Katharina Hechinger

Statistical Approaches towards Label Uncertainty in Machine Learning Models

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 22.07.2024

Katharina Hechinger

Statistical Approaches towards Label Uncertainty in Machine Learning Models

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 22.07.2024

Erster Berichterstatter: Prof. Dr. Göran Kauermann (LMU München)
Zweite Berichterstatterin: Prof. Dr. Xiao Xiang Zhu (TU München)
Dritte Berichterstatterin: Prof. Dr. Sylvia Frühwirth-Schnatter (WU Wien)

Tag der Disputation: 10.10.2024

Acknowledgments

First and foremost, I want to thank my supervisor, Prof. Dr. Göran Kauermann. His support, guidance, and encouraging words have helped me enormously throughout my academic journey. I would also like to thank Prof. Dr. Xiaoxiang Zhu, my second supervisor, for helpful insights and guidance along the way and Christoph Koller for the fruitful collaboration on our shared project. Furthermore, I want to thank Prof. Dr. Sylvia Frühwirth-Schnatter, who kindly agreed to serve as an external reviewer on the examination committee. Further thanks go to Prof. Dr. Christian Heumann and Prof. Dr. Helmut Küchenhoff for agreeing to steer the committee.

I also want to thank Dr. Matthias Aßenmacher and Prof. Dr. Barbara Plank for their advice and fruitful collaboration on our joint research project. Furthermore, I am grateful to the working group Uncertainty Quantification at the German Aerospace Center/TUM for the interesting paper reading sessions and associated discussions.

Furthermore, I am grateful to my colleagues from the chair for their support and encouragement, as well as for our canteen lunches together. In particular, I want to thank Giacomo De Nicola for answering all my organizational questions. A special thank you goes to Cornelia Gruber, not only for the amazing academic teamwork but also for having an open ear at all times, regardless of the topic. Thanks for the lovely conversations, laughs, and yoga classes together.

A special thanks also goes to Brigitte Maxa for her support in all organizational matters during all my years at the institute.

On a personal note, I want to thank my family and friends, especially my parents and my sister, for their unconditional support and understanding throughout the years. Finally, I want to thank Flo for his love, patience, and everything else on our journey together. Thanks for bearing with me during challenging and stressful times and encouraging me, no matter what.

Summary

This thesis addresses label or annotation uncertainty within supervised classification models, aiming to approach the problem from a statistical perspective. While variance and uncertainty, along with tools for their quantification, are deeply rooted in statistics, the machine learning community has only recently begun to explore them in complete depth. Generally, uncertainty is a rather vague concept and, hence, difficult to define, detect, and quantify. The uncertainty inherent in the ground truth labels used for training supervised classification models is often overlooked but of immense importance.

To put the contributing articles in a broader context, the first part of the thesis gives an overview of the two primary research domains. First, the focus lies on the fundamentals and the general idea of statistical modeling, subsequently introducing two specific model classes central to the contributing articles. Second, uncertainty in classification models is introduced and discussed. Within this context, particular emphasis is placed on the inherent uncertainty within ground truth labels, i.e., the label uncertainty. Analyzing this particular source of uncertainty requires multiple annotations per observation, a condition that only a few benchmark datasets meet. Central examples are, therefore, presented at the end of the first part.

The second part of the thesis builds on the assumption that each observation is associated with a latent one-dimensional ground truth label. In the first contribution, we aim to assess the various sources of label uncertainty based on this central assumption. A multinomial mixture distribution is employed to model multiple annotations and to gain insights into the true class affiliations. The model's parameters are estimated via a stochastic Expectation Maximization algorithm and analyzed to understand the "uncertainty drivers". Based on the analyses of a benchmark dataset in earth observation, this study concludes that the general class distinguishability, the heterogeneity of the annotators, and external influencing factors, like geographic information, contribute to the ambiguity in the final labels. The second contribution builds on this work and extends it in the context of natural language inference. Again, a multinomial mixture model is employed to model multiple annotations based on the assumption that a latent ground truth label exists. This work focuses on the stability of the estimation procedure and explores it in terms of the number of annotations and the number of observations. It concludes that sufficient annotations are a crucial building block for adequately analyzing the associated uncertainty.

The central assumption is extended toward a multi-dimensional representation of the ground truth label in the third part of this thesis. In many realistic applications, a singular label is unrealistic and overly restrictive. Therefore, the third contribution proposes a Dirichlet multinomial model to estimate multi-dimensional label embeddings from multiple annotations, expressing inherent classification difficulty and uncertainty. The model's versatility and relevance are demonstrated through its application to three benchmark datasets from various domains. The estimated embeddings not only provide valuable insights into the inherent uncertainty but also serve as a starting point for subsequent classification models.

Zusammenfassung

Diese Arbeit befasst sich mit der Unsicherheit von Annotationen in Klassifikationsmodellen und nähert sich dem Thema aus statistischer Richtung. In der Statistik sind Varianz und Unsicherheit zentrale Komponenten von jeglichen Modellen. Im Kontext des maschinellen Lernens hingegen wurden diese Konzepte lange eher vernachlässigt und sind erst in den letzten Jahren mehr in den Fokus der Wissenschaft gerückt. Im Allgemeinen ist Unsicherheit ein eher vages Konzept und daher schwer zu analysieren und zu quantifizieren. Viele verschiedene Aspekte tragen dazu bei, wie auch die Unsicherheit, die in den Annotationen steckt, die zum Training von Klassifikationsmodellen verwendet werden. Diese werden generell als „Ground Truth Labels“ angesehen, sind aber oft fehlerhaft oder uneindeutig und sollten daher in die Gesamtbewertung der Unsicherheit eines Modelles mit einfließen.
Der erste Teil der Arbeit gibt einen Überblick über die beiden primären Themen der angehängen Forschungsarbeiten. Zuerst liegt das Augenmerk auf den Grundlagen der statistischen Modellierung. Dabei werden zwei spezifische Modellklassen vorgestellt, die in den Beiträgen eine zentrale Rolle spielen, und Methoden zur Schätzung solcher Modelle diskutiert. Anschließend liegt der Fokus auf dem Konzept der Unsicherheit im Kontext von Machine Learning. In diesem Zusammenhang wird besonderes Augenmerk auf die Unsicherheit in den Labeln gelegt, auch bezeichnet als Annotationsunsicherheit. Zur Analyse dieser Unsicherheitsquelle sind mehrere Annotationen pro Instanz nötig. Beispiele für passende Datensätze werden am Ende des Einleitungsteils vorgestellt.

Der zweite Teil dieser Arbeit behandelt Forschungsarbeiten, unter der Annahme, dass jeder Beobachtung ein latentes eindimensionales Label zugeordnet werden kann. Auf dieser Grundlage werden im ersten Beitrag verschiedene Ursachen für Unsicherheit in Annotationen herausgearbeitet und analysiert. Dabei nimmt man eine multinomiale Mischverteilung für die Annotationen an, um Einblicke in die wahren Klassenzugehörigkeiten zu gewinnen. Die Parameter des Modells werden mittels eines stochastischen iterativen Algorithmus geschätzt. Basierend auf den Ergebnissen für einen Datensatz aus dem Bereich der Erdbeobachtung kommt diese Studie zu dem Schluss, dass die allgemeine Unterscheidbarkeit von Klassen, die Heterogenität der Annotatoren und externe Einflussfaktoren zu Unsicherheiten in den Annotationen beitragen. Der zweite Beitrag baut auf diese Arbeit auf und erweitert sie im Kontext von Natural Language Inference, ein Fachbereich, der sich mit der Klassifizierung von natürlicher Sprache befasst. Auch hier wird ein multinomiales Mischungsmodell verwendet, basierend auf der Annahme, dass ein latentes wahres Label existiert. Diese Arbeit konzentriert sich auf die Stabilität des Schätzverfahrens und untersucht sie im Hinblick auf die Anzahl der Annotationen und die Anzahl der Beobachtungen.

Im dritten Teil der Arbeit wird die zentrale Annahme auf eine mehrdimensionale Darstellung des wahren Labels erweitert. In vielen Anwendungen ist ein eindimensionales Label unrealistisch und zu restriktiv. Daher wird im dritten Forschungsbeitrag ein bayesianisches Modell basierend auf der Dirichlet-Multinomial Verteilung vorgeschlagen. Damit kann aus mehreren Annotationen pro Instanz eine mehrdimensionale Repräsentation des „Ground Truth Labels“ geschätzt werden, um Unsicherheiten in den Annotationen adäquat auszudrücken. Die Vielseitigkeit und Relevanz des Modells wird durch seine Anwendung auf drei verschiedenen Datensätzen demonstriert. Die Ergebnisse bieten nicht nur wertvolle Einblicke in die inhärente Unsicherheit, sondern dienen auch als Ausgangspunkt für nachfolgende Klassifikationsmodelle, um die Annotationsunsicherheit mitberücksichtigen zu können.

Contents

Part I.

Introduction and Background

1. Outline

Variance and uncertainty, along with the tools for quantifying them, are fundamental in the field of statistics. However, the recent increase in interest from the machine learning and deep learning communities in fully exploring these concepts is a rather novel development. Uncertainty, notoriously difficult to define, detect, or quantify, is recognized as of the utmost importance for drawing reliable insights from any data. One contributing but often neglected part is the uncertainty inherent in the ground truth labels used for supervised learning. Classification models, i.e., supervised models, are trained on labeled training data and rely on "gold" labels. The accuracy of such labels is a crucial requirement for building reliable models and obtaining valuable results. However, acquiring ground truth labels is often not straightforward and, in many applications, even impossible for various reasons, including errors and ambiguities. A common strategy is to gather multiple annotations per instance and then use the majority voting as a final label. This simplifies subsequent steps while, of course, discarding valuable information about the uncertainty inherent in the annotations. This thesis addresses this so-called **annotation or label uncertainty** from a statistical perspective. By leveraging the toolbox of statistical modeling, this work provides a starting point for further incorporating knowledge about label uncertainty into the machine learning pipeline.

The first part of this cumulative thesis provides the articles' scientific context and aims to connect the two main concepts: statistical modeling and uncertainty in machine learning.
First, the general idea of and theory behind statistical models is covered in Chapter 2. Therefore, two important model classes, Multinomial Mixture models and Dirichlet Multinomial models, are introduced in more depth as these are vital for the contributing articles. Additionally, strategies for estimating the parameters associated with the models, particularly from a Bayesian perspective, are covered. Lastly, the role of uncertainty and approaches for estimating it in the context of statistical models are discussed briefly.
This bridges the gap to the second main topic of this thesis, uncertainty, and particularly annotation uncertainty in machine learning models, covered in Chapter 3. Thereby, its definition and sources, as well as ways to quantify it, especially in the context of deep learning, are discussed. Subsequently, the focus is explicitly placed on the uncertainty inherent in the labels used for training supervised machine learning models or, alternatively, the uncertainty reflected by multiple annotations. Consequently, this work introduces specific multi-annotator datasets that motivated this work and are extensively analyzed in the contributions.
Chapter 4 concludes the introduction with a short discussion and gives an outlook on possible directions of future work, highlighting the relevance of the topic of annotation uncertainty beyond the articles included in this work.

The first part of this thesis is followed by three contributing articles. Part II deals with approaches towards label uncertainty building on the assumption of a one-dimensional ground truth, as specified further in Chapters 5 and 6. In contrast, Part III moves beyond this assumption. The contribution in Chapter 7 proposes a model framework that allows the estimation of multi-dimensional

representations of ground truth labels. Chapters 5, 6 and 7 contain the published versions of the articles, along with a description of the contributions of all authors involved. Note that the notation in the contributing articles differs from that in the introductory part if necessary.

2. Statistical Modeling

"All models are approximations. Essentially, all models are wrong, but some are useful."
— George E. P. Box
(* 1919, † 2013)

Parametric statistical modeling is one of the most fundamental concepts in statistics. This approach seeks to extract valuable information from sampled data and to analyze the data-generating process. In this context, the term "process" refers to a probabilistic model characterized by parameters, offering a structured framework to describe and understand the generating mechanisms behind the observed data. The textbooks by Davison (2003) or Kauermann et al. (2021) provide an introduction to and thematic classification of the topic. A formal mathematical definition of statistical models and the associated parameters is provided by McCullagh (1980). As stated by Konishi and Kitagawa (2008), statistical models serve three main purposes: prediction, extraction of information, and description of stochastic structures. Specifically in the field of machine learning and neural networks, the primary objective of models often was and still is the first one, i.e., the prediction of future events or outcomes based on observed data. However, it is also essential to highlight the significance of extracting information and patterns from data to understand underlying mechanisms better and describe and quantify stochastic structures that represent uncertainties. These aspects constitute crucial components of data analysis, and the sole focus on prediction can have severe and undesired consequences. In general, parametric statistical modeling provides a flexible and powerful framework for analyzing data. Its versatility is evident in numerous fields, including psychology, economics, biology, and engineering.

First, let us introduce the general framework of parametric statistical models and derive central functions for estimation. Formally, the random variable Y is assumed to originate from a parameterized probability model

$$Y \sim F(y; \boldsymbol{\theta}).$$

The term $F(y; \boldsymbol{\theta})$ refers to the cumulative distribution function of the random variable Y and can be defined as $F(y; \boldsymbol{\theta}) = \mathbb{P}(Y \leq y)$. It depends on a parameter vector $\boldsymbol{\theta}$. Hence, the form of the probability distribution function is commonly known up to some p-dimensional parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$. The true values of these parameters remain unknown and can only be approximated in practice based on the observed data sample. Standard parametric distributions include the Gaussian distribution for continuous data, the binomial distribution for count data with two possible outcomes, or the exponential distribution for modeling the time between two events. More examples and details can be found in the textbook by Kroese et al. (2014), for example. The behavior of random variables is typically described by their **density** function $f(.)$, defined as $f(y) = \mathbb{P}(Y = y)$ for discrete random variables or as an integrable function where $\mathbb{P}(a \leq Y \leq b) = \int_a^b f(y)dy$ for continuous random variables.
Given an actual data sample $y = (y_1, ..., y_n)$, i.e., realizations of the random variable Y, and

the density $f(y;\boldsymbol{\theta})$, the Bayes rule can be applied. The Bayes rule is a fundamental theorem in statistics and is used to decompose the conditional density:

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|y) = \frac{f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \cdot f_Y(y|\boldsymbol{\theta})}{f_Y(y)} \propto f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \cdot f_Y(y|\boldsymbol{\theta}). \tag{2.1}$$

This formula defines three central functions for the estimation of the parameter vector $\boldsymbol{\theta}$, namely the **posterior distribution** $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|y)$, the **prior distribution** $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and the **likelihood function** $f_Y(y|\boldsymbol{\theta})$, also denoted by $L(\boldsymbol{\theta};y)$. The likelihood function represents the probability of generating the current data sample, given the parameter $\boldsymbol{\theta}$, and relies exclusively on the observed data. On the other hand, the prior distribution conveys the knowledge about $\boldsymbol{\theta}$ before any data is observed. The posterior distribution merges both by updating the knowledge of the parameter after new data is observed. Ultimately, the goal is to find the "best" estimate for the true parameters $\boldsymbol{\theta}$ given the observed data sample $y_1, ..., y_n$, defined as $\hat{\boldsymbol{\theta}} = s(y_1, ..., y_n)$. The function $s(.)$, also known as statistic, calculates the optimal parameter values based on the available data. Numerous methods exist to obtain $\hat{\boldsymbol{\theta}}$, and the appropriate choice depends on several factors, such as the underlying model, the available data, and the researcher's preferences and objectives. Some of the most popular approaches will be shortly sketched here.
In frequentist statistics, the most common approach to derive optimal parameter estimates is **Maximum Likelihood** estimation. Returning to the decomposition in (2.1), this approach assumes a uniform prior, i.e., $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is constant. Then, the estimates can be obtained by directly maximizing the likelihood function,

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}; y_1, ..., y_n). \tag{2.2}$$

This approach assumes that the observed data are generated from a specific probability distribution, and the goal is to find the parameter values that make the observed data most likely. In the field of statistics, it is widely used due to its simplicity and asymptotic properties, such as consistency and efficiency. In practice, $L(.)$ is often replaced by its logarithmic version for equal results at less computation costs.
From a Bayesian viewpoint, however, the posterior distribution encapsulates all information about the true parameter $\boldsymbol{\theta}$. Therefore, a natural estimate for the true parameters is the posterior mean value, defined as

$$\hat{\boldsymbol{\theta}}_{PMean} = E_{\boldsymbol{\theta}}(\boldsymbol{\theta}|y_1, ..., y_n).$$

The posterior mean is obtained by averaging all possible parameter values weighted by their posterior probabilities. Hence, it incorporates both the uncertainty of the parameters and the information the data provides. Alternatively, the mode of the posterior distribution can be chosen as the optimal parameter value.

Connection to Machine Learning Machine learning (ML) models are also parametric statistical models that aim to find optimal parameter values. Neural networks, for example, can be conceptualized as flexible and powerful statistical models that are capable of learning complex mappings between input data and output predictions. Depending on the task and architecture, neural networks can exhibit characteristics of regression models, classification models, probabilistic models, and latent variable models. Their versatility allows for various applications of networks in ML and data analysis (Cheng and Titterington, 1994). However, the purpose of such complex models,

	1	2	$\cdots$	$P-1$	P
1	j_{11}	j_{12}	$\cdots$	j_{1P-1}	j_{1P}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
N	j_{N1}	j_{N2}	$\cdots$	j_{NP-1}	j_{NP}

Table 2.1.: Sketch of a database of multiple annotations for N observations categorized into P classes by J annotators, $j_{ip} \in [0, J]$ denotes the number of annotations for class p for observation i.

particularly of deep neural networks, is often restricted to predicting future outcomes. According to Shmueli (2010), models can be of explanatory or predictive nature. Both play a central role in many statistical models. In contrast, deep neural networks often focus exclusively on prediction, leading to accurate but complex black-box-like models. The term "black-box models" refers to models whose internal workings are not easily interpretable or explainable, making it difficult to understand final predictions. For many years, the primary metric for measuring and comparing the performance of deep neural networks was their **accuracy**, i.e., how well the predictions match the ground truth. From a statistical viewpoint, this is a minimal perception and should by far not be the sole performance criterion. Instead, factors like the explainability or uncertainty of the results are vivid to be considered upon evaluation. Due to the growing complexity of the networks, explaining the results is not straightforward. This issues a whole area of research, often referred to as explainable AI, see, e.g., the survey by Linardatos et al. (2020). The reliability of the obtained results is closely connected to this issue. While accurate predictions are desired, without a measure of uncertainty, they are often of limited use in practical and critical applications. Hence, uncertainty in ML models is a highly relevant topic and will be discussed in more depth in Chapter 3.
However, specific settings enable us to evaluate aspects of the overall uncertainty using tools or methodologies derived from classical statistical modeling. In particular, this work deals with the problem of label uncertainty. In situations where multiple annotations are available for each instance, potential ambiguity in the correct label is expressed via multiple annotations. Modeling these annotations using the framework of parametric statistical models allows to extract information about the amount or the sources of inherent label uncertainty. Annotations can thereby be interpreted as multinomial count data, i.e., J annotators categorize N instances into P classes. Table 2.1 presents an exemplary data structure. The two model classes discussed in the following are suitable for handling data of such form. The primary goal is to model multiple annotations with statistical tools and, hence, develop simple yet effective approaches to assess label uncertainty.

2.1. Multinomial Mixture Model

This section introduces the multinomial mixture model, which is frequently deployed to describe categorical variables originating from different underlying components. The central assumption is that each instance is associated with a random, possibly latent variable expressing affiliation to an original component. The resulting clustered structure might be artificial but nevertheless helpful for uncovering hidden structures and cluster-specific characteristics. Therefore, this model class is well-suited for analyzing annotations, as outlined in Table 2.1, and is explained and applied in depth in the contributions in Chapters 5 and 6. First, this subsection introduces general mixture

models and some of their peculiarities before focusing on multinomial mixture models in the context of modeling annotations.

