Learning Dependency Structures for Weak Supervision Models

Paroma Varma [*1] **Frederic Sala** [*2] **Ann He** [2] **Alexander Ratner** [2] **Christopher Ré** [2]

Abstract

Labeling training data is a key bottleneck in the modern machine learning pipeline. Recent *weak supervision* approaches combine labels from multiple noisy sources by estimating their accuracies without access to ground truth labels; however, estimating the dependencies among these sources is a critical challenge. We focus on a robust PCA-based algorithm for learning these dependency structures, establish improved theoretical recovery rates, and outperform existing methods on various real-world tasks. Under certain conditions, we show that the amount of unlabeled data needed can scale sublinearly or even logarithmically with the number of sources m, improving over previous efforts that ignore the sparsity pattern in the dependency structure and scale linearly in m. We provide an information-theoretic lower bound on the minimum sample complexity of the weak supervision setting. Our method outperforms weak supervision approaches that assume conditionally-independent sources by up to 4.64 F1 points and previous structure learning approaches by up to 4.41 F1 points on real-world relation extraction and image classification tasks.

1. Introduction

Supervised machine learning models have increasingly become dependent on a large amount of labeled training data. For most real-world applications, however, hand labeling such a large magnitude of data is a major bottleneck, especially when domain expertise is required. Recently, generative models have been used to combine noisy labels from *weak supervision* sources, such as user-defined heuristics or knowledge bases, to efficiently assign training labels by treating the true label as a latent variable (Alfonseca et al., 2012; Takamatsu et al., 2012; Roth & Klakow, 2013; Ratner et al., 2019). Once the labels from the multiple noisy sources are used to learn the parameters of a generative model, the distribution over the true labels is inferred and used to produce probabilistic training labels for the unlabeled data, which can then be used to train a downstream discriminative model.

Specifying how these weak supervision sources are correlated is essential to correctly estimating their accuracies. In practice, weak supervision sources often have strongly correlated outputs due to shared data sources or labeling strategies; for example, developers might contribute near-duplicate weak supervision sources. Manually enumerating these dependencies is a development bottleneck, while learning them statistically usually requires ground truth labels (Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006; Ravikumar et al., 2010; Loh & Wainwright, 2013). Recently, Bach et al. (2017) proposed a structure learning method for weak supervision that requires $\Omega(m \log m)$ samples given m sources and *does not exploit the sparsity of the structure of the associated model*. This high sample complexity may prevent it from identifying dependencies, thus affecting the downstream quality of training labels assigned by the generative model.

We propose using a structure learning technique for the weak supervision setting that exploits the sparsity of the model to achieve improved theoretical recovery rates. We decompose the inverse covariance matrix of the observable sources via robust principal component analysis (Candès et al., 2011; Chandrasekaran et al., 2011). The decomposition produces a sparse component encoding the underlying structure and a low-rank component due to marginalizing over the latent true label variable. We build on previous approaches using this technique (Chandrasekaran et al., 2011; Wu et al., 2017), but improve over their requirement of $\Omega(m)$ samples under common weak supervision conditions.

The key to obtaining tighter complexity estimates is characterizing the *effective rank* (Vershynin, 2012) of the covariance matrix in terms of the structural information associated with the weak supervision setting. The effective rank can be unboundedly smaller than the true rank. We show that under certain reasonable conditions on the effective rank,

[*]Equal contribution [1]Department of Electrical Engineering, Stanford University [2]Department of Computer Science, Stanford University. Correspondence to: Paroma Varma <paroma@stanford.edu>, Frederic Sala <fredsala@stanford.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

intuitively similar to the presence of a stronger dependency in each cluster of correlated sources, the sample complexity can be sublinear $\Omega(d^2 m^\tau)$ for $0 < \tau < 1$ and maximum dependency degree d. Under a stronger condition equivalent to the presence of a dominant cluster of correlated supervision sources, we obtain the rate $\Omega(d^2 \log m)$ that matches the *optimal supervised rate* (Santhanam & Wainwright, 2012). We further study the unsupervised setting through an information-theoretic lower bound on the sample complexity, yielding a characterization of the *additional cost of the weak supervision setting* compared to the supervised setting. We find that, although latent-variable structure learning may result in much higher sample complexity in general, the additional number of samples required is small in the weak supervision setting.

For a variety of real-world tasks from relation extraction to image classification, correlations often naturally arise among weak supervision sources like distant supervision via dictionaries and user-defined heuristics. We show that modeling dependencies recovered by our approach improves over assuming conditional independence among the weak supervision sources by up to 4.64 F1 points, and over existing structure learning approaches by up to 4.41 F1 points.

2. Background

Related Work Manually labeling training data can be expensive and time-consuming, especially when domain expertise is required. A common alternative to hand-labeling data is using weak supervision sources. Estimating the accuracies of these sources without ground truth labels is a classic problem (Dawid & Skene, 1979). Methods like crowdsourcing (Dalvi et al., 2013; Joglekar et al., 2015; Zhang et al., 2016), and boosting (Schapire & Freund, 2012) are common; however, we focus on the case in which *no labeled data* is required. Recently, generative models have been used to combine various sources of weak supervision in such settings (Alfonseca et al., 2012; Takamatsu et al., 2012; Roth & Klakow, 2013; Ratner et al., 2019).

Dependencies occur naturally among weak supervision sources for a variety of reasons: sources may operate over the same input (Varma et al., 2017b), distant supervision sources may refer to similar information from a single knowledge base (Mintz et al., 2009), and heuristics over ontologies may operate over the exact same subtree (Mallory et al., 2015). Not accounting for these dependencies in the generative model can lead to incorrectly estimating the accuracy of the weak supervision sources. Dependencies are difficult to specify manually in cases with hundreds of sources, potentially developed by many users. Therefore, there is a need to learn dependencies directly from the labels assigned by the weak supervision sources without using ground truth labels.

Structure learning has a rich history outside of the weak supervision setting. The supervised, fully observed setting includes node-wise and matrix-wise methods. Node-wise methods, like Ravikumar et al. (2010), use regression on a particular node to recover that node's neighborhood. Matrix-wise methods use the inverse covariance matrix to determine the structure (Friedman et al., 2008; Ravikumar et al., 2011; Loh & Wainwright, 2013). In the latent variable setting, works like Chandrasekaran et al. (2012); Meng et al. (2014); Wu et al. (2017) perform structure learning via robust-PCA like approaches. In contrast, we focus on the weak supervision setting, providing a tighter characterization that leads to improved rates, and provide further details in the Appendix.

The major work for structure learning in weak supervision is Bach et al. (2017), which uses a ℓ_1-regularized node-wise pseudo-likelihood method to obtain a sample complexity of $\Omega(m \log m)$. This expression does not depend on the maximum dependency degree d. Our approach fundamentally differs—we use a matrix-wise method that scales better with key parameters (like the sparsity of the graph d) and offers improved performance for several real-world tasks.

Problem Setup We formally describe our setup and the generative model to assign probabilistic training labels. given a set of noisy labels from weak supervision sources. $X \in \mathcal{X}$ is a data point, $Y \in \mathcal{Y}$ is a label with (X, Y) drawn i.i.d. from some distribution $\mathcal{D}$. In the weak supervision setting, we never have access to the true label Y; instead we rely on m weak supervision sources that produce noisy labels λ_i for $1 \leq i \leq m$.

Example 1. *In a text relation extraction setting, X could be be a tuple of two words, such as names of people, and $Y \in \{0, 1\}$ then represents whether the relation of interest exists between the two words, for example whether these two people are being described as married. Potential weak supervision sources can use information from the sentence, such as whether the word "married" appears between the two words, to heuristically—and thus noisily—assign a label for a data point X. An example of an erring label is produced by applying the heuristic to the sentence "Bob and Alice were meant to get married in 2018, but postponed the wedding by 3 years."*

We model the joint distribution of $\lambda_1, \lambda_2, \ldots, \lambda_m, Y$ via a Markov random field with associated graph $G = (V, E)$ with $V = \{\lambda_1, \ldots, \lambda_m\} \cup \{Y\}$. If λ_i is not independent of λ_j conditioned on Y and the other sources, then (λ_i, λ_j) is an edge in G. For simplicity, we assume $\mathcal{X}, \mathcal{Y} = \{0, 1\}$,

although our results easily extend. The density f_G is

$$f_G(\lambda_1, \ldots, \lambda_m, y) = \frac{1}{Z} \exp\left(\sum_{\lambda_i \in V} \theta_i \lambda_i + \sum_{(\lambda_i, \lambda_j) \in E} \theta_{i,j} \lambda_i \lambda_j + \theta_Y y + \sum_{\lambda_i \in V} \theta_{Y,i} y \lambda_i \right), \quad (1)$$

where Z is a partition function to ensure f_G is a normalized distribution, and θ_i and $\theta_{i,j}$ represent the canonical parameters associated with the sources. We can think of $\theta_{i,j}$ as the strength of the correlation between sources λ_i and λ_j, and $\theta_{Y,i}$ as a measure of accuracy of the source λ_i. Once these parameters are learned, the generative model assigns probabilistic training labels by computing $f_G(Y|\lambda_1, \ldots, \lambda_m)$ for each object X in the unlabeled training set, which can be used to train any downstream model.

In the conditionally independent model, $\theta_{i,j} = 0 \; \forall i, j$. In cases with dependencies, the structure of G is user-defined or inferred from source metadata. Our approach learns the dependency structure, then applies previous work that samples from the posterior of a graphical model directly (Ratner et al., 2016) or uses a matrix completion approach to solve for the source parameters (Ratner et al., 2019).

We also rely on the *singleton separator set* assumption (Ratner et al., 2019), which implies the sources form s fully-connected clusters. This is common for weak supervision settings, motivated by the intuition that groups of weak supervision sources may share common data resources.

3. Learning Structures in the Weak Supervision Regime

Our goal is to learn the dependency structure among weak supervision sources, i.e. graph G, directly from data, without observing the latent true label Y. We introduce this *latent structure learning* problem, which we focus on for the remainder of the paper, in Section 3.1. We provide background on robust PCA in Section 3.2, and describe our algorithm adapting it to weak supervision in Section 3.3.

