Comparing the Value of Labeled and Unlabeled Data in Method-of-Moments Latent Variable Estimation

Mayee F. Chen* Benjamin Cohen-Wang* Stephen Mussmann Frederic Sala Christopher Ré
Stanford University

Abstract

Labeling data for modern machine learning is expensive and time-consuming. Latent variable models can be used to infer labels from weaker, easier-to-acquire sources operating on unlabeled data. Such models can also be trained using labeled data, presenting a key question: should a user invest in few labeled or many unlabeled points? We answer this via a framework centered on model misspecification in method-of-moments latent variable estimation. Our core result is a bias-variance decomposition of the generalization error, which shows that the unlabeled-only approach incurs additional bias under misspecification. We then introduce a correction that provably removes this bias in certain cases. We apply our decomposition framework to three scenarios—well-specified, misspecified, and corrected models—to 1) choose between labeled and unlabeled data and 2) learn from their combination. We observe theoretically and with synthetic experiments that for well-specified models, labeled points are worth a constant factor more than unlabeled points. With misspecification, however, their relative value is higher due to the additional bias but can be reduced with correction. We also apply our approach to study real-world weak supervision techniques for dataset construction.

1 Introduction

A key challenge in data-driven fields is the quality of training data. A fixed data collection budget can provide a large amount of incomplete training data or a smaller but cleaner dataset. Given a choice between these two options, which should we select and which factors should determine this decision? This fundamental question is especially relevant to modern machine learning, where vast amounts of unlabeled data is available. To exploit this without extensive hand-labeling, powerful techniques relying on *latent variable models*—in particular, *method-of-moments*—have been developed to generate labels.

Latent variable method-of-moments has been used to learn topic models (Anandkumar et al., 2014) and parse trees (Hsu et al., 2012), to evaluate crowdworkers (Joglekar et al., 2013), and to generate training datasets (Ratner et al., 2019; Fu et al., 2020). In these models, the observable outputs of *sources* are used to infer the latent variable, the true label. The core challenge is to learn correlations (i.e., *accuracies*) between the sources and the latent variable, which parametrize the model used to infer labels. To learn the accuracies, the method of moments, which relies on decomposing multiple observable statistics based on independence among sources, is commonly used to produce simple, closed-form estimators (in contrast to the EM algorithm). When some labeled data is available, this setup also allows for the accuracy parameters to be directly estimated (Figure 1). Thus, given a limited budget, a principle for choosing between labeled and unlabeled data is crucial; this motivates a theoretical framework to understand their relative value.

However, unmodeled dependencies among sources—a form of model misspecification—are common and yield inconsistent accuracy estimates, which in turn yield poor inferred labels. This affects the value of data produced with latent variable methods, so misspecification must play a role in our framework. While the question of how to analyze misspecification has been studied in classical statistics and semi-supervised learning, the focus is typically on estimator asymptotics (Kleijn and van der Vaart, 2006, 2012; Yang and Priebe, 2011). Our main challenge, however, is to understand and address misspecification for both

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

*Equal contribution. Contact: `mfchen@stanford.edu`.

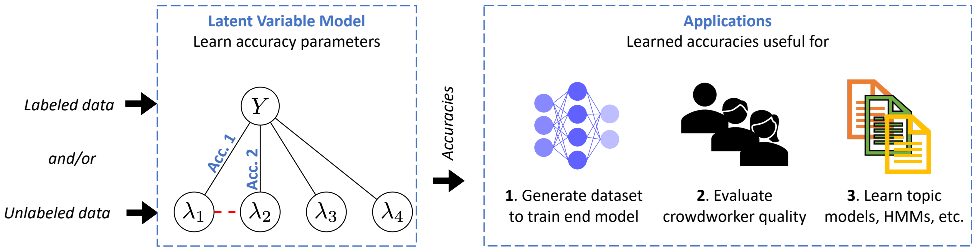

Latent Variable Model
Learn accuracy parameters
Labeled data
and/or
Unlabeled data
Y
Acc. 1
Acc. 2
λ_1
λ_2
λ_3
λ_4
Accuracies
Applications
Learned accuracies useful for
1. Generate dataset to train end model
2. Evaluate crowdworker quality
3. Learn topic models, HMMs, etc.

Figure 1: Latent variable methods (e.g., method-of-moments) can infer an unobserved variable (Y) by learning the accuracies of correlated sources ($\lambda_1, \ldots, \lambda_4$). This is done either from unlabeled data or directly from a small amount of labels; we seek a framework to explain the relative value of these choices. A major challenge are unmodeled dependencies between sources (red). Latent variable models have numerous applications.

parameter estimation and label inference in the finite sample setting.

We theoretically analyze the two alternatives in latent variable methods. In both cases, the output is a conditional distribution for the latent variable given observable sources. For the inputs, the choices are either n_L labeled and/or n_U unlabeled points (and the outputs of m sources per point). We examine misspecification in the form of unmodeled pairwise source dependencies, giving a generalization error analysis for method-of-moments latent variable model performance of the two alternatives. We present a bias-variance decomposition of the generalization error in Theorem 1, which for both the labeled and unlabeled data cases consists of (i) irreducible error, (ii) variance, and (iii) bias due to model misspecification at inference time. An important consequence is that for unlabeled data, we incur an additional (iv) standing bias due to inconsistent accuracy estimation that scales with the extent of misspecification, namely $\mathcal{O}(d/m)$ for m sources and d unmodeled dependencies among them.

Next, we turn to correcting this standing misspecification bias. In particular, a simple median-based approach is able to produce consistent estimators given $d = o(m^2)$ and sufficient amounts of unlabeled data. Therefore, in certain cases, the bias $\mathcal{O}(d/m)$ from misspecification can be completely eliminated (Proposition 1). This creates three scenarios to consider for our framework: well-specified (i.e. no unmodeled dependencies), misspecified, and corrected settings.