2.1.1. Mixture Models

Taking a closer look at finite mixture models in general is worthwhile, as they serve as a fundamental concept in statistical modeling. Models based on mixture distributions provide a flexible and convenient framework for representing complex data distributions by combining separate components. Unlike simpler models, which postulate a single distribution for the entire dataset, mixture models assume that the data is generated from a mixture of multiple component distributions. This is particularly valuable when dealing with complex and heterogeneous data, where observations are assumed to arise from different underlying processes or groups. Consequently, employing a single distribution is inadequate for capturing the entire complexity of the data in such situations. Figure 2.1 shows an exemplary case of simulated data originating from three underlying Gaussian distributions. By fitting only one distribution, it is impossible to capture the multi-modality of the data. Instead, the framework of mixture models allows the combination of multiple Gaussians to describe the data more accurately, as shown by the density plots.
Mixture models have been applied in various application areas and numerous domains in practice. Exemplary, this model class is popular for processing language in the form of text or speech (Reynolds and Rose, 1995) or genome classification (Allison et al., 2002).
Let us formally introduce the mixture model class, roughly following the notation used in the contributing articles in Chapter 5 and 6. The density of a finite mixture distribution with K mixing components of a random vector $\boldsymbol{Y}$ takes the form

$$f(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{k=1}^{K} \eta_k f_k(\boldsymbol{y}),$$

where $\eta_k, k = 1, ..., K$ denote the component-specific mixture weights, fulfilling $\sum_{k=1}^{K} \eta_k = 1$. Typically, the component densities $f_k(\boldsymbol{y})$ are known up to some component-specific parameter vector $\boldsymbol{\theta}_k$. They can, therefore, be written as $f(\boldsymbol{y}|\boldsymbol{\theta}_k)$, which refers to the density function of the k-th component distribution given the respective parameters. Hence, the probability of observing a data point $\boldsymbol{y}$ is calculated as the weighted sum of probabilities from different component distributions, each defined by its own set of parameters. In the remainder of this work and often also in practice, the proposition "finite" referring to a finite number of hidden components is omitted. However, infinite mixtures, i.e., $K \to \infty$, of distributions also exist, see, e.g., Rasmussen (1999), but will not be discussed further in this work.
For a recent general overview of mixture models, the reader is referred to McLachlan et al. (2019). Detailed information on the model class can also be found in various textbooks, including McLachlan and Basford (1988), Frühwirth-Schnatter (2006) and Mengersen et al. (2011).

Latent Variables Mixture models are closely connected to the concept of latent variables, i.e., variables that remain unobserved and, hence, are not part of the dataset. Data analysis in the presence of latent variables is a broad area of research, and numerous methods are available depending on the nature of the observed and unobserved variables. For example, if both the observed and the latent quantities are categorical, latent class models can be applied. More details and a broad introduction are given by, e.g., McCutcheon (1987). Generally, assuming that

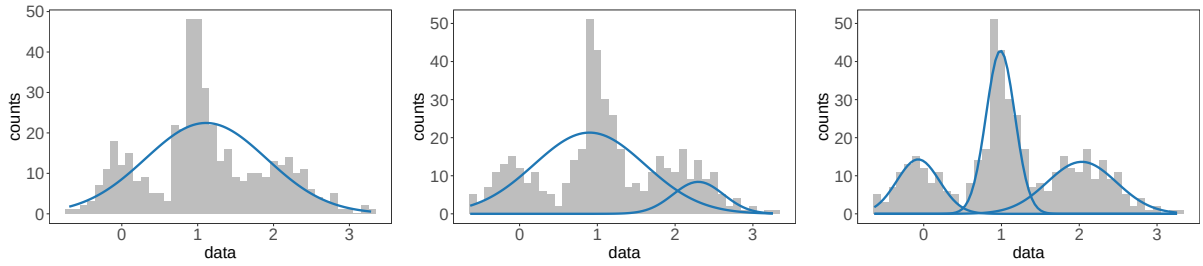


Figure 2.1.: The figure shows the histogram of a dataset simulated using three Gaussian distributions with different means and standard deviations. Fitting a mixture model with the help of the `R` package `mixtools` (Benaglia et al., 2010) allows us to estimate the underlying densities for $K = 1$, $K = 2$ and $K = 3$, which corresponds to the true number of underlying components.

the "full" relevant dataset also contains unobservable quantities allows the expression of complex distributions in simpler terms, similar to the core idea of mixture models. The distribution of the observed variables can be obtained by marginalizing the joint distribution. Formally, a latent variable model is a model over the set of observed variables $\boldsymbol{Y}$ and the set of unobserved variables $\boldsymbol{Z}$, following the joint density $f(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})$.
In mixture models with unknown allocations of the observations, the latent variables $\boldsymbol{Z}$ can be used to represent the underlying structure of the data $\boldsymbol{Y}$, i.e., the unknown affiliations to subgroups. Each observation is associated with a "source" component, which is often impossible to observe in practice and, therefore, latent. However, the observed variables depend on the latent affiliation to one of the components. When mixture models are employed in the context of latent variables, estimation of the optimal parameter values of the mixture distributions is not straightforward in the typical sense. To utilize an estimation procedure like maximum likelihood estimation (2.2), an iterative process has to be introduced. One widely popular method for this task is the Expectation Maximization (EM) algorithm, as proposed by Dempster et al. (1977), or any of its numerous variants. Details on the parameter estimation despite latent quantities and the iterative EM algorithm are covered in Section 2.3.
Latent variables are also closely connected to the general uncertainty of a model, as discussed further in Chapter 3. They contribute to the overall uncertainty of models in various ways if left untreated or unconsidered within the pipeline, see Gruber et al. (2023).

Number of Components In general, mixture models comprise a finite number of component distributions, denoted by $K < \infty$. It is worth noting that these models can handle an arbitrary number of components, i.e., the specification of K is not strictly necessary beforehand.
In many applications, however, K is assumed to be known in advance. Hence, the problem simplifies to allocating observations into the components $1, ..., K$. For a fixed and known value K, empty components can occur if categories are not represented in the observed sample, i.e., if $\eta_k = 0$ or if η_k is close to zero. Empty components can cause instabilities in the estimation of the parameters or degenerate solutions. Therefore, special attention should be paid to very small or even empty categories in the dataset. In certain cases, removing or consolidating tiny classes, depending on the specific application, may be inevitable to ensure stable estimation and interpretable results. In the contributing article in Chapter 5, it was necessary to omit one class due to its infrequent occurrence and the resulting instabilities in the estimation process.

However, the problem is even more prevalent if the actual number of components remains unknown. While this setting provides flexibility, which is desirable in some applications, it can also cause additional uncertainties and problems related to model specification. Choosing the number of components K to be larger than the actual number of clusters K^* leads to overfitting and irregular parameter estimation. In extreme cases, each data point could be represented by a separate component distribution, i.e., $K = N$. However, this issue also involves determining whether the observed data is better represented by a mixture of distributions rather than a single distribution, effectively questioning if $K > 1$ is appropriate. Overfitting the number of components is a common problem in mixture modeling. Hence, numerous approaches have been proposed to handle such model misspecifications and choose a suitable number of components. Frühwirth-Schnatter (2006) discusses approaches to guarantee identifiability via formal constraints on the parameter space. Simple examples are setting $\eta_k > 0, \forall k = 1, ..., K$, i.e., not allowing for empty components in general or imposing a nonequality constraint on the component parameters. When the number of components is unknown, Bayesian techniques, such as selecting a suitable prior distribution, can be beneficial. These methods can achieve identifiability by bounding the posterior distribution or applying shrinkage to the component parameters. The study by Rousseau and Mengersen (2011) analyzes the behavior of the posterior under overfitting conditions in depth.
Additionally, model selection criteria can be calculated for various models to identify the most suitable number of components. These measures are often based on the likelihood. A popular, rather heuristic, criterion is the Akaike Information Criterion (AIC) proposed by Akaike (1974). Commonly, model M_K, where K represents the number of components, is selected to minimize the quantity

$$AIC = -2 \log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_K, M_K) + 2d_K.$$

The parameter d_K refers to the dimensionality of the parameter vector, i.e., the model complexity. High-dimensional models are, therefore, penalized by the AIC. Alternatively, criteria like the BIC or the Schwarz Criterion can be used, which work similarly. Another likelihood-based method for model selection is hypothesis testing, i.e., employing a variant of the likelihood ratio test, where the goodness-of-fit of two model specifications is compared. Note that the usual regularity conditions do not hold, and therefore, one cannot assume that the test statistic follows a χ^2 distribution in this case. Instead, the respective p-values are often calculated via resampling approaches (McLachlan and Rathnayake, 2014). These measures commonly serve as model selection tools and can be computed for various model specifications to determine the best fit and, hence, the optimal number of components. Details on the methods and their tailoring towards mixture models are presented in Chapter 4 of Frühwirth-Schnatter (2006).

Label Switching A challenge specific to all variants of mixture models is the identifiability problem due to so-called "label switching", see Redner and Walker (1984). The components found by a mixture model do not have any inherent meaning or labeling and are, therefore, exchangeable. This phenomenon does not affect the likelihood or constitute an issue from the mathematical perspective. However, it can lead to problems related to the interpretation, specifically for the specified multinomial mixture model, as discussed in Chapters 5 and 6. In this case, latent clusters corresponding to given classes with inherent meaning and ordering should be recovered. To address this problem and to provide a meaningful interpretation of the components, strategies to either avoid label switching or resolve it after estimation need to be applied. For example, this can be achieved by introducing identifiability constraints to ensure unique labeling or by choosing

informative prior specifications that can break symmetries. Additionally, post-processing methods like relabeling algorithms or permutation-based approaches have been developed for this purpose. For an overview of the problem and possible solutions, see Celeux (1998) or Stephens (2000). However, if allocating the observations to the components is straightforward, heuristic methods specific to the application at hand suffice in most cases. In the contributing article in Chapter 5, we apply a post-processing step to the final parameter estimates to ensure that the labeling of the mixture components matches the labels provided by the original count observations. Therefore, a permutation σ was constructed on the numbers $\{1, \dots K\}$ such that $\sigma(C_k) = p$ expresses that the latent component C_k corresponds to the original label L_p. In this case, this can be accomplished by comparing the posterior probabilities of the latent classes based on the annotators' evaluations and matching them to the voted classes. If the posteriors are crisp, i.e., if classification is straightforward, it is easy to determine a distinct permutation. However, if the categories are ambiguous, a unique mapping cannot be guaranteed, and further constraints are necessary. Details on the specific algorithm applied to address label switching can be found in Chapter 5.

2.1.2. Multinomial Mixture Models

A specific type of mixture model is the multinomial mixture model. It is used for modeling categorical data. Hence, the data within the components follow separate multinomial distributions. This model class is useful for modeling data represented as counts or frequencies and for multiple annotations as sketched in Table 2.1. Furthermore, popular application areas of this specific model class include text mining, where documents are represented as word frequency vectors, or image segmentation based on the frequency of pixels. The contributions in Chapters 5 and 6 deal with latent class affiliations and employ a multinomial mixture model for modeling the categorical variables at hand. Therefore, K latent components are assumed. One defines the observed discrete counts as $\boldsymbol{Y}^{(i)} = (Y_1^{(i)}, ..., Y_P^{(i)})$, with P denoting the number of observed categories. The latent variable $Z^{(i)} \in \{1, \dots, K\}$ corresponds to the unknown component affiliations of the instances and can be re-written as $\boldsymbol{Z}^{(i)} = (\mathbb{1}\{Z^{(i)} = 1\}, ..., \mathbb{1}\{Z^{(i)} = K\})$ for notational simplicity. The multinomial mixture model assumes

$$\begin{aligned} \boldsymbol{Z}^{(i)} &\sim Multi(\boldsymbol{\eta}, 1) \text{ i.i.d.} \\ \boldsymbol{Y}^{(i)} | (Z^{(i)} = k) &\sim Multi(\boldsymbol{\pi}_k, J). \end{aligned} \tag{2.3}$$

The parameters of the first multinomial distribution $\boldsymbol{\eta} = (\eta_1, ..., \eta_K)$ describe the overall probability of each latent class and can therefore be interpreted as prior probability. The second distribution is parameterized by $\boldsymbol{\pi}_k = (\pi_{k1}, \dots, \pi_{kP})$, corresponding to the multinomial probability of the respective component, specified given $Z^{(i)} = k$. Quantity J denotes the number of draws from the multinomial distribution. The multinomial probabilities can be summarized into the matrix $\boldsymbol{C} = (\pi_{kp}, k = 1, ..., K; p = 1, ..., P)$, where index k refers to the true component and p to the observed class. The probability density function of the multinomial distribution is defined as

$$f(\boldsymbol{y}; \boldsymbol{\pi}_k) = \frac{J!}{y_1! \cdots y_P!} \pi_{k1}^{y_1} \cdots \pi_{kP}^{y_P}. \tag{2.4}$$

The likelihood contribution of image i is then obtained by calculating

$$\mathbb{P}(\boldsymbol{Y}^{(i)}; \boldsymbol{\eta}, \boldsymbol{C}) = \sum_{k=1}^{K} \eta_k \mathbb{P}(\boldsymbol{Y}^{(i)}; \boldsymbol{\pi}_k).$$

In the context of modeling annotations, the associated parameters can be interpreted in multiple ways. Given that $\boldsymbol{Z}$ is latent, this model provides insights into the true ground truth labels by estimating its parameters. The latent component corresponds to the true label of an observed instance. Hence, the mixing probabilities $\boldsymbol{\eta} = (\eta_1, ..., \eta_K)$ express the overall probability of observing an instance from the respective class. Additionally, the estimated matrix $\boldsymbol{C}$ contains the multinomial probabilities and can be interpreted as a confusion matrix. The entries $\pi_{kp}, k = 1, ..., K, p = 1, ..., K$ represent correct and incorrect classification probabilities. This corresponds to the probability of a correct ($p = k$) or incorrect annotation ($p \neq k$) given the true class k. Note that for the application of modeling multiple annotations and assessing the "true" ground truth label, the number of observed classes P equals the number of latent components K. In theory and without any uncertainty about the true affiliations, one would expect "perfect" annotation, i.e., the confusion matrix $\boldsymbol{C}$ would equal the unit matrix. Of course, this is not the practice case, and the entries of $\boldsymbol{C}$ usually differ from zero or one, respectively. The contribution in Chapter 5 analyzes this confusion matrix extensively and derives various insights into the labeling process. By inspecting the individual entries of the confusion matrix, it is possible to identify classes that are harder for the annotators to distinguish and, hence, should be treated with more caution. The associated posterior probabilities based on the estimated parameters are another central result of the multinomial mixture model. These are defined as

$$\tau_k^{(i)} = \mathbb{P}(Z^{(i)} = k | \boldsymbol{Y}^{(i)}; \boldsymbol{\eta}, \boldsymbol{C})$$

and result from the estimation with the EM algorithm, as described in Section 2.3. In the context of modeling annotations, these estimates represent the probability that an instance belongs to the true class k, given the observed annotations. If the labels are annotator-specific, the posterior probabilities allow for analyzing the effect of the individual labeler, see Chapter 5. Chapter 6 uses the posterior distributions of the true latent classes to define "decision borders" for the individual classes. This strategy allows insights into the labeling process itself and the confusion happening for ambiguous instances.

2.2. Dirichlet Multinomial Model

A natural extension of the model class presented in Section 2.1 is the Dirichlet multinomial model. While the multinomial probabilities were assumed to be fixed or estimated directly from the data beforehand, one can also take on a Bayesian perspective and assume a prior distribution for those parameters, which will be discussed in the following.

Dirichlet Distribution A popular choice for the prior distribution of multinomial parameters is the Dirichlet distribution. First, it is worthwhile to focus on the univariate version, the Beta distribution. The latter is a continuous distribution defined on the interval $[0, 1]$ and is, hence, often referred to as a probability distribution over probabilities. It is specified by two parameters, α and β, which control the shape of the distribution. The parameters act in a complementary manner, i.e., the higher α, the more the distribution is skewed towards 1, and vice versa, the opposite holds for β. If both parameters are large, the distribution will peak more with a smaller variance. Values equal to 1 lead to a uniform distribution. The Beta distribution is particularly popular in Bayesian statistics, as it is the conjugate prior to the binomial distribution. If prior and posterior belong to the same family of distributions, the prior is called a conjugate prior for the

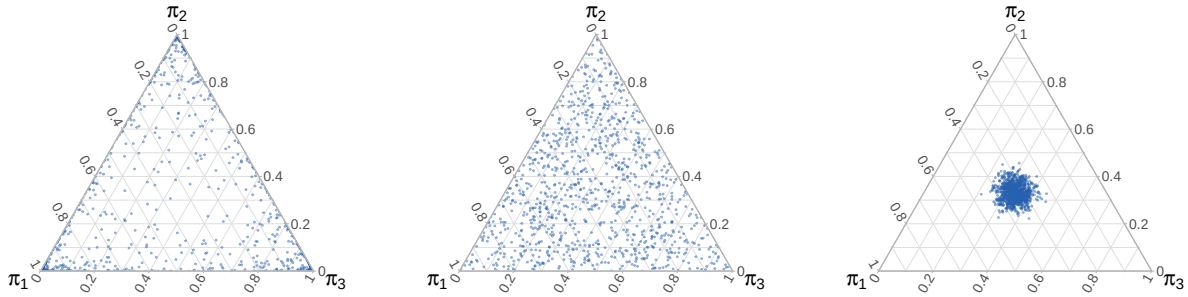
likelihood function. Returning to the Bayes theorem in Formula (2.1), the posterior, and hence all information about the parameter $\boldsymbol{\theta}$, can be obtained as a product of the likelihood and the prior. Conjugate priors are popular in Bayesian statistics as they allow for efficient and tractable updates of beliefs as new data is observed. Using a conjugate prior often allows to derive a closed-form expression for the posterior. Details can be found in various textbooks on Bayesian statistics, e.g., by Robert et al. (2007) or Bernardo and Smith (2009). Other popular examples are Poisson likelihood and Gamma prior or, in the case of continuous random variables, Gaussian likelihood and Gaussian prior. For binary variables, placing a Beta distribution on the parameters of a binomial distribution allows us to estimate the uncertainty in the binomial probability.
However, binary classification and the associated binomial distribution are insufficient in many scenarios. Therefore, the Dirichlet distribution serves as a multivariate extension of the Beta distribution (Minka, 2000). It constitutes the conjugate prior to the categorical and multinomial distribution and can again be interpreted as a probability distribution over probabilities for multiple classes. If a Dirichlet prior is assumed for the parameters in a Bayesian framework, the posterior also follows a Dirichlet distribution. Generally, it takes P non-negative parameters $\alpha_1, ..., \alpha_P$. It describes the random variables $Y_1, ..., Y_P, P \geq 2$ such that $y_p \in (0, 1)$ and $\sum_{p=1}^{P} y_p = 1$. The density function is defined as

$$f(\boldsymbol{y}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{p=1}^{P} y_p^{\alpha_p - 1}, \tag{2.5}$$

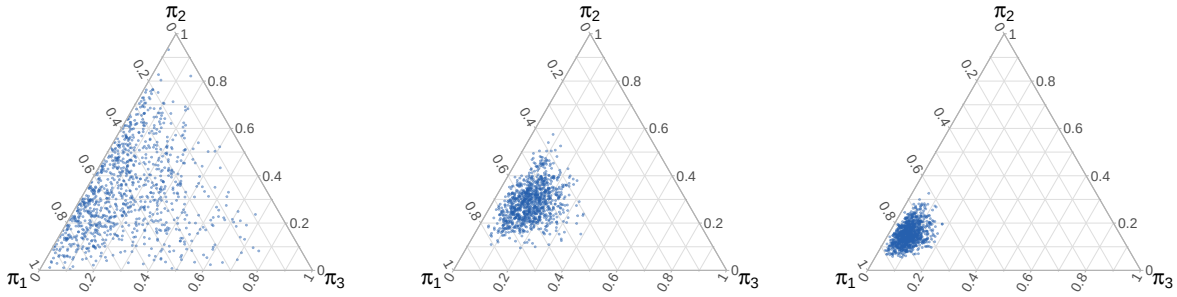
where $B(.)$ denotes the multivariate Beta function and is a normalizing constant. Figure 2.2 shows samples from Dirichlet distributions with varying parameter configurations. For $K = 3$, the samples can be visualized nicely in a ternary plot, spanned by three dimensions that sum up to 1, via the `R` package `ggtern` (Hamilton and Ferry, 2018).
Recently, the Dirichlet distribution has also been employed in the context of neural networks and, in particular, uncertainty quantification in the form of evidential learning (Sensoy et al., 2018, Ulmer et al., 2023). More details are provided in Section 3.2.