3.1. Structure Learning Objective

We want to learn the structure of graph G given access to noisy labels from m weak supervision sources and no ground truth labels. Let $O = \{\lambda_1, \ldots, \lambda_m\}$ be the observed labels from the weak supervision sources, and $S = \{Y\}$ be the unobserved latent variable. Then,

$$\textbf{Cov}\left[O \cup \mathcal{S}\right] := \Sigma = \begin{bmatrix} \Sigma_O & \Sigma_{O\mathcal{S}} \\ \Sigma_{O\mathcal{S}}^T & \Sigma_{\mathcal{S}} \end{bmatrix}.$$

We rely on a common weak supervision assumption that the graph is sparse, which implies the inverse covariance matrix

$$\Sigma^{-1} := K = \begin{bmatrix} K_O & K_{O\mathcal{S}} \\ K_{O\mathcal{S}}^T & K_{\mathcal{S}} \end{bmatrix}.$$

is *graph-structured*: there is no edge between λ_i and λ_j in G when the corresponding term in Σ^{-1} is 0, or, equivalently, λ_i and λ_j are independent conditioned on all of the other terms (Loh & Wainwright, 2013). However, a key difficulty is that we never know Y, so *we cannot observe the full covariance matrix Σ, or the graph-structured Σ^{-1}.*

However, we can take advantage of the fact that the sub-block of the inverse covariance matrix K_O is graph-structured. In turn, this implies that K_O^{-1} is a permutation of a block-diagonal matrix with s blocks corresponding to the s source clusters, where each block is no larger than $(d+1) \times (d+1)$, where d is the maximum dependency degree. From the block matrix inversion formula,

$$K_O = \Sigma_O^{-1} + c\Sigma_O^{-1}\Sigma_{O\mathcal{S}}\Sigma_{O\mathcal{S}}^T\Sigma_O^{-1}, \quad (2)$$

where $c = \left(\Sigma_{\mathcal{S}} - \Sigma_{O\mathcal{S}}^T\Sigma_O^{-1}\Sigma_{O\mathcal{S}}\right)^{-1} \in \mathbb{R}^+$. Let $z = \sqrt{c}\Sigma_O^{-1}\Sigma_{O\mathcal{S}}$; we can write (2) as

$$\Sigma_O^{-1} = K_O - zz^T.$$

While we cannot observe Σ since it contains the true label Y, we can observe Σ_O and thus Σ_O^{-1}. We form the empirical covariance matrix of observed labels $\Sigma_O^{(n)} \in \mathbb{R}^{m \times m}$:

$$\Sigma_O^{(n)} = \frac{1}{n}\Lambda\Lambda^T - vv^T,$$

where Λ represents the $m \times n$ matrix of labels from the weak supervision sources assigned to the unlabeled data, n represents the total number of datapoints, and $v \in \mathbb{R}^{m \times 1}$ is the average label assigned by the weak supervision sources.

Our goal is to calculate K_O, which is graph-structured and allows us to read off the structure of G from its entries. We therefore have to decompose the observable Σ_O^{-1} into K_O and zz^T, unknown sparse and low-rank components. This inspires the use of robust principal component analysis (Candès et al., 2011; Chandrasekaran et al., 2011).

3.2. Robust PCA

The robust PCA setup consists of a matrix $M \in \mathbb{R}^{m \times m}$ that is equal to the sum of a low-rank matrix and a sparse matrix, $M = L + S$, where $\text{rank}(L) = r$ and $|\text{supp}(S)| = k$. The name is inspired by the observation that although standard PCA recovers a low-dimensional subspace in the presence of bounded noise, it is not robust to gross corruptions (modeled by the entries of the sparse matrix). Note that the decomposition $M = L + S$ is not identifiable without additional conditions. For example, if $M = e_i e_j^T$, M is itself both

sparse and low-rank, and thus the pairs $(L, S) = (M, 0)$ and $(L, S) = (0, M)$ are equally valid solutions. Therefore, the fundamental question of robust PCA is to determine when a unique decomposition can be recovered.

The two seminal works on robust PCA (Candès et al., 2011; Chandrasekaran et al., 2011) studied *transversality* conditions for identifiability. In particular, the solution spaces L, S can only intersect at 0. For the sparse component, let

$$\Omega(S) = \{N \in \mathbb{R}^{m \times m} \mid \text{supp}(N) \subseteq \text{supp}(S)\}.$$

For the low-rank component, let $L = UDV^T$ be the SVD of L with rank r. Then, let

$$T(L) = \{UX^T + YV^T \mid X, Y \in \mathbb{R}^{m \times r}\}.$$

The key notion for identifiability in robust PCA problems is to ensure these subspaces are transverse—so that neither the low-rank components are too sparse, nor the sparse component too low-rank. We measure these notions via the the functions μ, ξ (Chandrasekaran et al., 2011):

$$\mu(\Omega(S)) = \max_{N \in \Omega(S), \|N\|_\infty = 1} \|N\|, \text{ and}$$
$$\xi(T(L)) = \max_{N \in T(L), \|N\| \leq 1} \|N\|_\infty.$$

These two quantities govern how well-aligned the sparse matrix S is with the coordinate axes and how spread out the low-rank matrix L is. For the decomposition of $M = L + S$ to be identifiable, the required condition is

$$\mu(\Omega(S))\xi(T(L)) < 1. \tag{3}$$

3.3. Adapting Robust PCA for Weak Supervision

We now adapt the robust PCA setting to our setup: $S = K_O$ and $L = zz^T$, a rank one matrix. First, we determine identifiability in the noiseless case: if we do not have identifiability even with the true Σ_O matrix, we have no hope of recovering structure in the sampled case $\Sigma_O^{(n)}$.

Let $a_{\min}, a_{\max}$ be the smallest and largest terms in Σ_{OS}, respectively. These represent the smallest and largest covariances between the true label Y and the weak supervision sources λ_i, which are the smallest and largest accuracies of the sources. Similarly, we let $c_{\min}, c_{\max}$ be the smallest and largest terms in Σ_O, respectively, representing the smallest and largest correlations among the sources. We can now write the identifiability condition in terms of the extreme values of the source accuracies and correlations.

Lemma 1. *Let K_O be the block of the inverse covariance matrix Σ^{-1} corresponding to the observed variables, and let $a_{\min}, a_{\max}, c_{\min}, c_{\max}$ be defined as above. Then,*

$$\mu(\Omega(K_O))\xi(T(zz^T)) \leq \frac{6.4d}{\sqrt{m}} \left(\frac{c_{\max}}{c_{\min}}\right) \left(\frac{a_{\max}}{a_{\min}}\right).$$

Algorithm 1 Weak Supervision Structure Learning

Input: Estimate of the covariance matrix $\hat{\Sigma}_O$, parameters λ_n, γ, threshold T, loss function $\mathcal{L}(\cdot, \cdot)$
Solve:

$$(\hat{S}, \hat{L}) = \text{argmin}_{(S,L)} \mathcal{L}(S - L, \Sigma_O^{(n)}) + \lambda_n(\gamma\|S\|_1 + \|L\|_*)$$

s.t. $S - L \succ 0, L \succeq 0$
$\hat{E} \leftarrow \{(i, j) : i < j, \hat{S}_{ij} > T\}$
Return: $\hat{G} = (V, \hat{E})$

Thus, for a fixed degree d, if we have access to

$$m \geq 40.96d^2[c_{\max}a_{\max}/c_{\min}a_{\min}]^2$$

weak supervision sources, then $\mu(\Omega(K_O))\xi(T(zz^T)) < 1$ and there is a unique solution to the decomposition of Σ_O^{-1}.

Implementation We use the loss function from Wu et al. (2017) for Robust PCA in Algorithm 1, which helps avoid inverting the covariance matrix:

$$\mathcal{L}(S - L, \Sigma_O^{(n)}) = \frac{1}{2}\text{tr}((S - L)\Sigma_O^{(n)}(S - L)) - \text{tr}(S - L).$$

We implement Algorithm 1 using standard convex solvers. The recovered sparse matrix $\hat{S}$ does not have entries that are perfectly 0. Therefore, a key choice is to set a threshold T to find the zeros in $\hat{S}$ such that

$$\tilde{S}_{ij} = \begin{cases} \hat{S}_{ij} & \text{if } \hat{S}_{ij} > T, \\ 0 & \text{if } \hat{S}_{ij} \leq T. \end{cases}$$

We can then pass the nonzero entries of $\tilde{S}$ as dependencies to the generative model described in Section 2.

4. Analysis

Our goal is to provide guarantees on the probability that Algorithm 1 successfully recovers the exact dependency structure. The critical quantity in establishing these guarantees is $\|\Sigma_O^{(n)} - \Sigma_O\|$, the spectral norm of the estimation error of the covariance matrix. We control it by characterizing the effective rank of the covariance matrix Σ_O in Section 4.1. We then introduce two different conditions on the effective rank, which enable us to derive our main result of improved sample complexities in Section 4.2.

4.1. Controlling the Covariance Estimation Error

Structure learning algorithms for the supervised case (Ravikumar et al., 2011; Loh & Wainwright, 2013) recover the structure with high probability given $\Omega(d^k \log m)$ samples, where $k \geq 2$ depends on the approach taken. The

unsupervised (latent variable) algorithms in Chandrasekaran et al. (2012); Wu et al. (2017) require $\Omega(m)$ samples.

The critical difference between these classes of algorithms is in their objectives. The objective function for Ravikumar et al. (2011); Loh & Wainwright (2013) contains the regularizer $\|\cdot\|_1$, while the algorithms in Chandrasekaran et al. (2012); Wu et al. (2017) instead have $\|\cdot\|_1 + \|\cdot\|_*$. The presence of the $\|\cdot\|_*$ norm in the objective for the latent settings is the key difference. Both classes of algorithms rely on the *primal-dual witness* approach for their proofs of consistency. The dual norm of $\|\cdot\|_*$ is the spectral norm $\|\cdot\|$. As a result, a bound on $\|\Sigma_O^{(n)} - \Sigma_O\|$ is necessary, while a simpler entry-wise bound is sufficient for the supervised case. To ensure high-probability recovery, the unsupervised approaches rely on matrix concentration inequalities bounding $\|\Sigma_O^{(n)} - \Sigma_O\|$ that require $\Omega(m)$ samples.

Characterizing the Effective Rank To reduce this sampling rate, we leverage a refined measure of rank, the *effective rank* (Vershynin, 2012), defined as

$$r_e(\Sigma_O) = \text{tr}(\Sigma_O)/\|\Sigma_O\|.$$

The effective rank may be much smaller than the true rank; the notion that data matrices are approximately low-rank is well-known (Udell & Townsend, 2018). Characterizing the effective rank in the weak supervision setting enables us to apply sharper concentration inequalities. We build on the analyses in Chandrasekaran et al. (2012); Wu et al. (2017) while providing improved rates. We note that Meng et al. (2014) also considered the effective rank for a related problem; we additionally cover a wider range of cases in the weak supervision setting and give a tighter characterization.

Recall that the structure of K_O^{-1} contains our key problem parameters—but Σ_O does not. We show that

$$r_e(\Sigma_O) \leq r_e(K_O^{-1}) + \frac{\|v\|^2}{\|K_O^{-1}\|}.$$

Therefore, the effective rank of Σ_O can be controlled via the effective rank of K_O^{-1}. We can then characterize $r_e(\Sigma_O)$ in terms of structural information about the supervision sources. More details on this process are in the Appendix.

4.2. Conditions on the Effective Rank & Main Results

We provide two separate conditions on the effective rank, which lead to two different improved regimes for recovery in Algorithm 1. Let $0 < \tau \leq 1$ be a constant and d the maximum dependency degree.