We give two applications of our theoretical framework for the three scenarios. First, we develop a criterion, *the data value ratio*, for choosing between labeled and unlabeled data, which is based on the relative minimum amount of labeled points needed to perform as well as a fixed amount of unlabeled points in terms of generalization error. For well-specified models, labeled data is a constant factor more valuable than unlabeled, but for misspecified models the value grows linearly in d and n_U. Furthermore, corrected models are able to *improve the value of unlabeled data*. Second, we combine the estimated parameters from the unlabeled approach, which are biased (and potentially inconsistent), with ones from the labeled approach—in certain cases outperforming either individually. We validate our framework with synthetic experiments, verify the scaling of our generalization error, data value ratio, and the performance of combined estimators across the three settings.

An important real-world application of our results on latent variable methods are weak supervision (WS) frameworks, in particular data programming (Ratner et al., 2016), used in a huge range of products and systems across industry and academia. WS frameworks construct datasets without ground-truth annotations by using unlabeled points and distant or weak sources, such as heuristics (Gupta and Manning, 2014), external knowledge bases (Mintz et al., 2009; Craven and Kumlien, 1999; Takamatsu et al., 2012), or noisy crowd-sourced labels (Karger et al., 2011; Dawid and Skene, 1979). Data programming encompasses many such prior approaches, and has shown excellent results with the method-of-moments approach (Fu et al., 2020). We perform a real-world WS case study, where ground-truth source dependencies are not known, but sources are likely to be correlated to some extent. We observe that the relative value of labeled data is large, but the value of unlabeled data can be increased via our median correction. With equal amounts of data, the F1-score of the WS model for constructing datasets with a baseline unlabeled approach is 64.81 and the score of a labeled approach is 71.79, but the score of an unlabeled approach with correction is 68.12. This suggests that our theoretical explanation of the effects of misspecification can account for some of the behavior of models on real data.

2 Related Work

Misspecification in Graphical Models The asymptotic effect of misspecification on parameter estimation is studied by Kleijn and van der Vaart (2012), extending the Bernstein-Von Mises theorem to cases where observed samples are not of the assumed parametric distribution. However, their main results do not fully extend to method-of-moments estimators. Other analyses of model misspecification directly examine distribution families, such as Jog and Loh (2015)'s lower bound on KL-separation of Gaussian graphical models. This bound is important for modeling errors in inference, but it does not illustrate our additional error in parameter estimation. More generally, works on misspecification either study a particular class of techniques (De Blasi and Walker, 2013) or a particular model and propose repairs (Grünwald et al., 2017), while we compare effects on data types.

Structure Learning One way to reduce misspecification is to produce a more refined model. Graphical model structure learning aims to do so in both the supervised (Ravikumar et al., 2011; Loh and Wainwright, 2013) and unsupervised cases (Chandrasekaran et al., 2012; Meng et al., 2014; Bach et al., 2017; Varma et al., 2019). However, these works present computational challenges, require (often strong) conditions to hold, and do not analyze the downstream impact of errors. Our approach instead focuses on understanding the impact of errors, but it is also applicable to partial recovery that often results from structure learning.

Semi-Supervised Learning involves learning from a small set of labeled points and a larger set of unlabeled points (Chapelle and Scholkopf, 2006; Zhu and Goldberg, 2009). There are several works on the relative value of labeled and unlabeled data in semi-supervised settings, typically requiring assumptions about the data distribution (e.g., cluster, manifold) (Castelli and Cover, 1996; Singh et al., 2008; Ben-David et al., 2008). In contrast, our work explicitly considers violations of model assumptions by quantifying how misspecification influences the relative value of labeled and unlabeled data. This direction has been explored by Yang and Priebe (2011), who study asymptotic performance degradation due to misspecification in semi-supervised maximum likelihood estimation; however, their results only describe the conditions under which degradation occurs. We further bound the extent of degradation, handle the finite sample case, and propose a way to mitigate misspecification.

Valuation of Data Several methods have been proposed for measuring the value of individual data points, often based on the Shapley value (Ghorbani and Zou, 2019; Jia et al., 2019). Such valuations can then be used to inform what additional data should be acquired to improve a model. Our goal in valuing labeled versus unlabeled data is similar, but we do not value individual data points and instead compare performance of classifiers trained on labeled data, unlabeled data, and on both.

3 Background and Problem Setting

We start with background on latent variable models and introduce the model we analyze. We explain the two stages—learning accuracies and inferring labels—for both the labeled and unlabeled cases, and conclude with how to evaluate the model.

Setup In latent variable models, a number of sources are observed and used to infer the latent variable. The input is usually n_U unlabeled data points, but in our setting we also consider a small *labeled* dataset of n_L samples. The output is a large, labeled dataset.

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y} = \{-1, 1\}$. We consider an unlabeled dataset $\boldsymbol{X}_U = \{x_i^U\}_{i=1}^{n_U}$ and a labeled dataset $(\boldsymbol{X}_L, \boldsymbol{Y}_L) = \{(x_i^L, y_i^L)\}_{i=1}^{n_L}$ drawn from the distribution of (X, Y). There are m sources, each outputting a value in $\{-1, +1\}$ via a deterministic function $\lambda_j : \mathcal{X} \to \mathcal{Y}$ for all $j \in [m]$. Our goal is to use the outputs of $\boldsymbol{\lambda}$, the vector of sources, to construct a model to infer Y.

To infer Y, we learn the model $\Pr(Y|\boldsymbol{\lambda})$ to produce soft labels $\widetilde{y}_i := 2\Pr(Y = 1|\boldsymbol{\lambda} = \boldsymbol{\lambda}(x_i)) - 1 \in [-1, 1]$ for each x_i by applying the m sources' functions to the datasets $\boldsymbol{X}_U$ or $(\boldsymbol{X}_L, \boldsymbol{Y}_L)$. The overall approach has two steps: (i) learn the latent variable model (using labeled or unlabeled data), and (ii) infer labels $\widetilde{y}_i$.

