Making Interpretable Discoveries from Unstructured Data: A High-Dimensional Multiple Hypothesis Testing Approach

Jacob Carlson*

October 31, 2025

Abstract

Social scientists are increasingly turning to unstructured datasets to unlock new empirical insights, e.g., estimating causal effects on text outcomes, measuring beliefs from open-ended survey responses. In such settings, unsupervised analysis is often of interest, in that the researcher does not want to pre-specify the objects of measurement or otherwise artificially delimit the space of measurable concepts; they are interested in "discovery." This paper proposes a general and flexible framework for pursuing discovery from unstructured data in a statistically principled way. The framework leverages recent methods from the literature on machine learning interpretability to map unstructured data points to high-dimensional, sparse, and interpretable "dictionaries" of concepts; computes (test) statistics of these dictionary entries; and then performs selective inference on them using a newly developed statistical procedure for high-dimensional exceedance control of the k-FWER under arbitrary dependence. The proposed framework has few researcher degrees of freedom, is fully replicable, and is cheap to implement—both in terms of financial cost and researcher time. Applications to recent descriptive and causal analyses of unstructured data in empirical economics are explored.

1 Introduction

Empowered by recent developments in machine learning and AI, researchers in the social sciences are increasingly leveraging sources of unstructured data—such as text, images, videos, and audio—in quantitative analyses. In economics, interest in unstructured data sources is especially widespread, ranging from using speech recordings from FOMC meetings to better understand monetary policy [Gorodnichenko et al., 2023]; to using videos of start-up pitches to study entrepreneurship and investment [Hu and Ma, 2025]; to using open-ended survey questions to probe the economic behavior and beliefs of individuals [Haaland et al., 2024]; to using mug shots to study judicial decisionmaking [Ludwig and Mullainathan, 2024];

*Email: jacob_carlson@g.harvard.edu.

to using visual art or written narratives to infer long-run living standards [Gorin et al., 2025, Lagakos et al., 2025]; to using interview transcripts to better understand the impacts of RCT treatments [Bergman et al., 2024, Krause et al., 2025].

The promise of these unstructured data, fundamentally, is that they offer new opportunities to measure social phenomena that were previously unmeasurable. This includes both measures that a researcher is capable of pre-specifying as well as those that a researcher is not, and instead would like to "discover." When modern AI/ML methods are brought to bear on these goals, the former falls under the heading of "supervised learning," whereby the researcher typically makes predictions of their ex-ante known quantity of interest using the unstructured data as features, and the latter falls under the heading of "unsupervised learning," which is less amenable to a prediction-based framework, and demands that AI/ML methods uncover latent structure in the unstructured data with minimal researcher input. Though new statistical and econometric frameworks have recently become available for interpretable and statistically principled analyses of unstructured data in the supervised learning setting (e.g., Angelopoulos et al. [2023], Ludwig et al. [2024], Carlson and Dell [2025]), it is an open question as to how best perform interpretable and statistically principled analyses of unstructured data in unsupervised learning settings that emphasize discovery. The framework proposed in this paper provides one possible answer to this open question.

Specifically, this paper proposes a general framework for conducting interpretable and rigorous statistical inference on unstructured data in the following way, starting from an unstructured dataset $\{Z_i\}_{i=1}^n$:

1. *Convert unstructured data to dictionary feature vectors:* The researcher converts the unstructured data instances $Z_i \in \mathcal{Z}$ into high-dimensional, sparse, binary vectors $Y_i \in \{0,1\}^p$ using recently developed methods in dictionary learning for mechanistic interpretability of large language models (LLMs) and other deep neural networks (DNNs), i.e., $Y_i := \texttt{Dict}(Z_i)$. Both theory and empirics corroborate that vector entries Y_{ij} correspond to the presence of a monosemantic feature (i.e., a feature with a single, interpretable meaning) or "concept" j in the given unstructured data instance i. There are $p \gg n$ dictionary features, e.g., we assume that n may be on the order of 10^2 or 10^3, and p is on the order of 10^4 or 10^5 (or larger). The breadth and depth of concepts spanned by these dictionary features is vast, as can be gleaned from the literature on dictionary learning for LLMs, and as will be observed concretely in the empirical applications of this paper. Intuitively, the dictionary should contain all possible (monosemantic) concepts that the LLM has learned for the purpose of next token prediction over a massive internet-scale corpus. However, due to sparsity imposed by design and suggested by machine learning theory, only a handful of these dictionary features will be present ("activate") for any given unstructured data point.[1]

2. *Compute test statistics from the dictionary feature vectors:* The researcher computes from this newly created dictionary feature indicator dataset $\{Y_i\}_{i=1}^n$ a vector of p test statistics, $T_n \in \mathbb{R}^p$. These test statistics are intended to test p hypotheses, one for each dictionary feature, and may be flexibly chosen by the researcher. Especially important

[1]For more discussion of what "monosemanticity" means and its importance for LLM interpretability, see, e.g., the discussion in Bricken et al. [2023].

and simple choices explored in this paper's empirical examples include hypotheses testing means and differences-in-means, feature by feature.

3. *Perform high-dimensional selective inference using the test statistics:* Finally, the researcher invokes a high-dimensional selective inference procedure newly developed in this paper to control the k-FWER of the selected set of rejected hypotheses ("discoveries"). This procedure is asymptotically valid for small k, requires only very mild assumptions on the data generating process (which are met in the applications of interest), and is well-powered in the empirical examples considered. This procedure may also be of independent interest in other high-dimensional selective inference settings.

The output of this framework is therefore a set of interpretable discoveries (rejections of dictionary feature derived tests) based on some set of unstructured data $\{Z_i\}_{i=1}^n$ with generalized familywise error rate guarantees, which may be relaxed or tightened as the researcher sees fit to permit more or fewer discoveries.

Two notable use cases for this framework explored in the empirical applications of this paper include:

1. Estimating average treatment effects on dictionary features in RCTs, or testing, for treatment indicator D_i and fixed probability of treatment π,

$$H_{0,j} : E[Y_{ij}(1) - Y_{ij}(0)] = E[H_i Y_{ij}] = 0, \quad H_i = \frac{D_i - \pi}{\pi(1-\pi)}.$$

2. Estimating the probability of the presence of various concepts (measured by monosemantic dictionary features) in the population from which the unstructured dataset was sampled, or testing

$$H_{0,j} : E[Y_{ij}] = 0$$

and then inverting these tests to form a "generalized" confidence set [Romano and Wolf, 2007].

This particular framework for discovery from unstructured data conveys several scientific benefits. The first is that of having *few researcher degrees of freedom.* In many settings featuring exploratory analysis of unstructured data, there is concern that researchers may cherry pick aspects of their data to measure or cherry pick choices of measurements in order to reverse engineer a specific conclusion. By design, however, the framework proposed by this paper has very little room for motivated data mining, as it requires no "human-in-the-loop" to direct what quantities are ultimately measured. Off-the-shelf, open-source dictionary learning models generate p features spanning a huge number of concepts that are well-cataloged and publicly documented, making it unlikely the researcher needs to define additional concepts to measure ad hoc, and making a failure to report hypotheses for certain features conspicuous. Moreover, prior to analysis, a researcher could simply preregister a choice of dictionary learning model, such as particular sparse autoencoder (SAE) for a particular LLM, limiting the possibility that a researcher applies this framework with many different dictionary learning methods to select on favorable results. The proposed framework is also *inexpensive* both in terms of financial cost and researcher burden; both empirical

examples in this paper were computed on Google Colab notebooks using a single A100 GPU, access to which only costs tens of dollars per month.[2] Further, this framework is *easy to implement*, requiring no human-in-the-loop to label unstructured data instances, no additional AI/ML model training, and no additional data sources; the framework is nearly automatic in execution, and takes only minutes to run using the above-mentioned hardware. This automaticity also lends the framework to fast and easy replication, another important scientific benefit.

Thus, the primary contributions of this paper are:

1. The development of a flexible, general-purpose framework for making interpretable as well as statistically and scientifically principled discoveries from unstructured data, leveraging state-of-the-art intepretability methods for LLMs and other DNNs.

2. The development of a novel high-dimensional multiple hypothesis testing procedure for controlling k-FWER exceedance, facilitated by extensions of high-dimensional central limit theory.

The remainder of this paper is structured as follows: Section 2 discusses related literature; Section 3 describes the framework in detail; Section 4 applies the framework to two recent papers making use of unstructured data for discovery [Bursztyn et al., 2023, Stantcheva, 2024]; and Section 5 concludes.

2 Related Literature

The framework proposed in this paper is related to recent works from literatures spanning economics, statistics, and computer science.

Econometric methods for unstructured data. Motivated by the capabilities of modern AI/ML methods for learning from unstructured economic data (see, e.g., Mullainathan and Spiess [2017], Gentzkow et al. [2019a], Dell [2025]), new econometric and statistical frameworks have been developed to facilitate principled statistical inference on low-dimensional features (predictions) learned from unstructured datasets under supervision, e.g., Angelopoulos et al. [2023], Ludwig et al. [2024], Carlson and Dell [2025], Rambachan et al. [2024]. This most recent wave of econometric literature emphasizes nonparametric frameworks compatible with black-box AI models coupled with debiasing methods, as opposed to model-based approaches, e.g., Gentzkow et al. [2019b], Battaglia et al. [2024]. Though the framework presented in this paper is concerned with principled inference on unstructured data without making parametric assumptions, it differs from this existing econometric literature in that its goal is primarily unsupervised discovery, as opposed to supervised detection, using AI/ML models.

Both the proposed framework and the recent econometric framework of Modarressi et al. [2025] support making causal inferences on text, with an emphasis on discovery. Modarressi et al. [2025] uses LLM reasoning as a tool for dimensionality reduction on the space of concepts present in text (i.e., as a low-dimensional hypothesis generation methodology) and

[2]This pricing description is valid as of October 2025.

incorporates human validation of LLM-selected hypotheses, with estimation implemented via sample splitting; in contrast, the framework proposed in this paper leverages recent advances in interpretability methods for LLMs in order to directly analyze the high-dimensional space of concepts present in text, and uses state-of-the-art autointerpretation methods to describe discovered concepts without further researcher intervention, and without requiring data splitting. As such, these frameworks offer different profiles of scientific benefits and researcher costs for the problem of causal inference on text.

Modarressi et al. [2025] also directly leverages the work of Ludwig et al. [2017] in their pipeline, which is similarly relevant to the proposed framework, in that both emphasize discovery from high-dimensional data, and can likewise both be applied to the problem of causal inference on unstructured data (as it is used in Modarressi et al. [2025]). However, the proposed framework tackles the high-dimensional discovery problem using a novel high-dimensional selective inference procedure that maintains a high degree of interpretability, whereas Ludwig et al. [2017] modifies the target of discovery to improve power, trading off interpretability to do so.