Dirichlet Multinomial Model The Dirichlet distribution is a popular prior choice for the parameters of a multinomial model $\pi_1, ..., \pi_P$, as introduced in Formula (2.3). Combining the model (2.3) with a Dirichlet prior results in the Dirichlet Multinomial model. The associated distribution, also known as the multivariate Pólya distribution, is popular in statistics and ML for modeling categorical variables under prior knowledge. Intuitively, the Bayesian view provides "smoothing" to the predictive distribution of count data and allows the inclusion of categories, even if events did not occur in the training data. This simplifies estimation and computation processes. The respective Bayesian model finds widespread use in modeling various forms of count data, for example, in the application areas of genetics (Nowicka and Robinson, 2016) or the clustering of documents (Elkan, 2006).
Formally, the Dirichlet multinomial model results where the data $\boldsymbol{Y}$ follow a multinomial distribution. The latter is parameterized by $\boldsymbol{\pi}$, distributed according to a Dirichlet distribution with hyperparameters $\boldsymbol{\alpha}$:

$$\begin{aligned} \boldsymbol{\pi} &\sim Dir(\alpha_1, ..., \alpha_P) \text{ with } \alpha_p > 0 \\ \boldsymbol{Y}|\boldsymbol{\pi} &\sim Multi(\boldsymbol{\pi}, J). \end{aligned}$$



(a) $\alpha_1 = 0.2, \alpha_2 = 0.2, \alpha_3 = 0.2$ (b) $\alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 1$ (c) $\alpha_1 = 50, \alpha_2 = 50, \alpha_3 = 50$



(d) $\alpha_1 = 3, \alpha_2 = 2, \alpha_3 = 1$ (e) $\alpha_1 = 20, \alpha_2 = 10, \alpha_3 = 5$ (f) $\alpha_1 = 50, \alpha_2 = 10, \alpha_3 = 5$

Figure 2.2.: Each subfigure shows $N = 1000$ random samples from a Dirichlet probability distribution with different shape parameters as scatter points on a three-dimensional simplex.

Again, J denotes the number of draws from the multinomial distribution. Combining the two probability functions (2.5) and (2.4) allows to integrate over the unknown parameter $\boldsymbol{\pi}$, leading to the probability density

$$
\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{\alpha}) &= \int_{\boldsymbol{\pi}} f(\boldsymbol{y}|\boldsymbol{\pi}) f(\boldsymbol{\pi}|\boldsymbol{\alpha}) d\boldsymbol{\pi} = \frac{J!}{y_1!...y_P!} \cdot \frac{1}{B(\boldsymbol{\alpha})} \cdot \int_{\boldsymbol{\pi}} \prod_p \pi_p^{y_p + \alpha_p - 1} d\boldsymbol{\pi} \\
&= \frac{J!}{y_1!...y_P!} \cdot \frac{\Gamma(\sum \alpha_p)}{\prod_p \Gamma(\alpha_p)} \cdot \frac{\prod_p (\Gamma(\alpha_p + y_p))}{\Gamma(\sum_p \alpha_p + y_p)} = \frac{J!}{y_1!...y_P!} \cdot \frac{\Gamma(\sum \alpha_k)}{\Gamma(\sum_p \alpha_p + y_p)} \cdot \prod_p \frac{\Gamma(\alpha_p + y_p)}{\Gamma(\alpha_p)} \\
&= \frac{J!}{y_1!...y_P!} \cdot \frac{B(\boldsymbol{\alpha} + \boldsymbol{y})}{B(\boldsymbol{\alpha})}.
\end{aligned}
$$

Additional details on the respective model can be found in the contributing article in Chapter 7. Assuming a Dirichlet prior for the parameters in the context of modeling multiple annotations allows one to express belief about the proportions of the correct categories. Again, the parameters $\boldsymbol{\alpha}$ control the shape of the distribution, see also Figure 2.2, and hence the expected probabilities of the categories. The Dirichlet distribution is symmetric for equal values of α_p, as the upper row of Figure 2.2 shows. In particular, $\alpha_p = 1 \forall p = 1, ..., P$ corresponds to a uniform distribution of the points, see 2.2b. High values of α_p express a higher weight of π_p, and hence, in the context of probabilities, a higher belief in category p. This is expressed in the bottom row of Figure 2.2, where a tendency towards class 1 is evident. The parameters α_p and $\alpha_{p'}$ can also be compared in size. A value twice as large indicates that category p is twice as prevalent as category p'.

Additionally, the sum $\alpha_0 = \sum_{p=1}^{P} \alpha_p$ expresses the overall confidence in the prior belief. Larger values, as in Figures 2.2c or 2.2f, indicate a strong a priori knowledge, while smaller sums reflect a larger influence of the observed data.
In the context of modeling multiple annotations, this model class allows the extension of the limiting assumption of a singular latent true label. Instead, one can estimate a P-dimensional label representation for each instance based on the Dirichlet parameters $\alpha_1, ..., \alpha_P$ expressing prior beliefs about the corresponding categories for the individual observations. Again, details are provided in Chapter 7.

Dirichlet Multinomial Mixtures A popular and even more flexible extension of the described model class combines mixture models and Dirichlet priors by incorporating a mixture of Dirichlet distributions over the category probabilities. Instead of a single distribution, this model class assumes that the category probabilities $\boldsymbol{\pi}$ are generated from a mixture of Q Dirichlet distributions, where Q denotes the number of components in the mixture. Again, the parameters $\eta_q, q = 1, ..., Q$ represent the mixing coefficients. For each component q, a separate Dirichlet distribution is defined with parameters $\alpha_{q1}, ..., \alpha_{qP}$. As the shape parameters vary across the components, each component can capture different data characteristics. Hence, the model class allows for more heterogeneity and provides a more flexible framework for modeling categorical data with prior beliefs. It is especially popular in the domain of biology and genomics. For example, Holmes et al. (2012) employed a Dirichlet multinomial mixture model in the context of modeling and clustering genomics data into meta-communities.

2.3. Parameter Inference

At the beginning of this chapter, some of the most popular straightforward approaches for estimating the parameters of a model were already quickly sketched. However, in the presence of latent variables, maximum likelihood estimation cannot be carried out in a straightforward manner and, instead, has to be done iteratively. The parameters in the model classes described previously are either fixed or follow some prior distribution specified by hyper-parameters that must be estimated. The following subsections discuss suitable estimation approaches to obtain optimal parameter estimates despite latent variables and in Bayesian settings.

2.3.1. (Stochastic) Expectation Maximization

When dealing with latent variables, as presented in Section 2.1, direct parameter estimation via maximum likelihood is impossible. This is where the Expectation Maximization (EM) algorithm steps in, as proposed in the seminal paper by Dempster et al. (1977). The EM algorithm is an iterative procedure that is crucial in determining parameter estimates despite missing or unobserved quantities. Its widespread popularity in the field of statistics underlines its importance.
To calculate the maximum likelihood estimates for the set of parameters, denoted by $\boldsymbol{\theta}$, one assumes a complete dataset $\boldsymbol{W}$, which consists of the observed and unobserved data, i.e., $\boldsymbol{W} = (\boldsymbol{Y}, \boldsymbol{Z})$. The interest lies in the log-likelihood of the complete data, which is defined as

$$\log L(\boldsymbol{\theta}) = \log f(\boldsymbol{w}|\boldsymbol{\theta}). \tag{2.6}$$

As a part of $\boldsymbol{w}$ remains unobserved, computation and maximization of the full-data likelihood is impossible. Therefore, the EM algorithm is applied. Intuitively, it finds an approximation for the full-data likelihood function first, then maximizes it to update the parameters and repeats the procedure until convergence. The latent quantity $\boldsymbol{Z}$ is a random variable. Hence, as a function of $\boldsymbol{Z}$, the full-data likelihood function (2.6) is again a random variable. Taking its expectation given the observed data $\boldsymbol{y}$ leads to the required approximation

$$\mathbb{E}\Big(\log L(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}\Big),$$

with $\boldsymbol{\theta}^{(t)}$ denoting the current parameter estimates. This is referred to as the expectation step. The following step consists of maximizing this quantity to get new parameter estimates as

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}\Big(\log L(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}\Big)$$

and is called the maximization step. The EM algorithm is an iterative procedure that is repeated until convergence or until a valid stopping criterion is met. This iterative nature is a key aspect of the algorithm.
It is a particularly convenient tool when working with mixture models. In this context, the unobserved component affiliations are replaced by their expected value, given the data and the current parameter estimates. Hence, the optimal parameter estimates can be obtained despite latent component affiliations by iterating between the abovementioned steps. Putting the procedure into concrete formulas for the multinomial mixture model, as introduced in Section 2.1, leads to the following steps:

1. **E-Step:** Based on the observed data $\boldsymbol{y}$ and current parameter estimates (or the initialized values) $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\pi}}_1, ..., \hat{\boldsymbol{\pi}}_K)$, one takes the expectation of the full-data likelihood

   $$\mathbb{E}\Big(\log L(\hat{\boldsymbol{\theta}})|\boldsymbol{y}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\pi}}_1, ..., \hat{\boldsymbol{\pi}}_K\Big).$$

   Using the Bayes rule and plugging in the specific density functions, the posterior probabilities can then be calculated as

   $$\hat{\tau}_k^{(i)} = \mathbb{P}(Z^{(i)} = k|\boldsymbol{Y}^{(i)}; \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\pi}}_1, ..., \hat{\boldsymbol{\pi}}_K) = \frac{\mathbb{P}(Z^{(i)} = k; \hat{\boldsymbol{\eta}})\mathbb{P}(\boldsymbol{Y}^{(i)}|Z^{(i)} = k; \hat{\boldsymbol{\pi}}_1, ..., \hat{\boldsymbol{\pi}}_K)}{\mathbb{P}(\boldsymbol{Y}^{(i)}; \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\pi}}_1, ..., \hat{\boldsymbol{\pi}}_K)}$$
   $$= \frac{\hat{\eta}_k \mathbb{P}(\boldsymbol{Y}^{(i)}; \hat{\boldsymbol{\pi}}_k)}{\sum_{k'=1}^{K} \hat{\eta}_{k'} \mathbb{P}(\boldsymbol{Y}^{(i)}; \hat{\boldsymbol{\pi}}_{k'})}.$$

   These values express the probability of instance i being associated with the true component k given the observed count vector $\boldsymbol{y}^{(i)}$.

2. **M-Step:** The parameters are updated based on the posterior probabilities. For the multinomial mixture model, closed-form maximum likelihood estimates are available:

   $$\hat{\eta}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_k^{(i)}$$
   $$\hat{\pi}_{pk} = \frac{\sum_{i=1}^{n} y_p^{(i)} \hat{\tau}_k^{(i)}}{n \cdot \sum_{i=1}^{n} \hat{\tau}_k^{(i)}}.$$

It is possible to prove that the algorithm delivers local maxima or saddle points, i.e., that the likelihood always increases. Wu (1983) provides mathematical details on the convergence properties. Nevertheless, a global optimum cannot be guaranteed, and suboptimal solutions may arise. Therefore, it is recommended to choose a suitable initialization strategy and conduct multiple runs with varying initialization of the parameters to improve the practical performance of the algorithm. Different strategies have been proposed for this purpose, depending on the specific application, see Baudry and Celeux (2015) for an overview.
Although the classical EM algorithm is a widely used estimation technique, it can also be very slow, numerically intense, or even infeasible when dealing with complex data. As a result, simulation-based variants have been developed. One of them is the **stochastic EM algorithm** (Celeux et al., 1996). In this version of the original procedure, the E-step is replaced by a simulation, called S-step. This leads to simulated true image classes, allowing for more straightforward estimation. For the multinomial mixture model, in particular, simulation can be conducted as follows:

x. **S-Step:** Calculate the estimated posterior probabilities $\hat{\boldsymbol{\tau}}^{(i)} = (\hat{\tau}_1^{(i)}, ..., \hat{\tau}_K^{(i)})$ based on the current estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_K$ and sample realisations of the latent variable:

$$Z^{(i)} \sim Multi(\hat{\boldsymbol{\tau}}^{(i)}, 1).$$

This step allows the obtainment of complete data with simulated true classes, which can be used to calculate new estimates based on the complete likelihood. Like the classical EM algorithm, the procedure iterates between the S- and M-steps until convergence to estimate the unknown parameters.
Not only is this variant numerically more straightforward and particularly useful in high-dimensional applications, but the estimation uncertainty in the parameters can be directly quantified. This is, of course, especially useful in the context of uncertainty assessment and quantification. To calculate the variance of the parameters, Rubin's formula resulting from multiple imputations can be applied, see Rubin (1976) and Little and Rubin (2002). In the example above, estimation uncertainty is specifically interesting for the misclassification probabilities $\boldsymbol{\pi}_k, k = 1, ..., K$. The variance of the respective estimates can be calculated as

$$Var(\hat{\boldsymbol{\pi}}_{k,-K}|\boldsymbol{Z}) = \frac{diag(\hat{\boldsymbol{\pi}}_{k,-K}) - \hat{\boldsymbol{\pi}}_{k,-K}^T \hat{\boldsymbol{\pi}}_{k,-K}}{\sum_{i=1}^n \mathbb{1}\{Z^{(i)} = k\}}.$$

Details on the derivation of the formula can be found in Chapter 5.

2.3.2. Prior Parameters via Empirical Bayes

Empirical Bayes is a technique that is often employed to estimate parameters from the observed data in a Bayesian framework. The approach can be interpreted as an approximation to a fully Bayesian hierarchical view and, therefore, constitutes a sweet spot between complex Bayesian models and traditional frequentist methods. Suppose $\boldsymbol{Y}$ follows a distributional model $f(\boldsymbol{y}|\boldsymbol{\theta})$ with a vector of unknown parameters $\boldsymbol{\theta}$. In a Bayesian setting, a prior distribution is assumed for $\boldsymbol{\theta}$, i.e.

$$\boldsymbol{\theta} \sim f_{\boldsymbol{\theta}}(\boldsymbol{\alpha}),$$

with hyperparameters $\boldsymbol{\alpha}$. If these hyperparameters are known, the Bayes rule allows to compute the posterior distribution as defined in Formula (2.1). However, $\boldsymbol{\alpha}$ generally remains unknown. Information about it is contained in the denominator, i.e., the marginal distribution of the data. In an empirical Bayes analysis, the latter is utilized to estimate the hyperparameters $\boldsymbol{\alpha}$ by $\hat{\boldsymbol{\alpha}}$ based on observed data, leading to an empirical prior distribution. Then, the posterior distribution can be easily obtained as $f(\boldsymbol{\theta}|y, \hat{\boldsymbol{\alpha}})$ to infer the parameters of interest. In contrast to fully Bayesian approaches, the prior parameters are estimated from the data, making the approach computationally efficient and practical in many applications. Note that additional hierarchies can be introduced as needed.
Empirical Bayes methods have been successfully employed for many years. An introduction to the topic is given by Casella (1985), and for theory and applications of parametric empirical Bayesian inference, the reader is referred to Morris (1983). Carlin and Louis (2000) give an overview of the development of empirical Bayes over the years. Examples of the application of empirical Bayes include Latent Dirichlet Allocation (LDA) for the analysis of documents (Blei et al., 2003), missing data analysis (Follmann and Wu, 1995), or classification of micro-biological tissues (Demichelis et al., 2006). In this thesis, all contributions in Chapters 5, 6 and 7 rely on empirical Bayes and estimate the parameters on one of the specified hierarchies based on the available dataset.

2.3.3. Approximation of the Posterior

In a Bayesian setting, specifying a suitable prior distribution is typically non-trivial and, therefore, often shifted to the next level. This hierarchical setup works well if information is available on various levels of observation. However, for more complex problems, particularly for non-conjugate priors, analytical calculation of the posterior and closed-form estimation of the parameters is no longer possible. The posterior distribution takes the form

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y};\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{\int f(\boldsymbol{y};\vartheta) f_{\boldsymbol{\theta}}(\vartheta) d\vartheta}$$

and if the prior distribution is known, the only unknown quantity in the equation is the denominator. Hence, one strategy for estimating the posterior is the approximation of the integral in the denominator. Conventional numerical integration algorithms suffice for univariate parameters. Alternatives such as Laplace approximation (Kass et al., 1991) or Monte Carlo approximation of the integral, employing diverse simulation techniques, have been suggested in more complex situations.
However, explicit approximation of the posterior runs into problems for multivariate parameters $\boldsymbol{\theta}$. Therefore, an alternative approach is based on the idea of directly sampling from the posterior distribution instead of calculating it explicitly, i.e.,

$$\boldsymbol{\theta}^* \sim f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{y}).$$

This is commonly done by sampling a Markov Chain of correlated estimates, i.e., the chain $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, ...$ is generated. The distribution of this chain will converge to the posterior distribution $f_{\boldsymbol{\theta}}(.|\boldsymbol{y})$. This approach is called **Markov Chain Monte Carlo (MCMC)**.
Of course, directly sampling from $f_{\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{y})$ is impossible, as the explicit value of the denominator is still unknown. Therefore, alternative strategies have to be employed. A popular MCMC-based method is the **Metropolis-Hastings algorithm**, as introduced first by Metropolis et al. (1953) and a few years later generalized by Hastings (1970). This algorithm exploits the fact that the

unknown quantity, i.e., the denominator, cancels out when comparing the density of two possible parameter values. First, an appropriate proposal distribution $g(\boldsymbol{\theta}|\boldsymbol{\theta}_t^*)$ needs to be selected. Commonly, g is specified as a Gaussian distribution centered around the previous sample, i.e., $N(\boldsymbol{\theta}_t^*, \Sigma)$. To construct a Markov Chain $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, ..., \boldsymbol{\theta}_t^*$, one applies the two following steps:

1. Draw a new value $\boldsymbol{\theta}^*$ from the proposal distribution $g(\boldsymbol{\theta}|\boldsymbol{\theta}_t^*)$ based on the current value $\boldsymbol{\theta}_t^*$:

$$\boldsymbol{\theta}^* \sim g(.|\boldsymbol{\theta}_t^*).$$

2. Calculate the acceptance probability for $\boldsymbol{\theta}^*$ as

$$A(\boldsymbol{\theta}_t^*, \boldsymbol{\theta}^*) = \min\left\{1, \frac{f_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*|\boldsymbol{y})}{f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_t^*|\boldsymbol{y})} \cdot \frac{g(\boldsymbol{\theta}_t^*|\boldsymbol{\theta}^*)}{g(\boldsymbol{\theta}^*|\boldsymbol{\theta}_t^*)}\right\}.$$

   If $\boldsymbol{\theta}^*$ is accepted, add the new value to the Markov chain, i.e., set $\boldsymbol{\theta}_{t+1}^* = \boldsymbol{\theta}^*$.

It can be proven that the sequence $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, ..., \boldsymbol{\theta}_t^*$ generated by the Metropolis Hastings algorithm follows the stationary distribution $f_{\boldsymbol{\theta}}(.|\boldsymbol{y})$, see e.g. Robert and Casella (2004).
For the procedure to work reasonably well in practice, some aspects should be considered. First, the produced samples are correlated due to the nature of a Markov Chain. This issue can be addressed by "thinning" the samples, i.e., only including every $x - th$ sample to retain a set of samples devoid of correlation. Second, several "burn-in" iterations are required before the algorithm reaches the desired stationary distribution, i.e., a number of initial samples are discarded. Third, if the parameter is high-dimensional, the acceptance rate might decrease drastically, impacting the performance of the Metropolis-Hastings algorithm and demanding extensions that can handle high-dimensional parameters, like, e.g., Gibbs sampling, as described by Geman and Geman (1984), or Hamilton Monte Carlo introduced by Duane et al. (1987). In general, MCMC algorithms are very popular in the field of Bayesian statistics, and various alternatives and adaptations to special cases have been proposed over the years. A comprehensive overview is given, e.g., by Congdon (2007) or Berg (2004).