Theoretical model We pick a simple model that captures many latent variable model settings and still presents all of the challenges for comparing between the types of data. We assume an Ising model for $\Pr(Y, \boldsymbol{\lambda})$; the only difference between the labeled and unlabeled setting is that Y is latent in the latter. The dependency graph is $G = (V, E)$, where $V = Y \cup \boldsymbol{\lambda}$ and E consists of edges from Y to the sources as well as the d edges among the sources, E_λ. The lack of an edge in G between a pair of variables indicates independence conditioned on a separator set (Lauritzen, 1996), so the true distribution can be modeled as

$$\Pr(Y, \boldsymbol{\lambda}; \theta) = \frac{1}{Z} \exp\Big(\theta_Y + \sum_{i=1}^{m} \theta_i \lambda_i Y + \sum_{(i,j) \in E_\lambda} \theta_{ij} \lambda_i \lambda_j\Big),$$

with cumulant function Z and the set of canonical parameters $\theta = \{\theta_Y, \theta_i\ \forall i, \theta_{ij}\ \forall (i,j) \in E_\lambda\}$. For cleaner presentation, we assume $\theta \geq 0$ (no sources that disagree with others or Y on average) and E_λ is sparse enough such that $\deg(\lambda_i) \leq 2$ for all λ_i (each source is conditionally dependent on at most one other source).

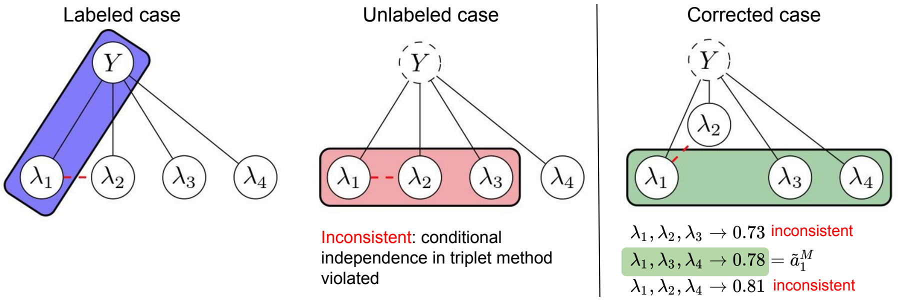

Labeled case
Unlabeled case
Corrected case
Y
λ_1
λ_2
λ_3
λ_4
Inconsistent: conditional independence in triplet method violated
$\lambda_1, \lambda_2, \lambda_3 \rightarrow 0.73$ inconsistent
$\lambda_1, \lambda_3, \lambda_4 \rightarrow 0.78 = \tilde{a}_1^M$
$\lambda_1, \lambda_2, \lambda_4 \rightarrow 0.81$ inconsistent

Figure 2: Estimating accuracy parameter of λ_1 with unmodeled dependency (red edge), leading to misspecification. Boxes indicate observable variables used for accuracy estimation. Left: model with access to label data. The accuracy parameter is directly estimated using Y and λ_1 and is not impacted by the unmodeled dependency. Center: latent model with unlabeled data and unobserved Y. The boxed triplet includes the unmodeled dependency, leading to inconsistent estimate $\widetilde{a}_1^U$. Right: Corrected model using medians. The boxed triplet, chosen as the median estimate among $\binom{m-1}{2}$ triplets, excludes the dependency, yielding a consistent $\widetilde{a}_1^M$.

Inference The label is computed using a naive Bayes approach that assumes all sources are conditionally independent with $E_\lambda = \emptyset$:

$$\begin{aligned}\widetilde{\Pr}(Y = 1 | \lambda = \lambda(X)) \\ = \frac{\prod_{i=1}^m \widetilde{\Pr}(\lambda_i = \lambda_i(X) | Y = 1) \Pr(Y = 1)}{\hat{\Pr}(\lambda = \lambda(X))}, \quad (1)\end{aligned}$$

where the class balance $\Pr(Y = 1)$ is assumed to be known, $\hat{\Pr}$ is an empirical probability , and $\widetilde{\Pr}$ indicates an estimated probability resulting from the parameter estimation step described below. In practice, the conditional independence assumptions required for (1) may not hold, but dependencies among sources are often unknown. Therefore, conditional independence is assumed, and we may suffer from misspecification in inferring our probabilistic labels.

Learning parameters with method-of-moments For the labeled dataset, we learn $\widetilde{\Pr}(\lambda_i = \lambda_i(X) | Y = 1)$ in (1) directly from samples, as Y is observed.

For the unlabeled dataset, we use the method-of-moments estimator from Fu et al. (2020) (described in Appendix B), which relies on the property that if $\lambda_i \perp\!\!\!\perp \lambda_j | Y$, then $\lambda_i Y \perp\!\!\!\perp \lambda_j Y$. This implies that $\mathbb{E}[\lambda_i Y] \cdot \mathbb{E}[\lambda_j Y] = \mathbb{E}\left[\lambda_i \lambda_j Y^2\right] = \mathbb{E}[\lambda_i \lambda_j]$, which is directly estimable. Define $a_i := \mathbb{E}[\lambda_i Y]$ as the unknown *accuracy* of λ_i. If we can introduce a third λ_k that is conditionally independent of λ_i and λ_j, we have a system of equations that can be solved using observable statistics. We use this *triplet method* to recover these accuracies: we choose two λ_j, λ_k at random for each λ_i and solve up to sign:

$$|\widetilde{a}_i^{(j,k)}| := \sqrt{\left| \frac{\hat{\mathbb{E}}[\lambda_i \lambda_j] \hat{\mathbb{E}}[\lambda_i \lambda_k]}{\hat{\mathbb{E}}[\lambda_j \lambda_k]} \right|}, \quad (2)$$

where $\hat{\mathbb{E}}$ is an empirical estimate of the expectation.

We use the estimated $\widetilde{a}_i^U := \widetilde{a}_i^{(j,k)}$ to directly compute $\widetilde{\Pr}(\lambda_i = \pm 1 | Y = 1)$ for (1). However, random λ_j and λ_k may not satisfy conditional independence, and thus we incur error in estimating accuracies due to misspecification in a way unique to the unlabeled setting. Figure 2 (left, center) describes how this misspecification impacts learning the accuracies in the labeled versus unlabeled cases. We aim to capture the role of this misspecification in our evaluation.