High-dimensional selective inference. There is an extensive literature on multiple hypothesis testing (or "selective inference" more broadly) in statistics, biostatistics, and economics (see, e.g., Romano et al. [2010] for a review in econometrics). The methods in this literature span asymptotic and finite sample valid frameworks, and low- and high-dimensional settings. In particular, for the problem considered in this paper, the literature on asymptotically valid, high-dimensional selective inference is most relevant, as such methods permit making very few assumptions about the data (c.f., assuming independence or PRDS of the p-values in the case of Benjamini and Hochberg [1995]) and readily incorporate resampling methods that improve power (relative to protecting against worst-case dependence, c.f., Bonferroni corrections or the method of Benjamini and Yekutieli [2001]). Specifically, the statistical procedure for high-dimensional exceedance control proposed in this paper most directly complements existing methods for high-dimensional control of the FWER [Belloni et al., 2018] and the FDR under weak dependence [Liu and Shao, 2014, Belloni et al., 2018]. This new procedure is facilitated by recent work on high-dimensional central limit theory for order statistics [Ding et al., 2025]. In fact, the proposed statistical procedure is the first that has been stated of its kind to enable (asymptotic) high-dimensional control of k-FWER exceedance in a way that adapts to arbitrary dependence between p-values or test statistics, without requiring any joint invariance under group transformations assumptions (c.f., Hemerik et al. [2019]).

Hypothesis generation. There is a growing literature in economics and computer science on "hypothesis generation" at the pre-scientific stage of empirical inquiry [Ludwig and Mullainathan, 2024]. From the perspective of this literature, the proposed framework may be viewed as leveraging dictionary learning as a high-dimensional hypothesis generator, on top of which discovery is performed at the "scientific" stage of inquiry using high-dimensional multiple hypothesis testing procedures. As such, the framework in this paper contributes to the literature on hypothesis generation by providing a statistically principled pathway from generation to evaluation—one which involves little to no researcher discretion about

what generated hypotheses should be tested. The recent framework of Movva et al. [2025] uses dictionary learned features (specifically, sparse autoencoder features) explicitly for the purposes of hypothesis generation, further motivating the use of dictionary learning methods as tools for hypothesis generation in the social sciences, both in the present framework and beyond.

Interpretability for LLMs and other DNNs. There is a large literature in computer science on machine learning interpretability (see, e.g., Doshi-Velez and Kim [2017] for a foundational agenda). The machine learning interpertability methods leveraged by this framework originate from a nascent though highly active literature known as "mechanistic interpretability," which seeks to develop methods for interpreting LLM behavior via quantitative analyses of model internals (e.g., activations, weights). In particular, the dictionary learning methods for DNN interpretability implemented in this framework were developed in Bricken et al. [2023], Templeton et al. [2024], i.e., sparse autoencoders (SAEs). SAEs are autoencoders attached to the residual streams (or other model internals) of a LLM, with hidden layers being many orders larger in dimension than the residual stream, and which are trained with sparsity-inducing penalization on reconstruction loss. Under the "linear representation hypothesis" and "superposition hypothesis" (see, e.g., Bricken et al. [2023] for more discussion), these sparse autoencoders are thought to act as an overcomplete basis of the space of concepts in text leveraged by a LLM to make next token predictions, encouraging learned features to be monosemantic. There has been much debate about the success of SAEs (e.g., Leask et al. [2025]), though most criticisms still support the notion that SAEs and other dictionary learning methods have a comparative advantage in discovering as opposed to detecting concepts of interest in text [Peng et al., 2025]. Other dictionary learning methods that have been proposed as competitors to SAEs (e.g., transcoders [Paulo et al., 2025]) are similarly compatible with the framework proposed in this paper.

SAEs and other dictionary learning methods generate features, but not natural language feature descriptions. Based on the popularity of these methods, a related literature on "autointerpretability" methods has become active, which seeks to coherently use LLMs to describe the features discovered by dictionary learning methods at scale. Recent important papers in this literature include: Bills et al. [2023], Shaham et al. [2024], Paulo et al. [2024], Rajamanoharan et al. [2024]. The framework proposed in this paper primarily uses insights from Paulo et al. [2024] to create "local interpretations" that are most relevant to the data generating process being considered, and otherwise leverages autointerpetations from Lieberum et al. [2024].

Though the empirical examples considered in this paper handle one of the most popular unstructured date types—text—for which dictionary learning-based interpretability methods are the most well-developed, dictionary learning techniques have been successfully applied to many other modalities, including audio and images [Abdulaal et al., 2024, Bhalla et al., 2024, Fry, 2024, Daujotas, 2024, Pluth et al., 2025].

3 Framework

3.1 Setup

Let $[n] := \{1, \ldots, n\}$, and let $x \mapsto \log x$ be the natural logarithm. The researcher has access to a dataset of size n, $\{(W_i, Z_i)\}_{i=1}^n$, which is sampled i.i.d. from some (super-)population of interest P. The $Z_i \in \mathcal{Z}$ are unstructured data instances and the $W_i \in \mathcal{W}$ are any other observed covariates of interest (e.g., for one empirical example considered in the following section, $W_i = D_i \in \mathcal{W} = \{0, 1\}$, a binary treatment status indicator). The space $\mathcal{Z}$ is typically high-dimensional and semantically poor, e.g., if each Z_i was a 244×244 pixel image, $\mathcal{Z}$ might then be the space of all $3 \times 244 \times 244$ arrays of RGB values, or if Z_i was text that was truncated up to some maximum length, $\mathcal{Z}$ might be the space of all binary matrices of a certain dimensionality, which are concatenated one-hot encodings for each word, term, or token in the text with respect to a specific pre-defined vocabulary.

The researcher has access to the function $\texttt{Dict} : \mathcal{Z} \to \{0, 1\}^p$, which maps an unstructured data observation to a a sparse binary vector. This function is computed by passing an unstructured data point to a DNN equipped with a dictionary learning model (e.g., a LLM equipped with pretrained SAE when Z_i are texts), recording the feature activations from the dictionary model, and then aggregating and binarizing these dictionary feature activations to form a single indicator for each feature in the dictionary $j \in [p]$. There is a large possible space of `Dict` functions the researcher could implement, though an especially straightforward choice for the use case of text data is implementing a `Dict` function based on the SAE activations at a single layer of the DNN, for which the j-th entry of the output is equal to 1 if the j-th dictionary feature activated on any token of the input text, and 0 otherwise:

$$\texttt{Dict}(z)_j := \texttt{Dict}(z)_j^l := \begin{cases} 1 \text{ if SAE feature } j \text{ at layer } l \text{ activates on any token of } z, \\ 0 \text{ otherwise.} \end{cases}$$

We denote $Y_i := \texttt{Dict}(Z_i) \in \{0, 1\}^p$ as the output of the dictionary transformation of the unstructured data. As such, each $Y_{ij} \in \{0, 1\}$ has the natural interpretation that a particular dictionary feature j activated for the unstructured data instance i, and, e.g., $\bar{Y}_{\bullet,j} := n^{-1} \sum_{i=1}^n Y_{ij}$ corresponds to an estimate of the probability of the presence of a feature j in the population P of interest.[3]

The crux of the proposed framework is that inference on functionals of Y_{ij} is desirable because:

1. Each Y_{ij} is monosemantic, and therefore functionals of Y_{ij} for a given j are *interpretable*, e.g., $E_P[Y_{ij}]$ is the probability of monosemantic feature j appearing in an unstructured data observation in population P.

[3]Dictionary feature activations are typically positively valued scalars, where magnitude is thought to correspond to some notion of intensity of activation; other works in mechanistic interpretability consider using these raw activation values directly, perhaps with max or average pooling across tokens. However, presently, magnitude of activation is not a well-understood or highly interpretable quantitative property of dictionary learning methods, and as such the default `Dict` function advocated for in the proposed framework does not incorporate this information. That said, researchers interested in extending this framework may find value in alternative `Dict` functions that leverage activation intensity.

2. Collectively, the p features in Y_{ij} for any given i are exhaustive of some vast space of concepts of interest to the researcher; intuitively, they are the set of all monosemantic features that a LLM needed to learn to perform well on next token prediction for a massive, internet-scale corpus of text (or other similarly massive unstructured data source). This means the researcher need not pre-specify any particular concepts of interest, reducing *researcher degrees of freedom* without limiting discovery.

However, inference on functionals of Y_{ij} for each j is also challenging, because:

1. The p is large, so inference on all $j \in [p]$ functionals is inherently a *multiple hypothesis testing* problem.

2. Not only is p large, but for most social science applications of interest $p \gg n$, so inference on Y_{ij} for each j is a *high-dimensional* multiple hypothesis testing problem.

3. Each estimator, test statistic, or p-value formed from the Y_{ij} for each $j \in [p]$ for the purposes of inference is plausibly statistically *dependent* on every other in a complicated way, ruling out multiple hypothesis testing approaches that assume independence or specific forms of dependence (e.g., PRDS).

4. Each Y_i is sparse, and the intent of the analysis is discovery, so the desired form of selective (familywise) error control *cannot be too conservative*.

What is needed, then, for principled inference on functionals of Y_{ij} is a high-dimensional selective inference procedure with control over a generalized error rate. To be as well-powered as possible, we want to focus on testing procedures that employ resampling methods for estimating the true covariance across test statistics, such that conservative protection against worst-case dependence is not required. To keep things as general as possible, we also want to allow for only asymptotically valid test statistics and p-values, as is common to much of econometric analysis. In the following sections, theory and corresponding statistical procedures are developed to achieve exactly these aims, and are stated as generally as possible to accommodate settings of interest even beyond analysis of unstructured data.

To pursue inference, consider defining $p \gg n$ hypotheses $\{H_{0,j}\}_{j\in[p]}$, where

$$H_{0,j} : \theta_j(P) := E_P[X_{ij}] \leq 0, \quad X_{ij} = h(W_i, Y_{ij}),$$

for some measurable function h.[4] Two choices of h considered in the empirical examples in this paper include $h(W_i, Y_{ij}) = Y_{ij}$ (testing probabilities of dictionary feature activation) and $h(W_i, Y_{ij}) = \frac{D_i - \pi}{\pi(1-\pi)} Y_{ij}$ (testing differences in probabilities of activation across groups under known, fixed group assignment), though many others are possible, lending to the flexibility of the framework. As shorthand, we will denote $X := \{X_i\}_{i=1}^n$.

We may form test statistics for such hypotheses as

$$T_{n,j} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{ij}, \quad T_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i.$$

[4]Unless otherwise specified, we assume that all $\theta_j(P)$ may be treated as fixed, i.e., are not growing or shrinking in n, p in the asymptotic thought experiment. This is a sufficient condition, however; $\theta_j(P)$ are still allowed to approach zero so long as they do so at a rate slower than $B_n\sqrt{\log p/n}$ (see the proof of Theorem 1 for more insight).

