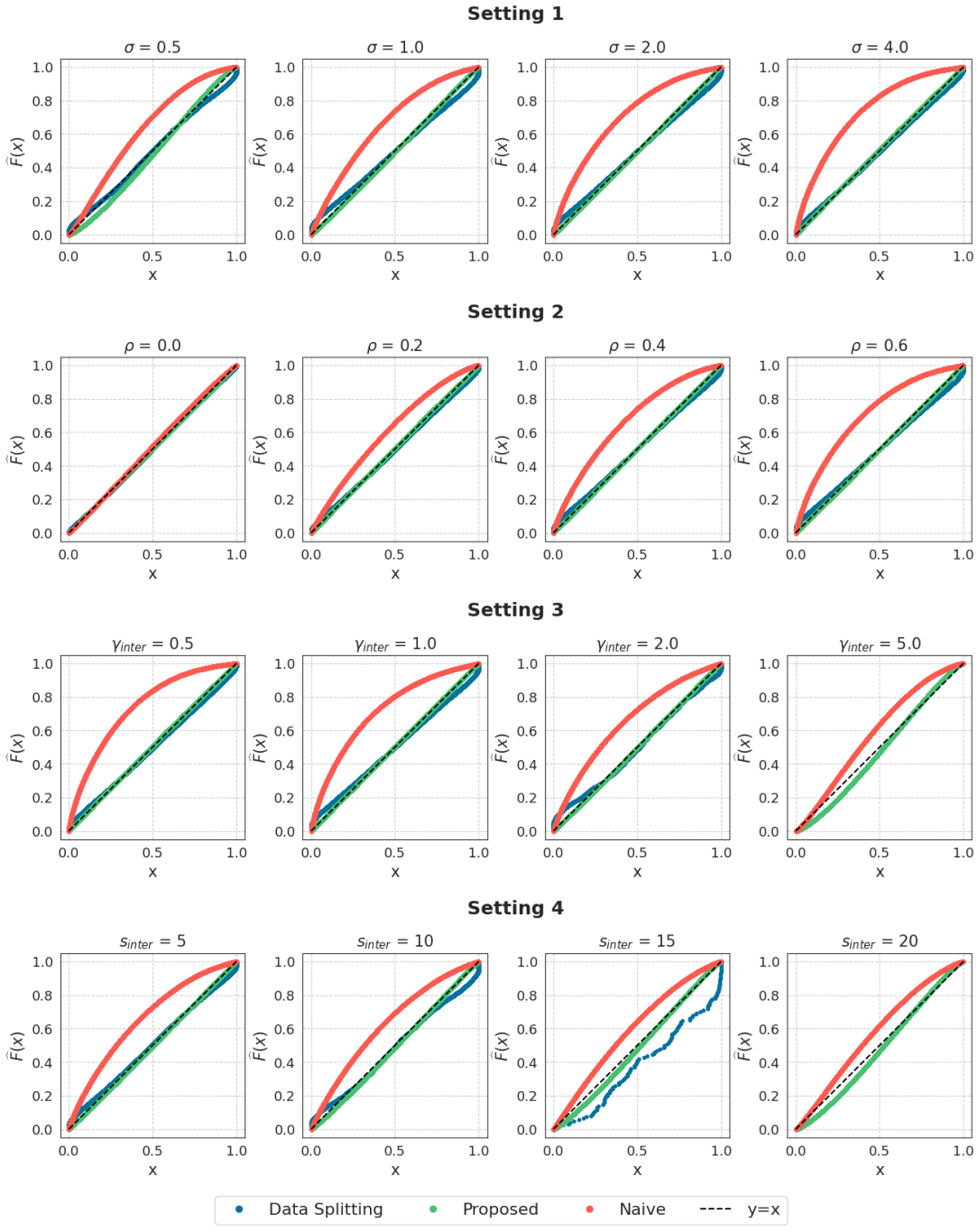


Figure 3: ECDF of the uniform pivots $\Phi(z_{jk})$ obtained by naive inference, data splitting, and the proposed method in all simulation settings

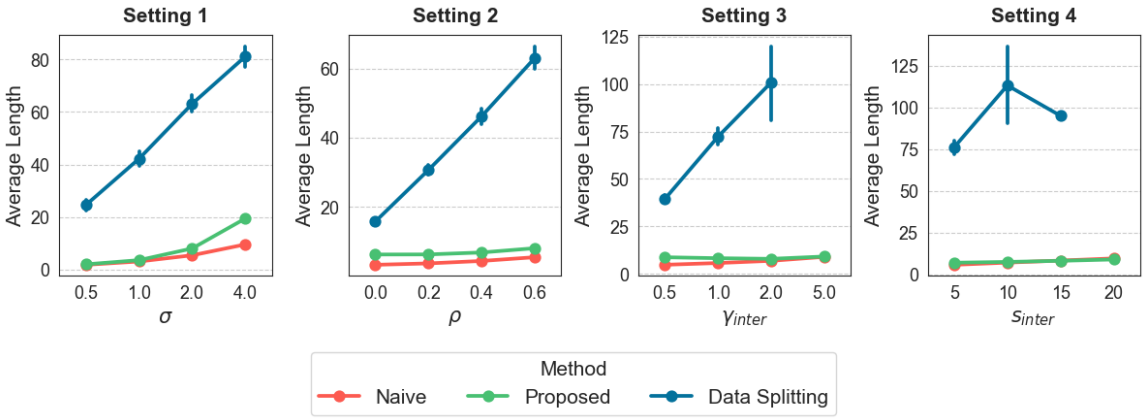


Figure 4: Average confidence interval lengths for θ^M_{jk} obtained by naive inference, data splitting, and the proposed method in all simulation settings

similar to our proposed method. However, this approach fails to deliver valid inference in certain settings, such as Setting 3 with larger interaction effects or Setting 4 with a larger number of interaction effects. This breakdown in inferential validity is due to either rank deficiency in the design matrix or an insufficient sample size in the holdout data reserved for inference, highlighting situations where data splitting may not be a feasible strategy for selective inference. (3) The naive approach consistently fails to produce valid inference across most settings. As expected, the empirical CDF of its pivots increasingly deviates from the uniform distribution as the noise level rises or as the collinearity between features increases, when selection bias due to main effect selection is much larger.

From Figure 4, we observe that the confidence intervals produced by our method are consistently shorter than those from the data splitting approach, which yields much wider intervals due to relying solely on holdout data for inference. In scenarios where the design matrix is rank-deficient or the holdout sample size is insufficient for reliable inference, the data-splitting method either yields extremely wide confidence intervals or fails to produce meaningful results. This issue can be noted for Settings 3 and 4, where the number of interaction effects is large or the magnitude of these effects is substantial. Furthermore, our intervals are only slightly longer than those from the naive method, which does not account for the selection of the main effects. This highlights that our method provides valid interval estimates for the interaction effects without incurring a substantial increase in interval length compared to the invalid naive approach.

Finally, from Figure 5, we observe the following trends: (1) The proposed method consistently achieves better or comparable F1 scores relative to both naive inference and data splitting, across the different settings, highlighting its strength in balancing false discovery rate (FDR) and power. (2) The data-splitting method consistently struggles with insufficiency of data when testing null hypotheses after the selection of main effects. As noted in earlier plots, this results in situations where inference cannot even be computed, in which case a F1 score of 0 is assigned. The interactions identified by the data-splitting method result in models with significantly lower F1 scores in these settings. (3) The only scenario