2.4. Uncertainty in Statistical Models

To bridge the gap to the second topic of this thesis, uncertainty in the area of ML, it is worthwhile to sketch the role of uncertainty in statistical models briefly. A major goal of traditional statistical modeling is to uncover underlying stochastic structures and, hence, to express the associated uncertainties. Typically, one distinguishes between aleatoric and epistemic uncertainty, which will be introduced in more detail in Chapter 3. However, a third type of uncertainty can be defined: statistical uncertainty (Hüllermeier and Waegeman, 2021). It is concerned with uncertainty due to the limited data basis used during any associated model's training phase. Assessing this uncertainty is crucial for obtaining reliable estimates, especially if the respective data basis is small. In the field of statistical modeling, numerous methods have been developed to appropriately express the uncertainties associated with parameter estimates.
The statistical uncertainty in $\hat{\boldsymbol{\theta}}$ or another quantity is commonly expressed via **confidence intervals**. For a parameter $\boldsymbol{\theta}$, the interval is defined as

$$\mathbb{P}(l(\boldsymbol{Y}) \leq \boldsymbol{\theta} \leq u(\boldsymbol{Y})) \geq 1 - \alpha.$$

The lower and upper boundaries $l(\boldsymbol{Y})$ and $u(\boldsymbol{Y})$ are thereby random variables based on $\boldsymbol{Y}$ chosen to minimize the resulting interval. Exact intervals can be obtained based on distributional assumptions and variance estimates. Generally, one is interested in quantifying the estimation variance in the parameter estimates $\hat{\boldsymbol{\theta}}$, expressing the associated uncertainty. For maximum likelihood estimates, a closed-form solution to calculate the variance exists in many cases. Returning to the example of the multinomial mixture model, the parameters are estimated via a stochastic EM algorithm. This setup allows quantifying the variance in the parameter estimates, as briefly discussed in Section 2.3. Details on theory and computation can be found in the contributing article in Chapter 5.
In a Bayesian context, methods based on MCMC sampling are popular and convenient with respect to estimation uncertainty. The variance of the resulting simulations from the posterior distribution is a natural way to express the uncertainty of the final parameter estimate and to construct confidence intervals if needed.
Another widespread universal approach for assessing estimation uncertainty is **bootstrapping**, a rather straightforward resampling method. While numerous variants of this approach exist today, Efron (1979) first proposed the empirical bootstrap. Based on an observed sample $(y_1, ..., y_n)$ multiple new samples $(y_1^{*b}, ..., y_n^{*b}), b = 1, ..., B$ are generated by drawing with replacement uniformly at random. Based on these samples, B estimates $\hat{\boldsymbol{\theta}}^{*b}, b = 1, ...B$ can be calculated and, hence, an empirical distribution of the desired estimate can be formed. This distribution allows the convenient computation of additional measures, like variance, standard deviation, bias, or confidence intervals. Hence, it provides insights into the (un-) certainty of the original parameter estimate without distributional assumptions. This fairly simple yet powerful approach is employed in the contribution in Chapter 5. By bootstrapping the negative log-likelihood, we derive insights into the heterogeneity of the annotation behavior of individual experts. Additionally, the stability of the general estimation procedure is evaluated via bootstrapping. The contributing article in Chapter 6 also employs this approach to assess the stability of the parameter estimates under varying amounts of observations and annotations.

However, decomposing the overall uncertainty into fixed components is controversial and often insufficient. Instead, all possible sources of uncertainty should be considered and appropriately analyzed. Label uncertainty, a crucial aspect, is central to this thesis. The toolbox of statistical modeling offers interesting possibilities for assessing such uncertainties. In particular, the models introduced in Sections 2.1 and 2.2 allow us to analyze label uncertainty based on multiple annotations. More details will be provided in Section 3.3 and in the respective contributing articles, see Chapters 5, 6 and 7.

3. (Label) Uncertainty in Machine Learning

"Scientific knowledge is a body of statements of varying degrees of certainty - some most unsure, some nearly sure, but none absolutely certain."
— Richard P. Feynman
(* 1918, † 1988)

As discussed in the previous section, uncertainty, and its quantification are central concepts in statistics. Providing variance estimates or confidence intervals to showcase the reliability of the proposed methods and their results is considered a fundamental step. In contrast, the predictions of ML algorithms, particularly of deep neural networks, are often solely evaluated in terms of accuracy and the relevance of uncertainty has been neglected for many years. With the increasing popularity and success of deep learning, the need for reliable predictions and models is apparent. This shift in focus has led to a rising interest in uncertainty and the need for quantification methods, marking a significant change in the field in recent years. While complex networks can achieve impressive prediction accuracy, often under optimal benchmarking conditions, they frequently lack the reliability crucial for deployment in safety-critical real-world applications. Given the sheer size and complexity of the models, there is an urgent need to develop appropriate methods for assessing and quantifying the associated uncertainty.
Extensive textbooks on uncertainty include, for example, Smith (2013) and Sullivan (2015). Several recent reviews and survey papers also shed light on the topic from various perspectives. Hüllermeier and Waegeman (2021) give an extensive overview of handling uncertainty in ML models. Gawlikowski et al. (2023) discuss its sources and types in depth and propose approaches for estimating uncertainty in deep neural networks. They also provide an overview of various benchmarks, implementations, and uncertainty estimation in real-world applications. Abdar et al. (2021) cover many existing methods for quantifying uncertainty, various application fields, and related open issues. Uncertainty in big data analytics is discussed by Hariri et al. (2019). Gruber et al. (2023) tackle the problem from a statistical perspective and focus on uncertainties arising from the data. The relevance and necessity of uncertainty estimates are apparent in many application areas. These include, for example, the analysis of medical images for diagnosis and treatment (Ching et al., 2018 or Kompa et al., 2021), remote sensing with satellite imagery (Zhu et al., 2017) or natural language processing (Xiao and Wang, 2019). Uncertainty- and risk-aware models are also crucial in the deployment of algorithms, for example, in autonomous driving (Shafaei et al., 2018 or Brechtel et al., 2014). This chapter introduces the general topic and popular directions for quantifying uncertainty for deep neural networks. Afterward, the focus is placed on supervised models and the uncertainty associated with the training data labels.

3.1. Definition and Types of Uncertainty

Typically, uncertainty was and still is decomposed into an **aleatoric** and an **epistemic** part in the context of ML, a distinction dating back to Hacking (1975) for describing probabilities from two different perspectives. Statistical uncertainty, as briefly described in the previous chapter, is often considered negligible in the age of big data and massive training datasets.
Epistemic uncertainty refers to the uncertainty of the model. It can be described as uncertainty that arises due to the lack of knowledge, e.g., arising from insufficient or inappropriate training data or misspecified models. Therefore, it can be reduced by acquiring additional data or developing better models, hence the description as "reducible". On the other hand, aleatoric uncertainty, also called data uncertainty, is the inherent stochasticity of the observations. It is often also referred to as "irreducible uncertainty". However, this categorization is, in fact, ambiguous and does not account for the multiple sources and types of uncertainty, as stated early on by Hora (1996) and explored in depth more recently by Gruber et al. (2023) or Baan et al. (2023). Following the statistical viewpoint adopted by Gruber et al. (2023), classical probability models can be used to disentangle the types of uncertainty. Given a set of input variables $\boldsymbol{X}$ and output variables $\boldsymbol{Y}$[1], aleatoric uncertainty can be described by

$$\boldsymbol{Y}|\boldsymbol{x} \sim f_{\boldsymbol{Y}|\boldsymbol{X}}(.|\boldsymbol{x}).$$

Aleatoric uncertainty originates from the stochastic relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ and, therefore, refers to the variance $Var(\boldsymbol{Y}|\boldsymbol{x})$. The remaining uncertainty can be defined as the epistemic part. Overall, uncertainty is heavily impacted by the data at hand. Misspecified models, i.e., model uncertainty, can arise from omitted quantities, errors in either the observed or the target variables, or violations of the *i.i.d.* assumption. Additionally, areas like survey design, treatment of missing data, and deployment settings heavily impact uncertainty. While the distinction between aleatoric and epistemic parts of uncertainty can be helpful to some extent, the common definition is too limiting in most applications. Instead, various possible sources must be considered for a comprehensive assessment of uncertainty and the role of the data at hand should not be underestimated (Gruber et al., 2023).

3.2. Uncertainty Quantification

Quantifying the inherent uncertainty of a model's outputs is crucial, particularly for deep neural networks. It is an essential step to address significant challenges, such as the lack of transparency and trust in their predictions, as well as their susceptibility to attacks. Therefore, understanding and quantifying associated uncertainties is key to building reliable models.
In recent years, numerous methods for quantifying uncertainty have been explored and are now increasingly implemented into the framework of deep networks, see Gawlikowski et al. (2023). Most methods can be classified into one of three major streams. First, **deterministic methods** provide uncertainty estimates based on one forward pass through a deterministic network. Already early on, prior networks were developed by Malinin and Gales (2018) for explicitly modeling

[1] Note that the common notation in ML is adopted here. However, defining $\boldsymbol{Y}$ as the output variables, or labels in the context of supervised models, is in line with $Y_1^{(i)}, ..., Y_p^{(i)}$ referring to the vector of annotations for instance i in the previous section. Summarizing $Y_1^{(i)}, ..., Y_p^{(i)}$ via some aggregation strategy results in output variables $Y^{(i)}$.

distributional uncertainties. This is achieved by introducing a prior distribution, e.g., a Dirichlet distribution, to represent the distributional uncertainty. Deterministic methods also include evidential neural networks, see Sensoy et al. (2018) or Ulmer et al. (2023) for a recent survey. These methods allow the output of uncertainty estimates by gathering "evidence" from the input data and, hence, determining the level of certainty or confidence in the associated predictions. The second stream of methods for uncertainty quantification can be described as **Bayesian approaches**, which include all kinds of stochastic neural networks. Similar to Bayesian methods known from classical statistics, the ultimate goal is to infer the posterior distribution via some approximation strategy. A popular related, though not strictly Bayesian, approach in this direction is Monte Carlo Dropout, as developed by Gal and Ghahramani (2016). Lastly, it is also possible to combine the predictions of multiple deterministic networks in **ensembles**, as demonstrated by Lakshminarayanan et al. (2017). These ensemble-based methods offer a relatively simple and efficient method for computing uncertainty estimates.
However, in most applications, the resulting uncertainty estimates alone are insufficient to express the overall uncertainty comprehensively. While low predictive uncertainty seems desirable, interpreting resulting estimates is not always straightforward. Additionally, many methods struggle to disentangle the multiple and interacting sources of uncertainty, further complicating interpretation (Valdenegro-Toro and Mori, 2022).

Uncertainty quantification techniques are commonly employed in combination with downstream tasks. For example, estimates of predictive uncertainties can be used to detect **out-of-distribution data** at test time. When models trained on specific datasets encounter samples from unknown classes or entirely different datasets, they cannot provide reliable predictions for these instances. This should be reflected in the associated uncertainty estimates. Another interesting use case is **Active Learning**. This learning regime is based on the idea that the required amount of training data can be tremendously reduced by letting the algorithm choose which observations to learn from. Settles (2009) describes this property as "curiosity" of the algorithm. In particular, a neural network is initially trained based on a small set of labeled data. Based on some query strategy, the network decides, which samples will be selected from a larger pool of unlabeled training data and labeled by an oracle, i.e., a human labeler. Hence, the training dataset is increased, and a new model is trained at each iteration. A common strategy for selecting new observations is uncertainty sampling (Lewis, 1995). Therefore, predictive uncertainties are calculated for the unlabeled pool. Choosing data points with high uncertainty estimates, i.e., instances that the current model is uncertain about, helps the model learn faster and more accurately with less training data, as shown by, e.g., Gal et al. (2017) or Nguyen et al. (2019).

Another concept closely related to estimating uncertainties is **calibration**. In the context of classification, it refers to the property that probabilities derived from a model adequately reflect the certainty in these predictions. Modern neural networks are often poorly calibrated and overconfident (Guo et al., 2017). However, to retrieve valuable predictions, not only a model's accuracy but also its confidence should be high. Guo et al. (2017) define those quantities as

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}(\hat{y}^{(i)} = y^{(i)})$$

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}^{(i)}$$

based on "bins" $B_m, m = 1, ..., M$, which split the data into M equally spaced regions. These values are commonly visualized in reliability diagrams to assess network calibration. Alternatively,

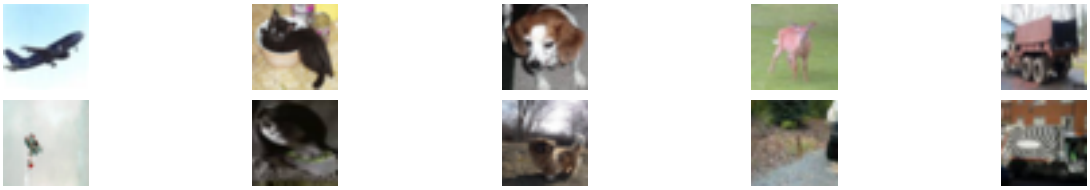
the difference in expectation between a model's confidence and accuracy, approximated by the expected calibration error (ECE), can be calculated. A lower value is desirable, as it indicates better calibration and a better match between the predicted and the actual probabilities.
Several post-processing methods are available to re-calibrate a model's prediction. A famous example is temperature scaling (Guo et al., 2017). The temperature parameter in the softmax function is adjusted for better calibration of the network's prediction and less overconfidence. Another popular approach is called label smoothing. Adding a small amount of uncertainty, i.e., smoothing, to the target labels during training leads to better generalization, see Szegedy et al. (2016).

3.3. Label Uncertainty

In supervised learning, training a model relies on large amounts of annotated training data. Here, uncertainty, in fact, is already present in the labeling process of the instances. Ambiguous annotations carry a non-neglectable amount of uncertainty that this work treats as a possible source or one layer of the overall uncertainty, termed **label or annotation uncertainty**. Effectively dealing with label uncertainty is essential for enhancing the reliability and accuracy of ML models. This issue must be addressed and incorporated into the final uncertainty estimates of classification models to assess its global uncertainty comprehensively.

3.3.1. Generation of Labels

Labeled data are the foundation of supervised ML models, and hence, huge efforts have been made in recent years to generate large labeled datasets for all kinds of domains. Classification algorithms commonly rely on the existence of a gold label associated with each training instance. However, obtaining such gold labels can be non-trivial depending on the complexity of the task at hand. A popular strategy is acquiring multiple annotations per instance in the first labeling stage to ensure consistent labeling and high quality. This is commonly done through human effort.
Crowdsourcing offers an inexpensive and efficient way to generate (multiple) annotations. However, while labeling by laypersons provides a cheap and relatively quick approach to gathering multiple evaluations for each instance via platforms like Amazon Mechanical Turk, one might sacrifice quality. The severity of this issue, of course, depends on the application at hand. While exploiting the crowd's wisdom has achieved promising results on some tasks, others require a more sophisticated labeling strategy. In some application areas, like the classification of medical X-ray scans (e.g., Dgani et al., 2018 or Zhou et al., 2021) or satellite images in the domain of earth observation (e.g., Zhu et al., 2020), annotation is non-trivial for layperson and untrained workers. Domain knowledge, detailed classification instructions, and training of the labelers are required for satisfactory annotation. Hence, the instances are assessed by experts instead of laypersons, which is costly and time-consuming. Therefore, multiply labeled training data are particularly rare in interesting application areas where annotation is not straightforward.
Three examples of datasets containing multiple annotations will be shortly introduced in Section 3.4.



(a) Plane (b) Cat (c) Dog (d) Deer (e) Truck

Figure 3.1.: The figure displays exemplary images from the dataset Cifar-10H, see Section 3.4. Each subfigure refers to one category and contains an unambiguous (top row) and an ambiguous (bottom row) example.

3.3.2. Label Problems

Uncertainty in the annotations can arise from numerous sources that must be treated in special ways. A common problem, particularly when dealing with large amounts of real-world data, is **label noise**. This term specifically refers to the lack of high-quality labels, see Frénay and Verleysen (2013) for an introduction. Some evidence exists that deep neural networks are not as sensitive to noisy labels due to their complexity (Rolnick et al., 2017). However, learning despite label noise and robust training are important tasks and calibration techniques can be used to minimize the negative effects (Lukasik et al., 2020). For an overview of the methodology, see Song et al. (2023) or Algan and Ulusoy (2021).
Furthermore, systematic **label bias** can be an issue related to annotation uncertainty. Exemplary sources are annotation guidelines, the expertise of the labelers, or the quality of data sources. Biased data can have negative impacts, especially related to the fairness of the subsequent model, and should be treated appropriately and with caution, see Jiang and Nachum (2020).
Another contributing issue to label uncertainty is **ambiguity**. In some classification scenarios, it is unclear which label to assign to a particular instance. Figure 3.1 shows exemplary cases for classifying tiny images into well-separated classes, see Section 3.4 for details on the respective dataset. Amongst other sources, ambiguity can arise due to subjectivity, the inherent complexity of the labeling task, inconsistent guidelines for annotators, or overlapping categories. A method for learning despite ambiguous labels is label distribution learning, as introduced by Geng et al. (2013), Geng (2016) or Gao et al. (2017). In this case, a probability distribution over possible labels is assigned instead of a single label, and the loss function is adapted accordingly.
However, ambiguity in labels is often inevitable. Suppose the annotation task at hand is complex and, hence, is associated with a complex classification problem. In that case, human labelers are likely to make mistakes but also to disagree with each other rightfully. The ambiguity in the resulting annotations reflects the ambiguity in the task itself. Hence, aggregating the annotations into a single label discards important information, and the question arises of how to represent the underlying ground truth adequately.

3.3.3. How to represent the ground truth?

Even if label quality generally benefits from acquiring multiple annotations per instance, the question of how to use the multivariate data structure for modeling and learning remains. Variation in the annotations is often undeservedly considered problematic and, hence, should be resolved.
The most straightforward approach is disregarding the annotation variation and compressing the

information into a singular label. This is commonly done via aggregation, for example, by majority voting or probabilistic methods. A comprehensive survey on annotation analysis is provided by Paun et al. (2022). Majority voting and other aggregation methods work well for simple annotation tasks, i.e., tasks with a high expected agreement. They can help filter spamming annotators or wrong answers by exploiting the crowd's wisdom compared to singular evaluations. However, this strategy is unfavorable for more complex and possibly subjective annotation tasks. Summarizing the annotations and, hence, "resolving" the label variation leads to a major loss of information. In some areas, disagreement in annotations is considered beneficial. Plank (2022) proposes the term "human label variation" in the context of natural language processing (NLP). It refers to inherent annotation disagreement arising from genuine disagreement or valid subjectivity. Figure 3.2 shows the label distribution for the exemplary tiny images from the category *deer* from Figure 3.1. Inspecting the annotations for the upper example shows that aggregation into a single label can indeed have positive filtering effects if the images are unambiguous. In contrast, choosing the majority-voted class as a final label for the bottom image is not only incorrect, but it also discards all information on the ambiguity of the image. Therefore, uncertainty in the respective final label is ignored completely.
In many real-world applications, classifying instances is complex and often ambiguous. In the domain of NLP, a growing body of work emphasizes the significance of inherent disagreements in labeling (Pavlick and Kwiatkowski, 2019, Nie et al., 2020 and Leonardelli et al., 2021). However, this issue is also prevalent in various image classification tasks, particularly with low-resolution or domain-specific images. Relevant example datasets will be introduced in Section 3.4. In such applications, the truth might indeed lie "somewhere in between" and a singular label cannot do justice to the complexity of the problem. Therefore, the variation in the annotation reflects the inherent complexity and provides valuable information. Neglecting and ignoring human label variation from the start harms the subsequent steps of the ML pipeline and induces additional uncertainty in the final predictions. Hence, interest in this prevalent issue has increased recently, and various strategies to approach variation in multiply labeled data were developed to preserve as much information as possible. A novel version of a famous benchmark dataset for image classification was introduced by Peterson et al. (2019) for incorporating human uncertainty via soft labels. The authors show that incorporating the annotation variation can lead to more robust classification models. Koller et al. (2024) explicitly integrate label uncertainty into the training process through distributional labels derived from multiple annotations. They show that incorporating this additional knowledge enhances generalization to unseen data and better performance in terms of calibration. A survey on different strategies of learning from disagreement in the annotations is given by Uma et al. (2021). However, acquiring multiple labels per instance, in general, remains a costly and labor-intensive process. This can be problematic for fixed annotation budgets. A framework that is often deployed to reduce the human labeling effort is active learning, as introduced shortly in the previous section. Recent work also explores the combination of active learning principles and label variation, mainly in the area of NLP (Baumler et al., 2023, Wang and Plank, 2023).