Definition 1 (SBD Condition)**.** *The matrix Σ_O satisfies the source block decay (SBD) condition if its effective rank $r_e(\Sigma_O)$ satisfies*

$$r_e(\Sigma_O) \leq \frac{m^\tau}{(1+\tau)\log m}$$

Cond.	$r_e(\Sigma_O)$	s	**Rate**
Bach	none	none	$\Omega(m \log m)$
Wu	none	none	$\Omega(d^2 m)$
SBD	$O(\frac{m^\tau}{\log m})$	$O\left(\left(\frac{m}{\log^2 m}\right)^{\tau/(2-\tau)}\right)$	$\Omega(d^2 m^\tau)$
SSB	$O(d)$	none	$\Omega(d^2 \log m)$

Table 1. Conditions and rates for latent variable structure learning.

and the number of clusters s satisfies

$$s \leq \frac{m^{\frac{\tau}{2-\tau}}}{((1+\tau)\log m)^{2/(2-\tau)}}.$$

This condition represents a mild assumption on the structure of Σ_O (and, equivalently K_O^{-1}). It corresponds to mild eigenvalue decay in the source blocks, and a condition limiting the total number of blocks. In the weak supervision setting, this translates to the strength of some of the correlations in a cluster differing. By exploiting this decay and controlling the total number of blocks s, we can obtain a sublinear sample complexity of $\Omega(d^2 m^\tau)$ for Algorithm 1.

Definition 2 (SSB Condition)**.** *The matrix Σ_O satisfies the strong source block (SSB) condition if its effective rank $r_e(\Sigma_O)$ satisfies $r_e(\Sigma_O) \leq cd$, where c is a constant.*

The alternate condition is equivalent to the presence of a cluster of sources that forms a strong voting block, dominating the other sources. With this condition, *we can retrieve the optimal rate of $\Omega(d^2(1+\tau)\log m)$ from the supervised case.* We provide a more precise characterization for the effective rank bounds in the proof of the theorem in the Appendix.

Additional Standard Conditions Next, we highlight the general conditions used by Chandrasekaran et al. (2012) and Wu et al. (2017) whose work we build on; we require these to hold in addition to the SBD or SSB conditions we define above. Specifically, we use a series of standard quantities that control transversality, introduced by Chandrasekaran et al. (2012) and Wu et al. (2017). Let

$$h_X(Y) = \frac{1}{2}(XY + YX).$$

Let $\mathcal{P}_S$ denote orthogonal projection onto subspace S. The following terms are used to control the behavior of $h_X(\cdot)$ on the spaces $\Omega(S)$ and $T(L)$. For convenience, we simply

use Ω and T to denote these spaces. Let

$$\alpha_\Omega = \min_{M\in\Omega,\|M\|_\infty=1} \|\mathcal{P}_\Omega h_{\Sigma_O}(M)\|_\infty,$$
$$\delta_\Omega = \min_{M\in\Omega,\|M\|_\infty=1} \|\mathcal{P}_{\Omega^\perp} h_{\Sigma_O}(M)\|_\infty,$$
$$\alpha_T = \min_{M\in T,\|M\|=1} \|\mathcal{P}_T h_{\Sigma_O}(M)\|,$$

$$\delta_T = \min_{M\in T,\|M\|=1} \|\mathcal{P}_{T^\perp} h_{\Sigma_O}(M)\|,$$
$$\beta_T = \max_{M\in T,\|M\|_\infty=1} \|h_{\Sigma_O}(M)\|_\infty,$$
$$\beta_\Omega = \max_{M\in\Omega,\|M\|=1} \|h_{\Sigma_O}(M)\|.$$

We set

$$\alpha = \min\{\alpha_\Omega, \alpha_T\},\ \beta = \max\{\beta_T, \beta_\Omega\},\ \delta = \max\{\delta_\Omega, \delta_T\}.$$

The following irrepresentability conditions are inherited from Wu et al. (2017) and are generalizations of standard conditions from the graphical model literature (Ravikumar et al., 2011; Zhang & Zou, 2014): there exists $\nu \in (0, 1/2)$

$$\delta/\alpha < 1 - 2\nu, \qquad \text{and}$$
$$\mu(\Omega)\xi(T) \le \frac{1}{2}\left(\frac{\nu\alpha}{(2-\nu)\beta}\right)^2.$$

Finally, let ψ_1 be the largest eigenvalue of Σ_O, ψ_m be the smallest, let $K_{O,\min}$ be the smallest non-zero entry in K_O, and σ be the nonzero eigenvalue of zz^T. We set

$$\gamma = \frac{\nu\alpha}{2d\beta(2-\nu)}.$$

Main Results We now present the formal result for the consistency of Algorithm 1. First, the SBD case:

Theorem 1 (Source Block Decay Case)**.** *Let $0 < \tau \le 1$ be a constant. Suppose that the standard conditions above and the SBD condition are met. Set*

$$\lambda_n = \max\{1, \gamma^{-1}\}\frac{(3-2\nu)c_1\psi_1\sqrt{m^\tau}}{\psi_m\sqrt{n}}.$$

Let

$$\rho_1 = \left[\frac{6c_2\beta(3-2\nu)(2-\nu)\psi_1}{\nu\alpha^2\psi_m}\max\{\frac{\gamma}{K_{O,\min}}, \sigma^{-1}, \frac{1}{\psi_m}\}\right]^2.$$

If the number of samples n satisfies

$$n > \rho_1 d^2 m^\tau,$$

and we run Algorithm 1, then, with probability at least $1 - m^{-\tau}$, we recover the exact structure G.

Next, the SSB case:

Theorem 1 (Strong Source Block Case)**.** *Suppose instead that in addition to the standard conditions, the SSB condition holds. Set*

$$\lambda_n = \max\{1, \gamma^{-1}\}\frac{(3-2\nu)c_4c_2\psi_1 d(1+\tau)\log(m)}{\psi_m n}.$$

Let

$$\rho_2 = \frac{6\beta c_2c_4(3-2\nu)(2-\nu)\psi_1}{\nu\alpha^2\psi_m}\max\left\{\frac{\gamma}{K_{O,\min}}, \sigma^{-1}, \frac{1}{\psi_m}\right\}.$$

If the number of samples n satisfies

$$n > \rho_2(1+\tau)d^2\log(m),$$

then, with probability at least $1 - m^{-\tau}$, we recover the exact structure G.

We provide a formal proof of Theorem 1 in the Appendix. The proof modifies the proof technique in Wu et al. (2017) by applying stronger concentration inequalities.

5. Information-Theoretic Lower Bound

So far, we analyzed a specific algorithm, showing that under the stronger of our two conditions, the sample complexity matches the optimal one of $\Omega(d^2 \log m)$ for supervised structure learning. Now we explore the general question of the fundamental limits of structure learning with latent variables. We derive the information-theoretic lower bounds on sample complexity: bounds that show that *for any* such algorithm, at least a certain number of samples is required to avoid incurring a particular probability of error.

First, we consider the general latent-variable case. We do not need to have Y connected to each of the λ_i source variables; Y may be connected to just some of these sources. Even if we ensure that the class of graphs we are working over is connected overall, there are graphs that cannot be distinguished, with any number of samples. One such example is shown in Figure 1. Here, we have two graphs, G_1 and G_2, where the only difference is that in one case, there is an edge between Y and λ_1, while in the other, there is an edge between Y and λ_2. By observing only λ_1 and λ_2, but not Y, we cannot distinguish between these two graphs.

Working in the fully-general latent structure learning setting leads to uninteresting results. Instead, we again work in the weak supervision setting where Y is connected to all of the λ_i's. We already know, from our algorithmic analysis, that in certain cases we can recover the structure with $\Omega(d^2 \log m)$ samples, and this quantity is optimal even in the supervised case. Certainly we expect that the presence of the latent variable Y will require more samples (in terms of lower bounds). In Theorem 2 we quantify this difference.

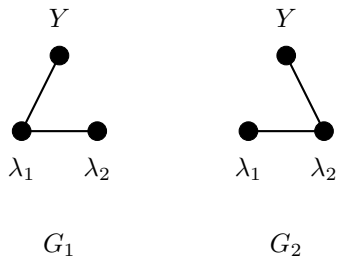

Figure 1. In these two graphs, Y is not connected to every source. Observing λ_1 and λ_2 does not allow us to establish which of $\{G_1, G_2\}$ is the true model, regardless of the number of samples.

The strategy used to derive information-theoretic lower bounds is to construct a collection of graphs along with a set of parameters and to use Fano's inequality (or related methods) that rely on a notion of distance between pairs of graphs in the collection. The smaller this distance, the larger the number of samples required to distinguish between a pair of graphs. Our approach is to consider a collection of graphs used to derive the $\Omega(d^2 \log m)$ lower bound, and to construct the equivalent collection in the latent-variable weak supervision case. We then compute how much larger the number of samples required for reliably selecting the correct graph is for the unsupervised versus supervised case.

Let $\mathcal{G}_{\text{ws}}$ be the class of graphs on $m+1$ nodes (m sources and 1 latent node connected to all of the other nodes) with maximum degree d, structured according to our exponential family model, restricted to the setting where only edge parameters are non-zero, and all such edge parameters are θ. Let $M = |\mathcal{G}_{\text{ws}}|$. Our main result is

Theorem 2. *Any decoding procedure to determine G from samples of $\lambda_1, \ldots, \lambda_m$ will have maximum probability of error at least $\delta - \frac{1}{\log M}$ if the number of samples n is upper-bounded as*

$$n < (1-\delta)\frac{\log(m(m-1)/2)}{2\theta(1-4(\exp(4\theta)+3)^{-1}-\tanh^2(\theta))}.$$

As expected, that the number of samples here is larger than supervised version, where the expression simply has a $2\theta \tanh\theta$ in the denominator (Santhanam & Wainwright, 2012). In particular, the number of additional samples n_Δ we need is given by

$$n_\Delta = \frac{(1-\delta)\log(m(m-1)/2)}{2\theta} \times \left[\frac{1}{1-4(\exp(4\theta)+3)^{-1}-\tanh^2(\theta)} - \frac{1}{\tanh\theta}\right].$$

This quantity characterizes the *cost in sample complexity due to the weak supervision setting*. We observe, however, that in the limit of $\theta \to 0$, this relative cost is not too high. This is the regime of interest for $d, m \to \infty$, where we require $\theta \to 0$ to avoid an exponential sample complexity (Santhanam & Wainwright, 2012). Then, the *relative* version of the cost above can be upper bounded by 2. That is, *we need no more than twice as many samples as in the supervised case* to avoid an unreliable encoder.

As expected, latent variable structure learning requires more samples than the fully-supervised version; potentially, infinitely more. However, the weak-supervision setting provides us with a tractable scenario, where the lower bounds are not much larger than the supervised equivalents.

We briefly comment on the approach to Theorem 2. The collection considered is takes the graphs where all of the λ_i's have no edge between them and adds a single edge between λ_s and λ_t, $s \neq t \in \{1, \ldots, m\}$. Thus there are $\binom{m}{2}$ such graphs in the collection. In the supervised setting, there is just one such edge per graph since there is no latent variable Y; in our setting, there are m additional edges between each λ_i and Y. Intuitively, the challenge when distinguishing between graphs is to ascertain whether a pair of nodes are connected by an edge; this is harder in our setting since all pairs of nodes as connected through Y.