All results to be discussed will hold for nulls of equality as well, with appropriate test statistics $|T_n|$. For the purposes of developing high-dimensional central limit theory later on, further define

$$S_{n,j} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_{ij} - E[X_{ij}]), \quad S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - E[X_i]).$$

We will denote $\Sigma := E\left[S_n S_n^{\intercal}\right] = n^{-1} \sum_{i=1}^{n} E[(X_i - E[X_i])(X_i - E[X_i])^{\intercal}]$, which simplifies to $\Sigma = E[(X_i - E[X_i])(X_i - E[X_i])^{\intercal}]$ under i.i.d. data. Further, for any subset of indices $K \subseteq [p]$, we will denote $S_{n,K} := (S_{n,j} : j \in K)^{\intercal}$.

3.2 High-Dimensional k-FWER Exceedance Control

We construct a method that provides high-dimensional k-FWER exceedance control asymptotically, or control of the probability of making k or more rejections of true nulls in the large sample limit when p is growing (much) faster than n in the relevant asymptotic thought experiment. Naturally, k-FWER control for $k = 1$ is control of the FWER, and larger choices of k permit more discoveries.

To proceed, we will adapt strategies from Romano and Wolf [2007] for achieving large sample k-FWER control to the high-dimensional setting. The testing procedures of Romano and Wolf [2007] rely on the asymptotic validity of bootstrap approximations of the distribution of the k-th largest coordinate (or "k-max") of S_n. Thus, a bootstrap procedure for the k-max statistic that is asymptotically valid in high-dimensions would allow one to run the algorithms of Romano and Wolf [2007] in the high-dimensional setting with their stated control under minimal modifications.

High-dimensional bootstraps must be supported by appropriate high-dimensional central limit theory. Fortunately, the recent work of Ding et al. [2025] provides high-dimensional central limit theory for the k-th largest coordinate of a scaled sum of centered independent random vectors so long as k is very small, i.e., k is fixed in the asymptotic thought experiment. We will build on these "small k" results. Assume $n \geq 3$ and $p \geq 3$. Let $b_1 > 0$ and $b_2 > 0$ be some constants such that $b_1 \leq b_2$, and let $B_n > 1$ be a sequence of constants, where it is possible that B_n diverges. Then we make the following assumptions.

Assumption M. For all $i \in [n], j \in [p]$, assume: (i) $E\left[\exp\left(|X_{ij}| / B_n\right)\right] \leq 2$; (ii) $b_1^2 \leq \frac{1}{n} \sum_{i=1}^{n} E[X_{ij}^2]$; and (iii) $\frac{1}{n} \sum_{i=1}^{n} E[X_{ij}^4] \leq B_n^2 b_2^2$.

These are mild conditions on the tails and moments of the data X_{ij}, which are stated in a way that accommodates data that are independent but not identically distributed (i.n.i.d.). Part (i) of Assumption M simply requires that X_{ij} be sub-exponential, or, equivalently, have an Orlicz norm in Ψ_1 bounded by $B_n < \infty$. The B_n is indexed by n to accommodate growing tail thickness within the sub-exponential regime as $p = p_n$ grows in the asymptotic thought experiment with i.n.i.d. data. Part (ii) of Assumption M insists that the second moments of the data be bounded away from zero, appropriately stated for i.n.i.d. data. Part (iii) insists on bounded fourth moments in a similar fashion. As such, for i.i.d. data, the case considered in this paper, we may simply require for some fixed $B < \infty$ and fixed b_1, b_2 with $b_1 \leq b_2$: (i) $E\left[\exp\left(|X_{ij}| / B\right)\right] \leq 2$; (ii) $b_1^2 \leq E[X_{ij}^2]$; and (iii) $E[X_{ij}^4] \leq B^2 b_2^2$. Note that, in many use

cases of interest for dictionary learned features, the X_{ij} are both bounded and studentized, aiding the plausibility of these assumptions.

In practice, some fraction of dictionary features may be "dead" for any given dataset, meaning they are degenerate at zero [Bricken et al., 2023]. Of course, these features are not of interest for the purposes of discovery. This phenomenon is tolerated by the framework so long as the count of degenerate features is smaller than any relevant $p - k$, in which case, in any finite sample, the k-th largest coordinate of all relevant statistics is invariant to whether or not the degenerate coordinates are dropped. As such, conceptually, one should think of j as ranging over the p non-degenerate features. In practice, to perform studentization, it is recommended that these degenerate coordinates are explicitly filtered out; studentization is discussed further in Section 3.3.

Assumption R. Assume that $B_n^2 \log^5(pn) = o(n)$.

This is the key rate condition needed for (sup-norm) Gaussian or bootstrap approximation error to go to zero asymptotically under the high-dimensional CLTs discussed in both Ding et al. [2025] and Chernozhuokov et al. [2022]. Rewritten, it says that

$$\frac{B_n^2 \log^5(pn)}{n} = o(1)$$

which permits p growing very fast with n in the asymptotic thought experiment. In fact, p may be growing nearly exponentially in n, e.g., $p = e^{n^{1/6}}$ for fixed B_n. That this rate condition permits $p \gg n$ is an important positive result in high-dimensional central limit theory [Chernozhukov et al., 2017, 2023].

We will focus on the Gaussian multiplier bootstrap as our high-dimensional bootstrap method. We define the Gaussian multiplier bootstrap quantity

$$S_n^B := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \left(X_i - \bar{X}_n \right),$$

where $\xi_i \overset{iid}{\sim} N(0,1)$ and $\bar{X}_n := n^{-1} \sum_{i=1}^n X_i$. We also introduce $x \mapsto x_{[k]}$ as notation for the function that selects the k-th largest coordinate of a vector x, i.e., $x_{[k]} = k\text{-max}(x)$ in the notation of Romano and Wolf [2007]. For the purposes of bootstrapping, we will denote the data-conditional probability measure $P^B(\cdot) := P(\cdot \mid X)$.

With these assumptions and notations in place, we now introduce a statistical procedure for controlling the k-FWER for small k, based on the step-wise algorithms of Romano and Wolf [2007].

Theorem 1 (High-dimensional k-FWER exceedance control for small k)**.** *Consider the method of Algorithm 2.1 or 2.2 in Romano and Wolf [2007] with test statistics T_n of hypotheses $\{H_{0,j}\}_{j \in [p]}$ and critical values $\hat{c}_{n,K}(1-\alpha, k)$ given by the $1-\alpha$ quantile of $S^B_{n,K,[k]}$ under P^B. Assume that k is fixed (i.e., not growing with n, p). Then under Assumptions* M *and* R*:*

(i) $\limsup_{n,p \to \infty} k\text{-}FWER_P \leq \alpha$.

(ii) If $H_{0,j}$ is false and $\theta_j(P) > 0$, then the probability that the step-down method rejects $H_{0,j}$ tends to 1.

Algorithm 2.1 or 2.2 of Romano and Wolf [2007] also apply for two-sided tests, using a completely analogous proof, and all high-dimensional central limit theory in Ding et al. [2025] also applies to the k-th largest coordinate of the absolute value of scaled sums of centered independent random vectors,[5] meaning that Theorem 1 holds for two-sided tests based on $|T_n|$ as well, using the quantiles of $|S^B_{n,K}|_{[k]}$ under P^B.

Importantly, the procedure of Theorem 1 is only valid for small k. To see this explicitly, note that the sup-norm bootstrap or Gaussian approximation error in the high-dimensional CLTs for the k-th largest coordinate introduced in Ding et al. [2025] only goes to zero if

$$\frac{k^8 B_n^2 \log^5(pn)}{n} = o(1),$$

meaning that k must be fixed or must grow incredibly slowly: for all practical purposes, k needs to be quite small. As such, the theory of Ding et al. [2025] is unsuitable for making progress on FDP exceedance control, which requires a k that may grow linearly with p. An important goal of future research would be to establish valid high-dimensional FDP exceedance control using appropriate high-dimensional central limit theory.

3.3 High-Dimensional CLT for k-max of Approximate Means

In practice, hypothesis testing with studentized statistics will be important for powering discoveries beyond just frequently occurring dictionary features (which are high variance features for Bernoulli random variables when $X_{ij} = Y_{ij} \in \{0, 1\}$). To facilitate studentization, we will need a high-dimensional CLT that permits statistics with small estimation errors. We therefore generalize the high-dimensional CLT of Ding et al. [2025] to handle "approximate means" in the parlance of Belloni et al. [2018], stated in the following lemmas.

Lemma 1 (High-dimensional CLT for the small k-max coordinate of approximate means)**.** *Let $\hat{S}_n := S_n + R_n$, and assume that $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$. Further assume that k is fixed (i.e., does not grow with n, p). If Assumptions* M *and* R *hold, then as $n, p \to \infty$*

$$\sup_{t\in\mathbb{R}} \left| P\left(\hat{S}_{n,[k]} \le t\right) - P\left(N(0,\Sigma)_{[k]} \le t\right)\right| \to 0.$$

Lemma 2 (High-dimensional bootstrap for the small k-max coordinate of approximate means)**.** *Let $\hat{S}^B_n := S^B_n + R_n$, and assume that $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$. Further assume that k is fixed (i.e., does not grow with n, p). If Assumptions* M *and* R *hold, then as $n, p \to \infty$*

$$\sup_{t\in\mathbb{R}} \left| P^B\left(\hat{S}^B_{n,[k]} \le t\right) - P\left(N(0,\Sigma)_{[k]} \le t\right)\right| \xrightarrow{P} 0.$$

Using these lemmas, we may now state two corollaries, which allow for studentization. First, define $\Lambda := \mathrm{diag}(\Sigma)$, as well as $\Sigma_0 := \Lambda^{-1/2}\Sigma\Lambda^{-1/2}$. Let $\widehat{\Sigma}_{jj} := n^{-1}\sum_{i=1}^n \left(X_{ij} - \bar{X}_{n,j}\right)^2$ and let $\widehat{\Lambda} := \mathrm{diag}\left\{\widehat{\Sigma}_{11}, \ldots, \widehat{\Sigma}_{pp}\right\}$ be an estimator of the asymptotic variances Λ.

[5] See Remark 2 of Ding et al. [2025].

Corollary 1 (High-dimensional CLT for the small k-max studentized coordinate)**.** *If Assumptions* M *and* R *hold, and* $B_n = O(1)$ *and* k *is fixed (i.e., does not grow with* n, p*), then, by application of Lemma 1, as* $n, p \to \infty$

$$\sup_{t \in \mathbb{R}} \left| P\left(\left(\hat{\Lambda}^{-1/2} S_n \right)_{[k]} \leq t \right) - P\left(N(0, \Sigma_0)_{[k]} \leq t \right) \right| \to 0.$$

Corollary 2 (High-dimensional bootstrap for the small k-max studentized coordinate)**.** *If Assumptions* M *and* R *hold, and* $B_n = O(1)$ *and* k *is fixed (i.e., does not grow with* n, p*), then, by application of Lemma 2, as* $n, p \to \infty$

$$\sup_{t \in \mathbb{R}} \left| P^B\left(\left(\hat{\Lambda}^{-1/2} S_n^B \right)_{[k]} \leq t \right) - P\left(N(0, \Sigma_0)_{[k]} \leq t \right) \right| \xrightarrow{P} 0.$$

With these corollaries, it is immediate that the bootstrap quantiles of studentized statistics well approximate the studentized version of S_n. Also note that, in the setting where the data is not just i.n.i.d. but i.i.d., as is true for the use case of interest based on dictionary features X_i, $B_n = O(1)$ is assured.