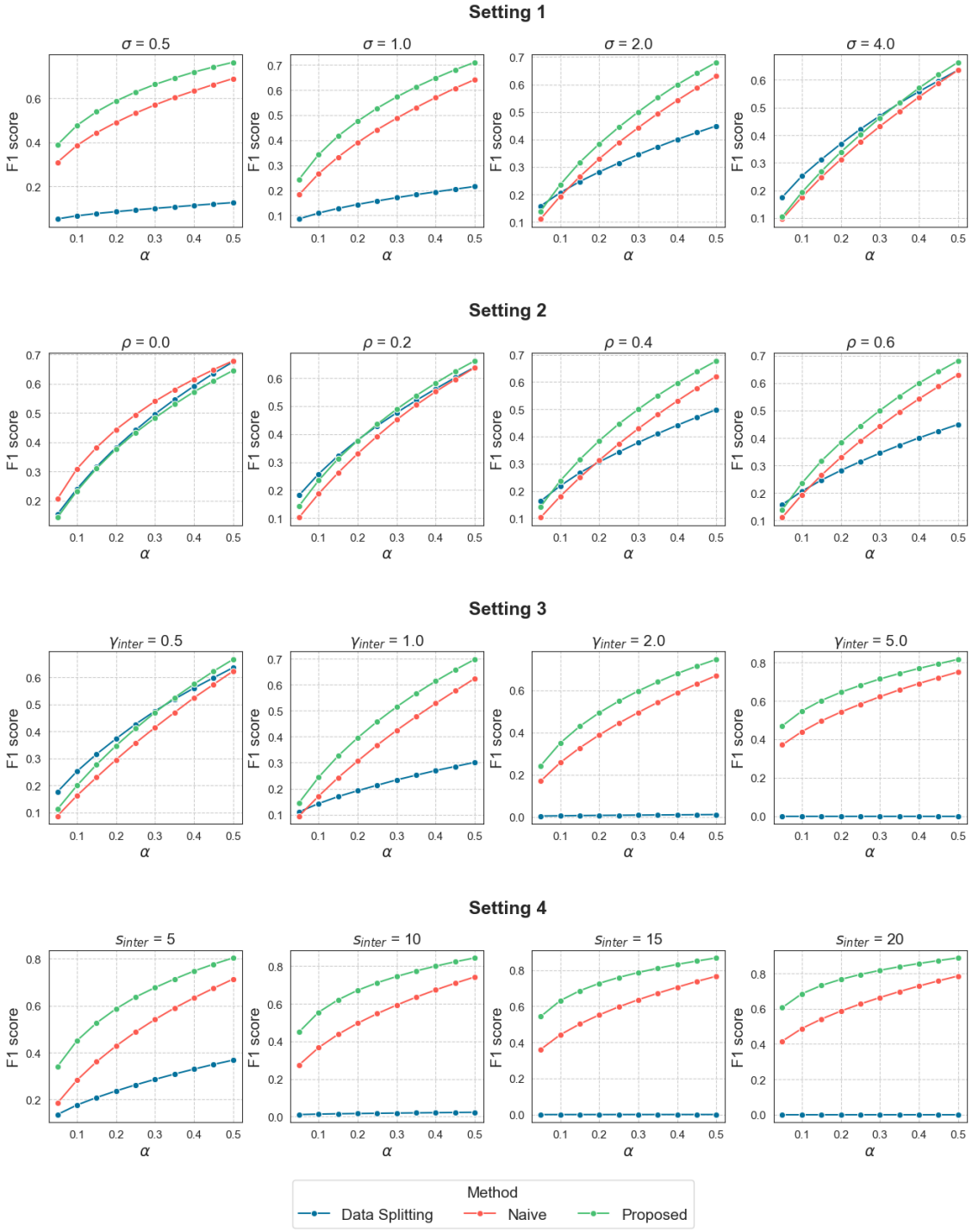


Figure 5: F1 scores for hypothesis testing of θ_{jk}^M obtained by naive inference, data splitting, and the proposed method in all simulation settings with varying α

where the proposed method yields a lower F1 score than naive inference is the special case of Setting 2 with $\rho_{cross} = 0$, where signal features and noise features are independent. In this case, naive inference produces valid pivots without requiring any selection adjustment. In all other scenarios, however, the proposed method outperforms naive inference, due to its ability to provide valid inference while reusing the full data for inference.

In the current simulations, we consider testing for interactions under a weak hierarchy principle. In Appendix C, we conduct an additional simulation to empirically evaluate the F1 scores and the false discovery rate (FDR) when testing all pairs of interactions using our p-values, both before and after applying the Benjamini–Hochberg (BH-q, Benjamini and Hochberg, 1995) correction to take into account the effect of multiple testing. As shown in the simulations, the BH-q correction consistently ensures valid FDR control, albeit at the expected cost of reduced power compared with the uncorrected p-values, as reflected in the F1 score plots.

5. Application: JFK Flight Data

In this section, we apply our method to a dataset consisting of flight information at the JFK airport in New York City during the time period from January 1, 2024 to January 30, 2024. The data can be accessed via the `R` package `anyflights` (Couch, 2025). The dataset contains information on individual flights departing from the JFK airport and includes three main components: (1) flight information, such as scheduled departure and arrival times, destination, distance to destination, departure and arrival delays, and carrier details; (2) aircraft information, including the number of seats and the year of manufacture; and (3) weather information, such as temperature, precipitation, wind speed, and visibility, leading to a dataset of size $n \times p = 7358 \times 16$.

We fit our model to this data to predict arrival delay, measured in minutes, using flight, aircraft, and weather information that is available at the time of scheduled departure. Future information, such as actual departure and arrival time, which is not available at the time of prediction and thus not useful for this task, is excluded from our model fit.

5.1 Reluctant Modeling of Interactions and Evaluation

To illustrate the accuracy of the reluctant model fitting approach with interactions, we first apply our method to a subsample containing about 10% of the samples in the full data ($n = 7358$), resulting in a subsample of size $n_{\text{sub}} = 735$, and use the remaining data as a test set to evaluate this model. For constructing high-quality nonlinear basis expansions, we model features with more than 40 unique values nonlinearly, indexed by $\mathcal{N} \subset [p]$, and treat the remaining features linearly, indexed by $\mathcal{L} \subset [p]$. A summary of the linear and nonlinear features is provided in Table 4 in Appendix D.1.

For each feature X_j modeled nonlinearly (i.e., $j \in \mathcal{N}$), we generate a B-spline basis expansion Ψ_j of degree 2 with 6 knots. This results in a final design matrix Ψ of dimension 735×21, including the intercept. We implement the proposed method using $\Omega = \widehat{\sigma}^2 \times \frac{1-r}{r} \Psi^\top \Psi$, with $r = 0.9$, as done in Section 4. Here, we use the plug-in estimate $\widehat{\sigma} = \sqrt{||y - \Psi(\Psi^\top \Psi)^{-1} \Psi^\top y||_2^2 / (n_{\text{sub}} - q)}$ in place of the unknown dispersion parameter σ. For the selection of main effects, both the proposed method and the naive approach for interac-

tion inference are implemented using the tuning parameter $\lambda_j = 0.5\widehat{\sigma}\sqrt{n_{\text{sub}}}\sqrt{B_j}\sqrt{2\log(q)}$, where the group size $B_j = 2$ for $j \in \mathcal{N}$ and $B_j = 1$ for $j \in \mathcal{L}$.

We now proceed to validate the discovery of significant θ^M_{jk}'s identified in the 10% subsample using the remaining 90% of the data, which was held out during the modeling of main effects and interactions. For each $(j,k) \in \mathcal{T}^M$, we fit an OLS model to the holdout 90% of the dataset using the specification in (3), and conduct a t-test for θ^M_{jk} on this holdout data. Given the large sample size of the test data, we treat the significant θ^M_{jk}'s identified in the holdout set as a proxy for the true set of discoveries. We then assess the quality of discoveries made on the 10% subsample using both the proposed method and naive inference by comparing them to this ground truth in terms of precision, recall, and F1 score. Put differently, we assess the degree of concordance between the findings from the holdout test data and those obtained from the 10% subsample.

Figure 6 summarizes the selected nonlinear main effects identified by the proposed method and by naive inference. Notably, the added randomization term in the proposed method does not lead to a substantially different model fit. The non-randomized naive inference selects only one additional nonlinear main effect "scheduled arrival hour" which exhibits a relatively flat fitted additive effect on arrival delay, at zero.

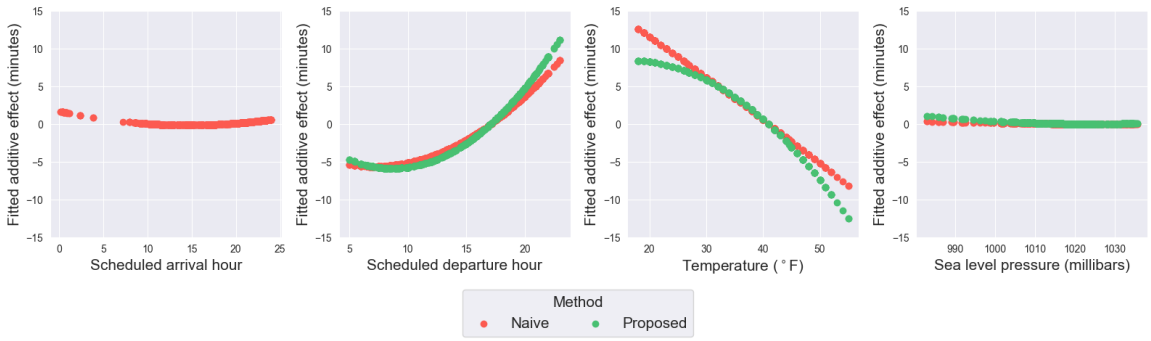


Figure 6: Fitted main effects using the proposed method and naive inference

After fitting the main effects models, we construct $\mathcal{T}^M$ using the weak hierarchy rule, where $\mathcal{T}^M = \{(j,k) \in [p]^2 : j \in M \vee k \in M\}$. Using a significance level of $\alpha = 0.1$, naive inference identifies 10 significant interaction terms, while the proposed method yields 13. Precision, recall, and F1 score based on the holdout validation set are reported in Table 2.

	Naive	Proposed
Precision	0.4	**0.615**
Recall	0.125	**0.308**
F1 score	0.190	**0.410**

Table 2: Precision, recall, and F1 score of naive inference and the proposed method evaluated on the holdout sample

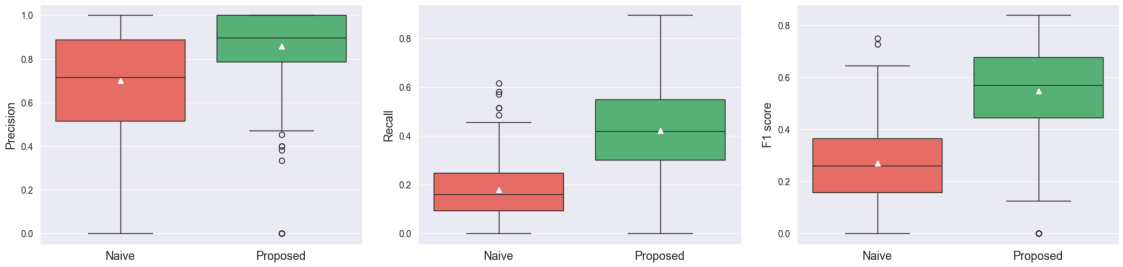


Figure 7: Precision, recall, and F1 scores of naive inference and the proposed method evaluated on the holdout sample, aggregated over 500 subsamples; empirical means are indicated by the white triangle

As shown in Table 2, the proposed method outperforms naive inference across all three metrics. In the following section, we evaluate the replicability of these results by performing repeated subsampling, repeating the above analysis 500 times.

5.2 Replication Analysis via Repeated Subsampling

We repeatedly draw 500 subsamples of size $n_{\text{sub}} = 735$ and in each repetition, we perform the analysis in Section 5.1.

As shown in Figure 7, the proposed method consistently achieves higher precision, recall, and F1 score in recovering interaction effects, compared to the naive inference method. These findings are consistent with the results observed in the simulation studies.