All contributing articles in the remainder of this thesis focus on analyzing label uncertainty based on multiple annotations. The overall goal is to assess and quantify the ambiguity in the "gold" labels of various datasets while also developing a more appropriate representation of the ground truth. Therefore, methodology from classical statistical modeling is applied, as introduced in Chapter 2.
In the contributions in Chapters 5 and 6, the central assumption is the existence of a one-

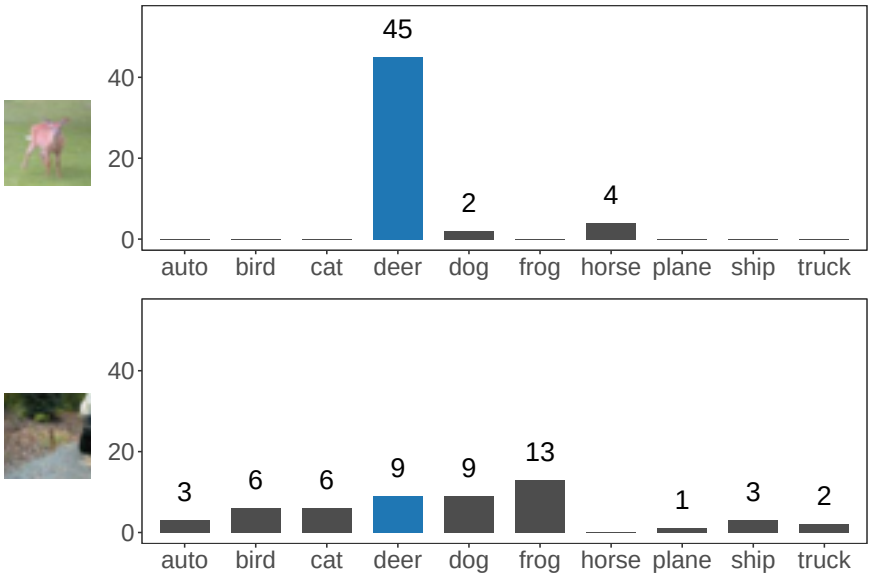


Figure 3.2.: The figure shows the label distribution for two images from the category *deer*. The top row displays the annotation results for a relatively unambiguous image, while the bottom row shows the annotations for an ambiguous image, where disagreement among the labelers is evident.

dimensional latent ground truth label for each observation. Hence, there is no ambiguity or uncertainty about the true class itself, only about the annotator's opinion regarding the correct class. This setup can be statistically described by a multinomial mixture model, as introduced in Section 2.1. This model enables us to retrieve information about the latent class affiliations and the uncertainty in the annotations. The model parameters provide insights into various factors that contribute to this uncertainty. Based on the conducted analyses in the article in Chapter 5, parts of it can be attributed to the general ambiguity of the classes in some applications, the annotators' heterogeneity, and the data's external distinguishing properties. The work in Chapter 6 applies the same model setup in the context of language classification, a subdomain of NLP. This contribution focuses on the stability of the estimation and emphasizes the importance of acquiring a sufficient amount of annotations, especially for uncertain observations.
Chapter 7 contains a subsequent article that modifies the central assumption of a one-dimensional ground truth label and instead assumes that the truth can be multi-dimensional. According to Plank (2022), human label variation generally does not constitute a problem but instead should be embraced. In many applications, the restriction of observations to a single class is overly strict and does not do justice to the complexity of the problem. Therefore, the contribution proposes to move away from a single ground truth label and instead estimate multi-dimensional ground truth representations for each observation based on the annotations. In other words, the observations are embedded into a P-dimensional space, with P denoting the number of classes. A Bayesian model framework using the Dirichlet multinomial model, as introduced in Section 2.2, is proposed to estimate suitable embeddings. The resulting embeddings provide insights into the correlation structure of the classes. Additionally, they express individual instances' classification difficulty and uncertainty and serve as a more appropriate representation of the underlying ground truth.

Dataset	Type	$\#N$	$\#J$	Type of Annotators	Annotator-ID	Source
So2Sat LCZ42	images	250k	11	Experts	yes	Zhu et al. (2020)
Cifar-10H	images	10k	[47,63]	MTurk	yes	Peterson et al. (2019)
Plankton	images	12k	[1,192]	Citizen Scientists	yes	Schmarje et al. (2022)
ChaosNLI	text	4.6k	100	MTurk (qualified)	no	Nie et al. (2020)
SNLI	text	570k	5	MTurk	no	Bowman et al. (2015)
MDAgreement	text	10k	5	MTurk	yes	Leonardelli et al. (2021)

Table 3.1.: Examples of openly available multi-annotation datasets.

3.4. Multi-Annotator Datasets

Naturally, appropriate assessment of label uncertainty requires observations with multiple labels. Most benchmark datasets for building classification models undergo thorough annotation processes to maintain a baseline level of label accuracy. Still, before the final dataset is released, these annotations are often streamlined. As the vast majority of models have been developed to predict one-dimensional target categories and are optimized and evaluated in this regard, the aggregation of annotations and, hence, the assumption of a single gold label per instance has been the standard practice for many years (Plank, 2022). However, as discussed in Section 3.3, this overly simplifying point of view can lead to serious uncertainty-related issues in the final predictions.
Analyzing and quantifying label uncertainty explicitly requires datasets containing the original annotations per instance, which are relatively rare but of immense importance. Existing datasets in this context differ depending on the application and task at hand. This refers to the size of the dataset ($\#N$), the number of available annotations per instance ($\#J$), the competence of the annotators, and the availability of annotator-specific information, amongst many more. Table 3.1 contains examples of multi-annotator datasets. The contributing articles in this thesis aim to assess label uncertainty mainly based on three specific multi-annotator datasets, which will be introduced shortly in the following.

In the domain of image classification, annotation uncertainty is common if the task at hand is complex, possibly due to ambiguous images or similar categories. A prominent example in this regard is the classification of satellite images into so-called local climate zones (LCZs), which is studied in depth in the articles in Chapter 5 and 7. The respective dataset for analyzing annotation uncertainty is a subset of the benchmark dataset **So2Sat LCZ42** for earth observation, developed by Zhu et al. (2020) and published within the framework of the work of Koller et al. (2024). The full version of the dataset serves as a benchmark for classifying satellite images into climatic zones to automatically develop climate maps and, therefore, monitor the development of urban climate. A control stage was introduced upon development to ensure label quality, where multiple earth observation experts annotated a subset of the original images from major European cities and some additional global cities ($N = 160K$). Figure 3.3 shows the respective classes, as well as satellite images and images taken from Google Earth. The classification scheme was developed by Stewart and Oke (2012) and today constitutes a gold standard in categorizing the earth's surface. In this case, the low resolution of the images themselves and the ambiguous nature of the classification scheme complicate the annotation process and demand special attention and skills. Labeling satellite images is challenging even for trained experts, leading to a major amount of disagreement and, hence, label uncertainty. As shown in Table 3.1, the number of annotations is relatively small. However, the provided annotations are expert-specific, which allows us to assess

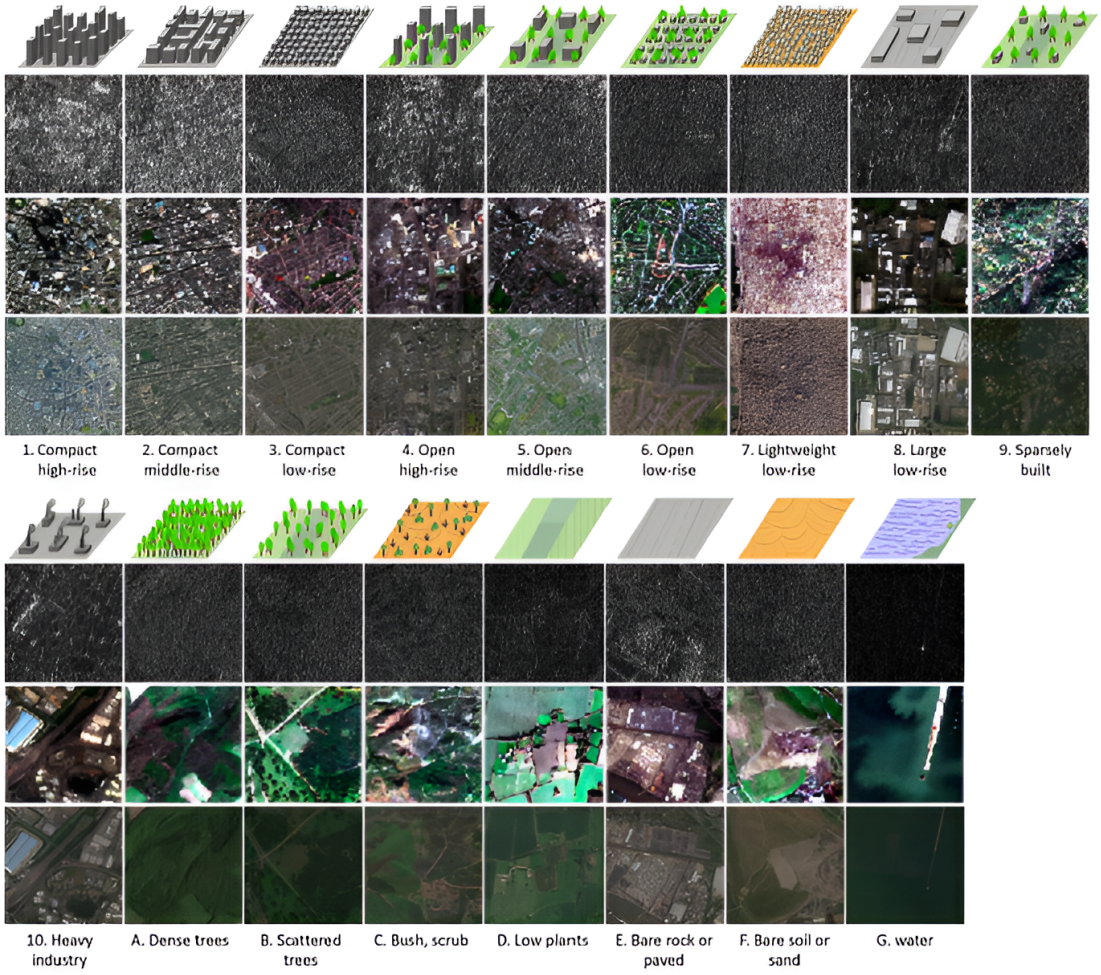


Figure 3.3.: Visualization of the 17 climatic classes in various formats (upper image: Sentinel-1, middle image: Sentinel-2, lower image: high-resolution Google Earth image), Source: Zhu et al. (2020).

the annotators' individual effects. The respective images also stem from highly heterogeneous sources, i.e., geographic locations, which can further complicate the assessment by annotators.

Another famous dataset in the context of human label uncertainty in image classification is **Cifar-10H**, as introduced by Peterson et al. (2019). The authors extend the test set of a famous benchmark dataset, Cifar-10 (Krizhevsky et al., 2009), with multiple annotations to explore the effects of incorporating human uncertainty into the classification model. While the classification scheme consists of unambiguous, well-separated classes, the size of the images still defies easy classification in some cases, see Figure 3.1 and Figure 3.2. Workers from Amazon Mechanical Turk annotated the images. On the one hand, the study reveals annotation errors, which can be resolved by aggregating the crowdsourced annotations. On the other hand, the dataset also contains a few ambiguous images, for which the correct class is not apparent from the annotations. The dataset provides annotator-specific labels and the original labels of the individual instances, which correspond to the search term used to find the respective image via web searches.

Context/Premise	Statement/Hypothesis	[E, N, C]
A man running a marathon talks to his friend.	There is a man running.	[100, 0, 0]
A live band on a lawn jamming out.	A band is practicing new tunes in the garage.	[0, 1, 99]
Two people riding bikes in the rain at skate park.	It is springtime outside.	[2, 98, 0]
A man in a black hat and jacket is sitting down.	A man is being arrested for breaking a window.	[0, 49, 51]
A woman holding a child in a purple shirt.	The woman is asleep at home.	[1, 53, 46]
An elderly woman crafts a design on a loom.	The woman is sewing.	[35, 31, 34]

Table 3.2.: Exemplary sentences of the dataset ChaosSNLI. Annotators were instructed as follows: "Given a context, a statement can be either: definitely correct (Entailment); or definitely incorrect (Contradiction); or neither (Neutral). Your goal is to choose the correct category for a given pair of context and statement.", Source: Nie et al. (2020)

Label uncertainty also plays a vital role in the context of NLP. A subfield of this prominent domain is natural language inference (NLI). NLI refers to the classification of language and is concerned with determining the inference relation between two sentences, which can be classified as entailment, contradiction, or neutral. Language is, of course, particularly prone to subjectivity and individual perception. Therefore, label variation is likely to arise due to high human disagreement. Multiple datasets have been introduced to assess label variation in this context. The article in Chapter 6 focuses on the dataset **ChaosSNLI**, a subset of the larger ChaosNLI dataset based on ambiguous examples from the Stanford Natural Language Inference (SNLI) corpus, as introduced by Nie et al. (2020). Upon constructing the dataset, one annotator developed entailing, neutral, or contradicting statements to an original premise. Hence, each resulting sentence pair is associated with a subjective ground truth. The goal of the remaining annotators is then to classify the pairs of statements into *entailment*, *neutral*, or *contradicting*, according to their relationship. However, language itself is highly ambiguous, and so is the individual perception of the human annotators. These issues undermine the actual meaning of a "ground truth". Instead, multiple plausible categories could be considered "correct", as emphasized by Nie et al. (2020). This is also obvious in the dataset, where a high rate of disagreement among the annotators can be observed despite the existence of a subjective ground truth label. Table 3.2 provides a schematic overview and examples. Qualified workers from Amazon Mechanical Turk carried out the annotation. To prohibit the skewing effect of spamming annotators, quality was controlled by tracking the performance of the individual annotators on unambiguous examples, see Nie et al. (2020) for details. Although the large number of annotations enables the exploration of interesting questions regarding labeling strategies, the dataset only contains aggregated label counts and does not provide information specific to individual annotators.

4. Concluding Remarks

This dissertation addresses label uncertainty in classification models from a statistical point of view. Uncertainty is a rather vague but crucial topic for building reliable and valuable models. While most supervised models for classification build on the assumption of available gold labels, one goal of this thesis is i) to question the validity of the category associated with each instance and ii) to re-assess the general assumption of singular ground truth labels for all kinds of classification problems. The contributing articles show that employing statistical models and analyzing the data-generating process allows for various insights into this specific source of uncertainty.

Contributions Part II focuses on modeling multiple annotations based on the central assumption of a single latent ground truth label for each observation. The contributing article in Chapter 5 aims to assess the inherent uncertainty in the annotations in the context of the classification of satellite imagery. By treating annotations as multinomial count data and employing a mixture model, the data-generating process can be modeled under the assumption of a latent true label. This approach provides insights into the confusion probabilities of the annotators and the posterior distributions of the true labels. Based on the resulting estimates, it is possible to assess the main factors contributing to label uncertainty, like annotator-specific effects or external properties of the instances. The proposed statistical model is highly adaptable and can be applied to different application areas, as shown in Chapter 6. Here, the mixture modeling approach is applied to a dataset from the domain of NLI, again providing insights into the latent posterior distributions. Additionally, this work assesses the stability of the proposed estimation procedure via a bootstrapping approach under various conditions. By systematically varying the number of annotations and observations, this work showcases the importance of sufficient annotations for ambiguous instances and, hence, aims to provide guidance for future annotation tasks. However, the question remains whether the assumption of a singular gold label is actually valid. A growing body of work, primarily in the field of NLP, see Plank (2022), Nie et al. (2020) and Zhang et al. (2021), suggests the opposite.
Therefore, Part III moves beyond this limiting assumption. In the contribution in Chapter 7, a latent label representation, so-called embedding, is estimated via a Bayesian model framework based on the Dirichlet-multinomial model. The resulting vector representations capture annotation ambiguities as well as actual classification difficulty of instances. The proposed universal approach can be easily applied to various multi-annotator settings if the assumption of a single gold label per instance is too limiting and unrealistic.

Discussion and Outlook While the proposed methods and models provide interesting and highly relevant insights, they serve more as a first step than a finished solution toward increasing awareness of annotation uncertainty and the need for a more appropriate representation of uncertain labels. This issue is highlighted by the present contributions in various domains and application

areas. Incorporating the results in future work is inevitable as part of the primary goal of reliable and uncertainty-aware classification models. Two possible future research directions will be sketched shortly to emphasize the topic's relevance.

First, the results of all contributions can be almost directly incorporated into the ML pipeline. Estimating the posterior distributions of the latent true label for each instance or multi-dimensional label embeddings can preserve information about uncertainty within the full set of annotations. Hence, they serve as more reliable label representations. Building ML models based on these representations incorporates the uncertainty of the original annotations and could, therefore, lead to more reliable and comprehensive uncertainty estimates. Koller et al. (2024) show that by employing approaches like label distribution learning, classification models already benefit from using the empirical label distribution instead of simple one-hot labels in terms of calibration. Incorporating the additional information in the form of the latent posteriors or latent Bayesian label embeddings could further enhance the results by allowing for a more accurate and realistic representation of a "ground truth".

Furthermore, the second contribution concludes that for ambiguous observations, a sufficient number of annotations per instance is more crucial than a large sample size. While this finding is reasonable from a data-generating viewpoint, quantitative measures and clear guidance on how and when to annotate observations still need to be explored. Hence, future work in this direction could concentrate on deploying strategies inspired by Active Learning to develop multi-annotator datasets without unnecessary human labeling effort. The primary goal is to ascertain whether obtaining an additional annotation for a given instance or collecting more instances proves more advantageous within a fixed annotation budget.

Generally, this thesis highlights the importance of incorporating label uncertainty into the overall pipeline of ML models and the utility of the statistical modeling toolbox for this purpose. However, to completely assess and quantify uncertainty, the collaboration of machine learners, statisticians, and domain experts is crucial and should be fostered further in the future. By leveraging their interdisciplinary expertise, more robust and reliable models can be developed, and uncertainties can be captured and addressed more accurately. Ultimately, collaborative efforts will lead to more trustworthy and effective ML applications.

References

Abdar, M., F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion 76*, 243–297.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control 19*(6), 716–723.

Algan, G. and I. Ulusoy (2021). Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems 215*, 106771.

Allison, D. B., G. L. Gadbury, M. Heo, J. R. Fernández, C.-K. Lee, T. A. Prolla, and R. Weindruch (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis 39*(1), 1–20.

Baan, J., N. Daheim, E. Ilia, D. Ulmer, H.-S. Li, R. Fernández, B. Plank, R. Sennrich, C. Zerva, and W. Aziz (2023). Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.

Baudry, J.-P. and G. Celeux (2015). Em for mixtures: Initialization requires special care. *Statistics and computing 25*, 713–726.

Baumler, C., A. Sotnikova, and H. Daumé III (2023). Which examples should be multiply annotated? active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10352–10371.

Benaglia, T., D. Chauveau, D. R. Hunter, and D. S. Young (2010). mixtools: an r package for analyzing mixture models. *Journal of statistical software 32*, 1–29.

Berg, B. A. (2004). *Markov chain Monte Carlo simulations and their statistical analysis: with web-based Fortran code*. World Scientific Publishing Company.

Bernardo, J. M. and A. F. Smith (2009). *Bayesian theory*, Volume 405. John Wiley & Sons.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*(Jan), 993–1022.

Bowman, S. R., G. Angeli, C. Potts, and C. D. Manning (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642. Association for Computational Linguistics.

Brechtel, S., T. Gindele, and R. Dillmann (2014). Probabilistic decision-making under uncertainty for autonomous driving using continuous pomdps. In *17th international IEEE conference on Intelligent Transportation Systems (ITSC)*, pp. 392–399. IEEE.

Carlin, B. P. and T. A. Louis (2000). Empirical bayes: Past, present and future. *Journal of the American Statistical Association 95*(452), 1286–1289.

Casella, G. (1985). An introduction to empirical bayes data analysis. *The American Statistician 39*(2), 83–87.

Celeux, G. (1998). Bayesian inference for mixture: The label switching problem. In *COMPSTAT: Proceedings in Computational Statistics 13th Symposium held in Bristol, Great Britain, 1998*, pp. 227–232. Springer.

Celeux, G., D. Chauveau, and J. Diebolt (1996). Stochastic versions of the EM algorithm: An experimental study in the mixture case. *Journal of Statistical Computation and Simulation 55*(4), 287–314.

Cheng, B. and D. M. Titterington (1994). Neural networks: A review from a statistical perspective. *Statistical science 9*(1), 2–30.

Ching, T., D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface 15*(141), 20170387.

Congdon, P. (2007). *Bayesian statistical modelling*. John Wiley & Sons.

Davison, A. C. (2003). *Statistical models*, Volume 11. Cambridge university press.

Demichelis, F., P. Magni, P. Piergiorgi, M. A. Rubin, and R. Bellazzi (2006). A hierarchical naive bayes model for handling sample heterogeneity in classification problems: an application to tissue microarrays. *BMC bioinformatics 7*, 1–12.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

Dgani, Y., H. Greenspan, and J. Goldberger (2018). Training a neural network based on unreliable human annotation of medical images. In *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*, pp. 39–42. IEEE.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid monte carlo. *Physics letters B 195*(2), 216–222.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1 – 26.