4 Empirical Applications

As an illustration of the framework proposed in this paper, we now reanalyze two recent works in empirical economics that pursue discovery from unstructured data, and show how new, principled, and interpretable discoveries may be made using the same exact data at low cost.

4.1 Bursztyn et al. [2023]

Bursztyn et al. [2023] study how the provision of "social cover" affects willingness to publicly dissent to socially stigmatized causes, and the perception of this dissent. As a key application of their theory, Bursztyn et al. [2023] run an information treatment experiment (Experiment 2) in which participants are told they have been matched with another (fictional) respondent that chose to join a campaign to defund the police (a plausible expression of dissent in liberal American politics at the time the experiment was conducted), and then show the participant a tweet that the matched respondent is said to have agreed to publicly post. This tweet has been randomized to either include social cover or no social cover, namely whether or not the tweet indicated that the matched respondent joined the campaign after reading an evidence-based article in support of it (the no social cover condition) or before (the social cover condition).[6]

An important outcome collected by this experiment is the participant's open-ended text response to the question "Why do you think your matched respondent chose to join the campaign to oppose defunding the police?" This open-ended text response is meant to capture the causal effect of social cover on the perception of dissent. As such, this is a

[6]See Online Appendix Figure B.3 of Bursztyn et al. [2023] for a schematic of the experimental design.

setting where discovery is of interest; ideally, to form a holistic understanding of how social cover affects perception of dissent, we do not want to pre-specify what aspects of perception we are interested in, and want to discover any interpretable systematic differences that exist in the text responses across treatment and control groups. In order to make progress on this analysis of unstructured data, Bursztyn et al. [2023] compute a Pearson's χ^2 statistic for all phrases of up to three words per Gentzkow and Shapiro [2010], which they use as an index to rank the phrases that are most differentially expressed in each condition's open-ended responses.[7] The interpretability of the results from this analysis, is, naturally, hindered by the coarseness of the featurization of the text as three word phrases, as well as the fact that there are no obvious estimands or inferential guarantees. The only qualitative conclusion gleaned by Bursztyn et al. [2023] from this quantitative exercise is: "we find that respondents in the Cover condition are more likely to use phrases related to the article or the associated evidence—for example, 'article,' 'read,' 'convincing,' or 'increase in crime.'"[8]

Can we make more (and more interpretable) discoveries with the framework of this paper? To investigate, we implement a `Dict` function based on the Gemma Scope autoencoders for Google's Gemma 2 2B [Lieberum et al., 2024], and specifically the SAE trained on the residual stream of layer 12.[9] We pull the Gemma Scope project's autointerpretation descriptions of these features from Neuronpedia,[10] though we also implement a custom, "local" autointerpretation pipeline based on the best practices of Paulo et al. [2024], discussed further in Appendix Section 6.1. These are mainstream choices for a SAE and LLM pairing in machine learning interpretability research.

This `Dict` function yields a dataset $\{X_{ij}\}_{i\in[n],j\in[p]}$, where $n = 1033$ and $p = 12{,}005$ (non-degenerate features). The hypotheses of interest are, for all $j \in [p]$,

$$H_{0,j} : E[Y_{ij}(1) - Y_{ij}(0)] = E[H_i Y_{ij}] := E[X_{ij}] = 0, \quad H_i = \frac{D_i - \pi}{\pi(1-\pi)}.$$

for $\pi = 0.5$, and where $\{D_i = 1\}$ indicates the social cover condition. As a baseline, we will implement a simple one-step version of the statistical procedure in Theorem 1, which will be a lower bound on the power of the step-wise procedure.

We may first consider whether or not, as anticipated, simply controlling the FWER would be too conservative for the purposes of discovery in this setting. To explore this, we set $k = 1$ and $\alpha = 0.05$ in the procedure based on Theorem 1 (equivalently, we implement the high-dimensional FWER controlling procedure of Belloni et al. [2018]), and we obtain only two significant discoveries: a 13.2 percentage point causal effect on the presence of dictionary feature 3518 (DF-3518) and a 13.4 percentage point causal effect on the presence of dictionary feature 3426 (DF-3426). The Gemma Scope autointerprations of these features

[7] Specifically, it is a Pearson's χ^2 statistic for a null hypothesis that the propensity to use a given phrase is equal across conditions, per Gentzkow and Shapiro [2010]. We interpret these statistics as simply forming an index, however, because the results of these hypothesis tests are not reported, and no multiple hypothesis testing corrections are implemented.

[8] The Online Appendix Table B.11 of Bursztyn et al. [2023] contains the top ten characteristic phrases in each condition based on the χ^2 index.

[9] It is recommended to work with SAEs trained on the middle layer of a LLM to capture coarser concepts and features, which is plausibly more suitable for analyzing broad themes in text.

[10] See `https://www.neuronpedia.org/gemma-scope` for more.

paint a picture that corroborates the original findings of Bursztyn et al. [2023]: DF-3426 is described as activating on "references to articles" and DF-3426 on "references to news articles and reports." However, we only discover two features out of over 12,000, and not much new insight relative to the original analysis—even if the interpretation of the features discovered was transparent and automatic, by construction.

In Table 1, we make discoveries based on the procedure of Theorem 1 now setting $k = 5$, still with $\alpha = 0.05$, and this time using local autointerpretation descriptions per the method Appendix Section 6.1.

Table 1: Discoveries for Experiment 2 of Bursztyn et al. (2023), 5-FWER Control at $\alpha = 5\%$

Feature	$\widehat{\text{ATE}}$	Description ("Feature activates highly on...")	t-stat
306	0.099	article or report claiming that defunding the police will increase violent crime	4.101
1392	0.118	mentions of article content or claims	3.908
1622	0.072	mentions of an authoritative source or author attribution	4.043
1992	0.095	article	4.284
3298	0.126	article about defunding the police and its impact on violent crime	4.221
3426	0.134	mentions of articles or sources discussing defunding the police and its impact on crime	4.801
3518	0.132	news article linking police defunding to increased violent crime	5.017
4320	0.081	article or article-related content	3.997
8486	0.085	policing and crime discourse	3.916
8902	0.079	reading an article that influences opinion	4.196
12287	-0.130	modal verbs expressing uncertainty or possibility	3.888

Note: Descriptions generated using the method described in Appendix Section 6.1, using GPT-OSS 20B [OpenAI, 2025]. For bootstrapping, 10,000 iterations are performed.

As can be seen, even with control of k-FWER at $k = 5$, we obtain 11 rejections of the null, or 11 discoveries. By guarantee of Theorem 1, the probability that 5 or more of these 11 discoveries is false is less than 5%. The previous features discovered with the FWER controlling procedure appear, as well as others that represent concepts both absent from the FWER controlled analysis and the original analysis in Bursztyn et al. [2023]. Though there are features that appear to activate on redundant concepts in the text (e.g., DF-306, DF-1392, DF-1992), there are others that contribute to entirely new characterizations of the causal effect of social cover on dissent. DF-1622 responds to "mentions of an authoritative source or author attribution," and increases by 7 percentage points in the treatment group, indicating that, as Bursztyn et al. [2023] posit, the social cover mechanism plausibly relies on the perceived credibility of the article cited. DF-12287 responds to "modal verbs expressing uncertainty or possibility," and its probability of activation is decreased by 13 percentage points in the treatment group. This supports the idea that the causal effect of not having social cover is, to first order, generating uncertainty and speculation about dissent, as opposed to an alternative where participants coalesce around a single narrative in the absence of cover. This insight is also congruous with the observation that few large negatively signed causal effects are discovered.

Importantly, all of this analysis was performed automatically, in a handful of minutes on a single A100 GPU in Google Colab. Though this analysis was not preregistered, it is possible to cheaply and automatically replicate these results, and assess their sensitivity to

different autointerpretation strategies, dictionary learning methods, and LLMs. Notably, the space of possible concepts was not pruned in any way prior to this analysis, nor were they weighted in some way towards topics related to articles; the analysis was automatic and without human intervention, with no room for motivated data mining.

4.2 Stantcheva [2024]

Stantcheva [2024] conducts surveys on representative samples of the United States population investigating attitudes towards inflation. These surveys include open-ended text responses to a variety of questions, with the goal of discovering attitudes and opinions that the researcher should not (for concern about priming respondents) or could not (for lack of imagination) pre-specify.[11]

An important open-ended prompt that Stantcheva [2024] solicit an answer to is "High inflation is caused by..." In the original analysis, Stantcheva [2024] code open-ended responses with a keyword-based method described in Ferrario and Stantcheva [2022], in which the researcher creates a list of topics and associated keywords in a discretionary way, ranging from "manual to semi-supervised or unsupervised" means. The result of applying this analysis to the open-ended responses to the above prompt yields Figure 3 in Stantcheva [2024], which finds, among other things, that mentions of "Biden and the administration," "Greed," "Monetary policy," "Fiscal policy," "War and foreign policy," "Demand vs supply," and "Supply-side mechanisms (other than input prices)" appear in more than 5% of all responses.

Do we make the same discoveries automatically if we apply the proposed framework instead? Do we make more and new interpretable discoveries? To find out, we use the same models as in Section 4.1, but reduce the space of features by half, filtering out all features that, in the corpus on which the Gemma Scope SAEs were trained, had greater than median empirical token activation frequency. This choice of dimensionality reduction is meant to screen out features that activate in many texts across domains, e.g., features related to grammatical aspects of text, which was handled implicitly in the previous causal analysis by virtue of differencing. (Ideally, such a dimensionality reduction choice would be preregistered, to prevent cherry-picking by filtering.) This yields a dataset $\{X_{ij}\}$ with $n = 503$ and $p = 3915$ non-degenerate features—still a very high-dimensional inference problem.

We test the $p = 3915$ hypotheses $H_{0,j} : E[X_{ij}] = E[Y_{ij}] = 0$, i.e., we test the probability of a given feature activating in the population of open-ended responses to the above prompt in the United States. Specifically, by again using 5-FWER control with $\alpha = 0.05$ based on the procedure of Theorem 1, we make 818 interpretable discoveries. In order to better present these discoveries, using test inversion we produce simultaneous "generalized" confidence sets [Romano and Wolf, 2007], which asymptotically guarantee that the probability that k or more estimands do not lie in the set is less than α.

