Crowdsourcing with Evolving Annotators

A thesis presented

by

Rohan Thavarajah

to

Institute of Applied Computational Science

in partial fulfillment of the requirements

for the degree of

Master of Engineering

in the subject of

Computational Science and Engineering

Harvard University

Cambridge, Massachusetts

May 2018

Thesis Co-Advisors:
Pavlos Protopapas
Guillermo Cabrera

Author:
Rohan Thavarajah

Crowdsourcing with Evolving Annotators

Abstract

Generating training data for a task like classification often involves surveying a panel of annotators. By obtaining multiple votes as to an item's true label, we hope to harness the collective wisdom of a crowd of individuals who are each individually subject to error. Probabilistic annotation models show that by simultaneously inferring each annotator's skill and weighting votes accordingly, one can recover true labels more accurately than by simple majority rule. This project extends annotation models by acknowledging that annotators are not only differentially skilled but are also time-variant. We claim that there are two classes of time variation. In chapter 1, we explore volume dependence where annotator change is induced by the gross number of times they have been queried. Instances of volume dependence include fatigue and practice. In chapter 2, we explore content dependence where annotator responses are influenced by patterns in their question history. An annotator's present response may alter if they see a run of objects of like-category (the frequency effect) or anchor themselves to a first impression (the halo effect). Our goal is to develop general models that can capture arbitrary kinds of volume and content dependence.

Contents

Acknowledgments

I am very grateful to Tribe Dynamics for funding my Masters research and to my advisors for their guidance and inspiration along the way. Thank you Pavlos and Guillermo, Ryan and Christian, my parents, brothers and (forgive me for the phrase) numerous others; there are too many to thank and too many thanks to owe.

1 Introduction

1.1 PROBLEM SETUP

Suppose you encounter an image of an unknown object. You query ten experts each of whom offers a vote as to the object's identity. What is your best guess of the object's category? It is reasonable to defer to a majority vote, accepting the most common response as ground truth.

Now suppose a second object arrives and you query the same ten experts again. What is your best guess of this new object's category? Though majority voting is reasonable, it is potentially wasteful to disregard what you have learned from the exercise of labeling the first object. For instance:

1. What if you know that annotators have diverse skill. Majority voting weighs each vote equally. If you are able to assess performance on the first object we might tier our trust according to skill when evaluating the second.

2. What if you know that the category of the second object depends on the category of the first. For instance, suppose your annotators are tagging parts-of-speech. If the first object is unanimously determined to be a preposition, we might be more inclined to believe a noun is to follow.

3. What if you can use the distribution of votes across all objects to identify and upweight votes associated with rare classes.

Annotation models refine majority voting by making claims about and formalizing conditional **dependencies**. These dependencies serve as gates that mediate how information spills from one annotation to the next.

The crux of many such models is the idea that annotators are coherent. We say that an annotator is coherent if when presented two similar objects, they consistently produce similar responses. In conjunction with the fact that surveys have a temporal dimension, coherency further demands that an annotator's response is independent of the time at which it is elicited; in other words coherence often goes hand-in-hand with an assumption that annotators are time-invariant.

In this paper we extend annotation models by allowing annotators to vary with time. We conceive of time dependence in two ways:.

1. **Volume dependence** where the path of credibility is a function of the gross number of responses elicited. For instance annotators may fatigue, their credibility eroding for queries posed at the survey tail or experience epiphany, becoming adept at the task after sufficient practice.

2. **Content dependence** where the path of credibility is affected by the content of the queries encountered so far. As seminal papers in the psychology of choice show, individuals facing a decision problem have preferences that may be sensitive to frame or perspective (Tversky and Kahneman (1981)). We argue that recent queries constitute different frames, which may alter an annotator's response. For instance an annotator who sees many instances of a class in quick succession, may become more frugal in offering that label in future.

We divide this paper in two parts in which we separately explore each kind of time dependence. Our goal is two-fold. We wish to both recover true labels more accurately and to parametrically quantify bias arising from frame and volume of responses elicited. By bettering the quality of our training sets we hope to better the performance of our classifiers. By parameterizing time dependence, we hope to afford survey designers a measure of control in understanding how the order in which they pose queries influences responses.

1.2 PROBLEM FORMULATION

Training sets serve classifiers. From the destination classifier, we inherit the twin objectives of maximizing the size of a dataset (volume) and the quality of its labels. However annotators are costly and, in concert with a budget constraint, these objectives compete. Collecting labels on a bundle of objects therefore presents itself as an exercise in resource allocation.

Whether we choose to prioritize volume or quality hinges on how the performance of this classifier behaves upon collection of a marginal data point. Initially, when our training set is small, we expect that the classifier will underfit so we are willing to forfeit some quality in favor of a larger dataset. As our training set grows the performance of our classifier begins to saturate with respect to volume and the noisiness of our labels becomes limiting.

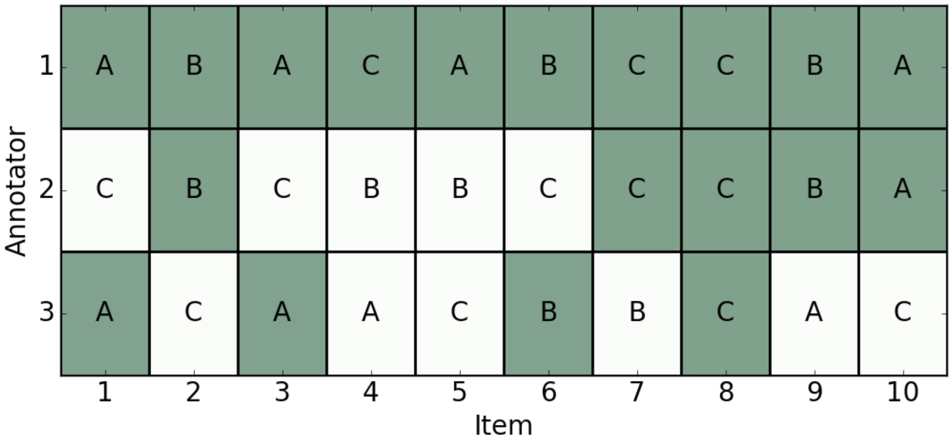
We navigate the volume-quality tradeoff with two dials:

1. If we know the skill of our annotators, we favor label quality by **annotator composition**, concentrating our resources on a few expensive experts of high skill.

2. If our most skilled annotators are unknown, prohibitively expensive or themselves inadequate, we favor quality by employing **repeated labeling**; we choose to collect multiple labels per data point at the expense of labeling a new object.

We seat this paper as an aid to researchers employing repeated labeling. In particular, having collected multiple labels per data point, we explore how to aggregate them to most accurately recover ground truth.

To be concrete we gear our models towards training data which resembles **Figure 1.1**. We have N items belonging to one of three classes {A,B,C}. In general we can have K classes but focus on queries whose responses are Yes/No. Each of the J annotators (or a subset thereof) provides a vote as to which class an item belongs. Critically, we expect annotators to disagree. For a given item, i, how do we consolidate repeated labels r_{ij} to better deduce the true label z_i?



Annotator	1	2	3	4	5	6	7	8	9	10
1	A	B	A	C	A	B	C	C	B	A
2	C	B	C	B	B	C	C	C	B	A
3	A	C	A	A	C	B	B	C	A	C

Figure 1.1: *An example of a dataset with repeated labels. If an annotator successfully identifies the true class, we shade the corresponding cell green. In this example, annotator 1 is highly skilled and annotator 2 gets better at the task after labeling 6 items.*

1.3 MAJORITY VOTING BASELINE

Label consolidation is non-trivial. Consider for instance the naive strategy of majority voting and suppose we are interested in consolidating votes from multiple annotators on a single object. The success of majority voting is predicated on three assumptions:

1. Votes are **identically distributed**. For a given object, a vote from an annotator may be represented as a draw from a common categorical distribution with class probabilities given by the vector θ.

2. Votes are **independent**. Voting behavior on one object does not reveal anything about behavior on a second.

3. Annotator ability to identify the **truth dominates systematic biases**. The class with the largest probability of being reported θ_{max} is the object's true class.

Let $\hat{\theta}$ represent the observed class proportions. Then majority voting corresponds to choosing the modal reported class, $\arg\max(\hat{\theta})$. This is sensible because by (1) and (2) and by invoking the law of large numbers, as the number of reported labels collected gets large,

$\hat{\theta}$ tends to θ. It follows that selecting the modal reported class $\arg\max(\hat{\theta})$ approximates $\arg\max(\theta)$, which by (3), is the object's true class.

The more assumptions we peel away, the more egregiously majority voting will fail. Much of the literature on label consolidation can be framed as developing strategies that are robust to the relaxation of a subset of these three assumptions (summarized in **Table 2.1**).

2 Literature Review

2.1 ANNOTATOR DIVERSITY

Annotators tend to be diversely skilled so that assumption (1), votes are identically distributed, is unlikely to hold. When annotators are diversely skilled, majority voting mixes good labels in with the bad. Prior work tends to address this in one of two ways:

1. If J is sufficiently large, the collective wisdom of a crowd of weak annotators can overpower that of a single expert. The first approach is to determine bounds on the number of annotators needed till a crowd becomes collectively wise.

2. To address annotator diversity more directly, it makes sense to weight votes according to the accuracy of the annotators that supplied them. The second approach is to construct a weighted voting scheme.