In Table 3, we report the main effects that are most frequently selected. Finally, in Figure 8, we depict the frequency of significance of interaction signals that were most frequently deemed significant across the 500 repetitions of the analysis. This is based on the union of the top 30 most frequently significant interaction signals from both naive inference and our approach.

Note that the proposed method identifies the top interaction signals far more frequently than the naive approach. In particular, the interactions on the left half of Figure 8 are identified by the proposed method more than twice as often, compared to naive inference. This further highlights the proposed method's ability to consistently recover key interaction signals, while the naive approach is likely susceptible to selection biases introduced during main effects selection, leading to an inferior model. A documentation of the features in Figure 8 is provided in Table 4 in Appendix D.1.

6. Conclusion

In this paper, we develop a selective inference method to assess whether there is statistical evidence for moving beyond an additive, non-linear model. This approach follows the principle of reluctance, favoring simpler and more interpretable models without interactions unless it is truly warranted. We cast this as a data-adaptive hypothesis testing problem, where the null hypothesis assumes a SPAM that is fitted to the data prior to testing. Our solution makes use of the recently developed selective inference techniques for selected groups

Feature	Naive	Proposed
Sea level pressure (millibars)	420	399
Temperature (°F)	400	324
Visibility (miles)	370	357
Scheduled departure hour	359	353
Wind speed (mph)	348	381

Table 3: Features that were selected more than 300 times by both naive inference and the proposed method, together with the number of times they are selected out of 500 replications

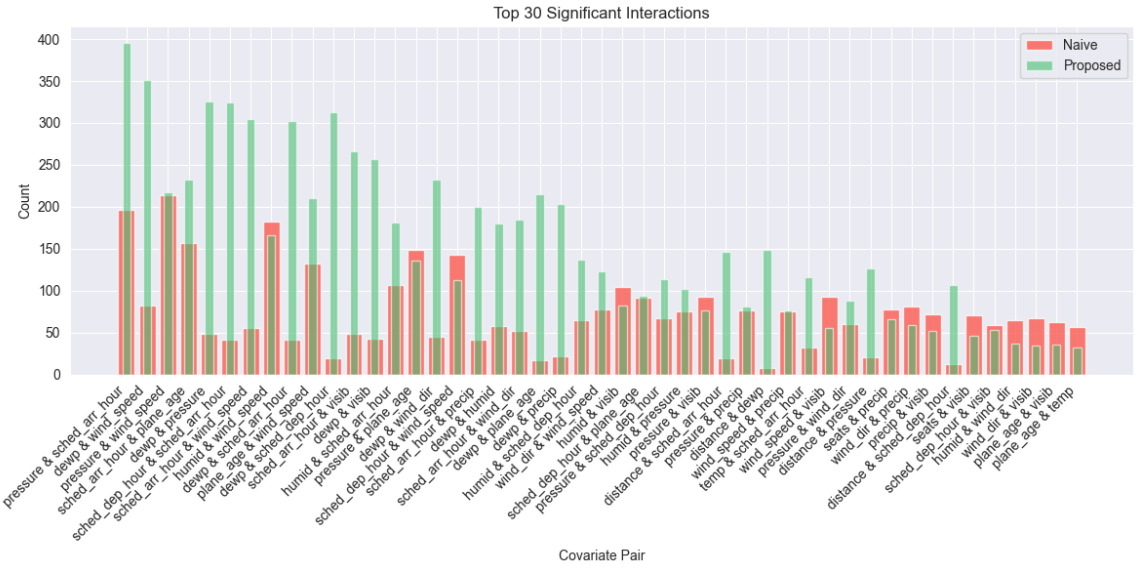


Figure 8: Frequency of significance of interaction signals that were most frequently deemed significant by naive inference and the proposed method, aggregated over 500 subsamples; interactions are ordered along the x-axis according to the aggregated counts by the two methods.

after solving a group lasso problem, yielding valid p-values for the hypotheses on interaction effects. Through extensive experiments on simulated and real data, we demonstrate that the resulting non-linear model—constructed via our rigorous, reluctant approach—achieves superior modeling accuracy compared to both naive strategies and existing selective inference methods such as data splitting.

While our primary focus in the paper is on testing for interactions, a natural by-product of our method is valid inference for the main effect variables included in the fitted SPAM, in the form of p-values and confidence intervals. Another immediate extension of our method is to test groups of interaction terms jointly, rather than one at a time, using a χ^2-test

adjusted for the SPAM fit to compute the corresponding p-value. Finally, as a direction for future work, we note that a maximum likelihood approach could be developed for data types beyond normal data by constructing an asymptotic selective likelihood, as developed in Huang et al. (2025); Guglielmini et al. (2025).

Acknowledgments

S. Panigrahi acknowledges the support of NSF CAREER Award DMS-2337882.

Appendix A. Proof of Theoretical Results

In this section, we provide proofs of the theoretical results presented in the main paper.

A.1 Marginal Distribution of the Key Statistics

Proof [Proof of Lemma 2] We first consider the sub-vector $\left(\widehat{\theta}^M_{jk}, (\widehat{\beta}^M_{jk})^\top\right)^\top$ in the first two components of $\widehat{S}^M_{jk}$. From (8), standard OLS theory establishes that

$$\begin{pmatrix}\widehat{\theta}^M_{jk}\\ \widehat{\beta}^M_{jk}\end{pmatrix} \sim N_{q_M+1}\left(\begin{pmatrix}\theta^M_{jk}\\ \beta^M_{jk}\end{pmatrix}, \overline{\Sigma}^M_{jk}\right),$$

under the model $y \sim N_n(\Psi_M \beta^M_{jk} + \mathcal{I}_{jk}\theta^M_{jk}, \sigma^2 I_n)$. Then,

$$\begin{aligned}\widehat{\mathcal{A}}^M_{jk} =& -\Psi^\top_{-M}\left(y - \overline{Z}^M_{jk}\begin{pmatrix}\widehat{\theta}^M_{jk}\\ \widehat{\beta}^M_{jk}\end{pmatrix}\right)\\ =& -\Psi^\top_{-M}\left(I_n - \overline{Z}^M_{jk}\left(\overline{Z}^M_{jk}\right)^+\right)y\\ =& -\Psi^\top_{-M}\left(I_n - \overline{Z}^M_{jk}\left(\overline{Z}^M_{jk}\right)^+\right)\left(\overline{Z}^M_{jk}\begin{pmatrix}\theta^M_{jk}\\ \beta^M_{jk}\end{pmatrix} + e\right)\\ =& -\Psi^\top_{-M}\left(I_n - \overline{Z}^M_{jk}\left(\overline{Z}^M_{jk}\right)^+\right)e,\end{aligned}$$

where $e \sim N_n(0_n, \sigma^2 I_n)$, which implies

$$\widehat{\mathcal{A}}^M_{jk} \sim N_{q-q_M}(0_{q-q_M}, \widetilde{\Sigma}^M_{jk}).$$

Then the result follows by establishing that

$$\mathrm{Cov}\left(\begin{pmatrix}\widehat{\theta}^M_{jk}\\ \widehat{\beta}^M_{jk}\end{pmatrix}, \widehat{\mathcal{A}}^M_{jk}\right) = 0_{(q_M+1)\times(q-q_M)},$$

while this follows from the orthogonality of the OLS coefficients and the residuals. ■

A.2 Developing the Conditional Distribution of the Key Statistics

Proof [Proof of Lemma 3] Recall (6), which expresses the KKT conditions as

$$\omega = -\Psi^\top y + \left(\Psi^\top\Psi + \epsilon I_q\right)\begin{pmatrix}(\widehat{\gamma}_j\widehat{u}_j)_{j\in M}\\ (0_{B_j})_{j\notin M}\end{pmatrix} + \begin{pmatrix}(\lambda_j\widehat{u}_j)_{j\in M}\\ (\lambda_j\widehat{z}_j)_{j\notin M}\end{pmatrix}. \tag{19}$$

We will show that we can write the right-hand side in terms of our key statistics $\widehat{S}^M_{jk}$ instead of y. First, by (9),

$$-\Psi^\top y = A^M_{jk}\begin{pmatrix}\widehat{\theta}^M_{jk}\\ \widehat{\beta}^M_{jk}\end{pmatrix} + \begin{pmatrix}0_{q_M}\\ \widehat{\mathcal{A}}^M_{jk}\end{pmatrix}.$$

Second, we observe that

$$\left(\Psi^\top\Psi + \epsilon I_q\right)\begin{pmatrix}(\widehat{\gamma}_j\widehat{u}_j)_{j\in M}\\ (0_{B_j})_{j\notin M}\end{pmatrix} = B(\widehat{U})\widehat{\gamma}.$$

Thus, defining $c(U, Z)$ as in the statement of the lemma, we get

$$\omega = \Pi_{\widehat{S}^M_{jk}}(\widehat{\gamma}, \widehat{U}, \widehat{Z}).$$

Inverting the mapping $\Pi_{\widehat{S}^M_{jk}}$ proves our claim. ■

A.2.1 Proof of Proposition 1

Proof [Proof of Proposition 1] We derive the conditional distribution

$$\left(\widehat{S}^M_{jk},\ \widehat{\gamma},\ \widehat{U},\ \widehat{Z}\right) \ \Big|\ \left\{\widehat{\gamma} \succ 0_{|M|},\ \widehat{U} \in \mathcal{U}_M,\ \widehat{Z} \in \mathcal{Z}_M\right\},$$

by truncating the marginal density of $\left(\widehat{S}^M_{jk},\ \widehat{\gamma},\ \widehat{U},\ \widehat{Z}\right)$ to the set $\mathbb{R}^{q+1} \times \mathbb{R}_+^{|M|} \times \mathcal{U}_M \times \mathcal{Z}_M$. Note that the density function of this distribution, at (S, γ, U, Z), is proportional to

$$f(S) \cdot h_S(\gamma, U, Z) \cdot 1_{\mathbb{R}_+^{|M|}}(\gamma) \cdot 1_{\mathcal{U}_M}(U) \cdot 1_{\mathcal{Z}_M}(Z), \tag{20}$$

where $f(\cdot)$ is the marginal density of the key statistics $\widehat{S}^M_{jk}$ and $h_S(\cdot)$ is the conditional density of $\left(\widehat{\gamma},\ \widehat{U},\ \widehat{Z}\right) \mid \widehat{S}^M_{jk} = S$.