In Table 2, we report the simultaneous confidence intervals for all discoveries with a lower confidence bound above 20%. Notably, we discover, automatically, with no manual pruning

[11] See, e.g., Haaland et al. [2024] for more discussion of the benefits of open-ended survey questions for understanding economic behavior.

or interpretation required, many sensible commonplace themes related to the topic at hand: DF-9804 responds to economic terms, DF-14304 responds to financial terminology, and DF-13447 responds to mentions of inflation. However, we also recover many of the topics from the original analysis without any human discretion or intervention: DF-4192 responds to discussion of monetary policy, DF-5719 responds to discussion of fiscal policy, and DF-14747 responds to discussion of supply-side issues. Undoubtedly, many other topics are recovered as we look deeper down the list of discoveries, beyond the top 11 out of 818. However, even in this top 11, we learn that there is a great deal of pure uncertainty expressed in these open-ended responses based on the activation of DF-4794; that macroeconomic indicators get frequently discussed by virtue of DF-11036; and that money supply and government spending and budgeting are often mentioned, per DF-104 and DF-8316, respectively.

Table 2: Largest Discoveries for "High inflation is caused by..." in Stantcheva (2024)

Feature	Description ("Feature activates highly on...")	Sim. CI, Lower	Sim. CI, Upper
104	excess money supply leading to inflation	0.208	0.353
4192	central bank monetary policy terms	0.217	0.363
4794	expressing uncertainty or personal opinion	0.210	0.355
5719	government spending and fiscal policy	0.843	0.943
8316	government spending and budgeting terms	0.279	0.433
9804	economic terms related to the economy	0.267	0.421
11036	economic indicators and macro-economic terms	0.212	0.357
13447	inflationary price rise	0.393	0.554
13574	negative economic sentiment	0.230	0.378
14304	Economic or financial terminology	0.300	0.456
14747	supply shortage or supply constraints	0.234	0.383

Note: Generalized confidence intervals based on k-FWER control at $k = 5$, $\alpha = 5\%$. Descriptions generated using the method described in Appendix Section 6.1, using GPT-OSS 20B [OpenAI, 2025]. For bootstrapping, 10,000 iterations are performed.

Once again, this procedure was cheap and fast to implement, and could be replicated and analyzed for sensitivity by any other researcher, quickly. Beyond filtering out greater than median activating dictionary features, no other choices were made to delimit the space of possible discoveries, or focus them in some way on the space of economic- or inflation-relevant concepts; the exact same models that yielded the results of Section 4.1 yielded the results in this section.

5 Conclusion

Existing literature in empirical economics and econometrics has long suggested the importance of open-ended discovery from high-dimensional or unstructured data. The framework proposed in this paper shows how new statistical procedures for high-dimensional multiple hypothesis testing, when combined with the latest innovations in interpretability methods for machine learning models, can facilitate open-ended, interpretable discovery at scale with higher practicality and higher fidelity than previously possible.

Importantly, the automaticity of the proposed framework makes it subject to very few researcher degrees of freedom, making it resilient to cherry-picking and motivated data-mining without compromising the purpose of discovery in the first place. This is especially true if one couples this framework with minimal preregistration efforts, e.g., simply publicly declaring what dictionary learning model, LLM, and automatic interpretation method are to be used for the analysis of a text, and if any of the features are to be filtered on principled grounds.

A reasonable possible criticism of the proposed framework is that the interpretations associated with dictionary learning methods are not credible. On this topic exactly, there has been an immense amount of attention from AI/ML research communities, which has, to date, primarily concluded that such dictionary features are useful and reliable for discovery, even if they do not uncover "atomic" or fundamental conceptual units in deep neural networks like LLMs [Leask et al., 2025, Peng et al., 2025]. The best practices and methods for moving from features to feature descriptions is an open and important question in the literature, and choices made on this front certainly affect interpretations of discoveries made by the proposed framework. However, as the empirical examples of the previous section make clear, straightforward automatic interpretation methods—which are completely data-driven, replicable, and not subject to motivated human reasoning—yield sensible outputs, scale well with the size of selected sets, can be further scrutinized by humans to assess their validity, and can be perturbed or modified cheaply to assess sensitivity to autointerpretation prompts. Future research related to this framework would seek to apply quantitative methods for validating feature descriptions at scale as well [Paulo et al., 2024, Movva et al., 2025].

The proposed framework is most naturally viewed as one tool of many in the empirical researcher's toolkit for making discoveries from unstructured data. Using this framework alongside others that researchers may already be implementing is entirely complementary, and would only serve to deepen insights into possible interpretations of inference on unstructured data.

References

Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. The Voice of Monetary Policy. *American Economic Review*, 113(2):548–584, February 2023. ISSN 0002-8282. doi: 10.1257/aer.20220129. URL https://pubs.aeaweb.org/doi/10.1257/aer.20220129.

Allen Hu and Song Ma. Persuading Investors: A Video-Based Study. *The Journal of Finance*, 80(5):2639–2688, October 2025. ISSN 0022-1082, 1540-6261. doi: 10.1111/jofi.13471. URL https://onlinelibrary.wiley.com/doi/10.1111/jofi.13471.

Ingar Haaland, Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart. Understanding Economic Behavior Using Open-ended Survey Data. Technical Report w32421, National Bureau of Economic Research, Cambridge, MA, May 2024. URL http://www.nber.org/papers/w32421.pdf.

Jens Ludwig and Sendhil Mullainathan. Machine Learning as a Tool for Hypothesis Generation. *The Quarterly Journal of Economics*, 139(2):751–827, March 2024. ISSN 0033-5533, 1531-4650. doi: 10.1093/qje/qjad055. URL https://academic.oup.com/qje/article/139/2/751/7515309.

Clément Gorin, Stephan Heblich, and Yanos Zylberberg. State of the Art: Economic Development Through the Lens of Paintings. Technical Report w33976, National Bureau of Economic Research, Cambridge, MA, June 2025. URL http://www.nber.org/papers/w33976.pdf.

David Lagakos, Stelios Michalopoulos, and Hans-Joachim Voth. American Life Histories. Technical Report w33373, National Bureau of Economic Research, Cambridge, MA, January 2025. URL http://www.nber.org/papers/w33373.pdf.

Peter Bergman, Raj Chetty, Stefanie DeLuca, Nathaniel Hendren, Lawrence F. Katz, and Christopher Palmer. Creating Moves to Opportunity: Experimental Evidence on Barriers to Neighborhood Choice. *American Economic Review*, 114(5):1281–1337, May 2024. ISSN 0002-8282. doi: 10.1257/aer.20200407. URL https://pubs.aeaweb.org/doi/10.1257/aer.20200407.

Patrick Krause, Elizabeth Rhodes, Sarah Miller, Alexander Bartik, David Broockman, and Eva Vivalt. The Impact of Unconditional Cash Transfers on Parenting and Children. Technical Report w34040, National Bureau of Economic Research, Cambridge, MA, July 2025. URL http://www.nber.org/papers/w34040.pdf.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, November 2023. doi: 10.1126/science.adi6000. URL https://www.science.org/doi/10.1126/science.adi6000. Publisher: American Association for the Advancement of Science.

Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Large Language Models: An Applied Econometric Framework, December 2024. URL http://arxiv.org/abs/2412.07031. arXiv:2412.07031 [econ].

Jacob Carlson and Melissa Dell. A Unifying Framework for Robust and Efficient Inference with Unstructured Data, 2025. URL https://arxiv.org/abs/2505.00282. Version Number: 2.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Joseph P. Romano and Michael Wolf. Control of generalized error rates in multiple testing. *The Annals of Statistics*, 35(4), August 2007. ISSN 0090-5364. doi: 10.1214/009053606000001622. URL https://projecteuclid.org/journals/annals-of-statistics/volume-35/issue-4/Control-of-generalized-error-rates-in-multiple-testing/10.1214/009053606000001622.full.

Leonardo Bursztyn, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth. Justifying Dissent. *The Quarterly Journal of Economics*, 138(3):1403–1451, June 2023. ISSN 0033-5533, 1531-4650. doi: 10.1093/qje/qjad007. URL https://academic.oup.com/qje/article/138/3/1403/7000850.

Stefanie Stantcheva. Why Do We Dislike Inflation? Technical Report w32300, National Bureau of Economic Research, Cambridge, MA, April 2024. URL http://www.nber.org/papers/w32300.pdf.

Sendhil Mullainathan and Jann Spiess. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106, May 2017. ISSN 0895-3309. doi: 10.1257/jep.31.2.87. URL https://pubs.aeaweb.org/doi/10.1257/jep.31.2.87.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as Data. *Journal of Economic Literature*, 57(3):535–574, September 2019a. ISSN 0022-0515. doi: 10.1257/jel.20181020. URL https://pubs.aeaweb.org/doi/10.1257/jel.20181020.

Melissa Dell. Deep Learning for Economists. *Journal of Economic Literature*, 63(1):5–58, March 2025. ISSN 0022-0515, 2328-8175. doi: 10.1257/jel.20241733. URL https://pubs.aeaweb.org/doi/10.1257/jel.20241733.

Ashesh Rambachan, Rahul Singh, and Davide Viviano. Program Evaluation with Remotely Sensed Outcomes, 2024. URL https://arxiv.org/abs/2411.10959. Version Number: 2.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340, 2019b. ISSN 0012-9682. doi: 10.3982/ECTA16566. URL https://www.econometricsociety.org/doi/10.3982/ECTA16566.

Laura Battaglia, Timothy Christensen, Stephen Hansen, and Szymon Sacher. Inference for Regression with Variables Generated by AI or Machine Learning, 2024. URL https://arxiv.org/abs/2402.15585. Version Number: 5.

Iman Modarressi, Jann Spiess, and Amar Venugopal. Causal Inference on Outcomes Learned from Text, 2025. URL https://arxiv.org/abs/2503.00725. Version Number: 1.

Jens Ludwig, Sendhil Mullainathan, and Jann Spiess. Machine-Learning Tests for Effects on Multiple Outcomes, 2017. URL https://arxiv.org/abs/1707.01473. Version Number: 2.

Joseph P. Romano, Azeem M. Shaikh, and Michael Wolf. Hypothesis Testing in Econometrics. *Annual Review of Economics*, 2(1):75–104, September 2010. ISSN 1941-1383, 1941-1391. doi: 10.1146/annurev.economics.102308.124342. URL https://www.annualreviews.org/doi/10.1146/annurev.economics.102308.124342.

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 (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101. Publisher: [Royal Statistical Society, Oxford University Press].

Yoav Benjamini and Daniel Yekutieli. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2674075. Publisher: Institute of Mathematical Statistics.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato. High-Dimensional Econometrics and Regularized GMM, 2018. URL https://arxiv.org/abs/1806.01888. Version Number: 2.

Weidong Liu and Qi-Man Shao. Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics*, 42 (5), October 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1249. URL https://projecteuclid.org/journals/annals-of-statistics/volume-42/issue-5/Phase-transition-and-regularized-bootstrap-in-large-scale-t-tests/10.1214/14-AOS1249.full.