6. Experimental Results

We evaluate our structure learning method on real-world applications ranging from medical image classification to relation extraction over text. We compare our performance to several common weak supervision baselines: an unweighted majority vote of the weak supervision source labels, a generative modeling approach that assumes independence among weak supervision sources (Ratner et al., 2016), and a generative model using dependency structure learned with an existing structure learning approach for weak supervision (Bach et al., 2017). We report performance of the discriminative model trained on labels from these generative models in Table 2. Finally, we run simulations to explore the performance of our method under two conditions from Section 3.

6.1. Real-World Tasks

Task Descriptions We describe the different weak supervision tasks, the associated weak supervision sources, and the discriminative model used to perform classification. The **Bone Tumor** task is to classify tumors in X-rays as aggressive or non-aggressive (Varma et al., 2017b). The discriminative model is a logistic regression model over hundreds of shape, texture, and intensity-based image features. The supervision sources are user-defined heuristics and decision trees over features extracted from the X-rays.

The **CDR** task is to detect relations among chemicals and disease mentions in PubMed abstracts (Bach et al.,

Application	m	(s, d)	MV	Indep.	Bach et al.	Ours	Improvement Over Indep.	Improvement Over Bach et al.
Bone Tumor	17	(2,3)	65.72	67.32	67.83	71.96	+4.64	+4.13
CDR	33	(22,14)	47.74	54.60	55.90	56.81	+2.21	+0.91
IMDb	5	(1,4)	55.21	58.80	60.23	62.71	+3.91	+2.48
MS-COCO	3	(1,2)	57.95	59.47	59.47	63.88	+4.41	+4.41

Table 2. Statistics for weak supervision tasks (m: number sources, s: number of cliques, d: max. degree of source). F1 scores of discriminative models trained on labels generated by majority vote (MV), a generative model with no dependencies (Indep.), a generative model with dependencies learned by a prior structure learning approach for weak supervision (Bach et al), and by our approach (Ours).

2017; Wei et al., 2015). The discriminative model is an LSTM (Graves & Schmidhuber, 2005) that takes as input sentences containing the mentions. The supervision sources are distant supervision from the Comparative Toxicogenomics Database (Davis et al., 2016) and user-defined heuristics. The **IMDb** task is to classify plot summaries as describing action or romantic movies (Varma et al., 2017c). The discriminative model is an LSTM that takes as input the entire plot summary. The supervision sources are user-defined heuristics that look for mentions of specific words. The **MS-COCO** task is to classify images as containing a person (Varma et al., 2017a). The discriminative model is GoogLeNet. The supervision sources are user-defined heuristics written over associated captions.

Performance Our method learns dependencies among the supervision sources for each of the tasks described above, which leads to an average improvement of 3.80 F1 points over the model that assumes independence. For the MS-COCO task, Bach et al. (2017) is unable to learn *any* dependencies while our method learns a single pairwise dependency, which improves performance by 4.41 F1 points. For the Bone Tumor task, our method identifies 2 cliques with 3 supervision sources. The first clique consists of heuristics that all rely on features related to edge sharpness along the lesion contour of the tumor, while the sources in the second clique rely on features describing the morphology of the tumor. Incorporating these dependencies in the generative model improves over Bach et al. (2017) by 4.13 F1 points. Finally, for the IMDb task, our method learns a clique involving 4 sources while Bach et al. (2017) only learns 3 pairwise dependencies among the same sources. Learning a clique improves performance by 2.48 F1 points.

6.2. Simulations

We also perform simulations over synthetic data using 200 weak supervision sources to explore how our performance compares to Bach et al. (2017) under the two conditions on effective rank described in Section 4, the SSB condition and the SBD condition. We define success as how often these methods are able to learn the true dependencies and plot our results in Figure 2. We first generate labels from supervision sources to match the SSB condition by ensuring there exists a single cluster of strongly correlated sources along with other more weakly correlated sources. We observe that our method performs significantly better than Bach et al. (2017), and is capable of recovery in the regime where n is roughly in the range $(\log m, m)$. Second, we simulate the SBD condition by generating multiple cliques of sources where a single dependency in each clique is stronger than the rest. We continue to perform better compared to Bach et al. (2017) under this condition and across all values of n.

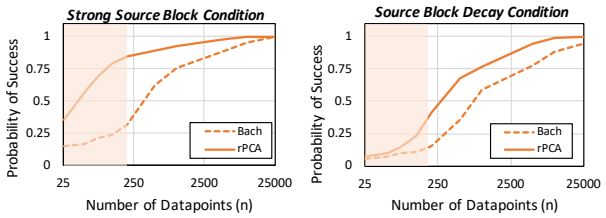

Figure 2. Shaded region shows where $n < m$. With the SSB condition, our method significantly outperforms existing method. Without this condition, neither methods work well when $n < m$.

7. Conclusion

The dependency structure of generative models significantly affects the quality of the generated labels. However, learning this structure without any ground truth labels is challenging. We present a structure learning method that relies on robust principal component analysis to estimate the dependencies among the different weak supervision sources. We prove that the amount of unlabeled data required to estimate the true structure can scale sublinearly or even logarithmically with the number of weak supervision sources, improving over the standard sample complexity, which is linear. Under certain conditions, we match the information-theoretic optimal lower bound in the supervised case. Empirically, this translates to our method outperforming traditional structure learning approaches by up to 4.41 F1 points and methods that assume independence by up to 4.64 F1 points.

Acknowledgements

We gratefully acknowledge the support of DARPA under Nos. FA87501720095 (D3M) and FA86501827865 (SDH), NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity) and CCF1563078 (Volume to Velocity), ONR under No. N000141712266 (Unifying Weak Supervision), the Moore Foundation, NXP, Xilinx, LETI-CEA, Intel, Google, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, the Okawa Foundation, and American Family Insurance, Google Cloud, Swiss Re, the National Science Foundation (NSF) Graduate Research Fellowship under No. DGE-114747, Joseph W. and Hon Mai Goodman Stanford Graduate Fellowship, and members of the Stanford DAWN project: Intel, Microsoft, Teradata, Facebook, Google, Ant Financial, NEC, SAP, VMWare, and Infosys. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of DARPA, NSF, NIH, ONR, or the U.S. Government.

References

Alfonseca, E., Filippova, K., Delort, J.-Y., and Garrido, G. Pattern learning for relation extraction with a hierarchical topic model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pp. 54–59. Association for Computational Linguistics, 2012.

Bach, S. H., He, B., Ratner, A., and Ré, C. Learning the structure of generative models without labeled data. In *ICML*, 2017.

Bunea, F. and Xiao, L. On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. *Bernoulli*, 21(5):1200–1230, 2015.

Bunescu, R. and Mooney, R. Learning to extract relations from the web using minimal supervision. In *ACL*, 2007.

Candes, E., Tao, T., et al. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM*, 58(11), 2011.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. Latent variable graphical model selection via convex optimization. *Annals of Statistics*, 40(4):1935–1967, 2012.

Craven, M., Kumlien, J., et al. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, pp. 77–86, 1999.

Dalvi, N., Dasgupta, A., Kumar, R., and Rastogi, V. Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 285–294. ACM, 2013.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., King, B. L., McMorran, R., Wiegers, J., Wiegers, T. C., and Mattingly, C. J. The comparative toxicogenomics database: update 2017. *Nucleic acids research*, 45(D1): D972–D978, 2016.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics*, pp. 20–28, 1979.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Graves, A. and Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

Hillar, C. J., Lin, S., and Wibisono, A. Tight bounds on the infinity norm of inverses of symmetric diagonally dominant positive matrices. 2014.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., and Weld, D. S. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 541–550. Association for Computational Linguistics, 2011.

Joglekar, M., Garcia-Molina, H., and Parameswaran, A. Comprehensive and reliable crowd assessment algorithms. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, pp. 195–206. IEEE, 2015.

Loh, P.-L. and Wainwright, M. J. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Annals of Statistics*, 41(6):3022–3049, 2013.

Lounici, K. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

Mallory, E. K., Zhang, C., Ré, C., and Altman, R. B. Large-scale extraction of gene interactions from full text literature using DeepDive. *Bioinformatics*, pp. btv476, 2015.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pp. 1436–1462, 2006.

Meng, Z., Eriksson, B., and III, A. O. H. Learning latent variable Gaussian graphical models. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, Beijing, China, 2014.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. Distant supervision for relation extraction without labeled data. In *Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011. Association for Computational Linguistics, 2009.

Qi, L. Some simple estimates for singular values of a matrix. *Linear Algebra and Its Applications*, 56:105–119, 1984.

Ratner, A. J., Sa, C. M. D., Wu, S., Selsam, D., and Ré, C. Data programming: Creating large training sets, quickly. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, 2016.

Ratner, A. J., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., and Ré, C. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, 2019.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Riedel, S., Yao, L., and McCallum, A. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 148–163. Springer, 2010.

Roth, B. and Klakow, D. Combining generative and discriminative model scores for distant supervision. In *EMNLP*, pp. 24–29, 2013.

Santhanam, N. P. and Wainwright, M. J. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

Schapire, R. E. and Freund, Y. *Boosting: Foundations and algorithms*. MIT press, 2012.

Shanmugam, K., Tandon, R., Dimakis, A. G., and Ravikumar, P. On the information theoretic limits of learning Ising models. In *Proceedings of the 27th Conference on Neural Information Processing Systems (NIPS 2016)*, Montreal, Canada, 2014.

Shin, J., Wu, S., Wang, F., De Sa, C., Zhang, C., and Ré, C. Incremental knowledge base construction using DeepDive. *Proceedings of the VLDB Endowment*, 8(11):1310–1321, 2015.

Takamatsu, S., Sato, I., and Nakagawa, H. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 721–729. Association for Computational Linguistics, 2012.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Udell, M. and Townsend, A. Why are big data matrices approximately low rank? *SIAM Mathematics of Data Science (SIMODS)*, 2018.

Varma, P., He, B., Iter, D., Xu, P., Yu, R., De Sa, C., and Ré, C. Socratic learning: Augmenting generative models to incorporate latent subsets in training data. *arXiv preprint arXiv:1610.08123*, 2017a.

Varma, P., He, B. D., Bajaj, P., Khandwala, N., Banerjee, I., Rubin, D., and Ré, C. Inferring generative model structure with static analysis. In *Advances in neural information processing systems*, pp. 240–250, 2017b.

Varma, P., Iter, D., De Sa, C., and Ré, C. Flipper: A systematic approach to debugging training sets. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, pp. 5. ACM, 2017c.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. In *n Compressed Sensing*, pp. 210–268. Cambridge Univ. Press, 2012.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Wei, C.-H., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., Wiegers, T. C., and Lu, Z. Overview of the BioCreative V chemical disease relation (CDR) task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, pp. 154–166, 2015.

Wu, C., Zhao, H., Fang, H., and Deng, M. Graphical model selection with latent variables. *Electronic Journal of Statistics*, 11:3485–3521, 2017.