Based on Lemma 2, we already know that $f(S) \equiv \phi\left(S; \nu^M_{jk}, \Sigma^M_{jk}\right)$. In what follows, we determine $h_S(\gamma, U, Z)$ by using $\Pi_{\widehat{S}^M_{jk}}$, defined in Lemma 3, as a change of variables mapping from the randomization ω in (1) to $(\widehat{\gamma},\ \widehat{U},\ \widehat{Z})$, while conditioning on the value of $\widehat{S}^M_{jk}$.

Since $\widehat{S}^M_{jk}$ depends only on y and not on the randomization variable ω, which is independent of y, the density of $\omega \mid \widehat{S}^M_{jk} = S$ is equal to $\phi(\omega; 0_q, \Omega)$. Now, we apply the change of variable mapping

$$(\widehat{\gamma}, \widehat{U}, \widehat{Z}) = \Pi^{-1}_{\widehat{S}^M_{jk}}(\omega),$$

where $\widehat{\gamma} \in \mathbb{R}_+^{|M|}$ and $\widehat{U} \in \mathcal{U}_M$. At $\widehat{S}^M_{jk} = S$, this leads to

$$h_S(\gamma, U, Z) \cdot 1_{\mathcal{U}_M}(U) = \phi\left(\Pi_S(\gamma, U, Z); 0_q, \Omega\right) \cdot \det D_{\Pi_S}(\gamma, U, Z) \cdot 1_{\mathcal{U}_M}(U),$$

where $D_{\Pi_S}(\gamma, U, Z)$ denotes the Jacobian matrix of the mapping $\Pi_S(\gamma, U, Z)$. Applying the same calculations as done in the proof of Theorem 2 in Panigrahi et al. (2023), we obtain

$$\det D_{\Pi_S}(\gamma, U, Z) \propto \det\left(\Gamma(\gamma) + \left(\mathcal{U}_\perp(U)\right)^\top Q_M^{-1} \Lambda \mathcal{U}_\perp(U)\right).$$

Thus, we conclude that the density in (20) is proportional to

$$\phi(S; \nu^M_{jk}, \Sigma^M_{jk}) \cdot \phi\left(\Pi_S(\gamma, U, Z); 0_q, \Omega\right) \cdot \det\left(\Gamma(\gamma) + \left(\mathcal{U}_\perp(U)\right)^\top Q_M^{-1} \Lambda \mathcal{U}_\perp(U)\right) \times 1_{\mathbb{R}_+^{|M|}}(\gamma) \cdot 1_{\mathcal{U}_M}(U) \cdot 1_{\mathcal{Z}_M}(Z).$$

■

A.2.2 Proof of Theorem 1

We first state the following lemma.

Lemma 4 (Exact joint likelihood)**.** *Consider a density function that is proportional to*

$$\phi\left(\begin{pmatrix} t \\ b \end{pmatrix}; \begin{pmatrix} \theta^M_{jk} \\ \beta^M_{jk} \end{pmatrix}, \overline{\Sigma}^M_{jk}\right) \cdot \phi\left(\Pi_S(g, U, Z); 0_q, \Omega\right) \cdot \det D_{\Pi_S}(g, U, Z),$$

when evaluated at (t, b, g). This density function admits an alternative representation and can also be expressed as being proportional to

$$\phi\left(\begin{pmatrix} t \\ b \end{pmatrix}; \overline{R}\begin{pmatrix} \theta^M_{jk} \\ \beta^M_{jk} \end{pmatrix} + \overline{s}, \overline{\Theta}\right) \cdot \phi\left(g; \overline{A}\begin{pmatrix} t \\ b \end{pmatrix} + \overline{b}, \overline{\Omega}\right) \cdot \det D_{\Pi_S}(g, U, Z).$$

Proof The proof follows from straightforward algebra, by expanding the quadratic forms in the exponents of density functions $\phi\left(\begin{pmatrix} t \\ b \end{pmatrix}; \begin{pmatrix} \theta^M_{jk} \\ \beta^M_{jk} \end{pmatrix}, \overline{\Sigma}^M_{jk}\right)$ and $\phi\left(\Pi_S(g, U, Z); 0_q, \Omega\right)$, and reordering the terms. A similar derivation can be found in the proof of Proposition 3.4 in Huang et al. (2025). ■

Now, we are ready to state the proof of Theorem 1.

Proof [Proof of Theorem 1] Recall from (10) that $\widehat{S}^M_{jk} = \left(\widehat{\theta}^M_{jk}, \left(\widehat{\beta}^M_{jk}\right)^\top, \left(\widehat{\mathcal{A}}^M_{jk}\right)^\top\right)^\top$. Using Proposition 1, and the independence between $\widehat{\mathcal{A}}^M_{jk}$ and $(\widehat{\theta}^M_{jk}, \widehat{\beta}^M_{jk})$, the conditional distribution

$$\left(\widehat{\theta}^M_{jk},\ \widehat{\beta}^M_{jk},\ \widehat{\gamma},\ \widehat{U},\ \widehat{Z}\right) \ \Big|\ \left\{\widehat{\gamma} \succ 0_{|M|},\ \widehat{U} \in \mathcal{U}_M,\ \widehat{Z} \in \mathcal{Z}_M,\ \widehat{\mathcal{A}}^M_{jk} = \mathcal{A}\right\}$$

evaluated at (t, b, g, U, Z) has density

$$\phi\left(\begin{pmatrix} t \\ b \end{pmatrix}; \begin{pmatrix} \theta^M_{jk} \\ \beta^M_{jk} \end{pmatrix}, \overline{\Sigma}^M_{jk}\right) \cdot \phi\left(\Pi_S(g, U, Z); 0_q, \Omega\right) \cdot \det D_{\Pi_S}(g, U, Z) \cdot 1_{\mathbb{R}^{|M|}_+}(g) \cdot 1_{\mathcal{U}_M}(U) \cdot 1_{\mathcal{Z}_M}(Z).$$

Further conditioning on $\{\widehat{U} = U\ \widehat{Z} = Z\}$, the density of

$$(\widehat{\theta}^M_{jk},\ \widehat{\beta}^M_{jk},\ \widehat{\gamma}) \mid \{\widehat{\gamma} \succ 0_{|M|},\ \widehat{U} = U,\ \widehat{Z} = Z,\ \widehat{\mathcal{A}}^M_{jk} = \mathcal{A}\}$$

evaluated at (t, b, g) is equal to

$$\frac{\phi\left(\begin{pmatrix} t \\ b \end{pmatrix}; \begin{pmatrix} \theta^M_{jk} \\ \beta^M_{jk} \end{pmatrix}, \overline{\Sigma}^M_{jk}\right) \cdot \phi\left(\Pi_S(g, U, Z); 0_q, \Omega\right) \cdot \det D_{\Pi_S}(g, U, Z) \cdot 1_{\mathbb{R}^{|M|}_+}(g)}{\int \phi\left(\begin{pmatrix} t' \\ b' \end{pmatrix}; \begin{pmatrix} \theta^M_{jk} \\ \beta^M_{jk} \end{pmatrix}, \overline{\Sigma}^M_{jk}\right) \cdot \phi\left(\Pi_S(g', U, Z); 0_q, \Omega\right) \cdot \det D_{\Pi_S}(g', U, Z) \cdot 1_{\mathbb{R}^{|M|}_+}(g') dg' dt' db'}.$$

Then, using Lemma 4, we observe that this density is further proportional to

$$\phi\left(\begin{pmatrix} t \\ b \end{pmatrix}; \overline{R}\begin{pmatrix} \theta^M_{jk} \\ \beta^M_{jk} \end{pmatrix} + \overline{s}, \overline{\Theta}\right) \cdot \phi\left(g; \overline{A}\begin{pmatrix} t \\ b \end{pmatrix} + \overline{b}, \overline{\Omega}\right) \cdot \det D_{\Pi_S}(g, U, Z) \cdot 1_{\mathbb{R}^{|M|}_+}(g).$$

Marginalizing over $\widehat{\gamma}$ yields the joint likelihood

$c(\theta^M_{jk}, \beta^M_{jk})^{-1} \times$

$$\phi\left(\begin{pmatrix} \widehat{\theta}^M_{jk} \\ \widehat{\beta}^M_{jk} \end{pmatrix}; \overline{R}\begin{pmatrix} \theta^M_{jk} \\ \beta^M_{jk} \end{pmatrix} + \overline{s}, \overline{\Theta}\right) \cdot \int_{\gamma \succ 0_{|M|}} \phi\left(\gamma; \overline{A}\begin{pmatrix} \widehat{\theta}^M_{jk} \\ \widehat{\beta}^M_{jk} \end{pmatrix} + \overline{b}, \overline{\Omega}\right) \cdot \det D_{\Pi_S}(\gamma, U, Z)\ d\gamma.$$

Dropping the integral that is independent of the parameters $\theta^M_{jk}, \beta^M_{jk}$ and taking the logarithm gives the desired result in (13). ■

Appendix B. Extension of the Inference Procedure to Simultaneous Screening of Interaction Terms

As described in Section 3, for marginal screening, we condition on the event in Lemma 1 to account for the SPAM fit in Step 1 of the pipeline; that is, the test is performed conditional on observing $\{\widehat{M} = M\}$. The key statistics in the simultaneous screening model (17), however, differ from those identified for the univariate screening of interactions. We begin by identifying these statistics and then characterize the conditional distribution of the relevant key statistics, which in turn leads to approximate maximum likelihood inference. These three steps are outlined in Sections B.1, B.2 and B.3 for the new setup.