Evaluating the model We define the model's generalization error as $R = \mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}[l(\widetilde{Y}, Y)]$ where expectation is taken over the distribution of $(Y, \boldsymbol{\lambda})$, $\mathcal{N}$ (the random dataset used), and τ (the algorithmic randomness if applicable, i.e. the triplets used in method-of-moments). $l(\cdot, \cdot)$ here is the cross entropy loss, $l(\widetilde{y}_i, y_i) = -\frac{1+y_i}{2} \log \widetilde{\Pr}(Y = 1 | \boldsymbol{\lambda} = \boldsymbol{\lambda}(x_i)) - \frac{1-y_i}{2} \log \widetilde{\Pr}(Y = -1 | \boldsymbol{\lambda} = \boldsymbol{\lambda}(x_i))$. Let R_U denote the error for the unlabeled dataset and R_L for labeled.

4 Theoretical Results

We theoretically analyze the quality of the latent variable model, taking into account the impact of misspecification when using unlabeled versus labeled data. In Section 4.1 we give an exact decomposition of the generalization error of the latent variable model, which demonstrates how misspecification is present in both the parameter learning and inference steps of the model when data is unlabeled and only present in the latter when data is labeled. In Section 4.2, we bound the generalization error using this framework to show how the unlabeled case has an additional standing bias of $\mathcal{O}(d/m)$. Given this standing bias, in Section 4.3 we introduce a simple method that can in some cases correct for dependency-based misspecification, and we analyze its impact on generalization error. In 4.4 we present synthetic experiments that verify our results.

4.1 Decomposition Framework

Our first result is a decomposition of the generalization error into four components. The last two components, the inference bias and parameter estimation error, reflect the role of misspecification.

Theorem 1. *The generalization error has the following decomposition:*

$$\mathbb{E}\left[l(\widetilde{Y},Y)\right] = \underbrace{H(Y|\boldsymbol{\lambda})}_{\text{Irreducible error}} - \underbrace{\mathbb{E}_{\mathcal{N}}\left[D_{\mathrm{KL}}(\Pr(\boldsymbol{\lambda})||\hat{\Pr}(\boldsymbol{\lambda}))\right]}_{\text{Observable sampling noise}} +$$

$$\sum_{(i,j)\in E_\lambda} \underbrace{I(\lambda_i;\lambda_j|Y)}_{\text{Inference bias}} + \sum_{i=1}^{m} \underbrace{\mathbb{E}_{Y,\mathcal{N},\tau}\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right]}_{\text{Parameter estimation error}},$$

where $I(\lambda_i;\lambda_j|Y)$ is the conditional mutual information between sources and $H(Y|\boldsymbol{\lambda})$ is conditional entropy. $\Pr$ *refers to the true data distribution, while* $\hat{\Pr}$ *and* $\widetilde{\Pr}$ *refer to the estimated probabilities in* (1)*.*

We now discuss each term above. The first two terms are independent of misspecification and are present in both the unlabeled and labeled cases:

- Irreducible error: an intrinsic property of the distribution of $(Y,\boldsymbol{\lambda})$, always present in bias-variance decomposition.
- Observable sampling noise: the expected KL divergence between the true marginal distribution of the observable sources and the empirical distribution. Particular to our inference approach, it is a common notion of sampling noise (Domingos, 2000; Yang et al., 2020) and approaches 0 asymptotically.

For the last two terms, misspecification plays a different role depending on the data type.

- Inference bias: the conditional mutual information among dependent sources. Particular to our inference approach, it is the approximation error of using marginal singleton probabilities rather than their product distributions. Therefore, it represents the role of misspecification at the inference step (1) and is present for both data types. It is independent of parameter estimation method.
- Parameter estimation error: the difference between the true and estimated distribution of $\lambda_i|Y$. For the labeled approach, this error corresponds to sampling noise and asymptotically approaches 0. For the unlabeled approach, it directly depends on the estimation error of accuracies in (2). However, these estimators are biased, as are many method-of-moments approaches. Furthermore, misspecification makes the estimators inconsistent when λ_i, λ_j, and λ_k used to produce $\widetilde{a}_i^{(j,k)}$ are not pairwise conditionally independent.

We now discuss in detail the scaling of these last two terms, which highlights the tradeoff between labeled and unlabeled data under misspecification.

4.2 Scaling of the Generalization Error

We bound the terms in Theorem 1 to understand the scaling of error due to misspecification in both the unlabeled and labeled cases. Since the irreducible error is always present, we bound *excess generalization error*, defined as $R_L^e = R_L - H(Y|\boldsymbol{\lambda})$ for labeled data and similarly R_U^e for unlabeled data. We use $\mathcal{B}_I = \sum I(\lambda_i;\lambda_j|Y)$ for the inference bias in these bounds since it is independent of our two cases, and while it scales in d, it is simply a measurement over the true data distribution. We present upper bounds here and lower asymptotic bounds in Appendix C.3.

We first bound R_L^e.

Theorem 2. *Suppose that there are $|E_\lambda| = d$ unmodeled dependencies. When we use the latent variable model described in section 3 with n_L labeled samples,*

$$R_L^e \leq \frac{m}{2n_L} + \mathcal{B}_I + o(1/n_L). \tag{3}$$

In this bound, $\frac{m}{2n_L}$ is an upper bound on parameter estimation error. It represents the sampling noise of $\widetilde{a}_i^L = \hat{\mathbb{E}}\left[\lambda_i Y\right]$, which asymptotically approaches 0. Therefore, the only standing bias is $\mathcal{B}_I$ due to inference approach. When the model is well-specified, the excess error is $\mathcal{O}(1/n_L)$, and thus for large n_L our generated labels eventually follow the true distribution $\Pr(Y|\boldsymbol{\lambda})$.