2.1.1 Determining When Crowds Become Wise

We say that a crowd is **wise** if majority voting yields a label of higher quality than any one member can achieve. The wisdom (or lack thereof) of a crowd may be characterized by its strongest and weakest member; the strongest annotator defines the highest quality label we

Table 2.1: *Related work and key assumptions*

	Identical annotators	Non-dominant systemic bias	Independence
Majority	×	×	×
Static (Dawid Skene, GLAD)		×	×
Static (ELICE, Surprisingly Popular)			×
Sequence Labeling (for NLP), Filtering Approaches		×	
Proposed Time Dependence Model			

can achieve through unilabeling whereas approximating a crowd of J diverse annotators with J duplicates of the weakest member defines a lower bound for the quality of a majority vote. In **Figure 2.1** we present iso-accuracy curves, the number of annotators that must be combined to achieve parity in label quality. A contour prescribes a conservative estimate of how many annotators are needed to ensure a crowd is consistently wise. For instance, if our strongest annotator has an accuracy of 90% and our weakest 75%, with more than five annotators, majority voting will beat unilabeling in expectation. A more elaborate analysis of the pitfalls of repeated labeling is conducted in Sheng *et al.* (2008) and Lin *et al.* (2014).

2.1.2 A Weighted Voting Scheme

In **Figure 2.1** we also skirt a scheme for attributing relative worth to labels arriving from annotators of **known** skill. Mechanically, when presented a bundle of votes, the exercise is to establish a ranking of which annotators to trust but this poses a dilemma; we need true labels to determine accuracies but we need accuracies to recover true labels. The most common approach in prior work is to employ a probabilistic model that simultaneously infers both.

The canonical probabilistic annotation model was developed by Dawid *et al.* (1979) who were interested in consolidating the reports of multiple physicians. At the heart of their model is the inference of confusion matrices that denote with what probability a physician will report class c when they encounter class k. Confusion matrices are annotator specific, so Dawid *et al.* (1979) allow annotators to be diverse, and through off-diagonal elements the authors are able to comment on moderate systemic biases of physicians that over- or underestimate patient symptoms.

The Dawid-Skene model assumes that all items within a class are of equal difficulty. The GLAD model, developed by Whitehill *et al.* (2009), incorporates a unique difficulty term per item and, to accommodate the many new parameters sacrifices some nuance in the structure of annotator skill. In place of J $K \times K$ matrices, the GLAD model learns a single skill parameter per annotator.

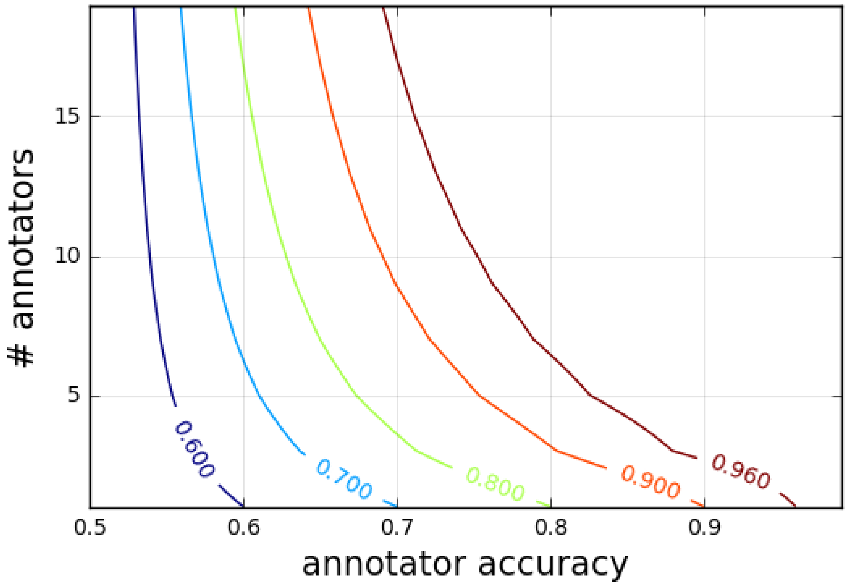


Figure 2.1: *Iso-accuracy curves. Tracing the* 0.9 *accuracy contour, we see that one annotator of* 90% *accuracy yields labels of equal quality to the majority vote of five annotators of* 75% *accuracy or sixteen annotators of* 65% *accuracy.*

2.1.3 Win Conditions of a Static Model

To gauge the potential gain from learning annotator skill we pose the following experiment. We draw J annotators of static credibility from an interval $[p_1, p_2]$ and simulate their reported labels on 100 objects. We continue in this way generating 150 datasets per configuration of accuracy interval and J. We investigate when a weighted approach, informed by an oracle that knows each annotator's skill, yields most significant gains over majority rule (**Figure 2.2**).

We observe:

- When annotators are systemically biased (with credibility less than 0.5), a weighted approach has greatest potential to outperform majority voting. In these cases, a weighted approach can recover information from annotators who consistently invert their labels.

- When we contrive that annotators are more similar or when we enlist many workers

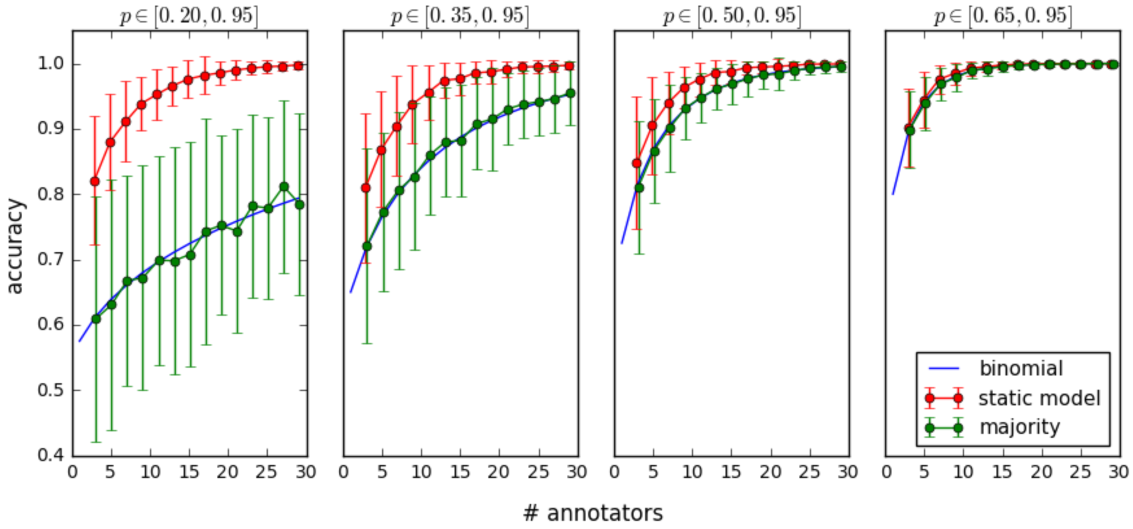


Figure 2.2: *Estimating the potential of a static model applied to stationary annotators. The weighted approach utilizes oracle-sourced information regarding annotator skill to determine a maximum likelihood estimate of true labels. The accuracy of majority voting may be roughly approximated by a binomial distribution parameterized by the mean skill of workers in the crowd.*

for the task, the spread between a weighted model and majority voting becomes thin. When annotators are chosen in the interval $[0.65, 0.95]$, the gain of a weighted approach over majority voting becomes negligible with as few as three annotators.

2.2 SYSTEMIC BIAS

Weighted voting schemes have the greatest **potential** when annotators are biased. However, if most annotators are systemically biased, we run the risk of failing to correctly learn the system. Strategies that redress this situation involve collecting accessory information. The ELICE model developed by Khattak and Salleb-Aouissi (2016) clamps a small set of labels around the reports of a few highly skilled annotators. This approach is found to successfully calibrate the credibilities of strongly biased annotators. Alternatively, in the "surprisingly popular" algorithm, Prelec *et al.* (2017) conduct surveys in which they ask respondents not only for their vote, but also for what they predict the most popular vote will be. By employing a heuristic that upweights the votes of individuals who self-identify as a minority, Prelec *et al.* (2017) are able to yield significant gains in accuracy over majority rule. Workers needn't be malevolent to be systemically biased. For instance spammers who gravitate to a prevalent class present as being systemically biased and are explored in Raykar and Yu (2012) and Passonneau and Carpenter (2014).

2.3 SYSTEM DEPENDENCIES

The assumption that votes (given an item) are independent may seem benign. With platforms like Mechanical Turks, annotators are unlikely to interact so that their answers are barred from influencing their peers. However, there may be other dependencies within the system. Sequence labeling accommodates dependencies between true labels for instance in the task of parts-of-speech tagging explored by Nguyen *et al.* (2017).

The novelty in this paper is to acknowledge that annotators may evolve. We allow that annotators be beholden to multiple credibilities over the course of a survey and reason that there exist dependencies in these credibilities across time.

In the case of volume dependence, we hope to capture arbitrary paths characteristic of traits like fatigue and epiphany. A simple demonstration of volume dependence is found in Serenko and Bontis (2013) who ask 379 researchers to rank 25 journals on a 7-point scale. They find that journals received consistently higher rankings when positioned earlier rather than later in the survey suggesting researchers fatigued and offered a disengaged response after losing interest in the task. Previous efforts to model volume dependence typically employ filtering (Jung *et al.* (2014) and Donmez *et al.* (2010)). Diverging from these texts, we leverage the probabilistic programming language `pymc3`s implementation of Hamiltonian Monte Carlo (Salvatier *et al.* (2016)). We incur computational burden, but are able to learn all key parameters (in cited texts some parameters are fixed arbitrarily), layer in additional complexity (e.g. addressing diverse class difficulties and injecting expert labels) and yield a framework that is readily adapted to other task representations (e.g. multiclass labeling).