Zhang, T. and Zou, H. Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika*, 101(1): 103–120, 2014.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(102):1–44, 2016.

Zhao, P. and Yu, B. On model selection consistency of lasso. *Journal of Machine learning research*, 7:2541–2563, 2006.

First, we provide a glossary of terms and notation that we use throughout this paper for easy summary. Afterwards, we provide an extended discussion of related work. We give the proofs of our main results (the lemma and the two theorems). Next, we include a discussion on other aspects of generating information-theoretic lower bounds for the weak supervision setting. We also consider extending robust PCA-based techniques for structure learning *without* the singleton separator set assumption. Finally, we give additional experimental details.

A. Glossary

The glossary can be found in Table 3 below.

Symbol	Used for
X	Data point, $X \in \mathcal{X}$
n	Number of data points
Y	Latent label
λ_i	Weak supervision sources output by the ith source for X
m	Number of sources
G	Source dependency graph, $G = (V, E)$, $V = \{\lambda_1, ..., \lambda_m\} \cup \{Y\}$
f_G	Density of weak supervision sources $\lambda_1 ... \lambda_m$ and latent variable Y
d	Maximum degree of weak supervision sources in G
s	Number of cliques of dependent weak supervision sourcess in G
O	The set of observable variables, i.e., weak supervision sources (but not the label Y)
$\mathcal{S}$	The set of unobserved variables, i.e., the latent label Y
Σ	Covariance matrix of $O \cup \mathcal{S}$, $\Sigma = \mathbf{Cov}\left[O \cup \mathcal{S}\right]$
K	The inverse covariance matrix $K = \Sigma^{-1}$
Σ_O	Covariance matrix of O, the observed variables. Neither Σ_O nor Σ_O^{-1} are graph structured
K_O	Sub-block of inverse covariance matrix K corresponding to observed variables
zz^T	Low rank matrix that encodes the parameters of the graph f_G such that $K_O = \Sigma_O^{-1} + zz^T$
$T(L)$	Tangent space for the low-rank component in robust PCA, $T(L) = \{UX^T + YV^T \mid Y_1, Y_2 \in \mathbb{R}^{m \times r}\}$
$\Omega(S)$	Tangent space for the sparse component, $\Omega(S) = \{X \in \mathbb{R}^{m \times m} \mid \text{supp}(N) \subseteq \text{supp}(S)\}$
$\xi(T(L))$	Measurement of diffuseness of the low-rank term, $\xi(T(L)) = \max_{N \in T(L), \|N\| \leq 1} \|N\|_\infty$
$\mu(\Omega(S))$	Measurement of sparsity, $\mu(\Omega(S)) = \max_{N \in \Omega(S), \|N\|_\infty = 1} \|N\|$
τ	Constant between 0 and 1 that controls the sampling rate and the error probability
ψ_1	Smallest eigenvalue of Σ_O
ψ_m	Largest eigenvalue of Σ_O
γ	Hyperparameter in Algorithm 1
λ_n	Positive eigenvalue of $L = zz^T$
σ	Constant related to λ_n that controls sample complexity of Algorithm
$K_{O,min}$	Smallest non-zero entry of $\|K_O\|$
α_Ω	$\min_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_\Omega h_{\Sigma_O}(M)\|_\infty$
δ_Ω	$\min_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_{\Omega^\perp} h_{\Sigma_O}(M)\|_\infty$
α_T	$\min_{M \in T, \|M\| = 1} \|\mathcal{P}_T h_{\Sigma_O}(M)\|$
δ_T	$\min_{M \in T, \|M\| = 1} \|\mathcal{P}_{T^\perp} h_{\Sigma_O}(M)\|$
β_T	$\max_{M \in T, \|M\|_\infty = 1} \|h_{\Sigma_O}(M)\|_\infty$
β_Ω	$\max_{M \in \Omega, \|M\| = 1} \|h_{\Sigma_O}(M)\|$
α	$\min\{\alpha_\Omega, \alpha_T\}$
β	$\max\{\beta_T, \beta_\Omega\}$
δ	$\max\{\delta_\Omega, \delta_T\}$

Table 3. Glossary of variables and symbols used in this paper.

B. Extended Related Work

A common alternative to hand labeling data is using weak supervision sources, such as distant supervision (Craven et al., 1999; Mintz et al., 2009), multi-instance learning (Riedel et al., 2010; Hoffmann et al., 2011) and heuristics (Bunescu & Mooney, 2007; Shin et al., 2015). Estimating the accuracies of weak supervision sources without ground truth labels is a classic problem (Dawid & Skene, 1979). Methods like crowdsourcing (Dalvi et al., 2013; Joglekar et al., 2015; Zhang et al., 2016), and boosting (Schapire & Freund, 2012) are common approaches; however, we focus on the case in which *no labeled data* is required.

Recently, generative models have been used to combine various sources of weak supervision (Alfonseca et al., 2012; Takamatsu et al., 2012; Roth & Klakow, 2013; **?**; Ratner et al., 2019). Most previous work assumes that the structure of these models is user-specified. Bach et al. (2017) recently showed that it is possible to learn dependencies with a sample complexity that scales quasilinearly with the number of sources. Varma et al. (2017b) inferred dependencies using the code used to define the weak supervision sources. Our method improves over Bach et al. (2017) by reducing the dependence on the number of sources to sublinear, and, under stronger conditions, logarithmic, and is able to learn dependencies that are not explicit in the code, thus improving over Varma et al. (2017b) as shown in Section 6.

Structure learning outside the context of weak supervision can be roughly divided into the supervised and unsupervised case, which require access to ground truth labels and not, respectively. Within these, we can further split the methods into node-wise and matrix-wise methods. Node-wise methods, like Bach et al. (2017), use regression on a particular node to recover that node's neighborhood (Candes et al., 2007; Ravikumar et al., 2010; Wainwright & Jordan, 2008; Tibshirani, 1996) and matrix-wise methods like ours use the inverse covariance matrix to determine the structure (Loh & Wainwright, 2013; Chandrasekaran et al., 2012). The canonical matrix-wise method in the supervised case is the graphical Lasso algorithm (GLASSO) (Friedman et al., 2008).

Analysis of the graphical lasso applied to sparse inverse covariance matrices was studied in Ravikumar et al. (2011), which achieves a sample complexity of $d^2 \log m$. The key question is then when the inverse covariance (precision) matrix of a random vector is sparse. In the classical, Gaussian case, the sparsity is governed by the graphical model associated with the vector: if a pair of variables are independent conditioned on all the other variables (i.e., there is no edge between the associated nodes in their graph), the precision matrix is 0 at the corresponding entry. This is not necessarily the case for non-Gaussian models. For discrete models, (Loh & Wainwright, 2013) characterizes the situations when the precision matrix is indeed graph-structured. In many cases, it is necessary to form a larger, *generalized* covariance matrix. However, for graphs with singleton separator sets, the normal precision matrix is graph-structured.

The idea of using robust PCA for separating the sparse part of the precision matrix (encoding the graph structure) from the low-rank matrix that captures the marginalizing effects dates back to the original papers on robust PCA (Chandrasekaran et al., 2012). For Gaussian graphical models with latent variables, Chandrasekaran et al. (2011) produced the seminal work Chandrasekaran et al. (2012). More recent work follows the same approach, but modifies the loss function (Wu et al., 2017) and relaxes the Gaussian assumptions. On the other hand, Wu et al. (2017) still requires a partially Gaussian model in order to induce sparsity, while our work operates in the discrete case entirely by leveraging the singleton separator set criteria. Both of these works lead to a $\Omega(m)$ rate. The work Meng et al. (2014) is closer to the spirit of our approach; one of their results also considers using the effective rank by applying a theorem from Lounici (2014). However, our work and theirs has key differences: they consider the Gaussian rather than discrete setting, they are interested in model estimation rather than selection. Finally, they work in a general setting; our work in the weak supervision setting enables us to more tightly characterize the effective rank in terms of key sparsity parameters, while they leave the effective rank as a parameter that can only be measured.

The major work for structure learning in the weak supervision regime is Bach et al. (2017). This is a fundamentally different approach, building on the node-wise methods and using a pseudo-likelihood to estimating the structure. The key requirement of Bach et al. (2017) is the maximum number of dependencies d. However, this maximum is taken over *all variables*, both observed and latent. In the weak supervision scenario, there is a dependency between each weak supervision sources and the latent true label, and therefore this degree d always takes the value m (Corollary 2 in Bach et al. (2017)), leading to a rate of $\Omega(m \log m)$. The key advantage of our work is that for us, d is taken over the observed variables only—and therefore can be much smaller than m. This enables us to obtain sublinear (and even logarithmic) rates in m, and to better scale with the sparsity of the model.

C. Proofs

Next, we give proofs of our results, starting with Lemma 1.

Proof. Our goal is to bound the product $\mu(K_O)\xi(zz^T)$. Bounding $\mu(K_O)$ is easy: we apply the simple bound $\mu(K_O) \leq d$ (Proposition 3 in Chandrasekaran et al. (2011)).

We must bound the $\xi(zz^T)$ term, which we do as follows. First, since zz^T is symmetric, it has the same row-space and column-space. Let this row-space be denoted $\text{rs}(zz^T)$. Define $\beta(S) := \max_i \|P_S e_i\|_2$, where P_S is projection onto the subspace S and e_i is the ith standard basis vector. Then, from Proposition 4 in Chandrasekaran et al. (2011)),

$$\xi(zz^T) \leq 2\beta(\text{rs}(zz^T)).$$

Since zz^T is rank-one, its row-space is simply spanned by z and $\beta(\text{rs}(zz^T)) = \|\bar{z}\|_\infty$, where $\bar{z}$ is $z/\|z\|$. Now, applying the definition of z,

$$\beta(\text{rs}(zz^T)) = \|\bar{z}\|_\infty = \frac{\|z\|_\infty}{\|z\|} = \frac{\|\Sigma_O^{-1}\Sigma_{OS}\|_\infty}{\|\Sigma_O^{-1}\Sigma_{OS}\|}. \tag{4}$$

Now, we can upper bound the numerator

$$\|\Sigma_O^{-1}\Sigma_{OS}\|_\infty \leq \|\Sigma_O^{-1}\|_\infty \|\Sigma_{OS}\|_\infty.$$

We lower bound the denominator as

$$\|\Sigma_O^{-1}\Sigma_{OS}\| \geq \sigma_{\min}(\Sigma_O^{-1})\|\Sigma_{OS}\| = \frac{\|\Sigma_{OS}\|}{\sigma_{\max}(\Sigma_O)}.$$

Using these bounds in (4), we have that

$$\beta(\text{rs}(zz^T)) \leq \left(\frac{\sigma_{\max}(\Sigma_O)\|\Sigma_O^{-1}\|_\infty \|\Sigma_{OS}\|_\infty}{\|\Sigma_{OS}\|}\right). \tag{5}$$

Recall that $a_{\min}, a_{\max}$ are the smallest and largest terms in Σ_{OS}, respectively. Similarly, recall that $c_{\min}, c_{\max}$ are the smallest and largest terms in Σ_O. We have that $\|\Sigma_{OS}\|_\infty = a_{\max}$. Also, $\|\Sigma_{OS}\| \geq \sqrt{m}a_{\min}$, so that

$$\frac{\|\Sigma_{OS}\|_\infty}{\|\Sigma_{OS}\|} \leq \frac{a_{\max}}{\sqrt{m}a_{\min}}.$$

Bounding $\|\Sigma_O^{-1}\|_\infty$ is slightly more challenging. Recall the definition of a symmetric diagonally dominant (SDD) matrix. A matrix $J \in \mathbb{R}^{m\times m}$ is SDD if it is symmetric and if $\Delta_i(J) := |J_{ii}| - \sum_{j\neq i}|J_{ij}| \geq 0$ for all $i = 1, \ldots, m$. It is often the case that covariance matrices are SDD (for example, this is the case for the covariances of Gaussian free field models). Even if Σ_O is not SDD, we can make it SDD by performing the operation $\Sigma_O \leftarrow \Sigma_O + \nu I$, for some ν satisfying $\nu \leq (m-1)c_{\max}$. This operation is equivalent to adding independent noise with variance ν to each of the supervision sources. Critically, this does not affect the off-diagonal entries of Σ, which are what we wish to recover.