Yixi Ding, Qizhai Li, Yuke Shi, and Liuquan Sun. Gaussian Multiplier Bootstrap Procedure for the $$th Largest Coordinate of High-Dimensional Statistics, 2025. URL https://arxiv.org/abs/2508.14400. Version Number: 1.

J Hemerik, A Solari, and J J Goeman. Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, 106(3):635–649, September 2019. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asz021. URL https://academic.oup.com/biomet/article/106/3/635/5527339.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse Autoencoders for Hypothesis Generation, 2025. URL https://arxiv.org/abs/2502.04382. Version Number: 3.

Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning, 2017. URL https://arxiv.org/abs/1702.08608. Version Number: 2.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse Autoencoders Do Not Find Canonical Units of Analysis, February 2025. URL http://arxiv.org/abs/2502.04878. arXiv:2502.04878 [cs].

Kenny Peng, Rajiv Movva, Jon Kleinberg, Emma Pierson, and Nikhil Garg. Use Sparse Autoencoders to Discover Unknown Concepts, Not to Act on Known Concepts, 2025. URL https://arxiv.org/abs/2506.23845. Version Number: 1.

Gonçalo Paulo, Stepan Shabalin, and Nora Belrose. Transcoders Beat Sparse Autoencoders for Interpretability, 2025. URL https://arxiv.org/abs/2501.18823. Version Number: 2.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A Multimodal Automated Interpretability Agent, 2024. URL https://arxiv.org/abs/2404.14394. Version Number: 2.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically Interpreting Millions of Features in Large Language Models, 2024. URL https://arxiv.org/abs/2410.13928. Version Number: 3.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders, 2024. URL https://arxiv.org/abs/2407.14435. Version Number: 3.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2, 2024. URL https://arxiv.org/abs/2408.05147. Version Number: 2.

Ahmed Abdulaal, Hugo Fry, Nina Montaña-Brown, Ayodeji Ijishakin, Jack Gao, Stephanie Hyland, Daniel C. Alexander, and Daniel C. Castro. An X-Ray Is Worth 15 Features: Sparse Autoencoders for Interpretable Radiology Report Generation, 2024. URL https://arxiv.org/abs/2410.03334. Version Number: 1.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE), 2024. URL https://arxiv.org/abs/2402.10376. Version Number: 2.

Hugo Fry. Towards multimodal interpretability: Learning sparse interpretable features in vision transformers, Apr 2024. URL https://www.lesswrong.com/posts/iYFuZo9BMvr6GgMs5/case-study-interpreting-manipulating-and-controlling-clip. Accessed: 2024-05-16.

Gytis Daujotas. Case study: Interpreting, manipulating, and controlling clip with sparse autoencoders, Aug 2024. URL https://www.lesswrong.com/posts/iYFuZo9BMvr6GgMs5/case-study-interpreting-manipulating-and-controlling-clip. Accessed: 2025-10-03.

Daniel Pluth, Yu Zhou, and Vijay K. Gurbani. Sparse Autoencoder Insights on Voice Embeddings, 2025. URL https://arxiv.org/abs/2502.00127. Version Number: 1.

Victor Chernozhuokov, Denis Chetverikov, Kengo Kato, and Yuta Koike. Improved central limit theorem and bootstrap approximations in high dimensions. *The Annals of Statistics*, 50(5), October 2022. ISSN 0090-5364. doi: 10.1214/22-AOS2193. URL https://projecteuclid.org/journals/annals-of-statistics/volume-50/issue-5/Improved-central-limit-theorem-and-bootstrap-approximations-in-high-dimensions/10.1214/22-AOS2193.full.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4), July 2017. ISSN 0091-1798. doi: 10.1214/16-AOP1113. URL https://projecteuclid.org/journals/annals-of-probability/volume-45/issue-4/Central-limit-theorems-and-bootstrap-in-high-dimensions/10.1214/16-AOP1113.full.

Victor Chernozhukov, Denis Chetverikov, Kengo Kato, and Yuta Koike. High-Dimensional Data Bootstrap. *Annual Review of Statistics and Its Application*, 10(1):427–449, March 2023. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-040120-022239. URL https://www.annualreviews.org/doi/10.1146/annurev-statistics-040120-022239.

Matthew Gentzkow and Jesse M. Shapiro. What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78(1):35–71, January 2010. ISSN 0012-9682. doi: 10.3982/ECTA7195. URL https://doi.org/10.3982/ECTA7195. Publisher: John Wiley & Sons, Ltd.

OpenAI. gpt-oss-120b gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Beatrice Ferrario and Stefanie Stantcheva. Eliciting People's First-Order Concerns: Text Analysis of Open-Ended Survey Questions. *AEA Papers and Proceedings*, 112:163–169, May 2022. ISSN 2574-0768, 2574-0776. doi: 10.1257/pandp.20221071. URL https://pubs.aeaweb.org/doi/10.1257/pandp.20221071.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 1 edition, September 2018. ISBN 978-1-108-23159-6 978-1-108-41519-4. doi: 10.1017/9781108231596. URL https://www.cambridge.org/core/product/identifier/9781108231596/type/book.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, December 2022. ISSN 2049-8772. doi: 10.1093/imaiai/iaac012. URL https://academic.oup.com/imaiai/article/11/4/1389/6612958.

6 Appendix

6.1 Local Autointerpretation

Building "autointerpretation" pipelines for scalable description of features produced from dictionary learning methods is an active area of AI/ML research. Most best practices rely on leveraging LLMs themselves to interpret the features, per the pioneering work of Bills et al. [2023].

Heuristically, most of these pipelines work by collecting the text samples on which a given feature activates the most in a particular corpus, and weave information about these activations—alongside the activating text—into prompts that LLMs are asked to interpret. Though autointerpretation descriptions are available for all features for the Gemma Scope models, they are learned based on the distribution of Gemma's pretraining data, and as such may be refined through an autointerpretation method "localized" to the unstructured dataset distribution P.

To accomplish this goal, we adapt the methods of Paulo et al. [2024] to generate autointerpretation prompts of the form:

```
{"role": "system", "content": """
    You are a meticulous AI researcher conducting an important
    investigation into patterns found in language.
"""
},
{"role": "user", "content": f"""
    When a corpus of texts was passed through a LLM, a particular neuron
    most activated on the following examples, and specifically on the text
    delimited << like this >>. Provide a single phrase description of what the
    neuron likely responds to (in any corpus, not just this one), and
    delimit it as [[your concise description here]]. Do not mention the marker
    tokens ($<<$ $>>$) in your interpretation. The examples are: {texts}
"""
}
```

where `texts` is a concatenation of L texts—modified with delimiters as the prompt suggests—associated with the tokens that activated most highly on a given feature in the researcher's unstructured dataset. We use $L = 10$ in the empirical examples of this paper, and the open-source reasoning model GPT-OSS 20B [OpenAI, 2025] to perform interpretations. GPT-OSS 20B's reasoning effort is set to "low" for faster generations.

Future work would seek to provide quantitative assessments of the quality of these interpretations, either using the methods of Paulo et al. [2024] or Movva et al. [2025].

6.2 Proofs

Before stating proofs of the results in the main text, we state three additional useful lemmas.

Lemma 3. *For a sequence of random variables U_n and for a deterministic sequence r_n,*

$$U_n = o_P(r_n) \iff P^B(|U_n/r_n| \geq \varepsilon) = o_P(1) \text{ for any } \varepsilon > 0.$$

Proof. Note that, by definition, we have that $U_n = o_P(r_n)$ if for any $\varepsilon > 0$ that

$$\lim_{n\to\infty} P(|U_n/r_n| \geq \varepsilon) = \lim_{n\to\infty} E[Z_{n,\varepsilon}] = 0$$

where $Z_{n,\varepsilon} := P(|U_n/r_n| \geq \varepsilon \mid X)$, as by the law of total expectation $P(|U_n/r_n| \geq \varepsilon) = E[Z_{n,\varepsilon}]$. By Markov's inequality, because $Z_{n,\varepsilon}$ is always positive, for any $\delta > 0$,

$$P(Z_{n,\varepsilon} \geq \delta) \leq \frac{E[Z_{n,\varepsilon}]}{\delta}.$$

Thus we have that

$$0 \leq \lim_{n\to\infty} P(Z_{n,\varepsilon} \geq \delta) \leq \frac{\lim_{n\to\infty} E[Z_{n,\varepsilon}]}{\delta} = 0$$

and $\lim_{n\to\infty} P(|Z_{n,\varepsilon}| \geq \delta) = 0$ for any $\varepsilon, \delta > 0$. As such, $P(|U_n/r_n| \geq \varepsilon \mid X) = P^B(|U_n/r_n| \geq \varepsilon) = o_P(1)$ for any $\varepsilon > 0$ if $U_n = o_P(r_n)$.

For the other direction, note further that if $P(|U_n/r_n| \geq \varepsilon \mid X) = o_P(1)$ then the boundedness of $Z_{n,\varepsilon}$ permits using the bounded convergence theorem to show

$$P(|U_n/r_n| \geq \varepsilon) = o(1).$$

□

Lemma 4. *For $\{X_i\}_{i\in[n]}$ independent sub-exponential random vectors of dimension p, with each $\|X_{ij}\|_{\psi_1} \leq B_n$ for $i \in [n]$ and $j \in [p]$, then, under Assumption* R*,*

$$\|S_n\|_\infty = \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - E[X_i]) \right\|_\infty = O_P\left(B_n \sqrt{\log p} \right).$$

Proof. We start by noting Corollary 2.9.2 of Vershynin [2018], which implies here that

$$P\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - E[X_{ij}]) \right| \geq t \right\} \leq 2 \exp\left[-c \min\left(\frac{nt^2}{B_n^2}, \frac{nt}{B_n} \right) \right].$$

Thus by the union bound,

$$P\left\{ \max_{j\in[p]} \left| \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - E[X_{ij}]) \right| \geq t \right\} \leq 2p \exp\left[-c \min\left(\frac{nt^2}{B_n^2}, \frac{nt}{B_n} \right) \right].$$

Defining that

$$\varepsilon := 2p \exp\left[-c \min\left(\frac{nt^2}{B_n^2}, \frac{nt}{B_n} \right) \right]$$

then we have that for $\tilde{c} = 1/\sqrt{c}$

$$\tilde{c} \frac{\log(2p/\varepsilon)}{n} = \min\left\{ t^2/B_n^2, t/B_n \right\}$$

and therefore

$$t = \tilde{c} B_n \frac{\log(2p/\varepsilon)}{n} \vee \tilde{c} B_n \sqrt{\frac{\log(2p/\varepsilon)}{n}},$$

meaning that

$$P\left\{\max_{j\in[p]}\left|\frac{1}{n}\sum_{i=1}^{n}(X_{ij}-E[X_{ij}])\right|\geq \tilde{c}B_n\left(\frac{\log(2p/\varepsilon)}{n}\vee\sqrt{\frac{\log(2p/\varepsilon)}{n}}\right)\right\}\leq\varepsilon.$$