In the case of content dependence, we strive towards a general recipe for debiasing annotators that are subject to order effects. Different patterns in the ordering of prior queries may alter an annotators present response. The patterns that bear this influence are diverse (e.g. a run of consecutively like labels or the positioning of a pivotal query). However, the mechanisms by which these patterns affect a present response are thematically similar. Surveys are not so dissimilar to conversations (Strack (1992)) in which speaker and respondent are obligated to optimize information exchange. **In pursuit of information exchange, a pattern in the order of past queries that arises due to chance may be mistaken by the annotator as arising due to an intrinsic dependency**. These falsely perceived dependencies corrupt the quality of their responses. We adopt a novel two-step procedure. In stage 1 we estimate when, in what direction and to what extent annotators have been biased by their label history. In stage 2 we use bias characteristics from stage 1 to weight reported labels and better deduce ground truth.

3 PART I Volume Dependence - Approach

3.1 GENERATIVE MODEL

Our proposed volume dependence model may be described as a GLAD variant with a Hidden Markov structure on annotator skill. We present the graphical model in **Figure 3.1** (for clarity we denote plates by shared subscripts) and pair it with the following generative process:

1. True labels
   - Shortlist items that may belong to class k, $\rho_k \sim$ Beta(1,1). If ρ_k is close to 1 then the shortlisting mechanism is very strict and every item shortlisted for k belongs to k. If no shortlisting occurs, ρ_k represents the distribution of classes in the population.
   - After shortlisting we are furnished with a set of item-class pairs, $z_{ik} \sim \text{Bernoulli}(\rho_k)$. z_{ik} is unobserved and equals 1 if item i truly belongs to class k and is 0 otherwise.
2. Reported labels
   - To approximate z_{ik}, we collect noisy labels from a crowd of annotators:

     $$r_{ijk} \sim \begin{cases} \text{Bernoulli}\left(\theta_k^{(jt)}\right) & \text{if } z_{ik} = 1 \\ \text{Bernoulli}\left(1 - \theta_k^{(jt)}\right) & \text{if } z_{ik} = 0 \end{cases}$$

     Those labels are a function of the latent label of the item being observed and the annotator's accuracy.
3. Accuracies

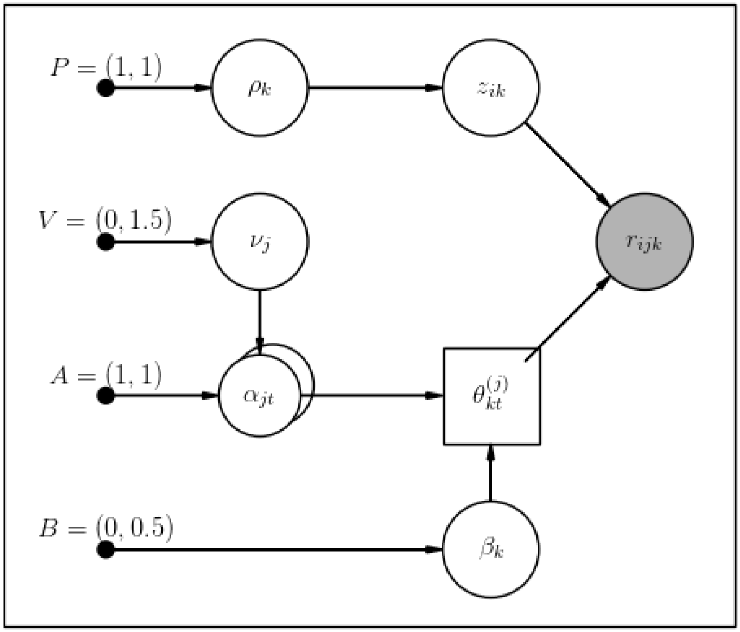


Figure 3.1: *Volume Dependence Graphical Model.*

- Annotator j's accuracy is a deterministic function of their skill and the brand's difficulty:

$$\theta_k^{(jt)} = \text{sigmoid}(\alpha_{jt}\beta_k)$$

  This function is plotted in **Figure 3.2**.

- Brand difficulty is time invariant and is non-negative. If $\beta_k = 0$, the task is extremely difficult and any labels obtained are nonsense. If $\beta_k = \infty$ the task is trivially easy. We impose the non-negativity constraint by exponentiation as follows.

$$\beta'_k \sim \text{Normal(0,0.5)}$$

  with $\beta_k = \exp(\beta'_k)$.

- Annotator skill varies with time. We say that an annotator's skill at time t is associated with their skill at time $t - 1$ and introduce the parameter, ν, to denote the annotator's volatility; how susceptible they are to change:

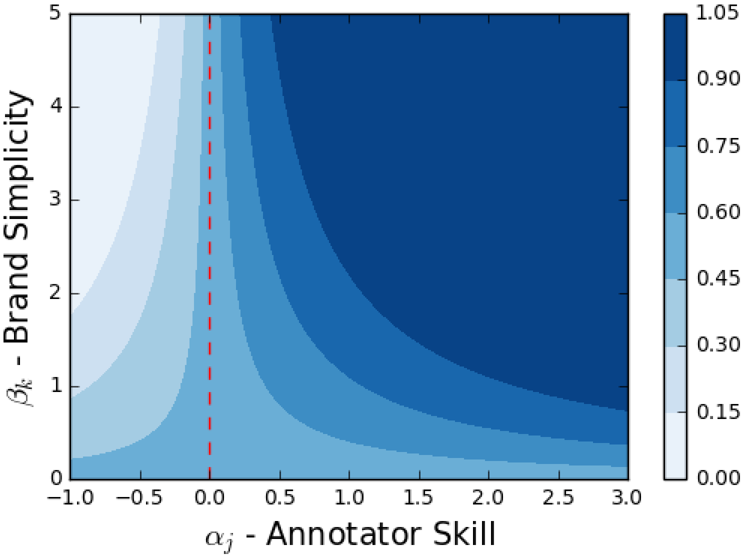


Figure 3.2: *The deterministic theta function. As skill and simplicity increase, credibility approaches 1.*

$$\alpha_{jt} \sim Normal(\alpha_{jt-1}, \nu_j)$$

In the graphical model, the shadow on α_{jt} is used to denote its dependency on its lagged state. If $\alpha_{jt} = \infty$, the annotator is of very high skill, if $\alpha_{jt} = 0$, they offer nonsense labels and if $\alpha_{jt} < 0$, the annotator is systemically biased.

3.2 DESIGN INTUITION

Though our performance metrics pay particular attention to the accuracy of recovered labels, a parametric model is particularly appealing because it lends itself to normative conclusions that may be applied to future iterations of the survey. For instance a survey designer may be interested in how aggressively respondents fatigue. If the model suggests annotators burned out early in the survey, a designer might shorten future iterations. Similarly we learn, how strict the shortlisting process is by class, how difficult classes are and who the strongest annotators are (it might be beneficial to pose the most difficult tasks to those best equipped to confront them). We construct the model in this way for a number of reasons:

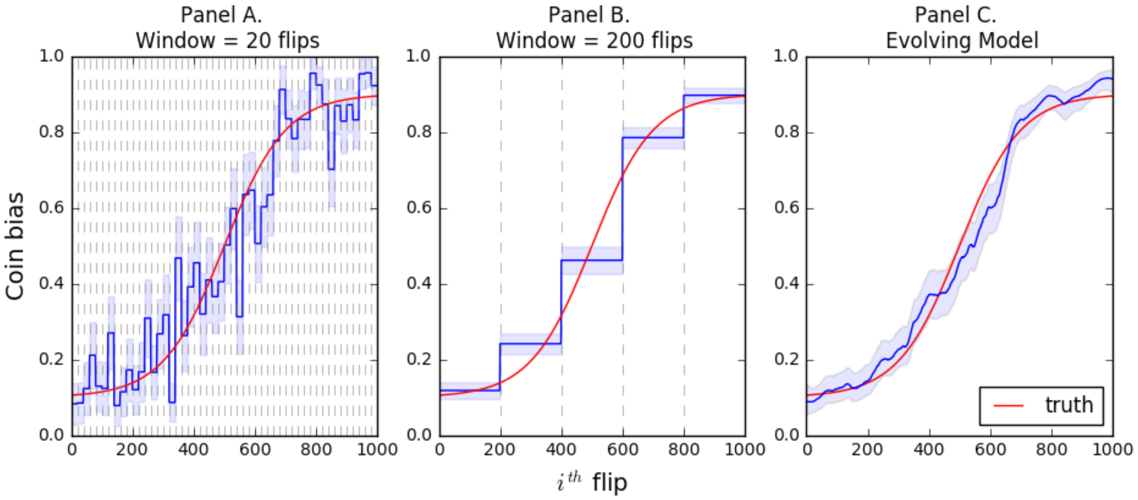


Figure 3.3: *The evolving model mirrors a static model estimated in windows. A coin is flipped* 1000 *times and its time-variant bias is estimated from the head counts. In Panel A, window size is small and we overfit to noise. In Panel B, window size is large and we miss nuance in the path. Panel C utilizes a random walk and by parameterizing the coins volatility, learns something in-between.*

3.2.1 A Single Structure for Learning Arbitrary Skill Paths

Skill paths may be diverse. Annotators may learn, fatigue or for instance label in sessions manifesting in piecewise switches. A coarse approach would be to apply a static variant of the model to windows of the data. However this introduces two challenges. First, a window should ideally contain (a) all of an annotator's labels that were reported close in time and (b) all labels reported on a constituent object. If objects are presented to annotators in different sequences, constructing windows that fulfill both of these criteria is non-trivial. Second, even if objects are presented in identical sequence (ill-advised by survey designers everywhere!), we must still grapple with choosing an appropriate window size. Choose a window too small and we do not have enough data to parse the signal from the noise (**Figure 3.3.A**). Choose a window too large and we pave over the change that motivated a time-variant model in the first place (**Panel 3.3.B**).