Elkan, C. (2006). Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution. In *Proceedings of the 23rd international conference on Machine learning*, pp. 289–296.

Follmann, D. and M. Wu (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics 51*(1), 151–168.

Frénay, B. and M. Verleysen (2013). Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems 25*(5), 845–869.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer.

Gal, Y. and Z. Ghahramani (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning*, pp. 1050–1059. PMLR.

Gal, Y., R. Islam, and Z. Ghahramani (2017). Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR.

Gao, B.-B., C. Xing, C.-W. Xie, J. Wu, and X. Geng (2017). Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing 26*(6), 2825–2838.

Gawlikowski, J., C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review 56*(1), 1513–1589.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6*(6), 721–741.

Geng, X. (2016). Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering 28*(7), 1734–1748.

Geng, X., C. Yin, and Z.-H. Zhou (2013). Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*(10), 2401–2412.

Gruber, C., P. O. Schenk, M. Schierholz, F. Kreuter, and G. Kauermann (2023). Sources of uncertainty in machine learning–a statisticians' view. *arXiv preprint arXiv:2305.16703*.

Guo, C., G. Pleiss, Y. Sun, and K. Q. Weinberger (2017). On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR.

Hacking, I. (1975). *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge University Press.

Hamilton, N. E. and M. Ferry (2018). ggtern: Ternary diagrams using ggplot2. *Journal of Statistical Software, Code Snippets 87*(3), 1–17.

Hariri, R. H., E. M. Fredericks, and K. M. Bowers (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data 6*(1), 1–16.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

Holmes, I., K. Harris, and C. Quince (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one 7*(2), e30126.

Hora, S. C. (1996). Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety 54*(2-3), 217–223.

Hüllermeier, E. and W. Waegeman (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning 110*, 457–506.

Jiang, H. and O. Nachum (2020). Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR.

Kass, R. E., L. Tierney, and J. B. Kadane (1991). Laplace's method in bayesian analysis. *Contemporary Mathematics 115*, 89–99.

Kauermann, G., H. Küchenhoff, and C. Heumann (2021). *Statistical Foundations, Reasoning and Inference*. Springer.

Koller, C., G. Kauermann, and X. X. Zhu (2024). Going beyond one-hot encoding in classification: Can human uncertainty improve model performance in earth observation? *IEEE Transactions on Geoscience and Remote Sensing 62*, 1–11.

Kompa, B., J. Snoek, and A. L. Beam (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine 4*(1), 4.

Konishi, S. and G. Kitagawa (2008). *Information criteria and statistical modeling*. Springer Science & Business Media.

Krizhevsky, A., G. Hinton, et al. (2009). Learning multiple layers of features from tiny images.

Kroese, D. P., J. C. Chan, et al. (2014). *Statistical modeling and computation*. Springer.

Lakshminarayanan, B., A. Pritzel, and C. Blundell (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Leonardelli, Elisa and. Menini, S., A. Palmero Aprosio, M. Guerini, and S. Tonelli (2021). Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10528–10539. Association for Computational Linguistics.

Lewis, D. D. (1995). A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, Volume 29, pp. 13–19. ACM New York, NY, USA.

Linardatos, P., V. Papastefanopoulos, and S. Kotsiantis (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy 23*(1), 18.

Little, R. J. and D. B. Rubin (2002). *Statistical analysis with missing data*, Volume 793. John Wiley & Sons.

Lukasik, M., S. Bhojanapalli, A. Menon, and S. Kumar (2020). Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pp. 6448–6458. PMLR.

Malinin, A. and M. Gales (2018). Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, Volume 31. Curran Associates, Inc.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological) 42*(2), 109–127.

McCutcheon, A. L. (1987). *Latent class analysis*, Volume 64. Sage.

McLachlan, G. J. and K. E. Basford (1988). *Mixture models: Inference and applications to clustering*, Volume 38. M. Dekker New York.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019). Finite mixture models. *Annual review of statistics and its application 6*, 355–378.

McLachlan, G. J. and S. Rathnayake (2014). On the number of components in a gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4*(5), 341–355.

Mengersen, K. L., C. Robert, and M. Titterington (2011). *Mixtures: estimation and applications*. John Wiley & Sons.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics 21*(6), 1087–1092.

Minka, T. (2000). Estimating a dirichlet distribution.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association 78*(381), 47–55.

Nguyen, V.-L., S. Destercke, and E. Hüllermeier (2019). Epistemic uncertainty sampling. In *Discovery Science: 22nd International Conference, DS 2019, Split, Croatia, October 28–30, 2019, Proceedings 22*, pp. 72–86. Springer.

Nie, Y., X. Zhou, and M. Bansal (2020). What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Nowicka, M. and M. D. Robinson (2016). Drimseq: a dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research 5*.

Paun, S., R. Artstein, and M. Poesio (2022). *Statistical methods for annotation analysis*. Springer Nature.

Pavlick, E. and T. Kwiatkowski (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics 7*, 677–694.

Peterson, J., R. Battleday, T. Griffiths, and O. Russakovsky (2019). Human uncertainty makes classification more robust. *Proceedings of the IEEE International Conference on Computer Vision*, 9616–9625.

Plank, B. (2022). The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10682. Association for Computational Linguistics.

Rasmussen, C. (1999). The infinite gaussian mixture model. In *Advances in Neural Information Processing Systems*, Volume 12. MIT Press.

Redner, R. A. and H. F. Walker (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review 26*(2), 195–239.

Reynolds, D. A. and R. C. Rose (1995). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE transactions on speech and audio processing 3*(1), 72–83.

Robert, C. P. et al. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Volume 2. Springer.

Robert, C. P. and G. Casella (2004). *The Metropolis—Hastings Algorithm*, pp. 267–320. Springer New York.

Rolnick, D., A. Veit, S. Belongie, and N. Shavit (2017). Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694* .

Rousseau, J. and K. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 73*(5), 689–710.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Schmarje, L., M. Santarossa, S.-M. Schröder, C. Zelenka, R. Kiko, J. Stracke, N. Volkmann, and R. Koch (2022). A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering. In *Computer Vision – ECCV 2022*, Cham, pp. 363–380. Springer Nature Switzerland.

Sensoy, M., L. Kaplan, and M. Kandemir (2018). Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, Volume 31. Curran Associates, Inc.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Shafaei, S., S. Kugele, M. H. Osman, and A. Knoll (2018). Uncertainty in machine learning: A safety perspective on autonomous driving. In *Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings 37*, pp. 458–464. Springer.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science 25*(3), 289 – 310.

Smith, R. C. (2013). *Uncertainty quantification: theory, implementation, and applications*, Volume 12. Siam.

Song, H., M. Kim, D. Park, Y. Shin, and J.-G. Lee (2023). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems 34*(11), 8135–8153.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 795–809.

Stewart, I. D. and T. R. Oke (2012). Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society 93*(12), 1879–1900.

Sullivan, T. J. (2015). *Introduction to uncertainty quantification*, Volume 63. Springer.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826.

Ulmer, D. T., C. Hardmeier, and J. Frellsen (2023). Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Transactions on Machine Learning Research*.

Uma, A. N., T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio (2021). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research 72*, 1385–1470.

Valdenegro-Toro, M. and D. S. Mori (2022). A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1508–1516. IEEE.

Wang, X. and B. Plank (2023). Actor: Active learning with annotator-specific classification heads to embrace human label variation. *arXiv preprint arXiv:2310.14979*.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics 11*(1), 95–103.

Xiao, Y. and W. Y. Wang (2019). Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI conference on artificial intelligence*, Volume 33, pp. 7322–7329.

Zhang, S., C. Gong, and E. Choi (2021). Learning with different amounts of annotation: From zero to many labels. *arXiv preprint arXiv:2109.04408*.

Zhou, S. K., H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE 109*(5), 820–838.

Zhu, X. X., J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Haberle, Y. Hua, R. Huang, L. Hughes, H. Li, Y. Sun, G. Zhang, S. Han, M. Schmitt, and Y. Wang (2020). So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine 8*(3), 76–89.

Zhu, X. X., D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine 5*(4), 8–36.

Part II.

One-dimensional Ground Truth

5. Categorising the world into local climate zones: towards quantifying labelling uncertainty for machine learning models

Contributing article:

Hechinger, K., Zhu, X.X. and Kauermann, G. (2024). Categorising the World into Local Climate Zones: Towards Quantifying Labelling Uncertainty for Machine Learning Models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(1), 143-161. https://doi.org/10.1093/jrsssc/qlad089.

Copyright information:

Code and data:

The supplementary code is publicly available on GitHub: https://github.com/katharinahech/label-uncertainty-so2sat.git

Author contributions:

The general idea of employing a mixture model with a latent variable originated from Göran Kauermann. He also proposed the stochastic expectation maximization algorithm for estimation. Katharina Hechinger is responsible for tailoring the approach to assess label uncertainty. The implementation of the model and estimation procedure in `Python`, including data preparation and visualization, was carried out by Katharina Hechinger. The manuscript was primarily written by Katharina Hechinger and Göran Kauermann. Xiao Xiang Zhu provided the dataset and offered helpful insights as a domain expert. All authors were involved in improving and proofreading the manuscript.

Further versions:

Hechinger, K., Zhu, X. X. & Kauermann, G. (2022). Categorizing the World into Local Climate Zones - Towards Quantifying Labelling Uncertainty for Machine Learning Models. Proceedings of the 36th International Workshop on Statistical Modelling, 1:193-197.

Journal of the Royal Statistical Society Series C: Applied Statistics, 2024, **73**, 143–161
https://doi.org/10.1093/jrsssc/qlad089
Advance access publication 11 October 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Categorising the world into local climate zones: towards quantifying labelling uncertainty for machine learning models

Katharina Hechinger[1], Xiao Xiang Zhu[2] and Göran Kauermann[1]

[1]Department of Statistics, Ludwig-Maximilians-University, Munich, Germany
[2]Data Science in Earth Observation, Technical University of Munich, Munich, Germany

Address for correspondence: Katharina Hechinger, Department of Statistics, Ludwig-Maximilians-University, Munich 80539, Germany. Email: Katharina.Hechinger@stat.uni-muenchen.de

Abstract

Image classification is often prone to labelling uncertainty. To generate suitable training data, images are labelled according to evaluations of human experts. This can result in ambiguities, which will affect subsequent models. In this work, we aim to model the labelling uncertainty in the context of remote sensing and the classification of satellite images. We construct a multinomial mixture model given the evaluations of multiple experts. This is based on the assumption that there is no ambiguity of the image class, but apparently in the experts' opinion about it. The model parameters can be estimated by a stochastic expectation maximisation algorithm. Analysing the estimates gives insights into sources of label uncertainty. Here, we focus on the general class ambiguity, the heterogeneity of experts, and the origin city of the images. The results are relevant for all machine learning applications where image classification is pursued and labelling is subject to humans.

Keywords: expert evaluations, labelling uncertainty, mixture models, multiple labellers, stochastic expectation maximisation

1 Introduction

Machine learning has achieved impressive standards in recent years. In particular, in image analysis and classification, deep learning has completely changed the way to approach image data. Today, machine learning is increasingly used for the classification of images, with applications for instance in medical image analysis, face recognition, machine vision, and many more. In this paper, we focus on satellite images and their use to classify the world into so-called local climate zones (LCZ) as a categorisation of the surface. The concept of LCZ, as proposed in Stewart and Oke (2012), has achieved a general standard in remote sensing and is based on the assumption that the structure of the landscape influences the local climate. The LCZ scheme categorises the surface of the world into 17 classes that are supposed to influence local climate behaviour. The classes differ in surface structure (e.g. related to the density or height of trees and buildings) and surface cover (unsealed or sealed). A schematic description and exemplary satellite images are shown in Figure 1. This categorisation serves as an international standard for the mapping and analysis of urban areas and massive effort has been spent in developing algorithms that transform satellite images into an LCZ map. For this purpose, deep learning offers promising solutions to achieve high-quality maps and has already proven its utility in this regard, see, e.g. Qiu et al. (2019) or Qiu et al. (2018). Zhu et al. (2022) combine earth observation data with deep learning and reveal detailed morphology of urban agglomerations across the globe. For an extensive overview of challenges, advances and resources of deep learning in the field of remote sensing, we refer to Zhu et al. (2017).

Received: July 26, 2022. **Accepted:** August 30, 2023

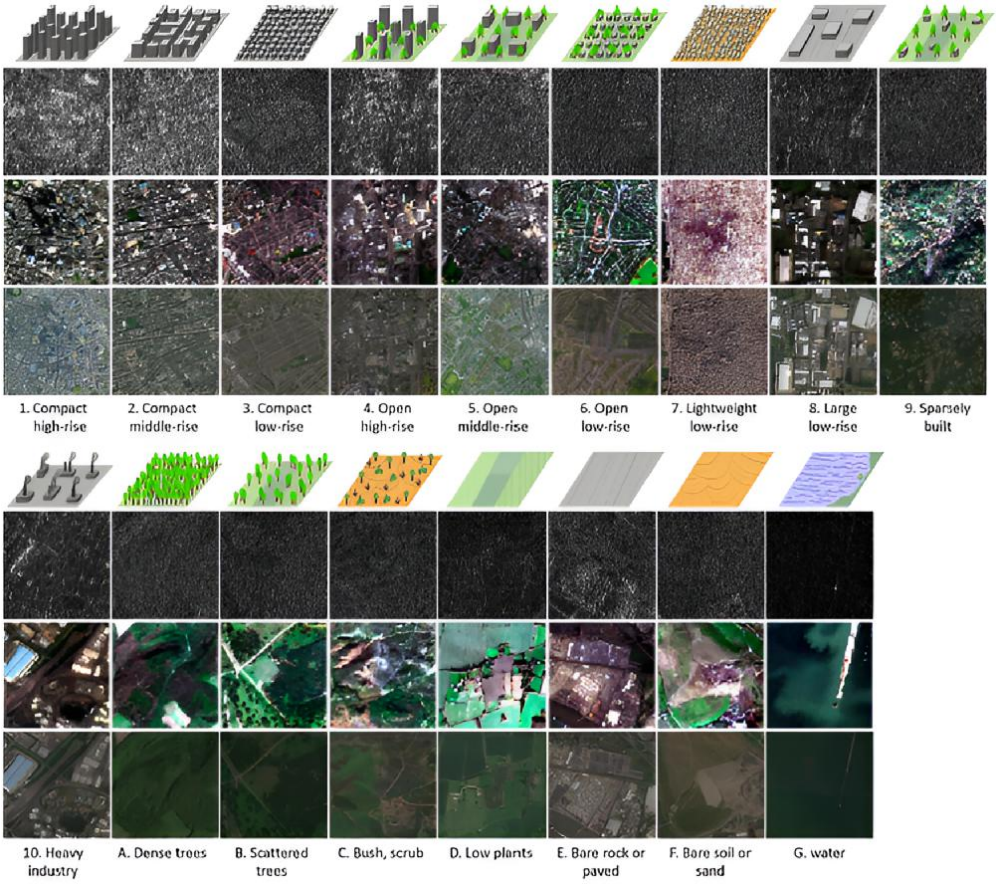


Figure 1. Example of image scenes of the 17 LCZ classes (upper image: Sentinel-1, middle image: Sentinel-2, lower image: high-resolution aerial image from Google Earth). Source: Zhu et al. (2020).

Machine learning is thereby based on labelled data, that is we are in the context of supervised learning, see, e.g. Friedman et al. (2001). In this context, the problem of acquiring labels is very common and often solved by crowdsourcing, as introduced by Estellés-Arolas and González-Ladrón-de Guevara (2012). A lot of effort has already been spent in analysing the quality of such labels, e.g. by Raykar and Yu (2011) or Karger et al. (2013) in the case of multi-class classification. Dawid and Skene (1979) also investigated the observed variation and its effect on the resulting measurements. More recently, Chang et al. (2017) developed a framework for improving the standard workflow of crowdsourcing by incorporating knowledge about labelling by experts. Northcutt et al. (2021) also proposed confident learning in large (crowdsourced) databases with mislabelling, also referred to as ontological uncertainty. Although promising results can be achieved by exploiting the wisdom of the crowd, it is not suitable for our area of application. The classification of satellite images into LCZ is non-trivial and relies on special knowledge and detailed classification instructions. Therefore, the insights from crowdsourcing theory are helpful to a certain degree but cannot be transferred directly for the classification of LCZ. In particular, our use case requires that experts label images by hand, classifying hundreds of images into one of the 17 categories. This process is apparently time-consuming and not without ambiguities. In fact, different experts come to different conclusions when classifying images. The quantification of uncertainty is therefore particularly crucial as data sources are highly inhomogeneous and labelled image data are rare, see Russwurm et al. (2020). All in all, classifying satellite images into their corresponding LCZs demands a complicated and time-consuming annotation process.

Using noisy or even deficient labels for the training of deep learning models leads to huge uncertainties and can result in serious challenges. Our problem concentrates on so-called label noise and

we refer to Frenay and Verleysen (2014) for an extensive survey. We are also faced with label ambiguity, where methods like label distribution learning have been introduced by, e.g. Geng (2016). Another approach is to incorporate the human component. Dgani et al. (2018) discuss methods of training neural networks despite unreliable human annotations and Peterson et al. (2019) incorporate this human uncertainty to increase the robustness of classification algorithms. Luo et al. (2021) investigate label distribution learning also in the particular field of remote sensing.

The problem of labelling uncertainty goes well beyond the particular problem considered here. It is found also, e.g. in medical image analysis as described in Zhang et al. (2020) or Ju et al. (2021), face identification (Kamar et al., 2012) or more generally in crowdsourcing areas (Phillips et al., 2018). In this work, we consider data, where each image has been classified by multiple experts but the true class remains unknown. This relates to the setting of latent class models, as introduced by Lazarsfeld (1950), where a set of observed variables is related to a set of latent variables. Goodman (1974) extended the original idea by using Maximum Likelihood methods and today, numerous variants of latent class analysis exist (Magidson et al., 2020). These methods are helpful in many applications where the goal is to uncover hidden groups or structures in observed data. In this work, we aim to quantify the uncertainty of the experts about some of the images by applying a classical finite mixture model. We refer to McLachlan and Peel (2000) or McLachlan et al. (2019) for a general description of the model class. See also Fraley and Raftery (2002) for the relation of mixture models and model-based clustering and Cadez et al. (2001) for an application to transaction data. To link the application to mixture models we assume a latent ground truth. To be specific, we employ a multinomial mixture model and our ultimate goal is to estimate the 'true' confusion matrix, i.e. without knowing the ground truth of an image. This will allow us to investigate the inevitable uncertainty in human image labelling. Moreover, we investigate if and how this uncertainty differs for images from different regions of the world, i.e. if and how the accuracy of annotation differs locally.

The quantification of uncertainty is receiving increasing interest in machine learning in recent years. We refer to Gawlikowski et al. (2023) or Hüllermeier and Waegeman (2021) for a general overview. Typically, uncertainty is decomposed into two parts: *aleatoric uncertainty* and *epistemic uncertainty*, sometimes also labelled as irreducible and reducible uncertainty. Such decompositions are not uniquely defined, and here we focus on an additional layer of uncertainty, which is often omitted, namely that the ground truth remains unknown. In our case, for each satellite image, we only have the annotations given by the human experts but the true LCZ is not given.

The paper is organised as follows. In Section 2, we give a detailed description of the data at hand and describe the annotation process that has been applied to generate the data. In Section 3, we introduce our statistical approach and the models used to quantify the labelling uncertainty. Section 4 discusses the results of the particular data set at hand. Section 5 concludes the paper.

2 Data

We will analyse label uncertainty based on the earth observation benchmark data set So2Sat LCZ42 (Zhu, 2021). For a detailed description of the full data set, we refer to Zhu et al. (2020). It comprises the LCZ labels of Sentinel-1 and Sentinel-2 image patches in 42 urban agglomerations across the globe. The images come in so-called patches, each covering an area of 320×320 m. Figure 1 shows an illustration of the LCZs, as well as examples of corresponding remote-sensing image patches. The data set was created by a complicated and labour-intensive labelling project. For selected cities, polygons of different sizes were extracted, delineated such that the surface was largely homogeneous within each polygon. Within these polygons, equidistant images were then selected and initially labelled by a panel of two experts, which also used auxiliary data such as high-resolution satellite images from Google Earth. This procedure resulted in 'clusters' of images, manually labelled by a larger panel of 11 experts. We here focus exclusively on these 11 votes per image. The produced labels do not serve as final labelling but rather as a validation stage. At this stage, one aims to assess overall labelling quality by comparing the opinion of the experts to the previously found label and possibly correcting it accordingly. A detailed layout of the labelling procedure and the validation stage can be found in Zhu et al. (2020). Finding the suitable LCZ for a polygon is impossible for a layperson and even labour-intense and non-trivial for trained experts. As the definition of LCZs is very vague in

Table 1. Sketch of the database

Image ID	1	2	3	4	5	6	7	8	9	10	A	B	C	D	E	F	G	City
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	Berlin
2	0	0	0	0	0	0	0	0	4	0	0	0	0	7	0	0	0	Berlin
...										...								...
3650	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	5	Zurich
...										...								...

its nature, a rigorous labelling workflow and decision rules had to be designed to ensure the highest labelling quality possible. Overall, we look at 159,581 images from 9 cities leading to a data structure as sketched in Table 1.