We next present an upper bound on the excess generalization error in the unlabeled case. Define $\varepsilon_{ij} = \mathbb{E}\left[\lambda_i\lambda_j\right] - \mathbb{E}\left[\lambda_i Y\right]\mathbb{E}\left[\lambda_j Y\right]$ as the extent of misspecification on a single pair of sources, and let $0 \leq \varepsilon_{\min} \leq \varepsilon_{ij} \leq \varepsilon_{\max}$ for all pairs (i,j) under our model assumptions in section 3. The exact value of ε_{ij} in terms of canonical parameters is in Appendix D.3.

Theorem 3. *Suppose that there are $|E_\lambda| = d$ dependencies. When we use the latent variable model described in section 3 using n_U unlabeled samples,*

$$\begin{aligned} R_U^e \leq & \varepsilon_{\max}\left(\frac{c_1 d}{m} + \frac{c_2}{\sqrt{n_U}} + \frac{c_3 d}{m n_U}\right) \\ & + \frac{c_4 m}{n_U} + \mathcal{B}_I + o(1/n_U), \end{aligned} \tag{4}$$

where c_1, c_2, c_3, and c_4 are constants depending on the intrinsic quality of the sources (Appendix D.3).

In this bound, we again have an observable sampling noise $\frac{c_4 m}{n_U}$, where the the constant term comes from estimating $\hat{\mathbb{E}}\left[\lambda_i\lambda_j\right]$ in (2) rather than $\hat{\mathbb{E}}\left[\lambda_i Y\right]$ in the labeled approach. However, here the parameter estimation error has an additional term $\mathcal{B}_{\text{est}} :=$

$\varepsilon_{\max}\left(\frac{c_1 d}{m} + \frac{c_2}{\sqrt{n_U}} + \frac{c_3 d}{m n_U}\right)$ which depends on misspecification. Therefore, asymptotically the unlabeled approach has a standing bias bounded by $\frac{c_1 d \varepsilon_{\max}}{m} + \mathcal{B}_I$ in comparison to the labeled case's $\mathcal{B}_I$, and the finite-sample regime contributes additional sampling noise for the unlabeled approach that scales in $\varepsilon_{\max}$. In the case the model is well-specified ($d = 0, \varepsilon_{\max} = 0$), the only term present is $\frac{c_4 m}{n_U}$, so our latent variable model would also approach the true distribution of $\Pr(Y|\boldsymbol{\lambda})$ but at a different rate than the labeled case.

Partial Recovery Our results hold almost exactly for the partial recovery case, where d' out of d dependencies are recovered (e.g. via structure learning) and our method in (2) avoids choosing known pairs of dependent sources. In particular, the additional estimation error now scales at rate $\frac{(d-d')\varepsilon_{\max}}{m-2d'}$.

4.3 Correcting for misspecification

How can we reduce the penalty for dealing with such unrecovered dependencies? We examine how to reduce misspecification for our estimator described in (2). Our correction can be applied to other method-of-moments approaches (Anandkumar et al., 2012; Chaganty and Liang, 2014), discussed in Appendix C.1.

In our estimation approach, if there exists an λ_i such that there are no λ_j, λ_k where all three sources are pairwise conditionally independent given Y, then it is not possible to learn a_i. In less demanding cases, we suggest an alternative approach based on *medians.* Recall that misspecification impacts accuracy estimation error because random triplets that violate pairwise conditional independence are selected to compute our $\widetilde{a}_i^U$. To reduce this impact, we estimate each a_i by computing the median accuracy over all pairs λ_j, λ_k using (2) a total of $\binom{m-1}{2}$ times, as shown in Figure 2 (right). The intuition behind this approach is that inconsistent estimates produced by dependent sources have more extreme values and thus may not impact the median.

Proposition 1. *Let* $\widetilde{a}_i^M = \text{median}(\{\widetilde{a}_i^{(j,k)}\ \forall\ j, k \neq i\})$. *Then* $\widetilde{a}_i^M$ *is not affected by misspecification and is thus a consistent estimator if* $m > 5$, $d < \frac{(m-1)(m-2)}{4}$, *and* $n_U \geq n_0$, *where* n_0 *is* $\omega(1/\varepsilon_{\min}^2)$.

Refer to $\rho_{n_U} = \max_i \mathbb{E}\left[(\widetilde{a}_i^M - a_i)^2\right]$ *as the maximum MSE for* $\widetilde{a}^M$. *Under these conditions, the excess generalization error* R_M^e *from using* n_U *unlabeled samples and a corrected model is, for constant* c_ρ,

$$R_M^e \leq c_\rho m \rho_{n_U} + \mathcal{B}_I + o(1/n_U). \qquad (5)$$

While ρ can be analyzed in detail as a variant of a medians-of-means estimator, we stress that $\lim_{n_U \to \infty} \rho_{n_U} = 0$. Thus the standing bias of order $\mathcal{O}(d/m)$ due to misspecification can be eliminated.

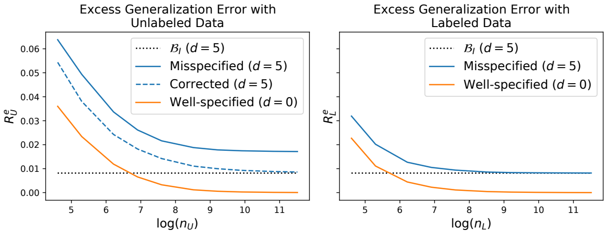


Figure 3: Excess generalization error vs. $\log(n)$ with different estimators for synthetic data. Left: comparison of unlabeled data performance under the three discussed settings. Right: comparison of labeled data performance for well-specified and misspecified models. A dashed line repesenting an empirical "$\mathcal{B}_I$" suggests how inference bias is present in both data cases.

This reduction has many implications for the value of labeled vs. unlabeled data in *corrected settings.*

4.4 Synthetic Experiments

We validate the fundamental principles of our theoretical framework using synthetic data. We measure the excess generalization error vs. $\log(n)$ in the well-specified, misspecified and corrected settings on synthetic data with $m = 10$ sources, accuracies drawn uniformly from $[.55, .75]$ and extent of misspecification fixed at $\varepsilon = 0.1$. To approximate expected excess generalization error for each n, we average results over 1000 samples. A more detailed protocol for synthetic experiments is available in Appendix F.1.