By leaning on the random walk construction we address these concerns:

1. We yield a model that is sufficiently general to capture diverse skill paths.

2. By applying a single model to the entire dataset, we unfetter ourselves from the risk that information useful to one window is insulated in another.

3. By estimating the volatility of annotator skill we mediate how much information is gleaned from performance on neighboring tasks. For instance, a low volatility suggests that, akin to choosing a large window size, we can dig deeper into the history of an annotator's performance to inform our estimate of their present skill.

3.2.2 Diverse Class Difficulties

We acknowledge that classes may vary in difficulty. If an annotator is subjected to a series of queries to do with a difficult class, their accuracy may drop but we argue that it is unfair to consider them less credible. By including β_k, we infer trends in skill that control for the order in which classes are presented.

3.2.3 Pooling

The deterministic theta function mediates how information is shared across annotators and classes. We reason that if an annotator is skilled at identifying classes 1 through $K-1$, then they will also be skilled at identifying items of class K. Similarly we reason that if a class is difficult for $J-1$ annotators, it will be difficult for the Jth annotator as well.

3.2.4 Missing Data

We allow z_{ik} to be partially observed. Using a few experts as oracles permits us to calibrate credibilities similarly to the ELICE model.

3.2.5 Complex Models Subsume Simpler Variants

If $\nu \approx 0$, then the volume dependence model resembles the static GLAD model. If in addition, we find skill (α_j) is identical across all annotators, we consolidate labels in a manner resembling majority voting. We choose priors such that in the absence of data, the evolving model cautiously settles in the neighborhood of majority voting.

4 PART I Volume Dependence - Results

4.1 EXPERIMENTS

We design experiments that investigate two questions (1) in what conditions does knowing the path of annotator skill yield greatest gains in accuracy? (2) how successfully can we recover latent skill paths?

4.1.1 Win Conditions of an Evolving Model

We gauge the potential gain from learning skill paths via simulation (**Figure 4.1**). We draw J annotators whose credibility follows a sigmoid that starts close to p_{LOW} and culminates near p_{HIGH}. We draw 100 items which are presented to annotators in random sequence. By shuffling the order of presentation we proxy diverse skill paths with about half of the labels being reported when p_{LOW} prevails and the rest when p_{HIGH} does. We plot accuracies of four consolidation strategies subject to different configurations of p_{LOW} and J.

- The evolving model is informed by an oracle that knows the skill of all annotators at every time step. This model reflects the potential gain to be had from successfully learning skill paths.
- A static model is applied to each simulated dataset. Recall that we violate the assumption that annotators are stationary. This demonstrates how static models behave when applied to time-variant data.
- The majority voting baseline is plotted as is a binomial approximation derived by assuming all J annotators have accuracy at the midpoint of p_{LOW} and p_{HIGH}.

We observe:

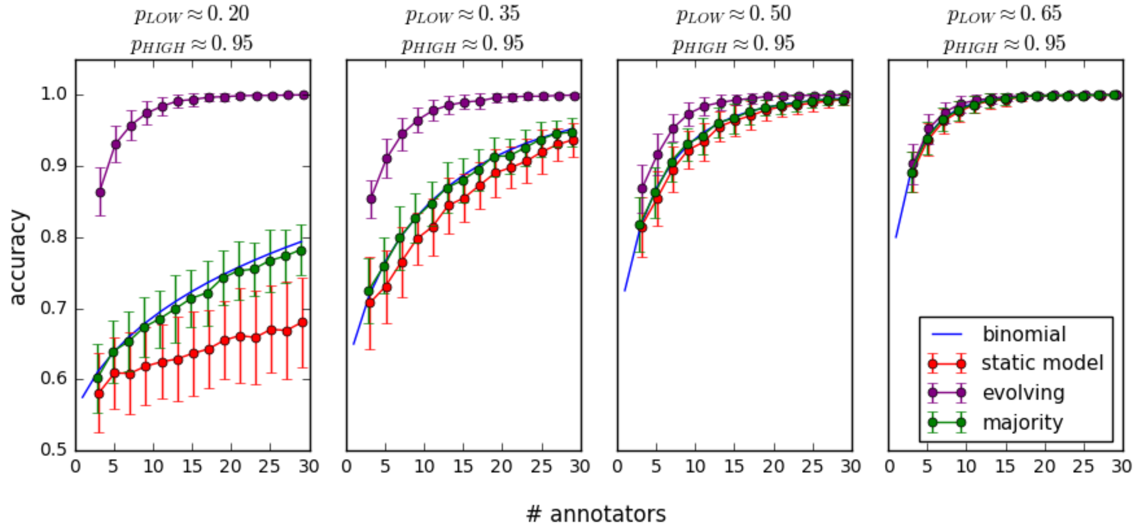


Figure 4.1: *Estimating the potential of an evolving model applied to non-stationary annotators.*

- That as was the case for the previous set of experiments, the evolving model exhibits greatest potential to outperform majority voting when annotators are systemically biased (credibility < 0.5) for a portion of their voting tenure.

- When the spread between p_{LOW} and p_{HIGH} diminishes or J becomes large, the spread in accuracy between all approaches thins. When p_{LOW} is drawn near 0.65, any gains to be had from modeling skill paths become negligible for J as low as three.

- When annotators exhibit periods of systemic bias, the static model fails to learn the system by definition and can fair worse than majority voting.

4.1.2 Learning Skill Paths

To test the evolving model's ability to capture non-stationary annotators we simulate labels from workers demonstrating stylized skill paths. We then occlude all but the reported labels and try to recover the path and true labels. In **Figure 4.2** we present the skill paths learned after one characteristic run. In general, when primed with a few expert labels, the evolving model consistently recovers trajectories characterizing fatigue, epiphany, session switches and episodes of spam-like behavior.

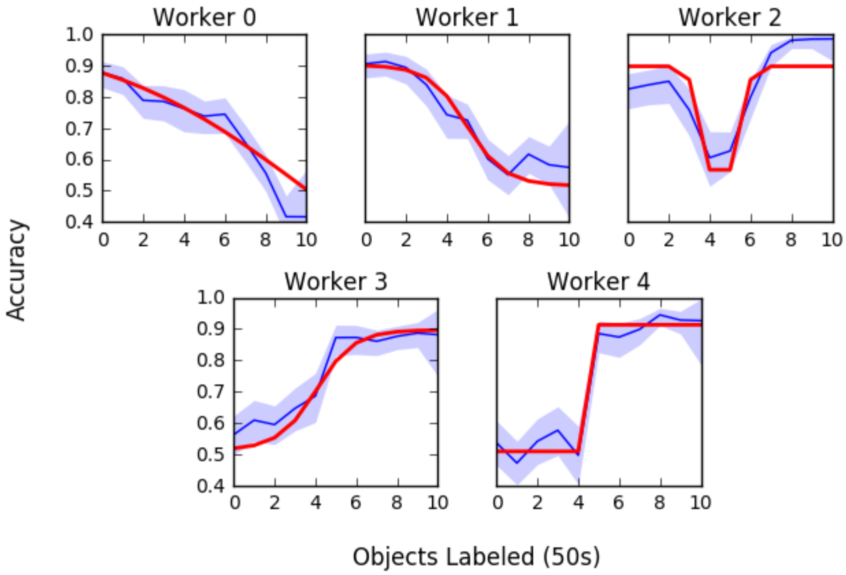


Figure 4.2: *Learning fatigue, epiphany, episodic spamming and session switches.*

4.2 REAL-WORLD APPLICATION

We apply the evolving model to data provided by Tribe Dynamics, a company that studies influencer marketing. Tribe Dynamics is interested in collecting all social media posts associated with a given brand. To do this the company uses the crowdsourcing platform Mechanical Turks. An example task is provided in **Table 4.1**. Posts are distributed among workers who affirm or reject brand affiliation. In **Table 4.2** we also summarize some key descriptive statistics. We apply the model to 16,107 posts that are each labeled by an average of 3.91 of the 120 total annotators. Tribe's crowdsourcing strategy dovetails with the previously described generative process as follows:

- First social media posts that may be associated with a brand of interest are shortlisted. The leniency of the preliminary screen is captured by ρ_k.
- There is a spectrum of brand difficulty. A brand whose namesake is a designer is easily identified; a brand with a pithy name like "fresh" or "clear", much less so.
- At the outset there is no strong intuition to suggest that annotators have been time-

Table 4.1: *An example task*

Brand Name	Text	Brand Affiliated
Mace Totally Fictitious Cosmetics Inc.	I picked up this centuries old **mace** at the Renaissance Festival. I can't wait to open jars of pickles with it.	?

Table 4.2: *Data description.*

	N	Mean	Min	Max
Number of posts (N)	16,107			
Number of annotators (J)	120			
Number of brands (K)	12			
Annotators per query		3.91	1	29
Queries per annotator		522.65	19	9,066
Queries per brand		1342.25	64	3,878

variant on this task. However, since the priors in the evolving model are structured to favor stationary skill, we argue that if annotators are time-invariant, the main risk we incur is in computation cost. Furthermore, posts are presented to workers in varying sequence so that the evolving model is preferred to a windowed approach.