B.1 The Key Statistic for Testing Multiple Interaction Terms

Denote by $\overline{Z}_{\mathcal{C}}^{M} = \left(\mathcal{I}_{\mathcal{C}} : \Psi_M\right) \in \mathbb{R}^{n\times(q_M+|\mathcal{C}|)}$, the interaction-augmented design matrix obtained by concatenating Ψ_M with $\mathcal{I}_{\mathcal{C}}$. Let

$$\begin{pmatrix}\widehat{\theta}_{\mathcal{C}}^{M}\\ \widehat{\beta}_{\mathcal{C}}^{M}\end{pmatrix} = \underset{\theta\in\mathbb{R}^{|\mathcal{C}|},\ \beta\in\mathbb{R}^{q_M}}{\arg\min} \|y - \Psi_M\beta - \mathcal{I}_{\mathcal{C}}\theta\|_2^2 = \left(\overline{Z}_{\mathcal{C}}^{M}\right)^{+} y, \tag{21}$$

be the least squares estimator when we regress y against the derived features in $\overline{Z}_{\mathcal{C}}^{M}$. Performing an orthogonal decomposition of y into the fitted value and the residuals from this regression, we can write:

$$y = \overline{Z}_{\mathcal{C}}^{M}\begin{pmatrix}\widehat{\theta}_{\mathcal{C}}^{M}\\ \widehat{\beta}_{\mathcal{C}}^{M}\end{pmatrix} + \left(y - \overline{Z}_{\mathcal{C}}^{M}\begin{pmatrix}\widehat{\theta}_{\mathcal{C}}^{M}\\ \widehat{\beta}_{\mathcal{C}}^{M}\end{pmatrix}\right).$$

Equivalently, we have a decomposition for $-\Psi^{\top}y$, given by:

$$-\Psi^{\top}y = -\Psi^{\top}\overline{Z}_{\mathcal{C}}^{M}\begin{pmatrix}\widehat{\theta}_{\mathcal{C}}^{M}\\ \widehat{\beta}_{\mathcal{C}}^{M}\end{pmatrix} + \begin{pmatrix}0_{q_M}\\ \widehat{\mathcal{A}}_{\mathcal{C}}^{M}\end{pmatrix}, \tag{22}$$

where $\widehat{\mathcal{A}}_{\mathcal{C}}^{M} = -\Psi_{-M}^{\top}\left(y - \overline{Z}_{\mathcal{C}}^{M}\begin{pmatrix}\widehat{\theta}_{\mathcal{C}}^{M}\\ \widehat{\beta}_{\mathcal{C}}^{M}\end{pmatrix}\right) \in \mathbb{R}^{q-q_M}$.

We define the collection of key statistics in our new setup with multiple interaction features as:

$$\widehat{S}_{\mathcal{C}}^{M} = \left(\widehat{\theta}_{\mathcal{C}}^{M},\ \left(\widehat{\beta}_{\mathcal{C}}^{M}\right)^{\top},\ \left(\widehat{\mathcal{A}}_{\mathcal{C}}^{M}\right)^{\top}\right)^{\top} \in \mathbb{R}^{q+|\mathcal{C}|}, \tag{23}$$

which coincides with the key statistics identified earlier in Section 3.2 when $|\mathcal{C}| = 1$.

Now, we state Lemma 5, the analog of Lemma 2 when we pose the model (17), which can be proved by a straightforward adaptation of the proof of Lemma 2.

Lemma 5 (Marginal distribution of the key statistics with multiple interaction terms)**.** *For any fixed set of additive main effects M and any fixed set of interaction indices $\mathcal{C} \subset \mathcal{T}^M$, consider the model*

$$y \sim N\left(\Psi_M\beta_{\mathcal{C}}^{M} + \mathcal{I}_{\mathcal{C}}\theta_{\mathcal{C}}^{M},\ \sigma^2 I_n\right).$$

In this model, $\widehat{S}_{\mathcal{C}}^{M} \sim N_{q+|\mathcal{C}|}\left(\nu_{\mathcal{C}}^{M}, \Sigma_{\mathcal{C}}^{M}\right)$ where

$$\nu_{\mathcal{C}}^{M} = \begin{pmatrix}\theta_{\mathcal{C}}^{M}\\ \beta_{\mathcal{C}}^{M}\\ 0_{q-q_M}\end{pmatrix}, \qquad \Sigma_{\mathcal{C}}^{M} = \begin{pmatrix}\overline{\Sigma}_{\mathcal{C}}^{M} & 0_{(q_M+|\mathcal{C}|)\times(q-q_M)}\\ 0_{(q-q_M)\times(q_M+|\mathcal{C}|)} & \widetilde{\Sigma}_{\mathcal{C}}^{M}\end{pmatrix},$$

$$\overline{\Sigma}_{\mathcal{C}}^{M} = \sigma^2 \left(\left(\overline{Z}_{\mathcal{C}}^{M} \right)^{\top} \overline{Z}_{\mathcal{C}}^{M} \right)^{-1} \in \mathbb{R}^{(q_M + |\mathcal{C}|) \times (q_M + |\mathcal{C}|)}, \text{ and}$$

$$\widetilde{\Sigma}_{\mathcal{C}}^{M} = \sigma^2 \Psi_{-M}^{\top} \left(I_n - \overline{Z}_{\mathcal{C}}^{M} \left(\overline{Z}_{\mathcal{C}}^{M} \right)^{+} \right) \Psi_{-M} \in \mathbb{R}^{(q - q_M) \times (q - q_M)}.$$

The distribution of the key statistics leads to naive inference, as it treats M and the set of interactions we aim to test, $\mathcal{C} \subset \mathcal{T}^M$, as fixed. In the next two sections, we derive a valid p-value for the null hypothesis in (16) by characterizing the conditional distribution of

$$\left(\widehat{\theta}_{\mathcal{C}}^{M}, (\widehat{\beta}_{\mathcal{C}}^{M})^{\top} \right)^{\top} \Big| \left\{ \widehat{\gamma} \succ 0_{|M|},\ \widehat{U} = U,\ \widehat{Z} = Z,\ \widehat{\mathcal{A}}_{\mathcal{C}}^{M} = \mathcal{A} \right\}, \tag{24}$$

based on which we obtain an approximate likelihood function in $\left(\theta_{\mathcal{C}}^{M}, (\beta_{\mathcal{C}}^{M})^{\top} \right)^{\top}$.

B.2 Conditional Distribution of the Key Statistics for Testing Multiple Interaction Terms

Similar to Section 3.3, we derive the appropriate change of variables mapping, as done previously in Lemma 3, which enables us to derive the conditional distribution in (24).

Lemma 6. *Define the mapping* $\Pi_{\widehat{S}_{\mathcal{C}}^{M}}(\cdot, \cdot, \cdot) : \mathbb{R}^{|M|} \times \mathcal{U}_M \times \mathbb{R}^{q - q_M} \to \mathbb{R}^q$ *by*

$$\Pi_{\widehat{S}_{\mathcal{C}}^{M}}(\gamma, U, Z) = A_{\mathcal{C}}^{M} \begin{pmatrix} \widehat{\theta}_{\mathcal{C}}^{M} \\ \widehat{\beta}_{\mathcal{C}}^{M} \end{pmatrix} + B(U)\gamma + c(U, Z),$$

where

$$A_{\mathcal{C}}^{M} = -\Psi^{\top} \overline{Z}_{\mathcal{C}}^{M},\ B(U) = \left(\Psi^{\top} \Psi_M + \begin{pmatrix} \epsilon I_{q_M} \\ 0_{(q - q_M) \times q_M} \end{pmatrix} \right) \mathrm{bd}(U),$$

$$c(U, Z) = \begin{pmatrix} 0_{q_M} \\ \widehat{\mathcal{A}}_{\mathcal{C}}^{M} \end{pmatrix} + \begin{pmatrix} (\lambda_j u_j)_{j \in M} \\ (\lambda_j z_j)_{j \notin M} \end{pmatrix}.$$

Consider the estimators $(\widehat{\gamma}, \widehat{U}, \widehat{Z})$ *that satisfy the relationship in* (6)*. Then, it holds that*

$$(\widehat{\gamma}, \widehat{U}, \widehat{Z}) = \Pi_{\widehat{S}_{\mathcal{C}}^{M}}^{-1}(\omega).$$

Similar to Lemma 3, Lemma 6 offers a mapping between ω and the key statistic $\widehat{S}_{\mathcal{C}}^{M}$ and the optimization variables $(\widehat{\gamma}, \widehat{U}, \widehat{Z})$. Here, the only term that differs from Lemma 3 is $A_{\mathcal{C}}^{M}$, while the other terms $B(U)$ and $c(U, Z)$ remain the same. This term differs from its counterpart A_{jk}^{M} because the new estimator $\begin{pmatrix} \widehat{\theta}_{\mathcal{C}}^{M} \\ \widehat{\beta}_{\mathcal{C}}^{M} \end{pmatrix}$ leads to a different orthogonal decomposition of $-\Psi^{\top} y$.

Lemma 6 offers a mapping between ω and $\left(\widehat{S}_{\mathcal{C}}^{M},\ \widehat{\gamma},\ \widehat{U},\ \widehat{Z} \right)$, and similar to Proposition 1, we can leverage this mapping to characterize the distribution of the randomized group lasso statistics $(\widehat{\gamma},\ \widehat{U},\ \widehat{Z})$ and the key statistics $\widehat{S}_{\mathcal{C}}^{M}$ through a change of variables applied to the known distribution of ω, under the model with multiple interaction terms. This characterization is provided in Proposition 2.