Our results are in Figure 3. With no misspecification ($d = 0$) the labeled and unlabeled estimators both tend towards zero in the two graphs. Under misspecification ($d = 5$), we see that learning from unlabeled data results in an additional standing bias that parallels $\mathcal{B}_{\text{est}}$. Median aggregation reduces this bias and results in error converging to roughly similar values, paralleling $\mathcal{B}_I$, in both the unlabeled and labeled cases in the two graphs. These observations are consistent with our theoretical findings.

5 Applications

Based on our generalization error framework, we now have a rigorous way to analyze misspecification in latent variable models. We examine two practical applications of our theoretical results in three settings—well-specified, misspecified, and corrected:

- **Understanding the value of labeled data:** we address our motivating question about the value of labeled data—-is a few labeled samples or many unlabeled samples better? This decision varies per setting, depending on the misspecification parameters (d, $\varepsilon_{\max}$), and n_U versus n_L.

- **Combining labeled and unlabeled data:** we show how simple linear combinations of the estimators can improve generalization error bounds over using one or the other. We also suggest a James-Stein type estimator from Green et al. (2005), which combines an unbiased estimator with biased information, to easily determine the weights of the linear combination.

We extend our upper bounds on the decomposition in Theorem 1 to these two applications of our framework, presenting theoretical results first and then verifying our results on synthetic data. In Appendix C.3, we comment on how lower bounds can be obtained and used for similar analysis as an avenue for future work.

5.1 Understanding the value of labeled data

We use our analysis from Section 4.2 to develop a criterion for deciding between labeled and unlabeled points. Compute

$$f(n_U) = \min_{n_L \in \mathbb{N}} \text{ s.t. } R^e_L(n_L) \leq R^e_U(n_U),$$

and define $V(n_U) = n_U / f(n_U)$ to be the *data value ratio.* The intuitive idea here is to compare, for some amount of unlabeled data n_U, what factor less labeled data we would require to produce an equivalent error bound. We consider an approximation of the data value ratio $\widetilde{V}(n_U)$ based on our upper bounds for excess generalization error. We examine the differences in $\widetilde{V}(n_U)$ for our three aforementioned settings:

- Well-specified setting: comparing excess risk when $d = 0$ and $\varepsilon_{\max} = 0$ reduces to examining $\frac{m}{2n_L}$ and $\frac{c_4 m}{n_U}$. Thus $\widetilde{V}(n_U) = 2c_4$ and our framework suggests that labeled data is only a *constant factor* more beneficial than unlabeled data.
- Misspecified setting: $\widetilde{V}(n_U)$ will capture the tradeoff between $\frac{m}{2n_L}$ and $\mathcal{B}_{\text{est}} + \frac{c_4 m}{n_U}$. We find that $\widetilde{V}(n_U) = 2\varepsilon_{\max}\left(\frac{c_1 d n_U}{m} + \frac{c_2\sqrt{n_U}}{m} + \frac{c_3 d}{m^2}\right) + 2c_4$. That is, the value of labeled data *increases linearly in the amount of unlabeled data and misspecification* due to the standing bias in the generalization error for the unlabeled approach.
- Corrected setting: under our conditions from Proposition 1, we examine the difference between $\frac{m}{2n_L}$ and $c_\rho m \rho_{n_U}$, and thus $\widetilde{V}(n_U) = 2n_U c_\rho \rho_{n_U}$. Since ρ_{n_U} converges to 0, $\widetilde{V}(n_U)$ is sublinear in n_U, showing that the *corrected model increases the relative value of unlabeled data.*

Synthetic Experiments We measure $V(n_U)$ in well-specified, misspecified and corrected settings on synthetic data with the same setup as discussed in 4.4. Our detailed protocol for approximating $V(n_U)$ is in Appendix F.1.

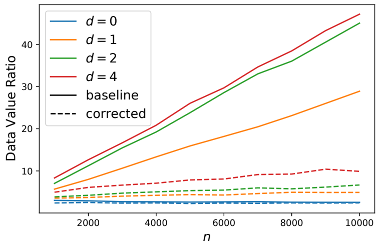


Figure 4: Data value ratio vs. n, using both the standard method-of-moments approach and the corrected approach, which aggregates results over triplets using medians. Note that $d = 0$ represents the well-specified setting.

We present the results in Figure 4. In the well-specified case ($d = 0$), $V(n)$ is small (less than 5) and roughly constant across n. Under misspecification however, the data value ratio grows with both d and n albeit much more slowly for the corrected setting, aligning with our theoretical findings.

5.2 Combining labeled and unlabeled data

While we now have a criterion to choose between datasets, how do we combine information from both? We examine ways to combine the accuracy parameters, namely $\widetilde{a}^U$ as defined in (2) for unlabeled data and an equivalent $\widetilde{a}^L := \hat{\mathbb{E}}\left[\boldsymbol{\lambda} Y\right]$ for labeled data. Recall that $\widetilde{a}^L$ is unbiased, while $\widetilde{a}^U$ is both biased and inconsistent if not corrected.

First, we consider a simple linear combination, $a^{\text{lin}}(\alpha) = \alpha\widetilde{a}^U + (1-\alpha)\widetilde{a}^L$ for some weight $\alpha \in [0,1]$. Using our framework in Theorem 1, we can derive similar upper bounds on excess generalization error when the estimator is $a^{\text{lin}}(\alpha)$. We summarize our findings across the three settings below, where we consider $\alpha\widetilde{a}^M + (1-\alpha)\widetilde{a}^L$ for the corrected setting.