Before we proceed we prospect the potential gains from employing an annotation model using the degree of consensus on each post. If the task is trivial, then all workers will agree and majority voting will perform too well to justify an annotation model. If the task is difficult and annotators are very diverse in skill, agreement with the consensus will be low and we risk labels being too noisy to successfully learn the system. For an annotation model to succeed, ideally annotator consensus strikes a middle ground. As we see in **Figure 4.3**, this labeling task errs on the side of simplicity though there is room for gain.

Table 4.3 details our results for varying combinations of two pivotal design decisions, evolving vs static and asymmetric vs symmetric skill. We say skill is asymmetric if annotators perform heterogeneously on acceptance vs rejection. Since we are working in an unsupervised setting validation is hard. We employ two validation sets. The easy validation

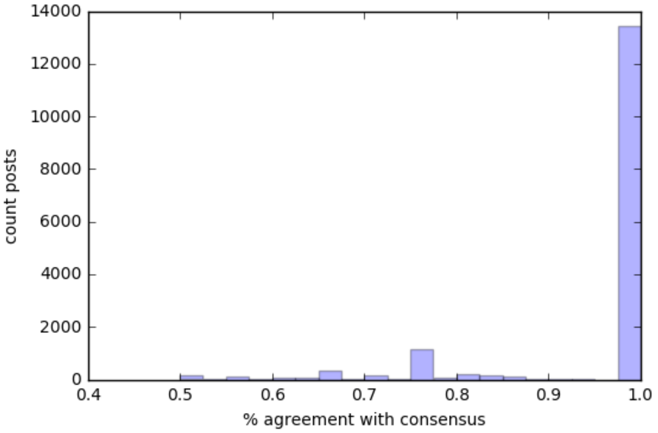


Figure 4.3: *Consensus is high which limits the gains to be had from employing an annotation model.*

set contains 168 posts that were pre-labeled by experts. However, the baseline accuracy of majority voting is high on this set (86.6%) so that it is difficult to resolve differences in the relative merit of alternative models. To furnish a hard validation set of more difficult tasks, we collect all cases when majority voting and our model conflict and resubmit them to the company to acquire labels of gold standard.

We observe that the evolving model outperforms the static model though only marginally so (0.840 vs 0.833 accuracy). We hypothesize that this boost in accuracy is muted for two reasons: (1) because annotator variation across time is not sufficiently strong (2) there is already a high degree of consensus in the data that bounds above the number of posts on which we can improve. Both models however succeed in identifying many cases in which majority voting fails; 84.0% of the time the best configured model identified a conflict, the model was correct and majority voting was wrong. Put differently, the accuracy of the majority voting baseline for these conflicts was 16.0% so that an annotation model proves highly beneficial. Finally, we infer that annotators are asymmetrically skilled so that it is useful to separately model their performance on acceptance vs rejection.

Over a hundred turkers are enlisted over the course of the project. In Figures 4.4 and 4.5 we focus on the top ten most prolific. In **Figure 4.4** we yield a rough picture of skill paths by assuming consensus coincides with ground truth and by calculating turker accuracy in

Table 4.3: *Grid search. The baseline accuracy of majority voting in the easy validation set is 86.6%. The hard validation set is defined as cases when the pertinent model and majority voting conflict. Therefore accuracy of majority voting in the hard set* $= 1 -$ *model accuracy.*

Evolving	Asymmetric Skill	Validation - Easy	Validation - Hard
×	×	0.915	0.840
	×	0.915	0.833
×		0.823	0.665
		0.793	0.616

a rolling window of 200 annotations. In **Figure 4.5** we plot the skills paths learned by the model. The resemblance of Figure 4.4 to Figure 4.5 serves as a sanity check that the model is learning something sensible. Inspecting the paths themselves, in Figure 4.5 we see that the most prolific workers unanimously perform better on posts warranting a negative response. The last two workers in particular appear to exhibit periods of systemic bias where their performance on affirmation ($z = 1$) dips below 0.5.

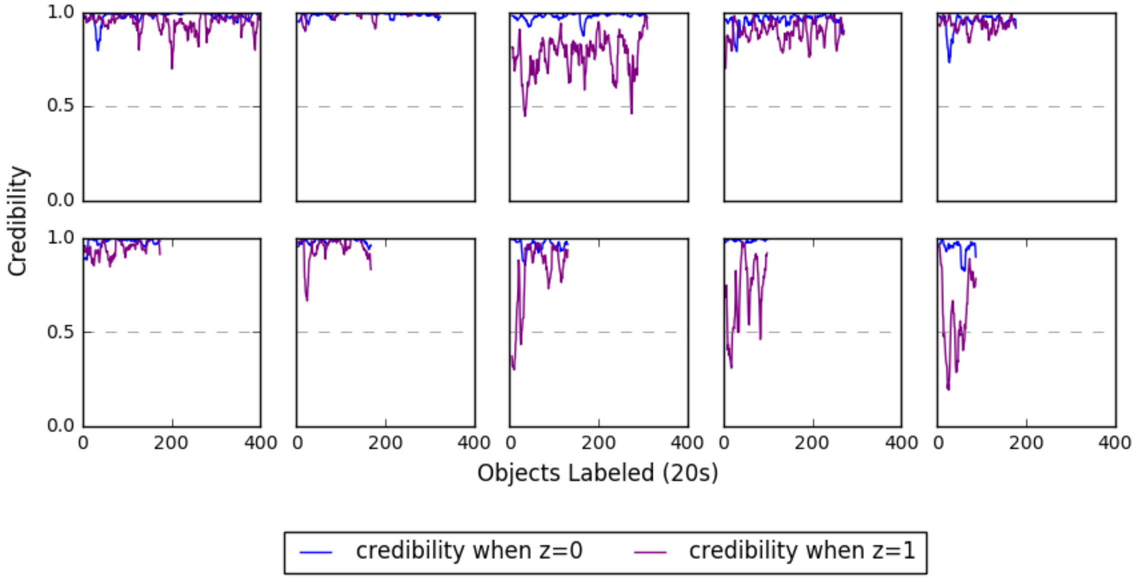


Figure 4.4: *Rough skill paths estimated by assuming majority voting yields ground truth and calculating accuracy in a rolling window of 200 annotations.*

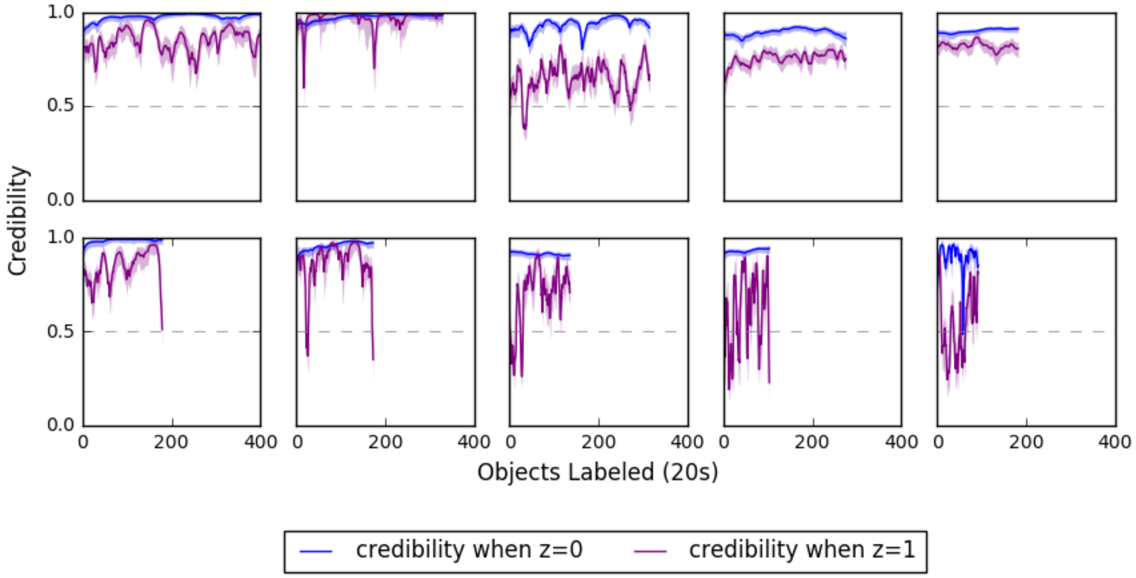


Figure 4.5: *Model derived skill paths on tasks of medium difficulty.*

5 PART II Content Dependence - Approach

5.1 MOTIVATION

In the previous chapter we studied how annotators change as a function of the volume of queries they have encountered. In the chapter to follow we study how they change depending on the **content** of those queries. In social science literature, biases associated with question order are often dubbed order effects. An umbrella narrative is that a particular **pattern in the order of past queries**, can alter an annotator's perspective on a **present** decision and in turn their response.

The catalog of biases induced by question order is extensive (Schwarz and Sudman (1992)). Often these biases are studied in experimental settings. The treatment group is exposed to a query pattern carefully curated such that the experimenter can render a sensible hypothesis as to which query will exhibit a bias, and in what direction. The difference between the responses of treatment and control gauges the magnitude of the bias and, provided this difference is statistically meaningful, is taken as evidence of its existence. Though this is useful from a psychological perspective, to practitioners the onus of debiasing annotators is borne at the **design stage**. Practitioners are advised to carefully judge their queries; is this query sensitive to preference reversal? Is that one prone to cause it? The search for these characteristics is a daunting enterprise owing to the sheer variety of cognitive biases that may prevail. Indeed, often the keenest insights into bias are only developed with the benefit of hindsight after the survey has been administered. Owing to the difficulty of judging queries a priori, it is common to adopt the coarse approach of randomizing the question order.