The voting data already suggests some degree of certainty for the experts. For 77.18 % of the images, the experts agree on one single LCZ. We observe so-called 'voting patterns' for the other images. Eleven experts voting for 17 classes leads to 11^{17} possible patterns, of which only 243 occur in the data set. This observation suggests that only a few classes are frequently confused, while others can always be distinguished.

The votes are not distributed evenly among the classes. As Figure 2 shows, the majority of votes were for classes A, D, and G, whereas other classes hardly occur in the votings. Looking at the different cities, there is a large difference in not only the number of patches per city but also in the distribution of votes within the cities, see Figure 2. We suspect a spatial correlation within the cities that influences the collection of votes. Looking at the plots, the distributions of votes in the different cities vary quite a lot. For example, class G dominates in London or Zurich, while it hardly occurs in Paris. One should also note that the number of images inspected by the experts is also different in each city, which might impact the quality of the voting process.

Another interesting aspect of the data is its clustered structure. The images were selected through cluster sampling of polygons including homogeneous areas. In Figure 3, we show exemplary the locations of the selected images in Berlin. The clustered structure is apparent. Other cities look comparable.

Finally, we look at the data from the voters' view. Figure 4 shows a histogram of the votes cast by each expert. We recognise heterogeneity among the voters, where the voting behaviour differs mostly for urban classes (1–10), while the distributions for the non-urban classes A–G are pretty similar. We will therefore also aim to question if and how the voters' classification differs.

3 Annotation uncertainty

3.1 Description of model

To achieve the goal of exploring labelling uncertainty, we look at the votes cast by earth observation experts. Each image patch i, $i = 1, \ldots, n$ is assessed by a set of experts indexed with j, $j = 1, \ldots, J$. The experts thereby classify each image individually into the LCZ k, where $k = 1, \ldots, K$. The corresponding vote of the expert is denoted by $V_j^{(i)} \in \{1, \ldots, K\}$. It is notationally helpful to rewrite this vote into the K dimensional indicator vector, which we denote in bold with $\boldsymbol{V}_j^{(i)} = (\mathbb{1}\{V_j^{(i)} = 1\}, \ldots, \mathbb{1}\{V_j^{(i)} = K\})$, with $\mathbb{1}\{\cdot\}$ as indicator function. This allows to accumulate the labellers' votes into the data points $\boldsymbol{Y}^{(i)} = (Y_1^{(i)}, \ldots, Y_K^{(i)})$ with $Y_k^{(i)} = \sum_{j=1}^{J} \mathbb{1}(V_j^{(i)} = k)$. This vector can be considered as the vote distribution for image i.

We assume further that each image comes from a single true class (=ground truth), which is a reasonable assumption based on the clustered data structure described above. Hence, we assume that there is no ambiguity in the image class, but apparently, there are ambiguities in the voters' opinions about this class. We denote with $Z^{(i)} \in \{1, \ldots, K\}$ the true class of image i, which apparently remains unknown. Like above, we can reformulate the true class as K dimensional index vector $\boldsymbol{Z}^{(i)} = (\mathbb{1}\{Z^{(i)} = 1\}, \ldots, \mathbb{1}\{Z^{(i)} = K\})$. Our intention is now to get information on $Z^{(i)}$ or $\boldsymbol{Z}^{(i)}$, respectively, given the voters' distribution $\boldsymbol{Y}^{(i)}$. We will therefore apply Bayesian reasoning using

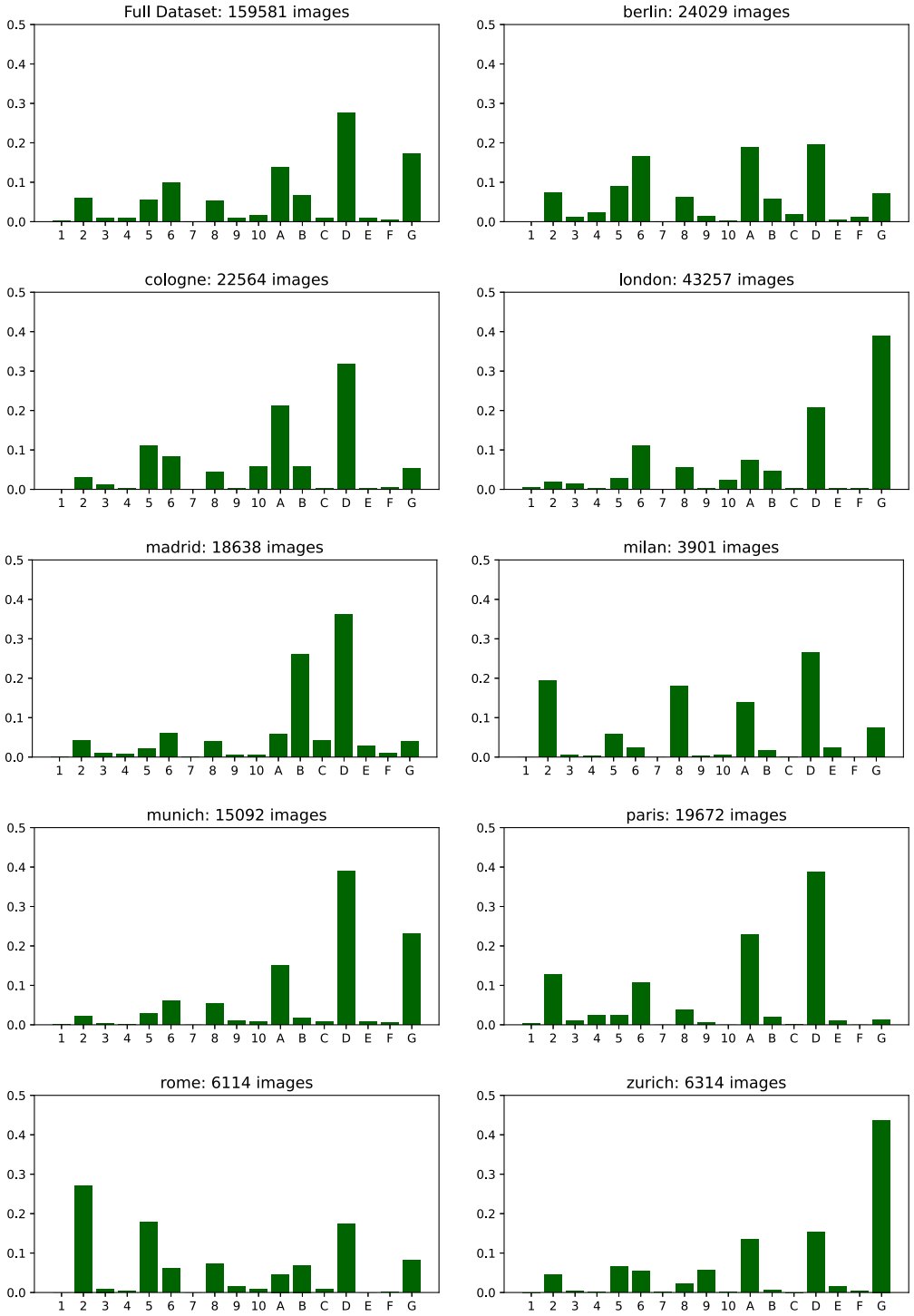


Figure 2. Class distribution of the votes per city, along with the number of image patches per city. The top left figure shows the distribution for all images, the remaining figures show the distribution in Berlin (row 1), Cologne and London (row 2), Madrid and Milan (row 3), Munich and Paris (row 4), Rome and Zurich (row 5).

a mixture modell approach, which requires formulating a distribution framework. For the true classes, we assume a multinomial distribution, also called prior distribution, i.e.

$$Z^{(i)} \sim Multi(\boldsymbol{\pi}, 1) \text{ i.i.d.},$$

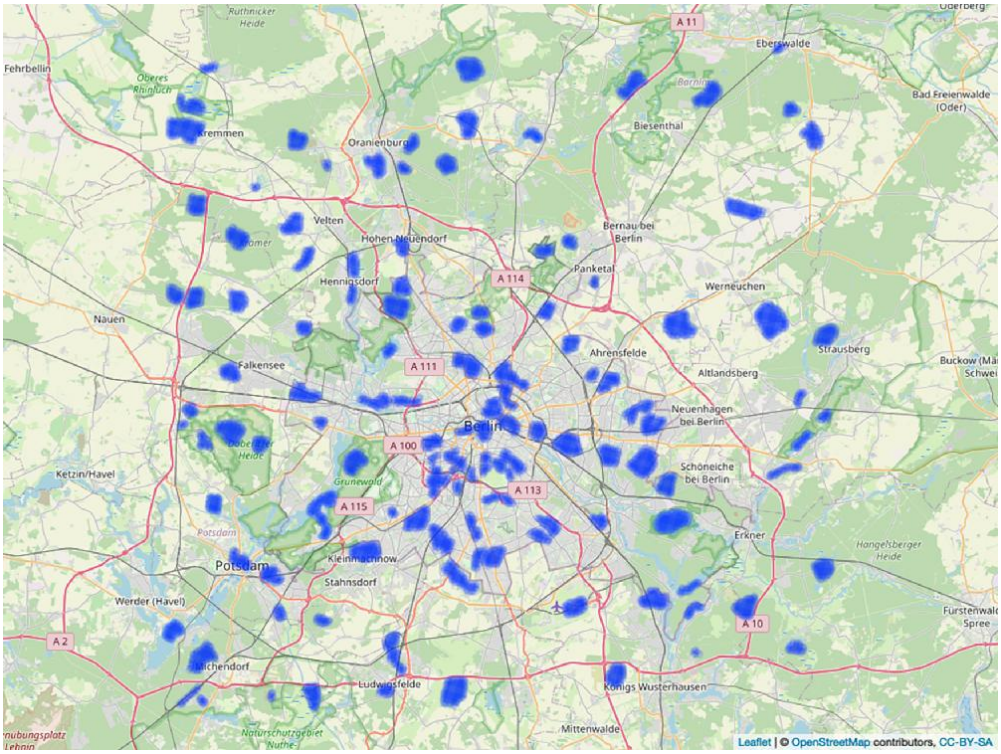


Figure 3. Spatial distribution of images and polygons across Berlin.

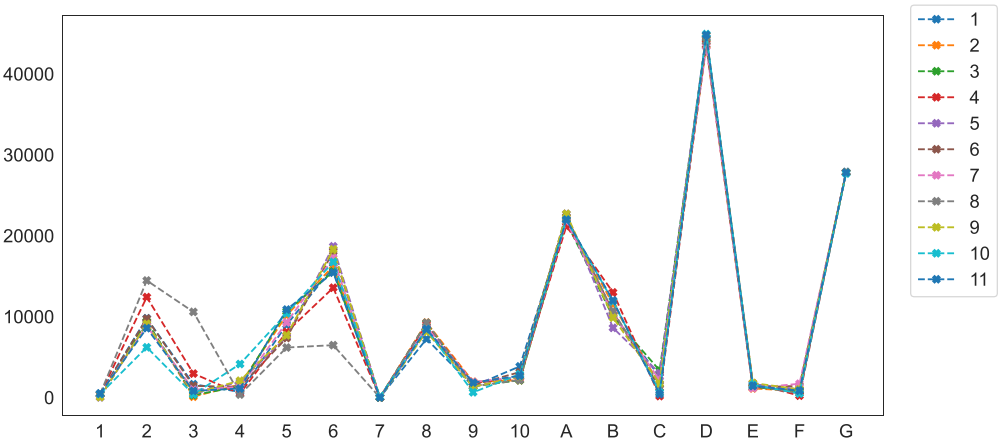


Figure 4. Distribution of votes, split by expert.

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ with π_k as so-called prior probability that image i is from the true LCZ k for $k = 1, \ldots, K$. Given the true class of the image, we further assume that the labellers' votes also follow a multinomial distribution, i.e.

$$\boldsymbol{Y}^{(i)} \mid Z^{(i)} \sim Multi(\boldsymbol{\theta}_{Z^{(i)}}, J), \tag{1}$$

where $\boldsymbol{\theta}_{Z^{(i)}} = (\theta_{Z^{(i)}1}, \ldots, \theta_{Z^{(i)}K})$. The parameters express the probabilities of voting for classes $1, \ldots, K$ given the true class is $Z^{(i)} \in \{1, \ldots, K\}$. We collect the coefficients into the matrix

$$\Theta = (\theta_{pk}, p, k = 1, \ldots K),$$

which will be estimated from the data. Again, θ_{pk} refers to the voting probabilities, given $Z^{(i)} = p$. We will also demonstrate how to estimate vector $\boldsymbol{\pi}$ if no prior knowledge about the general distribution of the LCZ is given or if any prior knowledge is intended to be ignored. Note that this approach corresponds to empirical Bayes estimation by estimating the prior to maximise the marginal likelihood, see, e.g. Robbins (1992).

Given that the true classes are unobserved, we are in the framework of mixture models. We obtain the likelihood contribution of the ith image by summing over all classes, that is

$$P(\boldsymbol{Y}^{(i)}, \boldsymbol{\pi}, \Theta) = \sum_{k=1}^{K} \pi_k P(\boldsymbol{Y}^{(i)}, \boldsymbol{\theta}_k), \tag{2}$$

where $\boldsymbol{\theta}_k$ is one column of the true confusion matrix and the probability in the sum results from the model (1). Apparently, this is getting clumsy, so we apply the expectation maximisation (EM) algorithm, or more precisely, a stochastic version of it.

3.2 Stochastic EM algorithm

The main idea of the EM algorithm, introduced by Dempster et al. (1977), is that the latent image class $Z^{(i)}$ is replaced by its expected value, given the data and the current estimates. For mathematical details, we refer to Appendix A. This gives complete data so that the above estimates can be easily derived. These steps are carried out iteratively. While the EM algorithm in general is a handy tool, it is also very slow and numerically intense. In fact, in our example, we would need to calculate the posterior expectation for all images. Instead, we make use of the stochastic EM algorithm (SEM) as proposed in Celeux et al. (1996). Here, the E-step is replaced by a simulation step, leading to simulated true image classes and hence allowing for simple estimation. Like the EM algorithm, one iterates between two steps to estimate the unknown parameters.

Let $\hat{\boldsymbol{\pi}}_{<t>}$ and $\hat{\Theta}_{<t>}$ be the estimates in the t-th iteration step of the algorithm. Taking these parameters, we can calculate the posterior probabilities

$$\begin{aligned}\tau^{(i)}_{<t>l} &= P(Z^{(i)} = l \mid \boldsymbol{Y}^{(i)}; \hat{\boldsymbol{\pi}}_{<t>}, \hat{\Theta}_{<t>}) = \frac{P(Z^{(i)} = l; \hat{\boldsymbol{\pi}}_{<t>}) P(\boldsymbol{Y}^{(i)} \mid Z^{(i)} = l; \hat{\Theta}_{<t>})}{P(\boldsymbol{Y}^{(i)}; \hat{\boldsymbol{\pi}}_{<t>}, \hat{\Theta}_{<t>})} \\ &= \frac{\hat{\pi}_{<t>l} P(\boldsymbol{Y}^{(i)}; \hat{\boldsymbol{\theta}}_{<t>l})}{\sum_{l'=1}^{K} \hat{\pi}_{<t>l'} P(\boldsymbol{Y}^{(i)}; \hat{\boldsymbol{\theta}}_{<t>l'})}.\end{aligned}$$

The simulation-based E-step is now carried out by drawing

$$Z^{(i)}_{<t>} \sim Multi(\boldsymbol{\tau}^{(i)}_{<t>}, 1),$$

where $\boldsymbol{\tau}^{(i)}_{<t>} = (\tau^{(i)}_{<t>1}, \ldots, \tau^{(i)}_{<t>K})$. We obtain complete data with these simulated true classes, leading to new estimates based on the complete likelihood.

Using this standard SEM procedure, we produce a chain of estimates (or simulated values) at each iteration, namely $(\hat{\Theta}_{<t>})_{t \geq 0}$ and $(\hat{\boldsymbol{\pi}}_{<t>})_{t \geq 0}$ and therewith $(\hat{\boldsymbol{\tau}}_{<t>})_{t \geq 0}$. The final estimate can then be calculated as the mean value of the produced estimates starting at iteration t_0, the end of the burn-in, i.e. the mean parameter resulting from the last $T - t_0$ iterations. For the parameter Θ describing the voting probabilities this results in

$$\hat{\Theta}_{final} = \frac{1}{T - t_0} \sum_{t=t_0}^{T} \hat{\Theta}_{<t>}.$$

The stochastic version of the EM has two advantages. First, it is numerically more straightforward, though it requires additional computation. Second, we can directly quantify the uncertainty of the estimates. We are interested in the estimation variance of the parameters, primarily of course in the

estimate of the (mis)classification matrix Θ. We refer to Rubin's formula resulting from multiple imputations, see Rubin (1976) and Little and Rubin (2002). Note that the matrix estimate of Θ does not have full rank since the rows sum up to one. We therefore drop the last column and write the matrix estimate into a vector $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\theta}}_{1,-K}^T, \ldots, \hat{\boldsymbol{\theta}}_{K,-K}^T)$, where $\hat{\boldsymbol{\theta}}_{l,-K}$ is the $K-1$ dimensional subvector resulting from the first $K-1$ columns of $\hat{\boldsymbol{\theta}}_l$. We obtain

$$Var(\hat{\boldsymbol{\vartheta}}) = E_Z(Var(\hat{\boldsymbol{\vartheta}} \mid Z)) + Var_Z(E(\hat{\boldsymbol{\vartheta}} \mid Z)), \tag{3}$$

where the subscript Z refers to expectation and variance with respect to the latent classes $Z^{(i)}$ for $i = 1, \ldots, n$. Note that for given Z we are in a complete data scenario and it is not difficult to show that in this case subvectors $\hat{\boldsymbol{\theta}}_{l,-K}$ and $\hat{\boldsymbol{\theta}}_{l',-K}$ of $\hat{\boldsymbol{\vartheta}}$ are independent for $l \neq l'$. This leads to the variance

$$Var(\hat{\boldsymbol{\theta}}_{l,-K} \mid Z) = \frac{\text{diag}(\boldsymbol{\theta}_{l,-K}) - \boldsymbol{\theta}_{l,-K}^T \boldsymbol{\theta}_{l,-K}}{\sum_{i=1}^n \mathbb{1}\{Z^{(i)} = l)},$$

which is estimated in the t-th iteration step by replacing $\boldsymbol{\theta}_{l,-K}$ through its estimate, i.e.

$$\widehat{Var}(\hat{\boldsymbol{\theta}}_{l,-K} \mid Z_{<t>}) = \frac{\text{diag}(\hat{\boldsymbol{\theta}}_{<t>l,-K}) - \hat{\boldsymbol{\theta}}_{<t>l,-K}^T \hat{\boldsymbol{\theta}}_{<t>l,-K}}{\sum_{i=1}^n \mathbb{1}\{Z_{<t>}^{(i)} = l\}}.$$

Replacing now the expectation in equation (3) through the simulated steps from the EM algorithm allows us to estimate the variance through

$$\begin{aligned}\widehat{Var}(\hat{\boldsymbol{\vartheta}}) = &\frac{1}{T - t_0} \sum_{t=t_0+1}^{T} \text{blockdiag}\Big(\widehat{Var}(\hat{\boldsymbol{\theta}}_{l,-K} \mid Z_{<t>})\Big) \\ &+ \frac{1}{T - t_0 - 1} \sum_{t=t_0+1}^{T} \big(\hat{\boldsymbol{\vartheta}}_{<t>} - \hat{\boldsymbol{\vartheta}}_{final}\big)\big(\hat{\boldsymbol{\vartheta}}_{<t>} - \hat{\boldsymbol{\vartheta}}_{final}\big)^T,\end{aligned}$$

with obvious definition of $\hat{\boldsymbol{\vartheta}}_{final}$.