- Well-specified setting: the upper bound on excess generalization error using a^{lin}, ignoring $\mathcal{B}_I$ and lower order terms, is $\alpha^2 \frac{c_4 m}{n_U} + (1-\alpha)^2 \frac{m}{2n_L}$. One can easily verify that there exists an $\alpha \in (0,1)$ that minimizes this upper bound. Since n_U is usually much larger than n_L, plugging in this optimal α shows that this new upper bound is roughly of the same order as the unlabeled case.
- Misspecified setting: the upper bound is a cubic polynomial in α. We find that the standing bias results in the optimal α weighting the labeled data's estimator more. This suggests that a combined estimator can yield an upper bound much smaller than that for the unlabeled case.
- Corrected setting: the upper bound now consists of $\alpha^2 c_\rho m \rho_{n_U} + (1-\alpha)^2 \frac{m}{2n_L}$. As a function of α, this differs from the well-specified setting's expression only in constant coefficients, so this again

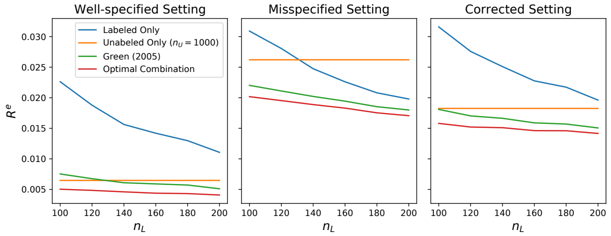


Figure 5: Excess generalization error for an optimally weighted combination of labeled and unlabeled estimators, and a combination weighted according to Green et al. (2005) across the well-specified (left), misspecified (center), and corrected (right) settings. The number of unlabeled points is fixed at $n_U = 1000$.

suggests an optimal $\alpha \in (0, 1)$ and performance roughly similar to the unlabeled case.

In practice, we do not know the exact α that optimizes generalization error. However, there is vast literature on combined estimators that dominate the MLE estimator $\widetilde{a}^L$. In particular, we suggest using an approach from Green et al. (2005), who propose a way of setting α given knowledge of an unbiased estimator with biased information.

Synthetic Experiments We investigate the empirical performance of estimators which combine labeled and unlabeled data in well-specified, misspecified and corrected settings. We measure both the error when using the fine-tuned α and the more practical approach of Green et al. (2005). We fix $n_U = 1000$ and vary n_L across a range of smaller values, aligning with the assumption that many more unlabeled than labeled points are typically available. Our results are in Figure 5. In the well-specified setting, the combined estimators perform roughly the same as just $\widetilde{a}^U$, matching up with our theoretical observations for large n_U. In the misspecified setting, both combined estimators result in lower excess risk than either estimator individually, and as n_L increases, the labeled estimator curve approaches those of the combined estimators, suggesting that the weight on $\widetilde{a}^L$ increases as more labeled data becomes available. Lastly, in the corrected setting both combined estimators perform better than $\widetilde{a}^U$, but not by much. The weights α are reported in Appendix F.1. The optimal weights for the well-specified and corrected settings are higher (i.e. more weight on the unlabeled estimator) than the misspecified setting, and these weights decrease with n_L.

6 Real-World Case Study: Weak Supervision

We validate our findings on real-world weak supervision dataset. Unlike our theoretical setting where we limit the number of dependencies d for simplicity, with real-world data we anticipate many small dependencies which cannot be completely corrected by the medians approach. We seek to answer the following key questions.

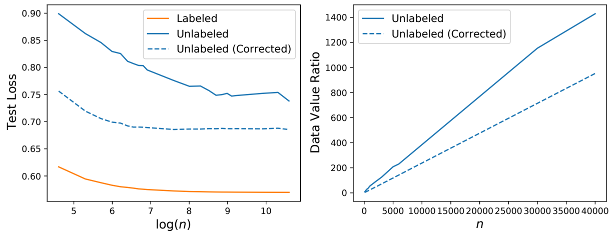
- What is the standing parameter estimation bias due to misspecification? To what extent does the corrected estimator, which only addresses unmodeled source dependencies, mitigate this bias?
- What is the data value ratio for misspecified and corrected settings?
- Can a combined estimator with access to a small amount of labeled data provide substantial benefits over using only unlabeled data?

Protocol Our real-world task is the sentiment analysis task of determining whether IMDB movie reviews are positive or negative (Maas et al., 2011). The dataset contains 50K movie reviews, which we split into a training set of 40K reviews and a test set of 10K reviews. Our weak supervision sources are simple heuristics that vote "yes" when positive words appear and "no" when negative words appear. We provide further details in Appendix F.2.

Unlike our theoretical model, where we assume that each source has a single accuracy parameter, we find that real-world sources have complex dependencies and can be better modeled with *class conditional* accuracies. The method-of-moments approach in this setting results in a quadratic version of the triplet method (Fu et al., 2020), the details of which we discuss in Appendix C.1. We use this version for our real-world case study, for which the same principles from our theoretical framework apply.

Standing bias and correction For our first real-world experiment, we measure the standing parameter estimation bias when learning from unlabeled data (paralleling $\mathcal{B}_{\text{est}}$), and measure the decrease in bias when using a corrected estimator. We compute the test cross entropy loss for a labeled model, a baseline unlabeled and an unlabeled model with correction while varying n and report results in Figure 6 (left, bottom). Losses appear to converge, with a large gap between the labeled and unlabeled models and a smaller gap between the labeled model and the unlabeled model with correction. These gaps in loss are reflected by gaps in F1-scores, computed using a threshold of .5.

Measuring the value of labeled data Next, we measure the data value ratio in the real-world setting. Since both the unlabeled model and the unlabeled model with correction have a standing bias compared to the labeled model, we anticipate that the data value ratio for both unlabeled approaches grows with



Model	Loss ($n = 40$K)	F1 ($n = 40$K)
Labeled	.570	71.79
Unlabeled	.740	64.81
Corrected	.686	68.12

Figure 6: We measure test losses and F1-scores for labeled, unlabeled and corrected models on the IMDB dataset. Top Left: losses vs. n; each model appears to flatten out by $n = 40,000$. Bottom: losses and F1-scores at $n = 40,000$, showing standing gaps in performance. Top Right: data value ratios for the two unlabeled models.

n_U	n_L	$\text{F1}_{\text{Unlabeled}}$	$\text{F1}_{\text{Labeled}}$	$\text{F1}_{\text{Combined}}$
40,000	40	68.12	64.70	67.06
40,000	80	68.12	67.65	68.81
40,000	120	68.12	68.92	69.64
40,000	200	68.12	69.97	70.41
40,000	400	68.12	70.81	71.04

Table 1: F1-scores for unlabeled, labeled and combined approaches on the IMDB dataset. We find that the combination generally outperforms either approach individually, and in particular both in cases where unlabeled only performs better and where labeled only performs better.

n, with the data value ratio for the baseline unlabeled model being higher. We report these results in Figure 6 (right).