As such, we have

$$P\left\{\max_{j\in[p]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_{ij}-E[X_{ij}])\right|\geq \tilde{c}B_n\left(\frac{\log(2p/\varepsilon)}{\sqrt{n}}\vee\sqrt{\log(2p/\varepsilon)}\right)\right\}\leq\varepsilon.$$

However, under Assumption R, $\log(p)/\sqrt{n}=o(1)$, so the Gaussian tail term dominates for large n, and thus for large n we may set $M_\varepsilon := \tilde{c}\sqrt{\frac{\log(2/\varepsilon)}{\log 3}+1}$ and observe that

$$P\left\{\frac{\max_{j\in[p]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_{ij}-E[X_{ij}])\right|}{B_n\sqrt{\log p}}\geq M_\varepsilon\right\}\leq\varepsilon.$$

We then conclude using the definition of stochastic boundedness that

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i-E[X_i])\right\|_\infty = O_P\left(B_n\sqrt{\log p}\right)$$

as claimed. □

Lemma 5. *For sequences of positive random variables U_n and V_n, if, uniformly in t, for $\delta_n=o(1),\nu_n=o(1)$,*

$$P(|P(U_n>t\mid X)-P(V_n>t)|\geq\delta_n)\leq\nu_n$$

then, uniformly in t,

$$|P(U_n>t\mid X)-P(V_n>t)|=o_P(1).$$

Proof. Define $d_n(X):=|P(U_n>t\mid X)-P(V_n>t)|$. We want to show that, for any fixed $\delta>0$,

$$\lim_{n\to\infty}P(d_n(X)\geq\delta)=0$$

(the definition of convergence in probability). Note that, given the definition $\delta_n=o(1)$, for some n^*, for all $n\geq n^*$ we have that $\delta_n<\delta$. Thus, define the events:

$$A_n:=\{d_n(X)\geq\delta\},\quad B_n:=\{d_n(X)\geq\delta_n\}$$

For all $n\geq n^*$, $A_n\subseteq B_n$. Thus, for all $n\geq n^*$

$$P\{d_n(X)\geq\delta\}\leq P\{d_n(X)\geq\delta_n\}.$$

Notice that then, using that $\nu_n=o(1)$,

$$0\leq\lim_{n\to\infty}P\{d_n(X)\geq\delta\}\leq\lim_{n\to\infty}P\{d_n(X)\geq\delta_n\}=0.$$

Because this holds for any choice of $\delta>0$ and any t, we have proven the stated result. □

Lemma 6. *Under Assumptions* M *and* R*,*

$$\|S_n^B\|_\infty = O_P(B_n\sqrt{\log p}).$$

Proof. Using Assumptions M and R, the proof of Lemma 4 implies that

$$P\left\{\|S_n\|_\infty \geq \tilde{c}B_n\left(\frac{\log(2p/\varepsilon)}{\sqrt{n}} \vee \sqrt{\log(2p/\varepsilon)}\right)\right\} \leq \varepsilon.$$

Using Lemma 4.6 of Chernozhuokov et al. [2022] and Lemma 5, notice that

$$b_n(X) := P\left(\|S_n^B\|_\infty > \tilde{c}B_n\left(\frac{\log(2p/\varepsilon)}{\sqrt{n}} \vee \sqrt{\log(2p/\varepsilon)}\right) \mid X\right) \leq \varepsilon + o_P(1).$$

Define the event $A_n := \{b_n(X) > \varepsilon + \eta\}$ for any $\eta > 0$. Then the earlier statement implies that $P(A_n) = o(1)$. Now, by the law of total expectation and the law of total probability,

$$\begin{aligned}
&P\left(\|S_n^B\|_\infty > \tilde{c}B_n\left(\frac{\log(2p/\varepsilon)}{\sqrt{n}} \vee \sqrt{\log(2p/\varepsilon)}\right)\right) \\
&= E[b_n(X)] \\
&= E[b_n(X) \mid A_n]P(A_n) + E[b_n(X) \mid A_n^c]P(A_n^c) \\
&= E[b_n(X) \mid A_n^c] + o(1) \\
&\leq \varepsilon + o(1).
\end{aligned} \tag{1}$$

It then follows, using the same logic as in Lemma 4, along with Assumption R, that

$$\|S_n^B\|_\infty = O_P(B_n\sqrt{\log p}).$$

□

We now prove the results from the main text.

Theorem 1 (High-dimensional k-FWER exceedance control for small k)**.** *Consider the method of Algorithm 2.1 or 2.2 in Romano and Wolf [2007] with test statistics T_n of hypotheses $\{H_{0,j}\}_{j\in[p]}$ and critical values $\hat{c}_{n,K}(1-\alpha,k)$ given by the $1-\alpha$ quantile of $S^B_{n,K,[k]}$ under P^B. Assume that k is fixed (i.e., not growing with n,p). Then under Assumptions* M *and* R*:*

(i) $\limsup_{n,p\to\infty} k\text{-}FWER_P \leq \alpha$.

(ii) If $H_{0,j}$ is false and $\theta_j(P) > 0$, then the probability that the step-down method rejects $H_{0,j}$ tends to 1.

Proof. We consider Algorithm 2.1 from Romano and Wolf [2007], which at each stage takes as inputs critical values $\hat{c}_{n,K}(1-\alpha,k)$ and a vector of test statistics T_n; we denote this as `RW-2.1`$(T_n, \hat{c}_{n,K}(1-\alpha,k))$.

The choice of critical values we will use is based on a bootstrap of the $1-\alpha$ quantile of the k-max:

$$\hat{c}_{n,K}(1-\alpha,k) := 1-\alpha \text{ quantile of } S^B_{n,K,[k]} \mid X.$$

As in Romano and Wolf [2007], define $I(P)$ to be the set of indices of true null hypotheses under P. Note that, for any $K \supset I(P)$,

$$\hat{c}_{n,K}(1-\alpha,k) \geq \hat{c}_{n,I(P)}(1-\alpha,k)$$

because for any $K \supset I(P)$, and under any distribution, $S^B_{n,K,[k]} \geq S^B_{n,I(P),[k]}$ almost surely (i.e., the k-max statistic can only get larger if we include more test statistics without dropping the others). As such, Theorem 2.1 (i) holds, and we conclude that `RW-2.1`$(T_n, \hat{c}_{n,K}(1-\alpha,k))$ delivers:

$$k\text{-FWER}_P \leq P\left\{k\text{-max}\left(T_{n,j} : j \in I(P)\right) > \hat{c}_{n,I(P)}(1-\alpha,k)\right\}.$$

Specifically, then, it is sufficient to show that

$$\limsup_{n,p} P\left\{k\text{-max}\left(T_{n,j} : j \in I(P)\right) > \hat{c}_{n,I(P)}(1-\alpha,k)\right\} \leq \alpha.$$

To continue to use the notation of Romano and Wolf [2007], let $\hat{\theta}_{n,j} := \bar{X}_{n,j}$. Since $\theta_j(P) \leq 0$ for $j \in I(P)$, it follows that, almost surely,

$$\begin{aligned} k\text{-max}\left(T_{n,j} : j \in I(P)\right) &= k\text{-max}\left(\sqrt{n}\hat{\theta}_{n,j} : j \in I(P)\right) \\ &\leq k\text{-max}\left(\sqrt{n}\left[\hat{\theta}_{n,j} - \theta_j(P)\right] : j \in I(P)\right) \\ &= k\text{-max}\left(S_{n,j} : j \in I(P)\right) \end{aligned}$$

and therefore

$$\begin{aligned} &P\left\{k\text{-max}\left(T_{n,j} : j \in I(P)\right) > \hat{c}_{n,I(P)}(1-\alpha,k)\right\} \\ \leq &P\left\{k\text{-max}\left(S_{n,j} : j \in I(P)\right) > \hat{c}_{n,I(P)}(1-\alpha,k)\right\}. \end{aligned}$$

If we can show that the limit of the quantity on the right-hand side is no greater than α, the proof is complete. However, Theorem 2.2 of Ding et al. [2025] delivers exactly that, for any K,

$$P\left\{S_{n,K,[k]} > \hat{c}_{n,K}\left(1-\alpha,k\right)\right\} \leq \alpha + o(1)$$

under Assumptions M and R. So, chaining inequalities, we note that

$$k\text{-FWER}_P \leq P\left\{S_{n,I(P),[k]} > \hat{c}_{n,I(P)}\left(1-\alpha,k\right)\right\} \leq \alpha + o(1)$$

and so

$$\limsup_{n,p\to\infty} k\text{-FWER}_P \leq \alpha$$

which is precisely what we wanted to show.

To prove the second statement, consider the $H_{0,j}$ corresponding to all $\theta_j(P) > 0$. Note that, for the Gaussian multiplier bootstrap, using Assumptions M and R, by the tail inequality Equation 1 in the proof of Lemma 6,

$$\hat{c}_{n,[p]}\left(1-\alpha,k\right) \leq \hat{c}_{n,[p]}\left(1-\alpha,1\right) = O\left(B_n\sqrt{\log p}\right).$$

Furthermore, each $S_{n,j} = \sqrt{n}\left[\hat{\theta}_{n,j} - \theta_j(P)\right]$ has a limiting distribution, so

$$T_{n,j} = \sqrt{n}\hat{\theta}_{n,j} = \sqrt{n}(\hat{\theta}_{n,j} - \theta_j(P)) + \sqrt{n}\theta_j(P) \xrightarrow{P} \infty.$$

However, because $\sqrt{n}$ grows faster than $B_n\sqrt{\log p}$ by Assumption R, we have that also

$$\frac{T_{n,j}}{B_n\sqrt{\log p}} \xrightarrow{P} \infty.$$

Therefore, with probability tending to one, $T_{n,j} > \hat{c}_{n,[p]}(1-\alpha, k)$, resulting in the rejection of $H_{0,j}$ in the first step of Algorithm 2.1, so long as $\theta_j(P)$ is fixed or approaches zero slower than a rate of $B_n\sqrt{\log p/n}$.

The asymptotic validity of the streamlined Algorithm 2.2 follows immediately from having proved this, as it does in the Romano and Wolf [2007] proof of Theorem 3.3, given fixed or slow shrinking $\theta_j(P)$, using the same logic as for the proof of the second statement, i.e., $\min(T_{n,j} : j \notin I(P))$ is diverging at rate $\sqrt{n}$, and if any $\theta_j(P) = 0$ then $\max(T_{n,j} : j \in I(P)) = O_P(B_n\sqrt{\log p})$. □

Lemma 1 (High-dimensional CLT for the small k-max coordinate of approximate means)**.** *Let $\hat{S}_n := S_n + R_n$, and assume that $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$. Further assume that k is fixed (i.e., does not grow with n, p). If Assumptions* M *and* R *hold, then as $n, p \to \infty$*

$$\sup_{t\in\mathbb{R}} \left| P\left(\hat{S}_{n,[k]} \leq t\right) - P\left(N(0,\Sigma)_{[k]} \leq t\right)\right| \to 0.$$

Proof. The proof of this lemma proceeds following the strategy of Chernozhukov et al. [2023].