Thus we take $\Sigma_O \leftarrow \Sigma_O + \nu I$ so that Σ_O is SDD. Then we apply the following tight bound on the ∞ norm of the inverse of a SDD matrix (Hillar et al., 2014):

$$\|\Sigma_O^{-1}\|_\infty \leq \frac{3m-4}{2c_{\min}(m-2)(m-1)} \leq \frac{8}{5c_{\min}m}.$$

Finally, we must bound the largest singular value of Σ_O. Here, we use a Gerschgorin-style bound (Qi, 1984): $\sigma_{\max}(\Sigma_O)$ is at most the largest row or column sum (excluding diagonal elements) plus the largest diagonal element. For us, $\sigma_{\max}(\Sigma_O) \leq \nu + mc_{\max} \leq (2m-1)c_{\max}$.

Putting these results into (5), we obtain

$$\begin{aligned}\beta(\text{rs}(zz^T)) \leq & \frac{8(2m-1)c_{\max}a_{\max}}{5m^{3/2}c_{\min}a_{\min}} \\ \leq & \frac{3.2}{\sqrt{m}}\frac{c_{\max}a_{\max}}{c_{\min}a_{\min}}.\end{aligned}$$

Thus,

$$\xi(zz^T) \leq \frac{6.4}{\sqrt{m}}\left(\frac{c_{\max}}{c_{\min}}\right)\left(\frac{a_{\max}}{a_{\min}}\right).$$

Multiplying by $\mu(K_O) \leq d$ gives the result. □

Proof of Theorem 1 Now we prove Theorem 1.

Proof. Our approach proceeds in two steps. First, we bound the *effective rank* (Vershynin, 2012) of our estimate of Σ_O, the covariance matrix of the observed sources. Next, we apply a pair of concentration bounds for estimating Σ_O. Afterwards, we show how to adapt the proof of Theorem 4.1 in Wu et al. (2017) to obtain the result in Theorem 1.

Effective Rank The effective rank of a matrix A is

$$r_e(A) = \frac{\text{tr}(A)}{\|A\|}.$$

This quantity can be far smaller than the actual rank. As we shall see, sharp concentration bounds for estimating Σ_O can be derived by exploiting $r_e(\Sigma_O)$. We begin by bounding this quantity in our setting. Applying the matrix inversion lemma, we have that

$$\begin{aligned}\Sigma_O &= K_O^{-1} + (K_S - K_{OS}^T K_O^{-1} K_{OS})^{-1} K_O^{-1} K_{OS}^T (K_O^{-1} K_{OS}^T)^T \\ &= K_O^{-1} + vv^T,\end{aligned}$$

where $v = (K_S - K_{OS}^T K_O^{-1} K_{OS})^{-\frac{1}{2}} K_O^{-1} K_{OS}^T$. Then,

$$\begin{aligned}r_e(\Sigma_O) &= \frac{\text{tr}(\Sigma_O)}{\|\Sigma_O\|} \\ &= \frac{\text{tr}(K_O^{-1} + vv^T)}{\|\Sigma_O\|} \\ &= \frac{\text{tr}(K_O^{-1}) + \text{tr}(vv^T)}{\|\Sigma_O\|} \\ &= (\text{tr}(K_O^{-1}) + \text{tr}(vv^T))(\lambda_{\min}(\Sigma_O^{-1})) \\ &= (\text{tr}(K_O^{-1}) + \text{tr}(vv^T))(\lambda_{\min}(K_O - zz^T)) \\ &\leq (\text{tr}(K_O^{-1}) + \text{tr}(vv^T))(\lambda_{\min}(K_O) + \lambda_{\max}(-zz^T)) \\ &= \frac{\text{tr}(K_O^{-1}) + \|v\|^2}{\|K_O^{-1}\|} \\ &= r_e(K_O^{-1}) + \frac{\|v\|^2}{\|K_O^{-1}\|}.\end{aligned}$$

Here, we upper bounded the effective rank of Σ_O in terms of the effective rank of K_O^{-1}. The motivation for doing so is that K_O^{-1} is more tractable to analyze with respect to our key quantities, such as d and s. Recall that K_O is sparse matrix. Moreover, it is (a permutation) of a block diagonal matrix. Then, K_O^{-1} is also block diagonal and sparse.

Next, we motivate the two conditions from Theorem 1. Recall that s is the number of cliques among our supervision sources. Let $C_1, C_2, \ldots, C_s$ be the cliques that correspond to the variables $\lambda_1, \ldots, \lambda_m$, with $\sum_{j=1}^s |C_j| = m$ and $|C_j| \leq d$ for all $1 \leq j \leq s$. With slightly abuse of notation, we also refer to C_j as the corresponding submatrix in K_O^{-1}.

For our first condition, note that in general $\text{tr}(C_i) \leq |C_i|\lambda_{\max}(K_O^{-1})$. We assume that $\text{tr}(C_i) \leq \frac{1}{2}|C_i|^{\tau/2}\lambda_{\max}(K_O^{-1}) \leq \lambda_{\max}(K_O^{-1})|C_i|^{\tau/2}$. Effectively, we are assuming eigenvalue decay in each clique of sources with rate $\tau/2$; this is reasonable, since these blocks behave like the adjacency graph of a complete graph. The largest eigenvalue of such an adjacency matrix is large, but all remaining eigenvalues are small. Now, under this assumption, we have, by Holder's inequality,

$$r_e(K_O^{-1}) = \frac{\sum_{j=1}^{s}\text{tr}(C_j)}{\lambda_{\max}(K_O^{-1})} \leq \frac{1}{2}\sum_{j=1}^{s}|C_i|^{\tau/2} \leq \frac{1}{2}s^{1-\tau/2}\left(\sum_{j=1}^{s}|C_i|\right)^{\tau/2} \leq \frac{1}{2}s^{1-\tau/2}m^{\tau/2}.$$

We have the following additional requirement:

$$s \leq \frac{m^{\frac{\tau}{2-\tau}}}{((1+\tau)\log m)^{2/(2-\tau)}}.$$

This condition controls the largest number of cliques; note that taking $\tau \to 1$ allows for nearly m cliques (this is thus close to the case where all the sources are conditionally independent on the true label). Now, with a little bit of algebra, we have that

$$\begin{aligned} r_e(K_O^{-1}) &\leq \frac{1}{2}s^{1-\tau/2}m^{\tau/2} \leq \frac{m^{\tau/2}}{(1+\tau)\log m}m^{\tau/2} \\ &= \frac{1}{2}\frac{m^{\tau}}{(1+\tau)\log m}. \end{aligned}$$

We will similarly require that $\|v\|$ is bounded by the expression above (that is, $O(\frac{1}{2}m^{\tau/2}/\log(m))$, so that

$$r_e(\Sigma_O) \leq \frac{m^{\tau}}{(1+\tau)\log m}. \tag{6}$$

Next, we have the alternative *strong source block* condition. Now, we assume that one of the blocks corresponding to a clique is dominant. Concretely, if C_i is dominant, we require that (i) $\text{tr}(C_i) \geq \sum_{j\neq i}\text{tr}(C_j)$, (ii) $\lambda_{\max}(C_i) \geq \lambda_{\max}(C_j)$ for all $j \neq i$, and (iii) $\|K_{OS}\|^2 \leq 2\|(K_{OS})_{C_i}\|^2$. In the latter term, $(K_{OS})_{C_i}$ is the subvector of K_{OS} corresponding to the variables in C_i.

Under these assumptions, we show that the effective rank is bounded by a constant times d. First,

$$r_e(K_O^{-1}) = \frac{\text{tr}(K_O^{-1})}{\|K_O^{-1}\|} = \frac{\text{tr}(K_O^{-1})}{\lambda_{\max}(C_i)} \leq \frac{2\text{tr}(C_i)}{\lambda_{\max}(C_i)} \leq \frac{2d\lambda_{\max}(C_i)}{\lambda_{\max}(C_i)} = 2d.$$

Next,

$$\begin{aligned} \|v\|^2/\|K_O^{-1}\| &= \|cK_O^{-1}K_{OS}^T\|^2/\|K_O^{-1}\| \leq c^2\|K_O^{-1}\|\|K_{OS}\|^2 \leq c^2\|K_O^{-1}\|2\|(K_{OS})_{C_i}\|^2 \\ &\leq 2c^2\|K_O^{-1}\|(d\|(K_{OS})_{C_i}\|_\infty^2) \leq c'd, \end{aligned}$$

for some constant c'.

Putting these together, we have that

$$r_e(\Sigma_O) \leq r_e(K_O^{-1}) + \|v\|^2/\|K_O^{-1}\| \leq 2d + c'd = c_4 d,$$

where $c_4 = 2 + c'$, as desired.

Concentration Inequality We use a very slightly-modified version of Theorem 2.2 from Bunea & Xiao (2015). Written in our notation, it states that with probability at least $1 - \exp(-t)$,

$$\|\Sigma_O^{(n)} - \Sigma_O\| \leq c_2\|\Sigma_O\|\max\left\{\sqrt{\frac{r_e(\Sigma_O)(t+\log m)}{n}}, \frac{r_e(\Sigma_O)(t+\log m)}{n}\right\}. \tag{7}$$

First, let us show why this result applies to our setting. Let λ be the observed random vector $(\lambda_1, \ldots, \lambda_m)$ and $\bar{\lambda}$ be its mean. Then $\lambda - \bar{\lambda}$ is a centered discrete random variable, and thus bounded. Our random vector is then sub-Gaussian. Next, we note that it satisfies Assumption 1 of Bunea & Xiao (2015), the only requirement for the concentration inequality we use (the assumption is a bound on the higher-order moments of λ as a function of the second moment); again, this property follows from boundedness.