Combining labeled and unlabeled data We finally measure the performance of the combined estimator from Green et al. (2005) in the setting where a small number of labeled points and many unlabeled points are available. We let $n_U = 40,000$ be the entire training set and vary n_L between 40 and 400. We use the corrected estimator for learning from unlabeled data. We report the F1-score using a threshold of .5. Results are in Table 1. We observe that the combined estimator outperforms either approach individually for $n_L > 40$.

7 Conclusion

Motivated by the practical tradeoff between acquiring large unlabeled datasets and small labeled datasets, we introduce a framework that aims to provide theoretically-grounded reasoning for using labeled versus unlabeled data in latent variable graphical models. We present three main technical contributions in this paper: a) a finite-sample decomposition for generalization error with labeled vs unlabeled input, focused on model misspecification; b) a correction approach for method-of-moments to reduce the impact of model misspecification; c) applications of this decomposition framework and correction, namely how to choose and combine the two data types. We show theoretically and validate empirically that labeled data is more valuable when models are misspecified, since learning from unlabeled data relies more heavily on structural assumptions that may be violated. Simple algorithmic corrections, however, can significantly improve the relative value of unlabeled data.

Acknowledgments

We gratefully acknowledge the support of NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and 1937301 (RTML); ONR under No. N000141712266 (Unifying Weak Supervision); the Moore Foundation, NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, the Okawa Foundation, American Family Insurance, Google Cloud, Swiss Re, Total, the HAI-AWS Cloud Credits for Research program, the Stanford Data Science Initiative (SDSI), and members of the Stanford DAWN project: Facebook, Google, and VMWare. The Mobilize Center is a Biomedical Technology Resource Center, funded by the NIH National Institute of Biomedical Imaging and Bioengineering through Grant P41EB027060. 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 NIH, ONR, or the U.S. Government.

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832.

Anandkumar, A., Hsu, D., and Kakade, S. M. (2012). A method of moments for mixture models and hidden markov models. volume 23 of *Proceedings of Machine Learning Research*, pages 33.1–33.34, Edinburgh, Scotland. JMLR Workshop and Conference Proceedings.

Bach, S. H., He, B., Ratner, A., and Ré, C. (2017).

Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 273–282. JMLR. org.

Ben-David, S., Lu, T., and Pál, D. (2008). Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *COLT*, pages 33–44.

Castelli, V. and Cover, T. M. (1996). The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on information theory*, 42(6):2102–2117.

Chaganty, A. T. and Liang, P. (2014). Estimating latent-variable graphical models using moments and likelihoods. In *International Conference on Machine Learning*, pages 1872–1880.

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

Chapelle, O.and Zien, A. and Scholkopf, B. (2006). *Semi-supervised learning*. MIT press.

Craven, M. and Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *International Conference on Intelligent Systems for Molecular Biology (ISMB)*.

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

De Blasi, P. and Walker, S. G. (2013). Bayesian asymptotics with misspecified models. *Statistica Sinica*, pages 169–187.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning*, pages 231–238.

Fu, D. Y., Chen, M. F., Sala, F., Hooper, S. M., Fatahalian, K., and Ré, C. (2020). Fast and three-rious: Speeding up weak supervision with triplet methods. In *Proceedings of the 37st International Conference on Machine Learning (ICML 2020)*.

Ghorbani, A. and Zou, J. (2019). Data shapley: Equitable valuation of data for machine learning.

Green, E. J., Strawderman, W. E., Amateis, R. L., and Reams, G. A. (2005). Improved Estimation for Multiple Means with Heterogeneous Variances. *Forest Science*, 51(1):1–6.

Grünwald, P., Van Ommen, T., et al. (2017). Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

Gupta, S. and Manning, C. D. (2014). Improved pattern learning for bootstrapped entity extraction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 98–108.

Hsu, D., Kakade, S. M., and Liang, P. (2012). Identifiability and unmixing of latent parse trees. In *Advances in Neural Information Processing Systems, (NIPS 2012)*.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., and Spanos, C. J. (2019). Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR.

Jog, V. and Loh, P. (2015). On model misspecification and KL separation for gaussian graphical models. *CoRR*, abs/1501.02320.

Joglekar, M., Garcia-Molina, H., and Parameswaran, A. (2013). Evaluating the crowd with confidence. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 686–694.

Karger, D. R., Oh, S., and Shah, D. (2011). Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pages 1953–1961.

Kleijn, B. and van der Vaart, A. (2012). The bernstein-von-mises theorem under misspecification. *Electron. J. Statist.*, 6:354–381.

Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional bayesian statistics. *Ann. Statist.*, 34(2):837–877.

Lauritzen, S. (1996). *Graphical Models*. Clarendon Press.

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

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

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

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings 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: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Ratner, A., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., and Ré, C. (2019). Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4763–4771.

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., and Ré, C. (2016). Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*, pages 3567–3575.

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

Singh, A., Nowak, R., and Zhu, J. (2008). Unlabeled data: Now it helps, now it doesn't. *Advances in neural information processing systems*, 21:1513–1520.

Takamatsu, S., Sato, I., and Nakagawa, H. (2012). Reducing wrong labels in distant supervision for relation extraction. In *Meeting of the Association for Computational Linguistics (ACL)*.

Varma, P., Sala, F., He, A., Ratner, A., and Ré, C. (2019). Learning dependency structures for weak supervision models. *arXiv preprint arXiv:1903.05844*.

Yang, T. and Priebe, C. E. (2011). The effect of model misspecification on semi-supervised classification. *IEEE transactions on pattern analysis and machine intelligence*, 33(10):2093–2103.

Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. (2020). Rethinking bias-variance trade-off for generalization of neural networks.

Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.