First, recall that the function $t \mapsto t_{[k]}$ is 1-Lipschitz wrt to the sup-norm, meaning almost surely

$$\left|\hat{S}_{n,[k]} - S_{n,[k]}\right| = \left|(S_n + R_n)_{[k]} - S_{n,[k]}\right| \leq \|R_n\|_\infty.$$

Consider the event $\{\|R_n\|_\infty \leq \epsilon\}$, as well as the event $\{(S_n + R_n)_{[k]} \leq t\}$. Then observe that

$$\{\|R_n\|_\infty \leq \epsilon\} \cap \{(S_n + R_n)_{[k]} \leq t\} \subseteq \{S_{n,[k]} \leq t + \epsilon\}$$

As such notice that

$$\{(S_n+R_n)_{[k]} \leq t\} = (\{(S_n+R_n)_{[k]} \leq t\} \cap \{\|R_n\|_\infty \leq \epsilon\}) \cup (\{(S_n+R_n)_{[k]} \leq t\} \cap \{\|R_n\|_\infty > \epsilon\})$$

and so

$$P(\hat{S}_{n,[k]} \leq t) = P((S_n + R_n)_{[k]} \leq t) \leq P(S_{n,[k]} \leq t + \epsilon) + P(\|R_n\|_\infty > \epsilon).$$

We then have, using Assumptions M and R in addition to the high-dimensional CLT for the k largest coordinate of Lemma A.6 in Ding et al. [2025] (which requires Assumption M):

$$\begin{aligned} P\left(\hat{S}_{n,[k]} \leq t\right) &\leq P\left(S_{n,[k]} \leq t + \epsilon\right) + P\left(\|R_n\|_\infty > \epsilon\right) \\ &= P\left(N(0,\Sigma)_{[k]} \leq t + \epsilon\right) + o(1) + P\left(\|R_n\|_\infty > \epsilon\right). \\ &\qquad \text{(Lemma A.6 + Assumption R)} \end{aligned}$$

We now need to apply an anti-concentration result, which can be found as Corollary A.1 of Ding et al. [2025]. Letting $G := N(0, \Sigma)$, it states, for any $\tilde{t}$,

$$P(\tilde{t} - \tilde{\epsilon} \leq G_{[k]} \leq \tilde{t} + \tilde{\epsilon}) \leq Ck\tilde{\epsilon}\sqrt{1 \vee \ln(p/\tilde{\epsilon})}.$$

Thus, letting $t := \tilde{t} - \tilde{\epsilon}$ and $\epsilon := 2\tilde{\epsilon}$,

$$P(t \leq G_{[k]} \leq t + \epsilon) \leq \frac{1}{2}Ck\epsilon\sqrt{1 \vee \ln(2p/\epsilon)}$$

and so because k is fixed,

$$P(G_{[k]} \leq t + \epsilon) = P(G_{[k]} \leq t) + O\left(\epsilon\sqrt{1 \vee \ln(2p/\epsilon)}\right)$$

meaning that

$$P\left(\hat{S}_{n,[k]} \leq t\right) \leq P\left(N(0, \Sigma)_{[k]} \leq t\right) + O(\epsilon\sqrt{1 \vee \log(2p/\epsilon)}) + o(1) + P\left(\|R_n\|_\infty > \epsilon\right). \tag{Corollary A.1}$$

Then we may choose $\epsilon = \epsilon_n = o(1/\sqrt{\log p})$ in such a way that we get that

$$P\left(\hat{S}_{n,[k]} \leq t\right) \leq P\left(N(0, \Sigma)_{[k]} \leq t\right) + o(1).$$

From the reverse direction, note that

$$\{\|R_n\|_\infty \leq \epsilon\} \cap \{(S_n + R_n)_{[k]} \leq t - \epsilon\} \subseteq \{S_{n,[k]} \leq t\}$$

so it also holds that similarly, partitioning $\{(S_n + R_n)_{[k]} \leq t - \epsilon\}$ using $\{\|R_n\|_\infty \leq \epsilon\}$ and its complement, that

$$P\left(\hat{S}_{n,[k]} \leq t\right) \geq P\left(S_{n,[k]} \leq t - \epsilon\right) - P\left(\|R_n\|_\infty > \epsilon\right),$$

and so using identical arguments as above we conclude that, uniformly in t,

$$\left|P\left(\hat{S}_{n,[k]} \leq t\right) - P\left(N(0, \Sigma)_{[k]} \leq t\right)\right| = o(1),$$

proving the stated result. □

Lemma 2 (High-dimensional bootstrap for the small k-max coordinate of approximate means)**.** *Let $\hat{S}_n^B := S_n^B + R_n$, and assume that $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$. Further assume that k is fixed (i.e., does not grow with n, p). If Assumptions* M *and* R *hold, then as $n, p \to \infty$*

$$\sup_{t \in \mathbb{R}} \left|P^B\left(\hat{S}_{n,[k]}^B \leq t\right) - P\left(N(0, \Sigma)_{[k]} \leq t\right)\right| \xrightarrow{P} 0.$$

Proof. The proof of this lemma proceeds following the strategy of Lemma 1. Using the same logic as the proof of Lemma 1, note that

$$P^B(\hat{S}_{n,[k]}^B \leq t) = P^B((S_n^B + R_n)_{[k]} \leq t) \leq P^B(S_{n,[k]}^B \leq t + \epsilon) + P^B(\|R_n\|_\infty > \epsilon).$$

Observe that, from Ding et al. [2025] Lemma A.8 and Lemma 5, using Assumptions M and R and the triangle inequality, that

$$\sup_{t\in\mathbb{R}} \left| P^B \left(S^B_{n,[k]} \leq t \right) - P\left(N(0,\Sigma)_{[k]} \leq t \right) \right| = o_P(1).$$

This equation then plays the role of Lemma A.6 of Ding et al. [2025] in the proof of Lemma 1; the stated result proceeds from this observation, continuing with identical logic as in Lemma 1, noting that if $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$ then $P^B(\sqrt{\log p}\|R_n\|_\infty > \epsilon) = o_P(1)$ for all $\epsilon > 0$ by Lemma 3. □

Corollary 1 (High-dimensional CLT for the small k-max studentized coordinate)**.** *If Assumptions* M *and* R *hold, and* $B_n = O(1)$ *and* k *is fixed (i.e., does not grow with* n, p*), then, by application of Lemma 1, as* $n, p \to \infty$

$$\sup_{t\in\mathbb{R}} \left| P\left(\left(\hat{\Lambda}^{-1/2} S_n \right)_{[k]} \leq t \right) - P\left(N(0,\Sigma_0)_{[k]} \leq t \right) \right| \to 0.$$

Proof. Let $\hat{S}_n := \hat{\Lambda}^{-1/2} S_n$. To show this corollary, note that

$$R_n = \hat{\Lambda}^{-1/2} S_n - \Lambda^{-1/2} S_n = (\hat{\Lambda}^{-1/2} - \Lambda^{-1/2}) S_n,$$

and thus we can use the machinery of Lemma 1 to prove the desired result so long as we can show that $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$.

Note that, by the sub-multiplicative induced matrix norm inequality,

$$\left\| \left(\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2} \right) S_n \right\|_\infty \leq \|\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2}\|_\infty \|S_n\|_\infty = \|\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2}\|_{\max} \|S_n\|_\infty$$

where the last equality follows from the fact that the max elementwise norm is equal to the induced operator ∞-norm for diagonal matrices.

To control the first term on the right-hand side, we may turn to Kuchibhotla and Chakrabortty [2022], Theorem 4.2, which shows that if X_{ij} are sub-Weibull with parameter $\alpha = 1$ (i.e., sub-exponential, granted by Assumption M), then Assumption R ensures that

$$\|\widehat{\Lambda} - \Lambda\|_{\max} = o_P(1/\log^2 p).$$

To see this, note that the condition discussed in Remark 4.1

$$(\log p)^{2/\alpha - 1/2} = o\left(\sqrt{n} (\log n)^{-2/\alpha} \right)$$

is satisfied under Assumption R, meaning that $\|\hat{\Sigma} - \Sigma\|_{\max} = O_P\left(\sqrt{\log p / n} \right)$ if $B_n = O(1)$ (where $\|\cdot\|_{\max}$ is the maximum elementwise norm). As a consequence, Assumption R also delivers that $\|\widehat{\Lambda} - \Lambda\|_{\max} \leq \|\hat{\Sigma} - \Sigma\|_{\max} = o_P\left(1/\log^2 p\right)$, and thus $\|\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2}\|_{\max} = o_P\left(1/\log^2 p\right)$ as well using a Taylor expansion argument and part (ii) of Assumption M.

For the second term, note that, using Lemma 4 via Assumptions M and R, that $\|S_n\|_\infty = O_P(B_n\sqrt{\log p}) = O_P(\sqrt{\log p})$, where the last equality follows because $B_n = O(1)$ by assumption. Putting everything together then, we conclude

$$\left\| \left(\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2} \right) S_n \right\|_\infty \leq o_P\left(1/\log^2 p\right) O_P(\sqrt{\log p}) = o_P(1/\log^{3/2}(p)) = o_P(1/\sqrt{\log p}).$$

□

Corollary 2 (High-dimensional bootstrap for the small k-max studentized coordinate)**.** *If Assumptions* M *and* R *hold, and* $B_n = O(1)$ *and* k *is fixed (i.e., does not grow with* n, p*), then, by application of Lemma 2, as* $n, p \to \infty$

$$\sup_{t \in \mathbb{R}} \left| P^B \left(\left(\hat{\Lambda}^{-1/2} S_n^B \right)_{[k]} \leq t \right) - P\left(N(0, \Sigma_0)_{[k]} \leq t \right) \right| \xrightarrow{P} 0.$$

Proof. The proof of this corollary proceeds just as the proof of Corollary 1. Let $\hat{S}_n^B := \hat{\Lambda}^{-1/2} S_n^B$ and

$$R_n = \hat{\Lambda}^{-1/2} S_n - \Lambda^{-1/2} S_n = (\hat{\Lambda}^{-1/2} - \Lambda^{-1/2}) S_n,$$

and thus we can use Lemma 2 to prove the desired result so long as we can show that, sufficiently, $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$.

Note then that almost surely, as in Corollary 1,

$$\left\| \left(\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2} \right) S_n^B \right\|_\infty \leq \|\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2}\|_{\max} \left\| S_n^B \right\|_\infty .$$

Using the same arguments as in the proof of Corollary 1, we conclude that $\|\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2}\|_\infty = o_P(1/\log^2 p)$. By Lemma 6 we have that $\|S_n^B\|_\infty = O_P(B_n \sqrt{\log p})$, and under $B_n = O(1)$ then $\|S_n^B\|_\infty = O_P(\sqrt{\log p})$. Thus, we complete the proof as in Corollary 1. □