Next, to produce (7), we combine Proposition A.2 and Proposition A.4, with the only change being the replacement of the norm in Proposition A.2 with $\|v - \bar{\lambda}\|^2$; here we recall that v is our estimate of the mean. Note that the variable $v - \bar{\lambda}$ is centered and sub-Gaussian; the remainder of the proof follows.

Now, if $r_e(\Sigma_O)(t + \log m)/n \leq 1$, or, equivalently, $r_e(\Sigma_O) \leq \frac{n}{t+\log m}$, the max on the right hand side in (7) takes the first value. We can then rewrite the bound as

$$\|\Sigma_O^{(n)} - \Sigma_O\| \leq c_2 \|\Sigma_O\| \sqrt{\frac{r_e(\Sigma_O)(t + \log m)}{n}}. \tag{8}$$

On the other hand, if $r_e(\Sigma_O) > \frac{n}{t+\log m}$, we have that the second term in the max is larger, so that we obtain

$$\|\Sigma_O^{(n)} - \Sigma_O\| \leq c_2 \|\Sigma_O\| \frac{r_e(\Sigma_O)(t + \log(m))}{n}. \tag{9}$$

Remainder of the Proof Now we tackle the proof of the main theorem, which adapts the proof of Theorem 4.1 in (Wu et al., 2017). The key is to replace Lemma D.1, which states (in our notation) that, for some constant C_K,

$$\Pr\left\{\|\Sigma_O^{(n)} - \Sigma_O\| \leq C_K \sqrt{\frac{m}{n}}\right\} \geq 1 - 2\exp(-m).$$

For the source block decay (SBD) assumption, we use (8) with $t = \tau \log m$. Then, as long as $n \geq m^\tau$, it is easy to verify that the condition

$$r_e(\Sigma_O) \leq \frac{m^\tau}{\log m(m^\tau)} \leq \frac{n}{(1+\tau)\log m}$$

is met by directly applying the bound on $r_e(\Sigma_O)$ from (6). Next,

$$\begin{aligned}\|\Sigma_O^{(n)} - \Sigma_O\| &\leq c_2 \|\Sigma_O\| \sqrt{\frac{r_e(\Sigma_O)(1+\tau)\log m}{n}} \\ &\leq c_2 \|\Sigma_O\| \sqrt{\frac{m^\tau}{n}}.\end{aligned}$$

We write

$$\mathcal{F}_{\text{sbd}}(n, m, \tau) = c_2 \|\Sigma_O\| \sqrt{\frac{m^\tau}{n}} = c_2 \psi_1 \sqrt{\frac{m^\tau}{n}}.$$

Next, we work with the strong source block (SSB) condition. Again, we wish to obtain a final error probability of at least $1 - m^{-\tau}$, so that we take $t = \tau \log m$. Applying our bound on $r_e(\Sigma_O)$, the tail bound (9) becomes

$$\|\Sigma_O^{(n)} - \Sigma_O\| \leq \frac{1}{n} c_2 c_4 \|\Sigma_O\| d(1+\tau)\log(m). \tag{10}$$

We set

$$\mathcal{F}_{\text{ssb}}(n, m, d) := \frac{c_2 c_4 \psi_1 d(1+\tau)\log(m)}{n}.$$

Now that we have our two tail functions $\mathcal{F}_{\text{sbd}}(n, m, \tau)$ and $\mathcal{F}_{\text{ssb}}(n, m, d)$, we will finish off the proof by adapting the proof of Wu et al. (2017). For the first condition, we replace the tail term $\sqrt{m/n}$ with a $\sqrt{m^\tau/n}$ term, so that our require number of samples is the sublinear m^τ (rather than m). For the second condition, we replace $\sqrt{m/n}$ in Wu et al. (2017) with a $(\log m)/n$ term, which produces a sampling rate in terms of $(\log m)$ instead of m.

Concretely, we replace the $C_K\sqrt{\frac{p}{n}}$ term in the proof of Theorem 4.1 in Wu et al. (2017) with $\mathcal{F}_{\text{sbd}}(n,m,\tau)$ and $\mathcal{F}_{\text{ssb}}(n,m,d)$. In particular, for the first case, we do this replacement in the following expression for the dual norm g_γ,

$$g_\gamma(\mathcal{A}^\dagger h_{\Sigma_O^{(n)}}(R^*)) \leq m\|\Sigma_O^{(n)}R^*\| \leq \frac{\gamma^{-1}}{\psi_m}\mathcal{F}_{\text{sbd}}(n,m,\tau),$$

in the step immediately preceding (D.4). Note that here, we replace $\max\{1,\gamma^{-1}\}$ with γ^{-1}, since in our regime of interest, $\gamma^{-1} \geq 1$. Indeed, this is equivalent to requiring that along with the condition on $\mu(\Omega)\xi(zz^T)$, we have $2d \geq \xi(zz^T)$. Note also that our notation for the smallest eigenvalue is slightly different.

The term then carries forward, with the final requirement being the selection of the regularization term λ_n at the end of Step 1 of the proof. Hence, we now require that

$$\lambda_n = \frac{(3-2\nu)\gamma^{-1}}{\psi_m}\mathcal{F}(n,m,\tau)_{\text{sbd}}.$$

For the second condition, we replace the $C_K\sqrt{\frac{p}{n}}$ term with $\mathcal{F}_{\text{ssb}}(n,m,d)$. We now need that

$$\lambda_n = \frac{(3-2\nu)\gamma^{-1}}{\psi_m}\mathcal{F}(n,m,d)_{\text{ssb}}.$$

All that is left is to ensure the three conditions in the statement of Theorem 4.1 in Wu et al. (2017) are met. These conditions are (in our notation)

$$\sigma > \frac{3}{\alpha}\lambda_n,$$
$$\frac{1}{\psi_m} > \frac{3\lambda_n}{\alpha},$$

and,

$$K_{O,\min} > \frac{3\gamma}{\alpha}\lambda_n.$$

Rewriting these, we have that

$$\frac{1}{\lambda_n} > \frac{3}{\alpha}\max\left\{\frac{1}{\psi_m}, \frac{\gamma}{K_{O,\min}}, \sigma^{-1}\right\}. \tag{11}$$

For our first condition, recalling that $\gamma = \frac{\nu\alpha}{2d\beta(2-\nu)}$,

$$\begin{aligned}\lambda_n &= \frac{(3-2\nu)\gamma^{-1}}{\psi_m}\mathcal{F}(n,m,\tau)_{\text{sbd}} \\ &= \frac{2d\beta(3-2\nu)(2-\nu)}{\nu\alpha\psi_m}\mathcal{F}(n,m,\tau)_{\text{sbd}} \\ &= \frac{2dc_2\beta(3-2\nu\psi_1)(2-\nu)}{\nu\alpha\psi_m}\sqrt{\frac{m^\tau}{n}}.\end{aligned}$$

Then, plugging this into (11), we get

$$n > \left[\frac{6c_2\beta(3-2\nu)(2-\nu)\psi_1}{\nu\alpha^2\psi_m}\max\left\{\frac{1}{\psi_m}, \frac{\gamma}{K_{O,\min}}, \sigma^{-1}\right\}\right]^2 d^2m^\tau.$$

This completes the first case of the theorem.

Now, for the second case,

$$\begin{aligned}\lambda_n &= \frac{(3-2\nu)\gamma^{-1}}{\psi_m}\mathcal{F}(n,m,\tau)_{\text{ssb}} \\ &= \frac{2d\beta(3-2\nu)(2-\nu)}{\nu\alpha\psi_m}\mathcal{F}(n,m,\tau)_{\text{ssb}} \\ &= \frac{2c_2c_4\beta(3-2\nu)(2-\nu)\psi_1}{\nu\alpha\psi_m}\frac{d^2(1+\tau)\log(m)}{n}.\end{aligned}$$

Again, we plug the latter expression into (11), getting

$$n > \frac{2c_2c_4\beta(3-2\nu)(2-\nu)\psi_1}{\nu\alpha^2\psi_m}\max\left\{\frac{1}{\psi_m}, \frac{\gamma}{K_{O,\min}}, \sigma^{-1}\right\}(1+\tau)d^2\log(m).$$

Now we are done. □

Proof of Theorem 2

Proof. The typical approach taken for minimax-style lower bounds is to construct an ensemble of hypotheses (in our case, graphs encoding the distribution) and to control the distance between these hypotheses. Concretely, Fano's lemma is used, which requires controlling the KL divergence between pairs of distributions. As in prior work (Santhanam & Wainwright, 2012; Shanmugam et al., 2014), we use the symmetric KL divergence S, which is defined by

$$S(f_G||f_{G'}) = D(f_G||f_{G'}) + D(f_{G'}||f_G),$$

with

$$D(f_G||f_{G'}) = \sum_{x\in\{0,1\}^r} f_G(x)\log\left(\frac{f_G(x)}{f_{G'}(x)}\right).$$

We use the following variant of Fano's lemma (Santhanam & Wainwright, 2012)

$$n < (1-\delta)\frac{\log M}{\frac{2}{M^2}\sum_{k=1}^{M}\sum_{\ell=k+1}^{M} S(f_{G_k}||f_{G_\ell})}. \tag{12}$$

Here, we have a class of graphs $G_1, G_2, \ldots, G_M$. The result states that if n is upper bounded as in (12), *no* structure learning procedure has a better maximum error probability (over the entire family) than $\delta - \frac{1}{\log M}$. Prior work on lower bounding the sample complexity for structure learning uses multiple choices of ensemble and takes the maximum over the resulting complexities. In particular, Santhanam & Wainwright (2012) (called SW from now on), considers three ensembles. The first of these takes a graph on m nodes with no edges, and then adds one edge to form $\binom{M}{2}$ graphs. We will use a similar construction, with the additional constraint that we are in the weak supervision setting, where we have the label node Y connected to all other nodes.

We start with full generality. Note that, from our model,

$$\begin{aligned}
f_G(\lambda_1,\ldots,\lambda_m) &= \sum_y f_G(\lambda_1,\ldots,\lambda_m,y)\\
&= \sum_y \frac{1}{Z}\exp\left(\sum_{\lambda_i\in V}\theta_i\lambda_i + \sum_{(\lambda_i,\lambda_j)\in E}\theta_{ij}\lambda_i\lambda_j + \theta_Y y + \sum_{\lambda_i\in V}\theta_{Y,i}y\lambda_i\right)\\
&= \frac{1}{Z}\exp\left(\sum_{\lambda_i\in V}\theta_i\lambda_i + \sum_{(\lambda_i,\lambda_j)\in E}\theta_{ij}\lambda_i\lambda_j\right)\left[\sum_y \exp\left(\theta_Y y + \sum_{\lambda_i\in V}\theta_{Y,i}y\lambda_i\right)\right].
\end{aligned}$$

Now we can start computing the symmetric KL divergence between a pair of graphs G, G' in our class of graphs:

$$S(G||G') = \mathbb{E}_G[\log f_G - \log f_{G'}] + \mathbb{E}_{G'}[\log f_{G'} - \log f_G] \tag{13}$$

$$\begin{aligned}
&= \mathbb{E}_G\left[\sum_{\lambda_i\in V}(\theta_i-\theta_i')\lambda_i + \sum_{(\lambda_i,\lambda_j)\in E}(\theta_{ij}-\theta_{ij}')\lambda_i\lambda_j + \log\frac{\sum_y\exp(\theta_Y y + \sum_{\lambda_i\in V}\theta_{Y,i}y\lambda_i)}{\sum_y\exp(\theta_Y' y + \sum_{\lambda_i\in V}\theta_{Y,i}'y\lambda_i)}\right]\\
&\quad + \mathbb{E}_{G'}\left[\sum_{\lambda_i\in V}(\theta_i'-\theta_i)\lambda_i + \sum_{(\lambda_i,\lambda_j)\in E'}(\theta_{ij}'-\theta_{ij})\lambda_i\lambda_j + \log\frac{\sum_y\exp(\theta_Y' y + \sum_{\lambda_i\in V}\theta_{Y,i}'y\lambda_i)}{\sum_y\exp(\theta_Y y + \sum_{\lambda_i\in V}\theta_{Y,i}y\lambda_i)}\right]
\end{aligned} \tag{14}$$

Here, the partition functions cancel out going from the first line to the second.