In this chapter we strive towards a general framework for debiasing the responses of annotators **post-collection**. We try to mimic the information available to the practitioner as opposed to the experimenter. The practitioner wishes to deduce the true labels of his queries not gather evidence of a cognitive bias. The practitioner knows that his respondents are vulnerable to cognitive biases but does not know which ones. As a result he neither knows which queries are particularly sensitive to question order nor what patterns to look for in an annotator's label history that might induce change. The goal of our framework is to allow the practitioner to identify which responses are most likely to have been biased by question order and to adjust them in pursuit of ground truth.

We explore three order effects as case-studies; the frequency effect, the halo effect and the priming effect. Each may be identified by the characteristic pattern in question order that causes them to manifest.

5.1.1 The Frequency Effect

Characteristic pattern: Consecutive queries of like kind.

Alters present response: via recalibration.

Suppose annotators are tasked with labeling images of food as healthy or unhealthy. The healthiness of some foods is particularly ambiguous and we consider granola an archetypal case. **Do you think granola is healthy?** Suppose instead that I had presented four images of vegetables to you first. Do you think granola is healthy now? And if I had presented four images of foods that would horrify your dentist; what about now (illustrated in **Figure 5.1**)?

We hypothesize that experiencing a run of healthy foods may render an annotator less willing to offer that label on neighboring images. Granola may seem healthy when presented in the context of ice cream, but this label may reverse when pitted against a frame of vegetables. Daamen and Bie (1992) might suggest that we are witnessing an example of the "frequency effect". When asked to categorize objects, annotators exhibit a bias towards seeing that each category is represented. When workers see many foods of a certain class, they recalibrate their definition of healthiness to encourage diversity in their responses.

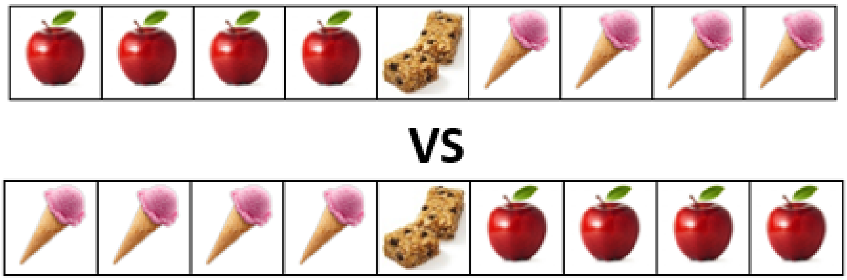


Figure 5.1: *The frequency effect. Claim that if you see multiple healthy objects, you underestimate the healthiness of granola.*

5.1.2 The Halo Effect

Characteristic pattern: The first rating for an entity.

Alters present response: via an anchoring to a general impression.

The halo effect arises when annotators are tasked with rating different traits of the same entity. Thorndike (1920) observed that when commanding officers were asked to rate their soldiers, seemingly independent traits were found to be highly correlated (stylized illustration in **Figure 5.2**). He reasoned, when tasked with evaluating multiple traits, an officer's evaluation of a pivoting trait begets a general impression (a halo) which colors subsequent ratings on the same soldier. This manifests in strong correlations between traits that should be independent or only weakly linked. We focus on the case when pivoting traits are the first evaluated by entity.

5.1.3 The Priming Effect

Characteristic pattern: Evidence of a questioner's intention.

Alters present response: via disambiguation.

Consider the "priming effect" studied by Strack (1992). The authors ambiguously asked students if they were opposed to an "educational contribution for students". However first they asked half the students to estimate support students received from the government and half to estimate how much they paid in tuition. Students favored educational contribution

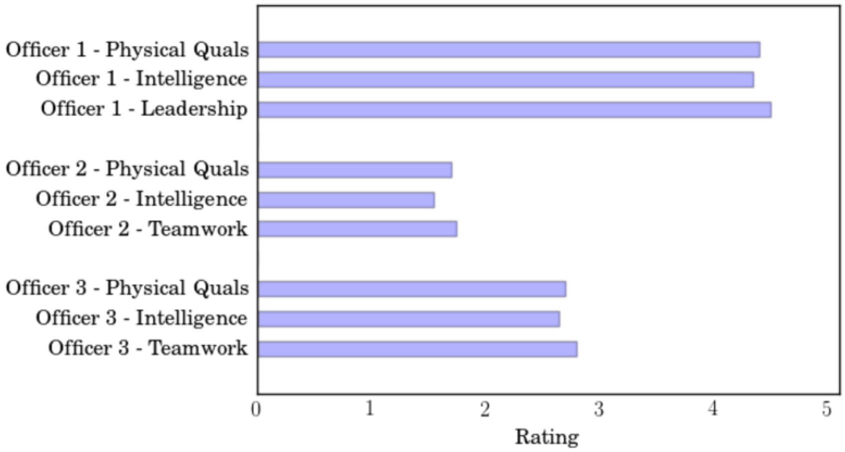


Figure 5.2: *The halo effect. Officers are first asked to rate the physical qualities of their soldiers forming a general impression. Ratings on subsequent traits (intelligence and leadership) cleave to this impression.*

when primed to believe that contribution referred to money received rather than paid. In general, an annotator may use an initial query to understand its successor and in the interests of coherency, offer correlated responses (**Figure 5.3**).

5.2 GENERATIVE MODEL

A reported label is a function of (1) the intrinsic **nature** of the object under consideration and (2) the annotator's **bias** arising from the order of past queries. We wish to disentangle these two components and in particular, most accurately comment on the nature of objects. For example if an annotator is biased towards reporting 1s we wish to place greater trust on their label when they report 0. In the previous chapter we encountered the dilemma of disentangling two unobserved predictors of a response, true labels and annotator credibilities. In this chapter, though the intrinsic nature of an object remains unobserved, an annotator's label history furnishes us with evidence of order effects. We present the graphical model in **Figure 5.4** and pair it with the following generative process:

1. Intrinsic Nature

    - Objects have a latent intrinsic nature, $z_i \sim \text{Normal}(\mu, \sigma^2)$

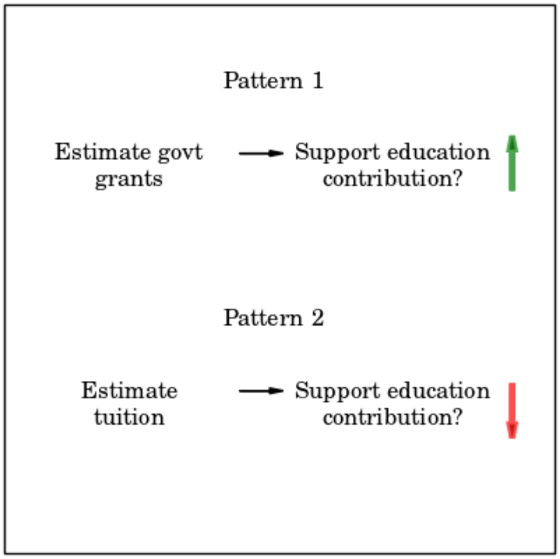

Pattern 1
Estimate govt grants
Support education contribution?
Pattern 2
Estimate tuition
Support education contribution?

Figure 5.3: *The priming effect. The term "contribution" is ambiguous. By priming students with initial questions on grants, students believe the direction of the contribution will be to rather than from them.*

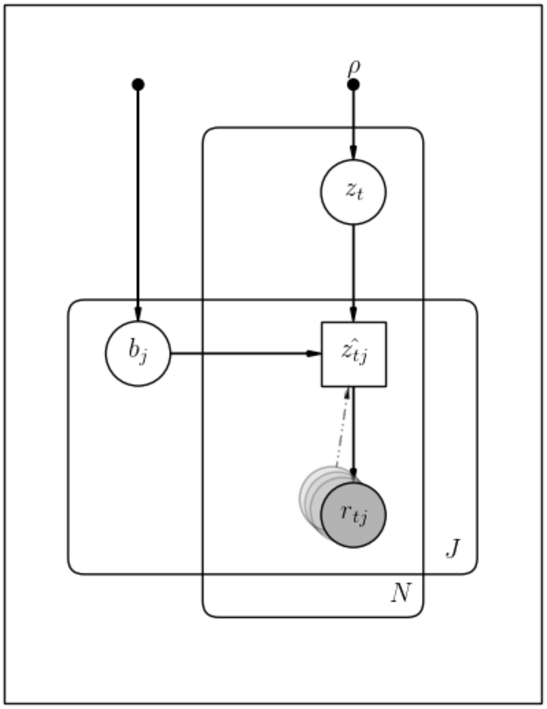


Figure 5.4: *Content Dependence Graphical Model. The shadow on r_{ij} reflects the influence of label history (h_{ij}) calculated outside the model.*

- Annotator's hold their own subjective estimate of the object's nature which we call $\hat{z_{ij}}$.