Proposition 2 (Extension of Proposition 1 to testing with multiple interaction terms)**.** *The density function of the conditional distribution of*

$$\left(\widehat{S}_{\mathcal{C}}^{M},\ \widehat{\gamma},\ \widehat{U},\ \widehat{Z} \right) \Big| \left\{ \widehat{\gamma} \succ 0_{|M|},\ \widehat{U} \in \mathcal{U}_M,\ \widehat{Z} \in \mathcal{Z}_M \right\}$$

when evaluated at (S, γ, U, Z) *is proportional to*

$$\phi(S; \nu_{\mathcal{C}}^M, \Sigma_{\mathcal{C}}^M) \cdot \phi\left(\Pi_S(\gamma, U, Z); 0_q, \Omega\right) \cdot \det D_{\Pi_S}(\gamma, U, Z) \cdot 1_{\mathbb{R}_+^{|M|}}(\gamma) \cdot 1_{\mathcal{U}_M}(U) \cdot 1_{\mathcal{Z}_M}(Z)$$

where $D_{\Pi_S}(\gamma, U, Z) = \Gamma(\gamma) + \left(\mathcal{U}_\perp(U)\right)^\top Q_M^{-1} \Lambda\left(\mathcal{U}_\perp(U)\right)$.

From the density in Proposition 2, we obtain the selective log-likelihood by conditioning further on the observed values of $\widehat{U}$, $\widehat{Z}$ and $\widehat{\mathcal{A}}_{\mathcal{C}}^M$. Theorem 3 provides the expression for this likelihood.

Define the following matrices:

$$\overline{\Omega} = \left(B(\widehat{U})^\top \Omega^{-1} B(\widehat{U})\right)^{-1}, \quad \overline{A} = -\overline{\Omega} B(\widehat{U})^\top \Omega^{-1} A_{\mathcal{C}}^M, \quad \overline{b} = -\overline{\Omega} B(\widehat{U})^\top \Omega^{-1} c(\widehat{U}, \widehat{Z}),$$

$$\overline{\Theta} = \left((\overline{\Sigma}_{\mathcal{C}}^M)^{-1} - \overline{A}^\top \overline{\Omega}^{-1} \overline{A} + (A_{\mathcal{C}}^M)^\top \Omega^{-1} A_{\mathcal{C}}^M\right)^{-1}, \quad \overline{R} = \overline{\Theta}(\overline{\Sigma}_{\mathcal{C}}^M)^{-1},$$

$$\overline{s} = \overline{\Theta}\left(\overline{A}^\top \overline{\Omega}^{-1} \overline{b} - (A_{\mathcal{C}}^M)^\top \Omega^{-1} c(\widehat{U}, \widehat{Z})\right),$$

where $A_{\mathcal{C}}^M$, $B(\widehat{U})$, and $c(\widehat{U}, \widehat{Z})$ are as defined in Lemma 6.

While we adopt the notation $\overline{\Omega}$, $\overline{A}$, $\overline{b}$, $\overline{\Theta}$, $\overline{R}$, and $\overline{s}$ from Proposition 1, we emphasize that the matrices $\overline{A}$, $\overline{\Theta}$, $\overline{R}$, and $\overline{s}$ are different from their counterparts under marginal screening in Section 3.

Additionally, let

$$c(\theta_{\mathcal{C}}^M, \beta_{\mathcal{C}}^M) = \int \int_{g' \succ 0_{|M|}} \phi\left(b'; \overline{R}\begin{pmatrix}\theta_{\mathcal{C}}^M \\ \beta_{\mathcal{C}}^M\end{pmatrix} + \overline{s}, \overline{\Theta}\right) \cdot \phi\left(g'; \overline{A}b' + \overline{b}, \overline{\Omega}\right) \cdot \det D_{\Pi_S}(g', U, Z)\, dg'\, db',$$

denote the normalizing constant in the selective log-likelihood, as seen in the next result.

Theorem 3 (Extension of selective log-likelihood to testing with multiple interaction terms.)**.** *The log-likelihood of* $\left(\widehat{\theta}_{\mathcal{C}}^M, (\widehat{\beta}_{\mathcal{C}}^M)^\top\right)^\top$ *conditional on*

$$\left\{\widehat{\gamma} \succ 0_{|M|},\ \widehat{U} = U,\ \widehat{Z}\ = Z,\ \widehat{\mathcal{A}}_{\mathcal{C}}^M = \mathcal{A}\right\},$$

is equal to

$$\ell_M(\theta_{\mathcal{C}}^M,\ \beta_{\mathcal{C}}^M; U, Z, \mathcal{A}) = -\log c(\theta_{\mathcal{C}}^M, \beta_{\mathcal{C}}^M) + \log \phi\left(\begin{pmatrix}\widehat{\theta}_{\mathcal{C}}^M \\ \widehat{\beta}_{\mathcal{C}}^M\end{pmatrix}; \overline{R}\begin{pmatrix}\theta_{\mathcal{C}}^M \\ \beta_{\mathcal{C}}^M\end{pmatrix} + \overline{s}, \overline{\Theta}\right). \tag{25}$$

Theorem 3 is an extension of Theorem 1, which derives the selective log-likelihood under the scenario of testing multiple interaction terms.

B.3 Maximum Likelihood Inference for Testing Multiple Interaction Terms

Applying the Laplace approximation to the normalizing factor, with

$$\begin{aligned}\log \widehat{c}(\theta_{\mathcal{C}}^M, \beta_{\mathcal{C}}^M) = -\inf_{b', g'} \Bigg\{ &\frac{1}{2}\left(b' - \overline{R}\begin{pmatrix}\theta_{\mathcal{C}}^M \\ \beta_{\mathcal{C}}^M\end{pmatrix} - \overline{s}\right)^\top \overline{\Theta}^{-1} \left(b' - \overline{R}\begin{pmatrix}\theta_{\mathcal{C}}^M \\ \beta_{\mathcal{C}}^M\end{pmatrix} - \overline{s}\right) \\ &+ \frac{1}{2}\left(g' - \overline{A}b' - \overline{b}\right)^\top \overline{\Omega}^{-1}\left(g' - \overline{A}b' - \overline{b}\right) \\ &- \log\det D_{\Pi_S}(g', U, Z) + \mathrm{Barr}(g')\Bigg\},\end{aligned}$$

we arrive at the approximate selective log-likelihood

$$\log \widehat{\mathcal{L}}_M(\theta_{\mathcal{C}}^M,\ \beta_{\mathcal{C}}^M; U, Z, \mathcal{A}) = \log \phi\left(\begin{pmatrix}\widehat{\theta}_{\mathcal{C}}^M \\ \widehat{\beta}_{\mathcal{C}}^M\end{pmatrix}; \overline{R}\begin{pmatrix}\theta_{\mathcal{C}}^M \\ \beta_{\mathcal{C}}^M\end{pmatrix} + \overline{s}, \overline{\Theta}\right) - \log \widehat{c}(\theta_{\mathcal{C}}^M, \beta_{\mathcal{C}}^M). \tag{26}$$

Note that the approximate selective likelihood expression in (14) can be viewed as a special case of this more general setup; i.e., the above-stated expression reduces to the likelihood function derived earlier in (14) when $|\mathcal{C}| = 1$.

Building on this generalized version of the approximate selective log-likelihood, we obtain the selective MLE $(\widehat{\theta}^{\mathcal{C}}_{mle}, \widehat{\beta}^{\mathcal{C}}_{mle})$ of $(\theta^M_{\mathcal{C}}, \beta^M_{\mathcal{C}})$ as before.

Theorem 4. *Consider solving the $|M|$-dimensional optimization problem* $g^*(\widehat{\theta}^M_{\mathcal{C}}, \widehat{\beta}^M_{\mathcal{C}}) =$

$$\arg\min_g \frac{1}{2}\left(g - \overline{A}\begin{pmatrix}\widehat{\theta}^M_{\mathcal{C}}\\ \widehat{\beta}^M_{\mathcal{C}}\end{pmatrix} - \overline{b}\right)^\top \overline{\Omega}^{-1}\left(g - \overline{A}\begin{pmatrix}\widehat{\theta}^M_{\mathcal{C}}\\ \widehat{\beta}^M_{\mathcal{C}}\end{pmatrix} - \overline{b}\right) - \log\det D_{\Pi_S}(g, U, Z) + \mathrm{Barr}(g). \tag{27}$$

Then, the joint maximum likelihood estimator for $\begin{pmatrix}\theta^M_{\mathcal{C}}\\ \beta^M_{\mathcal{C}}\end{pmatrix}$ *is obtained as*

$$\begin{pmatrix}\widehat{\theta}^{\mathcal{C}}_{mle}\\ \widehat{\beta}^{\mathcal{C}}_{mle}\end{pmatrix} = \overline{R}^{-1}\begin{pmatrix}\widehat{\theta}^M_{\mathcal{C}}\\ \widehat{\beta}^M_{\mathcal{C}}\end{pmatrix} - \overline{R}^{-1}\overline{s} + \overline{\Sigma}^M_{\mathcal{C}}\overline{A}^\top\overline{\Omega}^{-1}\left(\overline{A}\begin{pmatrix}\widehat{\theta}^M_{\mathcal{C}}\\ \widehat{\beta}^M_{\mathcal{C}}\end{pmatrix} + \overline{b} - g^*(\widehat{\theta}^M_{\mathcal{C}}, \widehat{\beta}^M_{\mathcal{C}})\right).$$

Define the $(q_M + |\mathcal{C}|) \times (q_M + |\mathcal{C}|)$ *matrix*

$$\begin{aligned} K =& \overline{\Theta}^{-1} + \overline{A}^\top\overline{\Omega}^{-1}\overline{A} \\ & - \overline{A}^\top\overline{\Omega}^{-1}\left[\overline{\Omega}^{-1} - \nabla^2\log\det D_{\Pi_S}(g^*(\widehat{\theta}^M_{\mathcal{C}}, \widehat{\beta}^M_{\mathcal{C}}), U, Z) + \nabla^2\mathrm{Barr}(g^*(\widehat{\theta}^M_{\mathcal{C}}, \widehat{\beta}^M_{\mathcal{C}}))\right]^{-1}\overline{\Omega}^{-1}\overline{A}. \end{aligned}$$

The observed Fisher information matrix based on the approximate selective log-likelihood in (26) *is equal to* $I_{mle} = (\overline{\Sigma}^M_{\mathcal{C}})^{-1}K^{-1}(\overline{\Sigma}^M_{\mathcal{C}})^{-1}$.