Now we build our ensemble. Let $G^{st} = (V, E)$, with $V = \{\lambda_1, \ldots, \lambda_m, Y\}$. We set

$$E = \{(\lambda_s, \lambda_t), (\lambda_1, Y), (\lambda_2, Y), \ldots, (\lambda_m, Y)\}.$$

Note that the edges consist of the latent label node connected to all other nodes, and the sole additional edge (λ_s, λ_t).

For this class of models, we consider only edge potentials, all with parameter θ (of course, the non-edges have a parameter of 0). With this setting, for two graphs G^{st}, G^{uv}, where the edge sets are E and E', respectively, (14) reduces to,

$$\begin{aligned} S(G^{st}||G^{uv}) = \mathbb{E}_{G^{st}}\left[\sum_{(\lambda_i,\lambda_j)\in E, \notin E'} \theta\lambda_i\lambda_j - \sum_{(\lambda_i,\lambda_j)\in E', \notin E} \theta\lambda_i\lambda_j + \log\frac{\sum_y \exp(\sum_{\lambda_i\in V}\theta y\lambda_i)}{\sum_y \exp(\sum_{\lambda_i\in V}\theta y\lambda_i)}\right] \\ + \mathbb{E}_{G^{uv}}\left[\sum_{(\lambda_i,\lambda_j)\in E' \notin E} \theta\lambda_i\lambda_j - \sum_{(\lambda_i,\lambda_j)\in E, \notin E'} \theta\lambda_i\lambda_j + \log\frac{\sum_y \exp(\sum_{\lambda_i\in V}\theta y\lambda_i)}{\sum_y \exp(\sum_{\lambda_i\in V}\theta y\lambda_i)}\right] \end{aligned}$$

Note that the fraction inside the log is equal to 1—this is because there is an edge between λ_i and Y for all i. As a result, the log term is 0. Note also that there is only one edge that differs in each graph, so that the above reduces further to

$$S(G^{st}||G^{uv}) = \theta(\mathbb{E}_{G^{st}}[\lambda_s\lambda_t] - \mathbb{E}_{G^{uv}}[\lambda_s\lambda_t]) + \theta(\mathbb{E}_{G^{uv}}[\lambda_u\lambda_v] - \mathbb{E}_{G^{st}}[\lambda_u\lambda_v]).$$

By symmetry, this is simply

$$S(G^{st}||G^{uv}) = 2\theta(\mathbb{E}_{G^{st}}[\lambda_s\lambda_t] - \mathbb{E}_{G^{uv}}[\lambda_s\lambda_t]) \tag{15}$$

In the supervised case, in the above expression there is no path connecting λ_s to λ_t in G^{uv}, so that $\mathbb{E}_{G^{uv}}[\lambda_s\lambda_t] = 0$ and the result further reduces to $2\theta(\mathbb{E}_{G^{st}}[\lambda_s\lambda_t])$. In particular, a simple computation shows that this is equal to $2\theta\tanh\theta$. However, in the unsupervised case, $\mathbb{E}_{G^{uv}}[\lambda_s\lambda_t] \neq 0$, since *there is a path* between λ_s and λ_t despite the fact that in G^{uv} there is no edge joining them! The path is through the latent variable Y: $\lambda_s - Y - \lambda_t$. As a result, the $\mathbb{E}_{G^{uv}}[\lambda_s\lambda_t] > 0$ and this term *reduces* the overall distance between our graphs. In turn, this means that we require *more* samples for the weak supervision case compared to the supervised setting. We make these notions concrete in the following.

We compute $\mathbb{E}_{G^{st}}[\lambda_s\lambda_t]$ and $\mathbb{E}_{G^{uv}}[\lambda_s\lambda_t]$. This is a simple calculation. Note that since we marginalize over the latent Y, the vertices λ_w for $w \notin \{\lambda_s, \lambda_t\}$ do not contribute anything. Then, we have that

$$\mathbb{E}_{G^{st}}[\lambda_s\lambda_t] = \frac{\exp(3\theta) - \exp(-\theta)}{\exp(3\theta) + 3\exp(-\theta)},$$

and

$$\mathbb{E}_{G^{uv}}[\lambda_s\lambda_t] = \frac{\exp(2\theta) + \exp(-2\theta) - 2}{\exp(2\theta) + \exp(-2\theta) + 2} = \tanh^2(\theta),$$

A small simplification to the first term and plugging this into (15) yields

$$S(G^{st}||G^{uv}) = 2\theta(1 - 4(\exp(4\theta) + 3)^{-1} - \tanh^2(\theta)).$$

Finally, we are ready to apply Fano (12). Since we have $\binom{m}{2}$ choices for which edge to choose, and since $S(G^{st}||G^{uv})$ is the same for all choices, we obtain the bound

$$n < (1-\delta)\frac{\log(m(m-1)/2)}{2\theta(1 - 4(\exp(4\theta) + 3)^{-1} - \tanh^2(\theta))}.$$

This completes the proof. □

Conjecture on Singleton Separator Set Dense Ensemble Our proof above used a *sparse ensemble* to derive a lower bound: such ensembles use few edges between the supervision sources, and add an edge. Note that our construction satisfied the singleton separator set property. The other approach is to consider *dense ensembles*, as in the second ensemble in SW.

This particular ensemble involves dividing up the m nodes evenly into cliques of d nodes each. Then, a single edge is removed from one of these cliques. However, this ensemble *does not* satisfy the singleton separator set assumption. To see why, suppose $\lambda_1, \ldots, \lambda_d$ is such a clique, and remove the edge between λ_i, λ_j for $1 \leq i, j \leq d$. Now, we have two maximal cliques: $\{\lambda_1, \ldots, \lambda_d\} \setminus \{\lambda_i\}$ and $\{\lambda_1, \ldots, \lambda_d\} \setminus \{\lambda_j\}$. The separator set is the intersection of these two cliques: $\{\lambda_1, \ldots, \lambda_d\} \setminus \{\lambda_i, \lambda_j\}$, which is not a singleton for $d > 3$.

For the second ensemble, the idea is to ensure that $\mathbb{E}[\lambda_i \lambda_j]$ is very close to 1, e.g., at least $1 - \exp(d\theta)/d$. We conjecture that the following ensemble, which *does* satisfy the singleton separator set property, also has the same behavior. Specifically, take two complete graphs $K_{d+1} \cup K_{d+1}$. Let us name the vertices as $\{\lambda_i\} \cup \{\lambda_{d+i}\}$ for $1 \leq i \leq d$. Now, in addition to the two complete cliques, we also add cross-edges (acting as ribs), $(\lambda_i, \lambda_{d+i})$ for $1 \leq i \leq d$. Now we remove a single edge, say $(\lambda_s, \lambda_{d+s})$ to form G^s. Note that G^s does satisfy singleton separator set: the maximal cliques are the two complete subgraphs, along with each of the cross edges. The intersections here are the single nodes that connect the clique with the cross-edge.

We conjecture that using G^s as in Ensemble 2 in SW will give us a similar amount of control over $\mathbb{E}[\lambda_s, \lambda_{d+s}]$, providing us with a similar bound for the more restricted singleton separator set ensemble. Deriving such a bound would present another bounding regime, sharpening our result in Theorem 2.

Beyond Singleton Separator Sets There is a further remarkable application of robust PCA. Say we wish to perform structure learning (in the fully supervised setting, where we see Σ) by using a covariance matrix-based approach, but our graph G *does not* satisfy the singleton separator set assumption. Then, Σ^{-1} is not graph-structured, but an enlarged *generalized covariance matrix* Σ_{aug} is, where this matrix is augmented with variables in the separator set S (Loh & Wainwright, 2013). Then robust PCA can recover the structure with only Σ^{-1} as input.

More concretely, suppose that G is a graph where all the separator sets are singleton, with the exception of one set $\{\lambda_s, \lambda_t\}$. Then, the generalized covariance matrix contains all variables and an additional row corresponding to the entry $\lambda_s \lambda_t$. Note that here we observe all the λ_i's, but, as we do not know the structure, we cannot select $\lambda_s \lambda_t$ to form the generalized covariance matrix. However, if we treat this variable as *latent*, taking the role of Y in our analysis, we can use Algorithm 1. In particular, if our second condition is met, our sample complexity is again $\Omega(d^2 \log m)$, which extends the result in Loh & Wainwright (2013) to the non-singleton separator set graph class.

D. Extended Experimental Details

For Algorithm 1, cases where we have direct access to the inverse matrix Σ_O^{-1}, such as in (Ratner et al., 2019), where it is computed for the parameter estimation algorithm, the loss function term in Algorithm 1 is not needed, and we simply run robust PCA. We now provide additional comparisons to a structure inference method for weak supervision, Varma et al. (2017b). The method uses the source code that defines different weak supervision sources to infer dependencies among them by looking at what features of the data the sources rely on. This approach has an advantage over statistical methods since it does not require *any* data to infer partial structure. However, this method is unable to infer any structure for CDR and IMDb, performing up to 3.91 F1 points worse than our method. For the other tasks, our method is able to learn dependencies that Varma et al. (2017b) infers *and* additional dependencies that are implicit for the Bone Tumor and MS-COCO tasks. This leads to an improvement of up to 1.60 F1 points as shown in Table 4.

								Improvement Over	
Application	m	(s, d)	**MV**	**Indep.**	**Bach et al.**	**Partial**	**Ours**	**Bach et al.**	**Partial**
Bone Tumor	17	(2,3)	65.72	67.32	67.83	70.79	71.96	+4.13	+1.17
CDR	33	(22,14)	47.74	54.60	55.90	54.60	56.81	+0.91	+2.21
IMDb	5	(1,4)	55.21	58.80	60.23	58.80	62.71	+2.48	+3.91
MS-COCO	3	(1,2)	57.95	59.47	59.47	62.28	63.88	+4.41	+1.60

Table 4. Extended results table with comparison to the structure inference method (Partial).