2. Reported Labels

   - Annotators evaluate whether an object exhibits a trait.

$$r_{ij} \sim \text{Bernoulli}\left(\text{sigmoid}(\hat{z_{ij}})\right)$$

   - If $\hat{z_{ij}}$ is strongly positive the annotator perceives the object as trait positive. Conversely if $\hat{z_{ij}}$ is strongly negative the annotator perceives the object as trait negative. The closer $\hat{z_{ij}}$ is to zero, the more difficult it is for the annotator to render a judgment.
   - Subjective assessments are a deterministic function of the object's true nature and the annotator's bias characteristics.

$$\hat{z_{ij}} = z_i + b_j^{sys} + b_j^{ord} \cdot h_{ij}$$

3. Bias

   - b_j^{sys} describes an annotator's systemic bias. If b_j^{sys} is strongly positive then (ceteris paribus) $\hat{z_{ij}} \gg z_i \;\; \forall i$, and we say that the annotator consistently overestimates the presence of a trait across all objects they encounter.
   - The term $b_j^{ord} \cdot h_{ij}$ jointly describes the effect of question order on annotator j's response to object i. Order effects can be decomposed into survey versus annotator properties.
   - h_{ij} is a feature calculated outside the model and captures the survey-specific component. The magnitude of h_{ij} describes the propensity for question order to bias **any** annotator's present response having experienced annotator j's question history. The sign of h_{ij} describes the bias direction. For instance if $h_{ij} = -5$, then we say j's question order induces an arbitrary annotator to underestimate z_i.

Table 5.1: *Sanity checks.*

Accuracy Spread (RF - Naive)	Action	Interpretation
0	Stop	Order effects are weak or not captured by the current set of features.
> 0	Proceed	Evidence of order effects.
$\gg 0$	Inspect features	If the accuracy of the random forest is close to one the chosen features may be leaking information about the next object.

- $b_j^{ord}(> 0)$ captures the annotator-specific component and describes annotator j's sensitivity to question order. For instance, a question order may have a high propensity to induce underestimation ($h_{ij} = -5$) but an annotator with low b_j^{ord} will be unaffected relative to more sensitive peers.

5.2.1 Recipe for h_{ij}

Assume questions are presented in random order so that $z_t \perp\!\!\!\perp z_{:t-1} \mid \mu, \sigma$. We employ a random forest to predict an annotator's next response given only their history and in particular no information about the object to come. We claim that if the random forest has greater accuracy than a naive model that predicts the modal class, then there is evidence of order effects.

We propose this recipe for generating h_{ij}.

1. Check for sanity **Table 5.1**.

2. Calculate:

$$\Delta p_{ij} = \hat{p_{ij}} - \bar{p}$$

$\hat{p_{ij}}$ = predicted probability of class 1 from the random forest.

$\bar{p}$ = proportion of objects reported to be class 1.

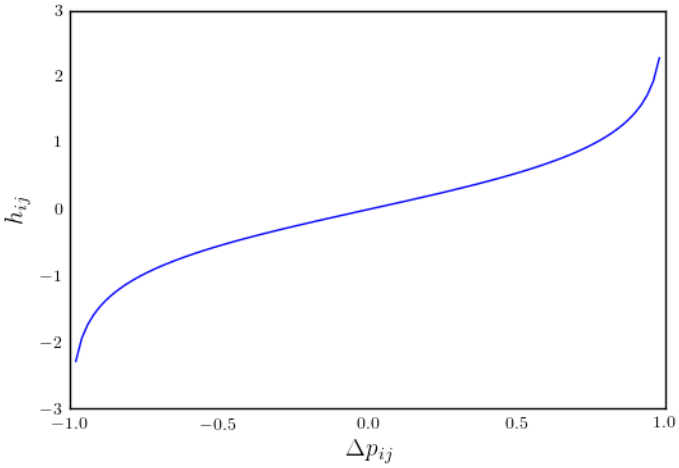


Figure 5.5: *Zero out values of Δp_{ij} close to zero using tanh^{-1}.*

The magnitude of Δp_{ij} gauges the propensity of annotator j's question history at the time they label object i to bias an arbitrary annotator. The sign of Δp_{ij} indicates if this bias renders that annotator an over- or underestimator (respectively if it is positive or negative).

For instance suppose respondents are tagging foods as healthy or unhealthy. We look across the body of reported labels and find 65% have been reported healthy, $\bar{p} = 0.65$. Suppose I have a new query and before I pose it, the random forest predicts an annotator's next response has a 95% probability to be healthy, $\hat{p_{ij}} = 0.95$. Then $\Delta p_{ij} = \hat{p_{ij}} - \bar{p} = 0.3$. We say that an arbitrary annotator with j's history is prone to overestimate the healthiness of i.

3. By definition $\Delta p_{ij} \in [-1, 1]$. However we expect that if Δp_{ij} is close to zero, this reflects more noise than signal. Therefore we transform Δp_{ij} using the inverse tanh function depicted in **Figure 5.5**.

$$h_{ij} = \tanh^{-1}(\Delta p_{ij})$$

5.3 DESIGN INTUITION

Consider a fully parametric strategy that frames the problem as a logistic regression. We wish to predict a response $p(r_{ij} = 1)$:

$$\ln\left(\frac{p}{1-p}\right) = z_1\mathbb{1}_1 + \cdots z_N\mathbb{1}_N + \beta_{\text{halo}}x_{\text{halo}} + \cdots + \beta_{\text{freq}}x_{\text{freq}}$$

The linear predictor consists of two sets of terms:

$z_1\mathbb{1}_1 + \cdots z_N\mathbb{1}_N$ = Fixed effects dummies for each object. z_i corresponds to object i's intrinsic nature. Note that unlike the traditional scenario where fixed effects are controls and of lesser interest, here z_i is the focal point of the model.

$\beta_{\text{halo}}x_{\text{halo}} + \cdots + \beta_{\text{freq}}x_{\text{freq}}$ = Order effects. For each order effect we create a variable to flag if a characteristic pattern is observed in an annotator's history. For instance $x_{\text{freq}} = 1$ if the annotator experiences a run of consecutive labels and is 0 otherwise. Our model adopts a similar structure but extends it to address two key weaknesses.

5.3.1 Side-stepping manual feature engineering

A fully parametric strategy necessitates that we step through the catalog of order effects to engineer features that flag characteristic patterns. Recall however that practitioners rarely know which of an extensive catalog of order effects are present in their survey. Manually engineering features is an endeavor doomed to be over-specified with respect to some forms of bias and incomplete with respect to others. By using a random forest as an accessory model, we are able to take very simple representations of question order and learn what high-order patterns are of interest.

5.3.2 Pooling

Question sequences are often randomized. Although an order effect may be very influential, randomization means that an annotator may encounter it too few times for the effect to be accurately estimated. Ideally we would wish to look across other annotators who have experienced a characteristic pattern to discern if it bears influence. This is the purview of the random forest. Furthermore, ideally for a single annotator, we would like to look across the cocktail of order effects they have been exposed to to assess their sensitivity to order effects in general. The random forest combines the influence of all order effects into h_{ij} thereby putting them on the same implicit scale. By learning a parameter b_j^{ord}, this allows us to look across biases and comment on an annotator's general sensitivity.

6 PART II Content Dependence - Results

6.1 Experiments

As was our approach in chapter 1, we test the content dependence model by injecting order effects into simulated data, occluding all but reported labels and evaluating how well we recover ground truth. We simulate datasets exhibiting the frequency, halo and priming effects discussed previously (confirming simulated responses exhibit characteristic behavior in **Figure 6.1**). Our recipe for injecting an order effect is always the same. Recall that in our generative model we define an annotator's perception of an object to be:

$$\hat{z_{ij}} = z_i + b_j^{sys} + b_j^{ord} \cdot h_{ij}$$

To inject an order effect we replace $b_j^{ord} \cdot h_{ij}$ with a variable x_{ij} that flags a characteristic patten in an annotator's history and a coefficient that describes the direction and extent of influence. For the frequency effect x_{ij} is 1 if $\sum(r_{t-1}, r_{t-2}, r_{t-3}) = 1$, is -1 if the sum is 0 and is 0 otherwise. For the halo effect we have K traits for each of the N individuals. We set x_{ij} equal to j's first impression (z_i at $t = 0$) of individual i. Finally for the priming effect we have a pair of questions where one labeled at time t, mimics its sibling provided the latter is labeled at time $t - 1$. Note that though x_{ij} is constructed from simple building blocks, the composition of these blocks to characterize order effects is non-trivial. For the frequency effect we needed to track the previous three responses and evaluate if they were identical. For halo, we needed to marry entity with information about the annotator's first query. Finally the priming effect necessitated we track each annotator's question schedule and previous response.

The appeal of a non-parametric approach is that the burden of pattern-finding is shifted to the model. Once a critical mass of building blocks has been cataloged, we hope that many

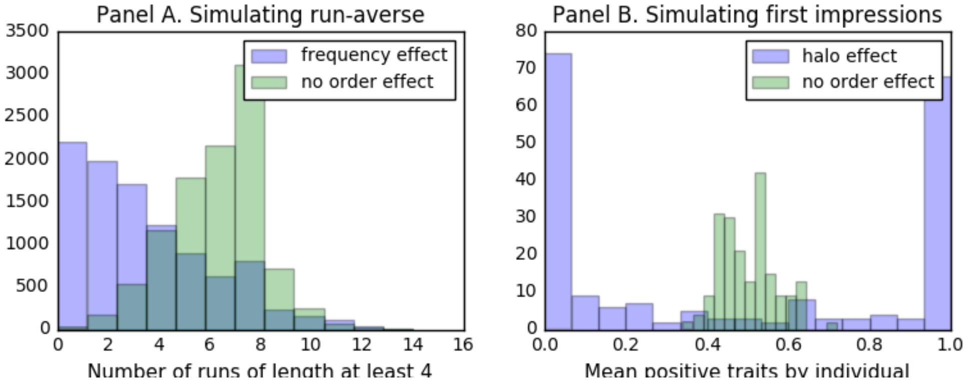


Figure 6.1: *In Panel A. we simulate annotators who are run averse (the frequency effect). We simulate 10,000 datasets of a 100 coin flips. If three heads or three tails are encountered the next flip is biased so as to break the run. For each run-averse coin we count the runs of at least length four and compare it to an unbiased coin. We confirm that the biased coin exhibits fewer runs.*

In Panel B. we simulate annotators who anchor to a first impression (the halo effect). For simplicity we have one annotator label 50 traits for 200 individuals. We calculate the mean number of positive responses by individual and compare it to responses drawn independently from a Bernoulli. We confirm strong within-entity correlation.