To conclude this discussion, we show how the MLE and the observed Fisher information matrix can be used to obtain a p-value for the null hypothesis in (16), which is equivalent to

$$H^{\mathcal{C}}_{0'} : \|\theta^M_{\mathcal{C}}\|_2 = 0.$$

Under the null, we have that

$$X^2_{\mathcal{C}} = (\widehat{\theta}^{\mathcal{C}}_{mle})^\top\left(I^{-1}_{mle}[\mathcal{C}]\right)^{-1}\widehat{\theta}^{\mathcal{C}}_{mle}$$

approximately follows the $\chi^2_{|\mathcal{C}|}$ distribution, where $I^{-1}_{mle}[\mathcal{C}]$ is the top-left $|\mathcal{C}|\times|\mathcal{C}|$-block of I^{-1}_{mle}. This is because the distribution of $\widehat{\theta}^{\mathcal{C}}_{mle}$ is approximated as $N_{|\mathcal{C}|}\left(\theta^M_{\mathcal{C}}, I^{-1}_{mle}[\mathcal{C}]\right)$. Therefore, a valid p-value for the global null $H^{\mathcal{C}}_{0'}$ is given by:

$$\widehat{p}^M_{\mathcal{C}} = \mathbb{P}\left(\chi^2_{|\mathcal{C}|} > X^2_{\mathcal{C}}\right).$$

Appendix C. Additional Simulation Studies

We conduct a simulation study to evaluate the performance of the proposed method when testing all pairs of interactions, i.e., $\mathcal{T}^M = \{(j,k) \in [p]^2\}$, rather than restricting interactions based on the weak hierarchy rule presented in Section 4. Data are generated under the four settings described in Section 4.

First, we report the false discovery rate (FDR) of the p-values produced by the proposed method, data splitting, and naive inference, both before and after applying the Benjamini–Hochberg correction (BH-q), in Figure 9. Second, we present the F1 scores for testing the hypotheses on θ^M_{jk} for all three methods, again before and after applying BH-q, in Figure 10.

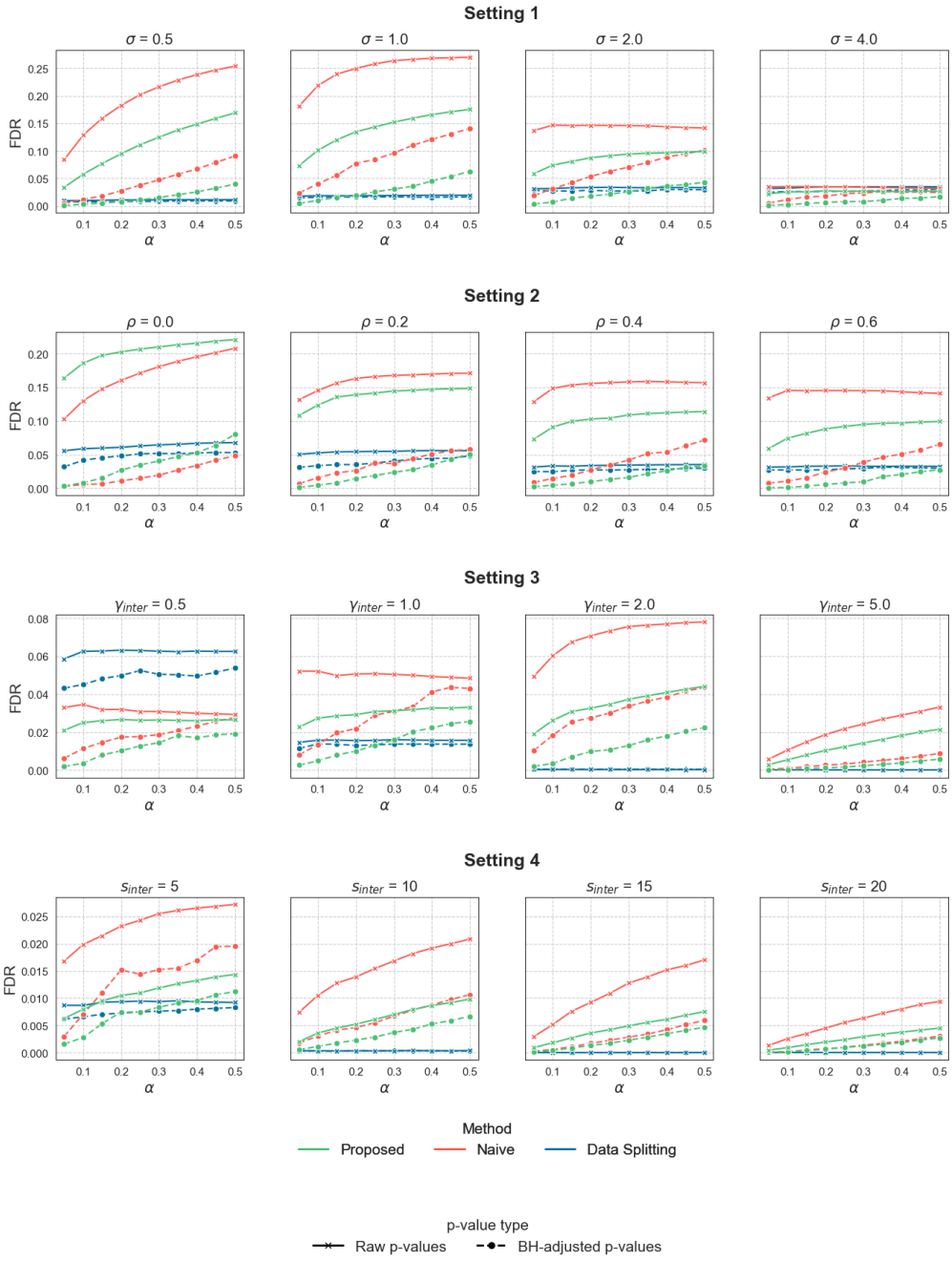


Figure 9: False discovery rate (FDR) of hypothesis testing of original and BH-adjusted p-values obtained by naive inference, data splitting, and the proposed method in all simulation settings with varying α, and all possible p^2 interactions tested

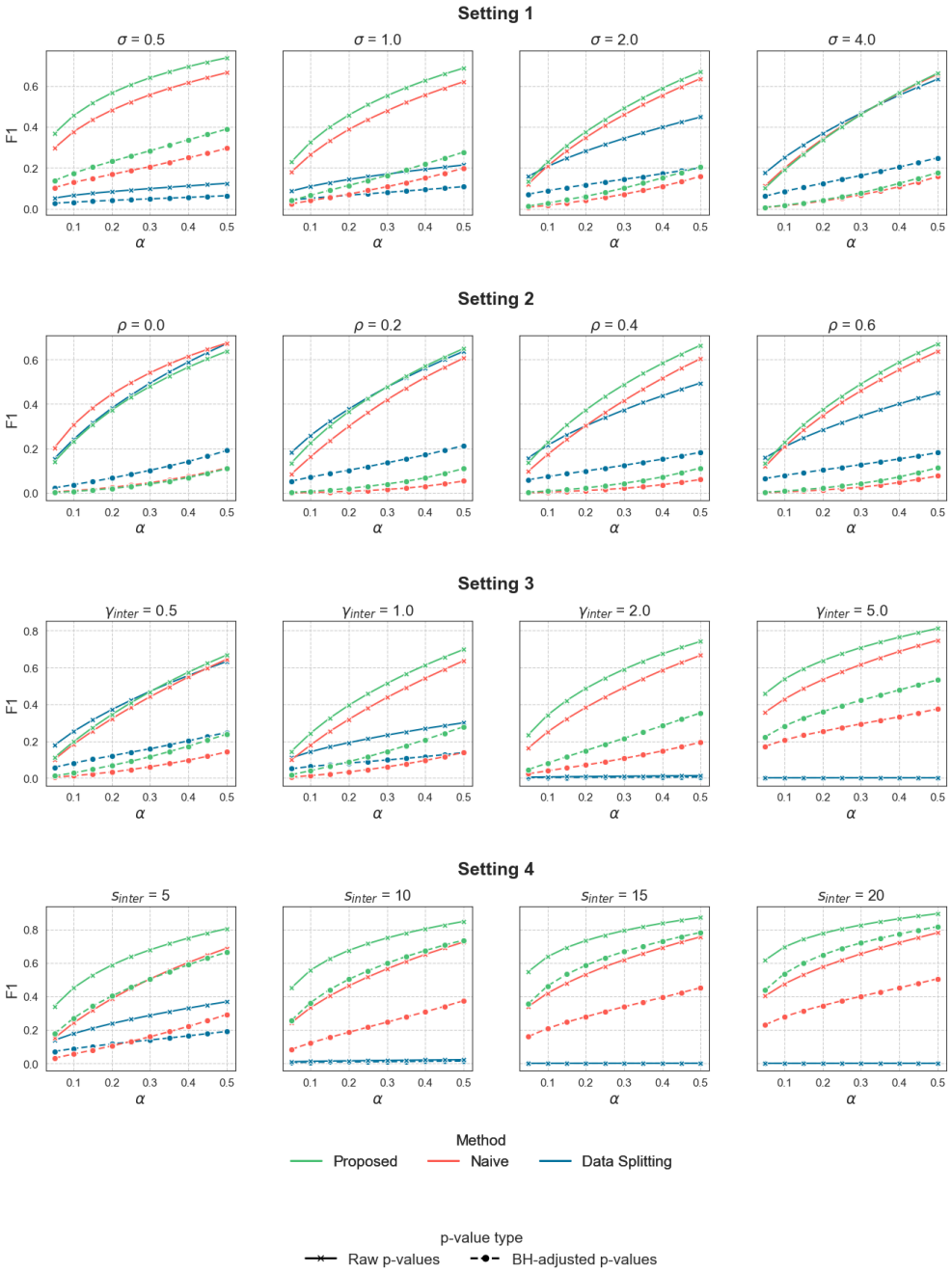


Figure 10: F1 scores for hypothesis testing of θ^M_{jk} of original and BH-adjusted p-values obtained by naive inference, data splitting, and the proposed method in all simulation settings with varying α, and all possible p^2 interactions tested