Table 6.1: *Mean accuracy spread between random forest and naive model on 100 simulations per case study.*

No Order Effect	Frequency	Halo	Priming
0.02	0.09	0.10	0.46

order effects may be characterized. We use the sanity checklist (**Table 5.1**) to assess whether a random forest is able to recover order effects from simple building blocks in simulations. We simulate 100 datasets for each of the four case studies and track the mean spread between a random forest that predicts the next response using only history and a naive model that predicts the modal class (**Table 6.1**). A random forest appears to successfully recover signal from annotators influenced by the frequency and halo effects. As expected, label history is meaningless in predicting the next response when no order effects prevail. Finally the high accuracy when studying the priming effect suggests we revisit features as it is very likely that information about the intrinsic nature of the next object has leaked in. Indeed this is confirmed by inspection; if I provide a question schedule and the identity of the last query, it is trivial to deduce the identity of the present one.

We apply the probabilistic model to the frequency and halo datasets to test how well we harvest information gleaned by the random forest. First we convert the continuous measure of an object's nature to a "true" label using $z_i > 0$. Then, as in chapter 1, we strive to outperform majority voting. On 100 simulations of the frequency effect, the model yielded a 2.1% [-3.0, 8.9] mean lift in accuracy over majority voting. For the halo effect, the model yielded a 3.2% [-6.0,13.5] lift. Though in both cases, the 95% confidence interval contains zero we are encouraged as (1) as we have shown previously, majority voting is a strong opponent. A more disciplined grid search may reveal the model is successful but that we have picked a configuration of N and J on which we are doomed to fail (2) the success of the random forest over a naive model in **Table 6.1** suggests that a non-parametric strategy was able to glean signal from very simple representations of question history (3) there are some coarse assumptions and clear areas on which we can improve.

6.2 Discussion

The content dependence model captures three key characteristics of the problem:

1. We care most about learning the intrinsic natures of our objects; we wish to be parametric with respect to z_i.
2. Features characterizing order bias are compositions of simple building blocks; we wish to be non-parametric with respect to question history.
3. We expect the intrinsic nature of our objects and question history to be uncorrelated after we control for class imbalance. Together (1) and (2) motivate a semi-parametric approach and, since the parametric and non-parametric components are uncorrelated, one hypothesizes the problem lends itself to being decomposed.

However the results on simulated data suggest some of our assumptions are limiting. There are two key weakness: First, the manner in which we combine the parametric and non-parametric components is coarse. We apply an approach inspired by two-stage regression models like 2SLS. However, non-linearities induced by having to convert continuous variables to 0,1 responses sullies the assumptions on which two-stage regression models are built. We take a strong signal from the random forest, adjust it for class imbalance (by calculating $\hat{p_{ij}} - \bar{p}$) and then massage it into the model as a feature. We need to be more strategic in mitigating the effect non-linearity has on muddying the signal along the way.

Second, we should experiment with other non-parametric models. Random forests were an appealing first choice due to how conveniently they handle categorical predictors (a common format for descriptors of label history). However, predicted probabilities are not finely estimated by random forests; after a grid search that found the optimal number of trees, many probabilities were still estimated to be 0 or 1 thwarting any hope of applying a logit transform and working in the space of the linear predictor. Recurrent neural networks are a clear next step.

7 Conclusion

We have investigated non-stationarity in annotator credibilities through the lens of volume and content dependence. In the case of volume dependence we explored the configurations of redundancy and extent of change in which such a model might thrive. We then applied it to a crowdsourced application labeling social media posts and confirmed a (marginal) lift in performance.

One of the main impediments of the model is that inference is conducted using Hamiltonian Monte Carlo and may be computationally too unwieldy for a practitioner (e.g. this model is ill-suited to inform online decision-making). To mitigate computational burden, it is common to focus on point estimates of annotator skill and frequently expectation-maximization is employed towards this end. Problematically, EM is sensitive to initialization which manifests in its performance being inconsistent. Zhang *et al.* (2014) guide exploration to the region of the solution using spectral decomposition to initialize credibility matrices. Alternatively, variational inference is explored by Liu *et al.* (2012), which would forfeit some accuracy but mitigate computational strain. Going forward it may be useful to explore how a cheaper approximation fares.

In the case of content dependence we proposed a semi-parametric approach that estimated the latent nature of objects while allowing the model to explore the high-order patterns in question history that might induce bias. We tested the model on simulated data and found that though we were able to beat majority voting on average, we didn't do so consistently enough to be reflected in a 95% confidence interval. It is however, encouraging that after injecting an order effect, the random forest was able to meaningfully predict the next response using only question history. We conclude that order effects can be deduced from simple building blocks and there is room for gain if we refine how we carry information from label history to the encompassing probabilistic model.

References

Banachewicz Konrad, Lucas André and Van Der Vaart Aad (2008). Modelling Portfolio Defaults Using Hidden Markov Models with Covariates. *The Econometrics Journal*, **11** (1), 155–171.

Daamen, D. D. L. and Bie, S. E. d. (1992). Serial Context Effects in Survey Interviews. In *Context Effects in Social and Psychological Research*, Springer, New York, NY, pp. 97–113.

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

DeRuiter, S. L., Langrock, R., Skirbutas, T., Goldbogen, J. A., Chalambokidis, J., Friedlaender, A. S. and Southall, B. L. (2016). A multivariate mixed hidden Markov model to analyze blue whale diving behaviour during controlled sound exposures. *arXiv:1602.06570 [q-bio, stat]*, arXiv: 1602.06570.

Donmez, P., Carbonell, J. and Schneider, J. (2010). A Probabilistic Framework to Learn from Multiple Annotators with Time-Varying Accuracy. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, Proceedings, Society for Industrial and Applied Mathematics, pp. 826–837.

Jung, H. J., Park, Y. and Lease, M. (2014). Predicting Next Label Quality: A Time-Series Model of Crowdwork. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Khattak, F. K. and Salleb-Aouissi, A. (2016). Toward a Robust Crowd-labeling Framework using Expert Evaluation and Pairwise Comparison. *arXiv:1607.02174 [cs]*, arXiv: 1607.02174.

Krosnick Jon A., Narayan Sowmya and Smith Wendy R. (2004). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, **1996** (70), 29–44.

Lin, C. H., Mausam and Weld, D. S. (2014). To Re(label), or Not To Re(label). In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Liu, Q., Peng, J. and Ihler, A. (2012). Variational Inference for Crowdsourcing. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, USA: Curran Associates Inc., pp. 692–700.

Meligkotsidou, L. and Dellaportas, P. (2011). Forecasting with non-homogeneous hidden Markov models. *Statistics and Computing*, **21** (3), 439–449.

Nguyen, A. T., Wallace, B. C., Li, J. J., Nenkova, A. and Lease, M. (2017). Aggregating and Predicting Sequence Labels from Crowd Annotations. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, **2017**, 299–309.

Passonneau, R. J. and Carpenter, B. (2014). The Benefits of a Model of Annotation. *Transactions of the Association for Computational Linguistics*, **2** (0), 311–326.

Prelec, D., Seung, H. S. and McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, **541** (7638), 532–535.

Raman, K. and Joachims, T. (2014). Methods for Ordinal Peer Grading. *arXiv:1404.3656 [cs]*, arXiv: 1404.3656.

Raykar, V. C. and Yu, S. (2012). Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *Journal of Machine Learning Research*, **13** (Feb), 491–518.

Ruiz, P., Besler, E., Molina, R. and Katsaggelos, A. K. (2016). Variational Gaussian process for missing label crowdsourcing classification problems. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6.

Salvatier, J., Wiecki, T. V. and Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, **2**, e55.

Schwarz, N. and Sudman, S. (eds.) (1992). *Context effects in social and psychological research*. Springer.

Serenko, A. and Bontis, N. (2013). First in, best dressed: The presence of order-effect bias in journal ranking surveys. *Journal of Informetrics*, **7** (1), 138–144.

Sheng, V. S., Provost, F. and Ipeirotis, P. G. (2008). Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, New York, NY, USA: ACM, pp. 614–622.

Strack, F. (1992). "Order Effects" in Survey Research: Activation and Information Functions of Preceding Questions. In *Context Effects in Social and Psychological Research*, Springer, New York, NY, pp. 23–34.

Thorndike, E. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, **4** (1), 25–29.

Tversky, A. and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, **211** (4481), 453–458.

Welinder, P., Branson, S., Perona, P. and Belongie, S. J. (2010). The Multidimensional Wisdom of Crowds. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta (eds.), *Advances in Neural Information Processing Systems 23*, Curran Associates, Inc., pp. 2424–2432.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R. and Ruvolo, P. L. (2009). Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta (eds.), *Advances in Neural Information Processing Systems 22*, Curran Associates, Inc., pp. 2035–2043.

Zhang, Y., Chen, X., Zhou, D. and Jordan, M. I. (2014). Spectral Methods meet EM: A Provably Optimal Algorithm for Crowdsourcing. *arXiv:1406.3824 [stat]*, arXiv: 1406.3824.