As shown in Figure 9, applying the BH-q correction consistently yields valid FDR control. However, as illustrated in Figure 10, BH-q correction leads to a uniform deterioration in selection quality for all three methods, due to the loss of power after correcting for multiplicity. When comparing methods based on BH-corrected p-values, the proposed method achieves better selection performance than data splitting and naive inference in most settings (1, 3, and 4), and performs only slightly worse than data splitting in Setting 2.

We also conduct additional simulation studies to investigate how the proposed method scales with the size of the selected model. In Setting 4 of the simulations in Section 4, we record the computational time required for inference on each θ_{jk}^M together with the size $|M|$ of the selected model M. As shown in Figure 11, for fixed sample size n and design dimension q, the computational time for inference on θ_{jk}^M grows linearly with $|M|$, i.e., $O(|M|)$.

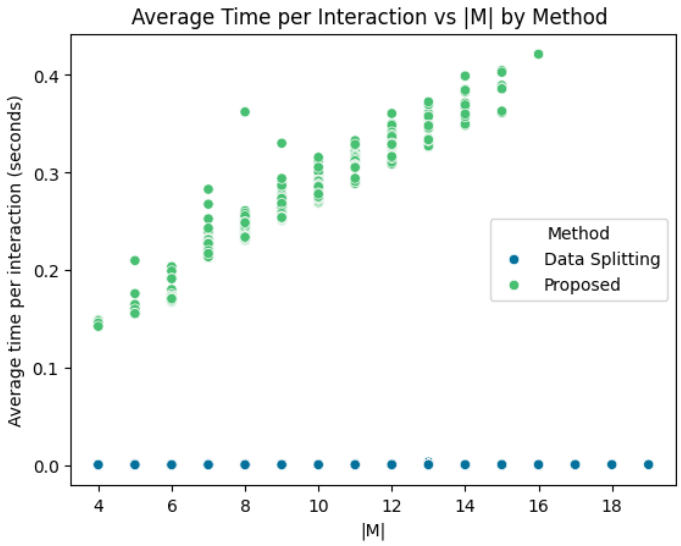


Figure 11: Average computation time (in seconds) for each interaction term as a function of $|M|$

Appendix D. Additional Details of the JFK Flight Data Analysis

In this section, we provide additional details of the JFK flight data analysis in Section 5 of the main paper.

D.1 Documentation of Features in Section 5

Feature name	Meaning	Source	Type
distance	Distance between airports (miles)	Flight	Nonlinear
sched_dep_hour	Scheduled departure hour	Flight	Nonlinear
sched_arr_hour	Scheduled arrival hour	Flight	Nonlinear
temp	Temperature (°F)	Weather	Nonlinear
dewp	Dewpoint (°F)	Weather	Nonlinear
humid	Relative humidity	Weather	Nonlinear
pressure	Sea level pressure (millibars)	Weather	Nonlinear
wind_dir	Wind direction (degrees)	Weather	Linear
wind_speed	Wind speed (mph)	Weather	Linear
precip	Precipitation (inches)	Weather	Linear
visib	Visibility (miles)	Weather	Linear
plane_age	Plane age (years)	Aircraft	Linear
seats	Number of seats on plane	Aircraft	Linear

Table 4: Feature names, meaning, source, and whether modeled linearly/nonlinearly in Section 5.1

References

Soham Bakshi, Yiling Huang, Snigdha Panigrahi, and Walter Dempsey. Inference with randomized regression trees. *arXiv preprint arXiv:2412.20535*, 2024.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *Annals of Statistics*, 41(2):802–837, 2013.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of Statistics*, 41(3):1111–1141, 2013.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, 2015.

Alexandra Chouldechova and Trevor Hastie. Generalized additive model selection. *arXiv preprint arXiv:1506.03850*, 2015.

Simon P. Couch. *anyflights: Query 'nycflights13'-Like Air Travel Data for Given Years and Airports*, 2025. URL `https://CRAN.R-project.org/package=anyflights`. R package version 0.3.5.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Ameer Dharamshi, Anna Neufeld, Keshav Motwani, Lucy L Gao, Daniela Witten, and Jacob Bien. Generalized data thinning using sufficient statistics. *Journal of the American Statistical Association*, 120(549):511–523, 2025.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Sofia Guglielmini, Gerda Claeskens, and Snigdha Panigrahi. Selective inference in graphical models via maximum likelihood. *arXiv preprint arXiv:2503.24311*, 2025.

Asad Haris, Daniela Witten, and Noah Simon. Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, 25(4):981–1004, 2016.

Asad Haris, Noah Simon, and Ali Shojaie. Generalized sparse additive models. *Journal of Machine Learning Research*, 23(70):1–56, 2022.

Hussein Hazimeh and Rahul Mazumder. Learning hierarchical interactions at scale: A convex optimization approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1833–1843. PMLR, 2020.

Yiling Huang, Sarah Pirenne, Snigdha Panigrahi, and Gerda Claeskens. Selective inference using randomized group lasso estimators for general models. *Electronic Journal of Statistics*, 19(2):3489–3531, 2025.

Danijel Kivaranovic and Hannes Leeb. A (tight) upper bound for the length of confidence intervals with conditional coverage. *Electronic Journal of Statistics*, 18(1):1677–1701, 2024.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927, 2016.

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: splitting a single data point. *Journal of the American Statistical Association*, 120(549):135–146, 2025.

Yi Lin and Hao Helen Zhang. Component selection and smoothing in smoothing spline analysis of variance models. *Annals of Statistics*, 34(5):2272–2297, 2006.

Yin Lou, Jacob Bien, Rich Caruana, and Johannes Gehrke. Sparse partially linear additive models. *Journal of Computational and Graphical Statistics*, 25(4):1126–1140, 2016.

Adam McCloskey. Hybrid confidence intervals for informative uniform asymptotic inference after model selection. *Biometrika*, 111(1):109–127, 2024.

John A Nelder. A reformulation of linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 140(1):48–63, 1977.

Anna Neufeld, Ameer Dharamshi, Lucy L Gao, and Daniela Witten. Data thinning for convolution-closed distributions. *Journal of Machine Learning Research*, 25(57):1–35, 2024.

Anna C Neufeld, Lucy L Gao, and Daniela M Witten. Tree-values: selective inference for regression trees. *Journal of Machine Learning Research*, 23(305):1–43, 2022.

Snigdha Panigrahi and Jonathan Taylor. Approximate selective inference via maximum likelihood. *Journal of the American Statistical Association*, 118(544):2810–2820, 2023.

Snigdha Panigrahi, Jelena Markovic, and Jonathan Taylor. An MCMC-free approach to post-selective inference. *arXiv preprint arXiv:1703.06154*, 2017.

Snigdha Panigrahi, Jonathan Taylor, and Asaf Weinstein. Integrative methods for post-selection inference under convex constraints. *Annals of Statistics*, 49(5):2803–2824, 2021.

Snigdha Panigrahi, Peter W MacDonald, and Daniel Kessler. Approximate post-selective inference for regression with the group lasso. *Journal of Machine Learning Research*, 24 (79):1–49, 2023.

Snigdha Panigrahi, Kevin Fry, and Jonathan Taylor. Exact selective inference with randomization. *Biometrika*, 111(4):1109–1127, 2024.

Julio L Peixoto. Hierarchical variable selection in polynomial regression models. *The American Statistician*, 41(4):311–313, 1987.

D García Rasines and G Alastair Young. Splitting strategies for post-selection inference. *Biometrika*, 110(3):597–614, 2023.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030, 2009.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022. doi: 10.1214/21-SS133. URL `https://doi.org/10.1214/21-SS133`.

Shinya Suzumura, Kazuya Nakagawa, Yuta Umezu, Koji Tsuda, and Ichiro Takeuchi. Selective inference for sparse high-order interaction models. In *International Conference on Machine Learning*, pages 3338–3347. PMLR, 2017.

J Kenneth Tay and Robert Tibshirani. Reluctant generalised additive modelling. *International Statistical Review*, 88:S205–S224, 2020.

Guo Yu, Jacob Bien, and Ryan Tibshirani. Reluctant interaction modeling. *arXiv:1907.08414*, 2019.

Ming Yuan, V Roshan Joseph, and Hui Zou. Structured variable selection and estimation. *The Annals of Applied Statistics*, 3(4):1738–1757, 2009.

Tijana Zrnic and William Fithian. Locally simultaneous inference. *Annals of Statistics*, 52 (3):1227–1253, 2024.