3.3 Label switching

Like in every mixture model, the resulting classes are subject to label switching, i.e. the numbering of the resulting classes does not match the original numbering of the LCZs. In other words, while the classes labelled by the voters have explicit meaning and therefore an interpretable order, the latent classes are subject to permutation and have no explicit interpretation. For the mixture model, we assume 17 true classes which are ordered at convergence as $C = \{C_1, C_2, \ldots\}$. On the observation side the experts categorise the satellite images into 17 classes denoted by $L = \{L_1, L_2, \ldots\}$. We now need to match the latent classes C to the labelled classes L. It is important to note that the labels of the clusters returned by the algorithm are unidentifiable. Therefore, to ensure a clear assignment, we need a bijective function going from the cluster labels C to the voter labels L. Or putting it differently, we need to construct a permutation $\sigma()$ on the numbers $\{1, \ldots K\}$ such that $\sigma(C_l) = k$ means that the latent class C_l corresponds to the LCZ L_k. This could be achieved by looking at the posterior probability of the latent classes given the voters' opinions. Note that for a single vote $V^{(i)}$ we obtain $P(Z^{(i)} = l \mid V^{(i)} = k) \propto P(V^{(i)} = k \mid Z^{(i)} = l)\pi_l$ or written in matrix form

$$P(Z^{(i)} = \{1, \ldots, K\}^T \mid V^{(i)} = \{1, \ldots, K\}) \propto \text{diag}(\boldsymbol{\pi})\Theta^T.$$

This suggests constructing the permutation $\sigma()$ such that its inverse fulfils

$$\sigma^{-1}(k) = \arg\max_l(\text{diag}(\boldsymbol{\pi})\Theta^T)). \tag{4}$$

This still might not lead to a unique definition. We, therefore, apply rule (4) in descending order of the relative frequency of the labellers' votes and choose the arg max from the not-allocated classes only. A detailed layout of the algorithm is provided in the Appendix B. After finding the correct permutation of the numbers, we rename the original clusters according to the respective local climate zone label.

4 Sources of uncertainty in the votes

We are now in the position to approach different questions related to human annotation of satellite images. These are:

1. How distinguishable are the LCZs in general, that is can we estimate the 'true' confusion matrix?
2. Is there an expert bias, that is are the experts heterogeneous or homogeneous with respect to the labelling?
3. Is the voting behaviour influenced by geographic differences, that is does the 'true' confusion matrix differ in the different cities where the data come from?

All three questions are tackled subsequently.

4.1 Ambiguity of LCZs

A very general aspect for quantifying the uncertainty in the voting data set are the LCZs themselves. By looking at the definition and characterisation of the classes, it is obvious that some are very similar and might not be easily distinguishable, even for experts. This is for example the case for classes 3 and 7, which both describe urban low-rise environments. Contrarily, there are LCZs that are easy to discover on images and that are likely to be never confused by humans, e.g. class 17 covering water areas. In general, it is presumably more difficult to distinguish urban classes, i.e. LCZs 1–10 than non-urban classes which are LCZs A–G.

The parameter of main interest is the confusion matrix $\hat{\Theta}$, i.e. the estimated true confusion matrix of the classes.

To obtain a stable estimation and interpretable results, we restricted the estimation procedure described in the previous section to $K = 16$ instead of $K = 17$, omitting LCZ 7 (lightweight low-rise building types). This is reasonable, not only due to the semantic interpretation of 'slums', which are very unlikely to occur at all in European cities but also necessary due to the lacking data basis. As mixture models are generally able to handle any arbitrary number of classes, including a class without sufficient observations or votes in this case, this will lead to instability and confusion in the estimation results.

Figure 5 shows the resulting estimate based on the full data. The entries on the diagonal contain the probability of correctly classifying images. In contrast, entries $\boldsymbol{\theta}_{lk}$ describe the probability of classifying an image that truly belongs to class l into LCZ k instead. Looking at the diagonal of the matrix, it is obvious that correct classification is highly dependent on the number of votes a class received from the experts. Apparently, classes 2, 6, 8, 10, A, B, D, E, and G are very well separable, whereas classes 3, 4, 9, and C are often not detected correctly. Note that said classes received a very small number of votes in the labelling process so the misclassification should not be over-interpreted. This can be seen from the frequency distribution indicated at the bottom of the plot and also the estimated prior distributions (right-hand side of the matrix), which show a strong tendency of the voters for classes A, D, and G. Apart from correct classification probabilities, we can also detect classes that are not that easy to distinguish, like, e.g. class 3, where the voting probabilities are distributed among classes 2, 3, and 6.

Generally, our results depend on the input votes as the algorithm can only detect classes where the data basis is sufficient. Furthermore, it should be mentioned here our true confusion matrix is subject to the implemented label-switching process. As the multinomial mixture model produces 'meaningless' clusters that must be assigned to LCZs afterwards, the resulting estimates and their interpretation are based on the assignment strategy, which might not be unambiguous. Generally, however, we obtain interpretable insight into the inevitable ambiguity when classifying LCZs.

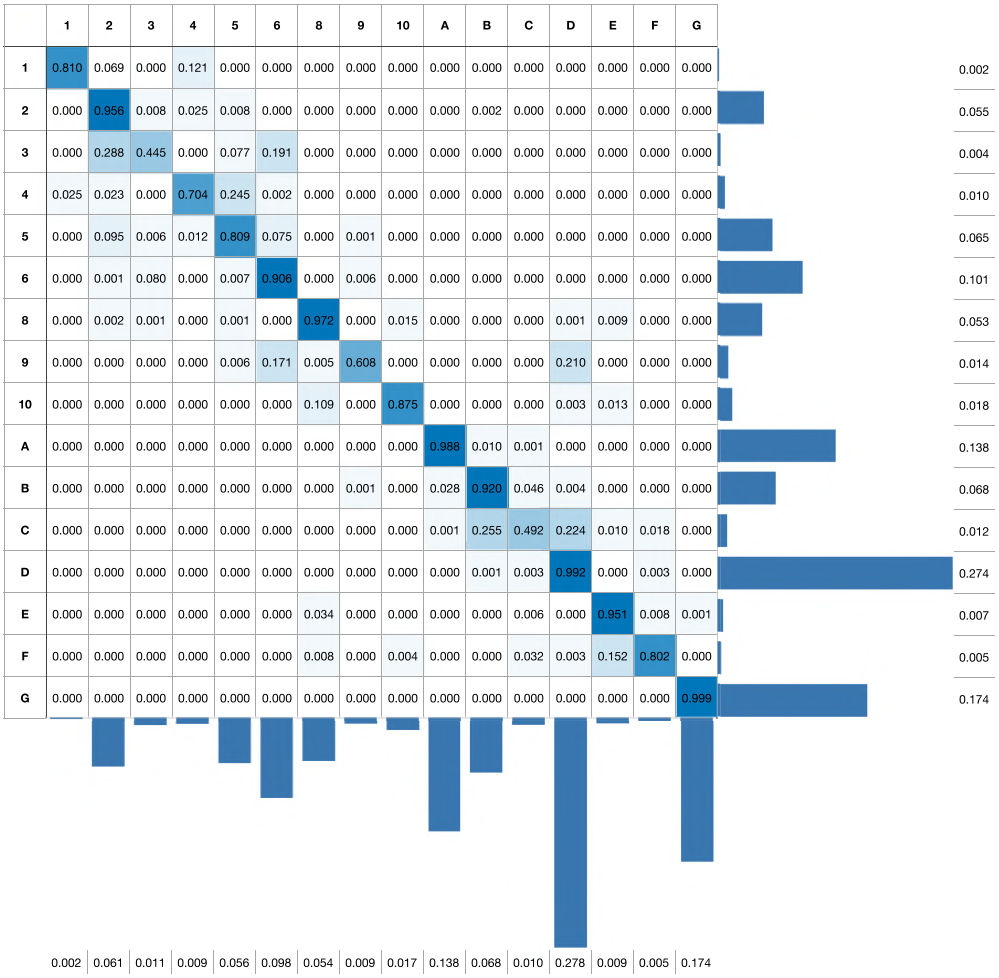


Figure 5. The matrix shows the true confusion of the voted local climate zones (columns) with the true classes (rows), along with the estimated prior probabilities on the right side and the relative vote frequency on the bottom of the plot.

4.2 Expert heterogeneity

As explained in the beginning, the task of classifying is not trivial, even for trained earth observation experts. Therefore, it is obvious that the human assessment causes some confusion and uncertainty within the data. The experts are assumed to be subject to some bias, that might impact or even skew the results. The model described in the previous section allows for assessing the impact of each individual expert and their heterogeneity. As shown in Section 2, the distribution of votes for each expert can differ quite a lot, in particular for urban classes. Therefore, it is worthwhile to investigate the individual impact of the voters further. If experts were homogeneous, their voting behaviour would not differ, and dropping the votes of one expert at a time should not change the final estimated distribution.

Looking at Figure 4, we already get a general overview of the observed voting behaviour of the experts. While the distribution of votes is similar for all experts for the non-urban LCZs (A–G), the distribution varies noticeably for the urban classes (1–10). Therefore, it is worth further examining the voting behaviours and their impact on the results.

The parameter of interest here is $\hat{\boldsymbol{\tau}}_l^{(i)}$, expressing the posterior probability of image i to belong to the true class l according to the applied model and algorithm. In order to analyse the difference between the voting behaviours, we calculate the posterior probabilities excluding a single expert. This leads to 11 estimates $\hat{\boldsymbol{\tau}}_{(-j)l}^{(i)}$, where the bracketed index $-j$ refers to the excluded voter j.

A very straightforward way of quantifying heterogeneity is calculating the log-likelihood. Here, we assume that the distribution of the categorical variable is the same for all subgroups. In this particular application, the subgroups consist of 11 experts, which classify images into 16 LCZs. Assuming a multinomial distribution the j-th expert contributes to the negative log-likelihood through

$$\Lambda_j = -\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{1}\{V_j^{(i)} = k\} \cdot \log(\tau_k^{(i)}), \tag{5}$$

where $0 \cdot \log(0)$ is defined as 0. We replace $\tau_k^{(i)}$ by the estimate excluding the j labeller and define the resulting negative log-likelihood as

$$\hat{\Lambda}_j = -\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{1}\{V_j^{(i)} = k\} \cdot \log(\hat{\tau}_{(-j)k}^{(i)}). \tag{6}$$

Statistic $\hat{\Lambda}_j$ is a random quantity, which we will now explore through bootstrapping. We, therefore, draw N images with replacement and denote with $V_j^{(i*)}$ the vote of the j-th labeller for the bootstrapped image $i*$. This leads to the bootstrap quantity

$$\hat{\Lambda}_j^* = -\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{1}\{V_j^{(i*)} = k\} \cdot \log(\tau_{(-j)k}^{(i*)}). \tag{7}$$

We repeat this step B times to obtain $\hat{\Lambda}_j^{*b}$ for $b = 1, \ldots, B$. To put the magnitude of these bootstrapped values into perspective, we look at the absolute differences to the overall mean and define

$$D_j := \left| \hat{\Lambda}_j - \frac{1}{J-1} \sum_{l=1, l\neq j}^{J} \hat{\Lambda}_l \right|$$

and the resulting bootstrapped versions

$$D_j^{*b} := \left| \hat{\Lambda}_j^{*b} - \frac{1}{J-1} \sum_{l=1, l\neq j}^{J} \hat{\Lambda}_l^{*b} \right|.$$

If the experts are homogeneous, then D_j should be in the range of 0, while if experts are heterogeneous, the range of D_j will vary. We, therefore, look at the densities of the bootstrapped distances as shown on the left side of Figure 6. The plots show a clear difference between the bootstrapped differences and therefore a clear separation of the experts regarding voting behaviour. To confirm this we run an additional bootstrap under the assumption that experts are homogeneous. To do so we conduct a randomised version of the former analysis by replacing the votes of expert j in equation (7) with a random vote. In other words, for each image we randomly permute the experts. This leads to the densities displayed on the right side of Figure 6, showing homogeneous curves for randomised votes.

We can investigate this further and check whether the experts' voting behaviour and their homogeneity differ for different cities. In Figure 7, we show the same bootstrapped densities as described above for each city separately. It appears that the experts conduct their voting differently in comparison to the rest and this varies over the different cities. For some cities, the voting behaviour seems more homogeneous, which might however also mirror smaller sample sizes as in cities like Milan, Rome, or Zurich. Overall, the analysis shows that while some of the experts align well with the overall voting behaviour, others conduct a more heterogeneous and individual voting, leading to confusion and an ambiguous majority voting.

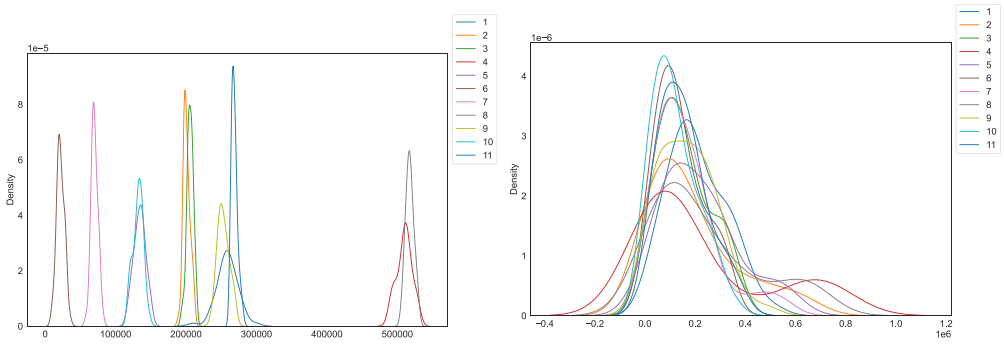


Figure 6. On the left-hand side, the densities of the differences of the bootstrapped goodness-of-fit statistics for each expert are shown. On the right-hand side, the densities belong to randomised versions, showing a clear difference between the actually observed and random votes.

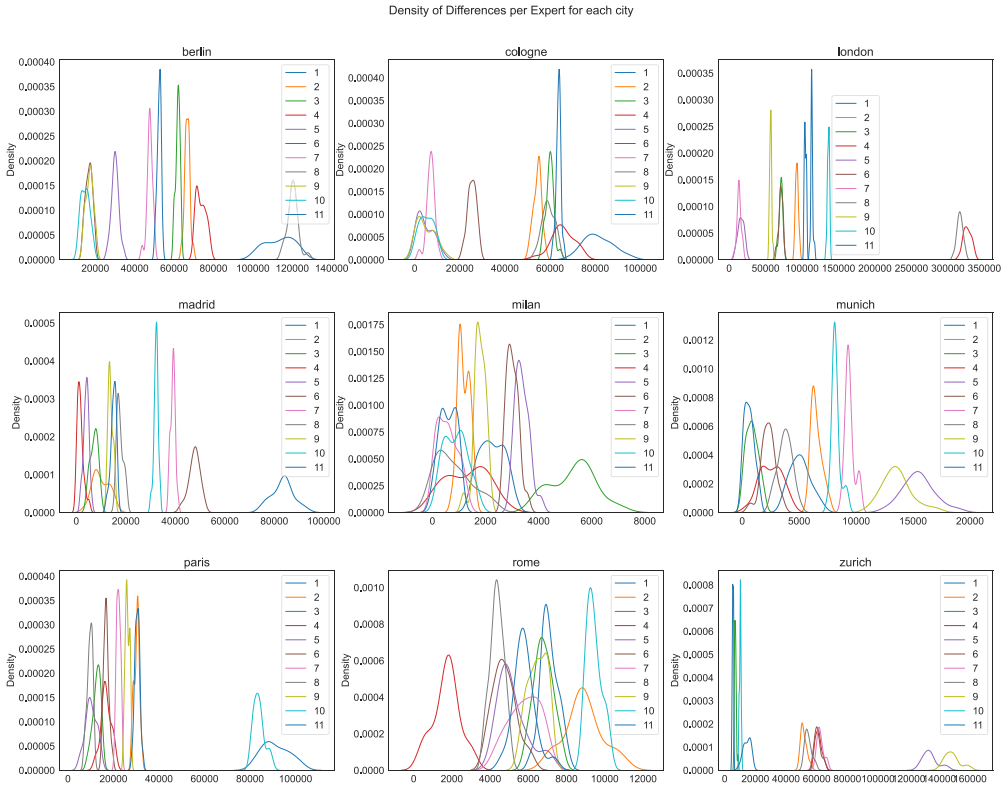


Figure 7. The plot shows the densities of the differences of the bootstrapped goodness-of-fit statistics for each expert, split up per city.

Depending on the application at hand, expert heterogeneity is often not only accepted but also desired. In this particular case, experts received the same training on how to conduct classification of satellite images and are generally assumed to produce homogeneous votings.

4.3 Geographic differences

The last question we want to consider is geographic variation, as an external influencing factor. The polygons used for the voting procedure come from nine different European cities, which are known to be quite diverse in terms of structure and architecture. While this might be intended to cover all LCZs as well as possible, it complicates the assessment of the images. The question is

whether earth observation experts have difficulties in assigning certain images to certain climate zones, depending on the respective region. Looking at Figure 2, we see differences in the sample sizes and also the vote distributions in different cities. Additionally, Figure 7 shows that the voting behaviour of experts is subject to the location of the images. This brings us to the question of whether certain LCZs are harder to identify in some cities than in others. Apparently, this might have a relevant impact on the assessment by the experts and influence the voting behaviour, which leads to uncertainty in the training data set. However, one has to note here that the voting distribution always depends on the initial draw of images or polygons in each city. The pursued strategy might lead to imbalanced labels and therefore a bias in the voting probabilities. We here are however interested in the confusion matrix and whether this matrix differs in the different cities. Maps with the location of the used images are shown in Appendix C.

Following this idea, one assumes that the distribution of Z in different cities can differ, i.e. the parameters $\pi_1, \ldots, \pi_K$ of the multinomial model are different for different regions. The model should be able to represent those differences in terms of different estimated posterior probabilities. But the crucial aspect here are the misclassification probabilities matrices Θ. Does the voting behaviour of experts look different in different regions? In order to answer this question, we will further investigate the differences in voting probabilities in various locations. This can be achieved by calculating the matrix for all regions separately:

$$\Theta(s) \text{for } s = \{1, \ldots, S\}.$$

To illustrate the problem, we focus on three regions first: Berlin, London, and Munich. Figure 8 shows these regional confusion matrices Θ_B, Θ_L, and Θ_M for the example cities. The values on the diagonal refer to the probability of correctly classifying an image to its corresponding LCZ. The three cities were chosen as examples as they differ in their general structure and have a large number of labelled patches. As the plots and the values of the matrices show, the voting probabilities and therefore the probability of correct classification of certain classes are different. This supports the claim that some classes might be more difficult to spot in some cities than in others.

Looking at the general setting and comparing all pairs of cities, we can conduct statistical tests to assess the geographic differences in the voting behaviour of experts. After estimating $\Theta(s)$ and $\Theta(s')$ separately, we want to test the hypothesis $H_0 : \Theta(s) = \Theta(s')$. Therefore, we make use of the vectorisations of Θ and the variance estimates of the parameters, as described in Section 3. The difference between the vectorisations of the confusion matrices is denoted by

$$\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\vartheta}}(s) - \hat{\boldsymbol{\vartheta}}(s'),$$

with $E(\hat{\boldsymbol{\delta}}) = 0$ if H_0 holds. Additionally, we know that $Var(\hat{\boldsymbol{\delta}}) = Var(\hat{\boldsymbol{\vartheta}}(s)) + Var(\hat{\boldsymbol{\vartheta}}(s'))$. The test statistic is constructed as

$$T = Var(\hat{\boldsymbol{\delta}})^{-1/2}\hat{\boldsymbol{\delta}},$$

where $Var(\hat{\boldsymbol{\delta}})^{-1/2}$ is calculated by using a singular value decomposition. Under H_0 it holds

$$T \overset{a}{\sim} N(0, \boldsymbol{I}),$$

which suggests using ordinary one-sample t-tests to test the null hypothesis of equal confusion matrices for cities s and s'. Repeating this procedure for all pairs of cities leads to the p-values reported in Table 2. On a significant level of 0.05, we can assume that the confusion matrices are different for 11 of 34 city pairs. Returning to the exemplary matrices in Figure 8, we conclude that Θ_B is significantly different from Θ_L, but not from Θ_M. In other words, we can demonstrate that misclassification probabilities differ in the different cities.

It should be noted that we omitted estimation variability in the derivation of the p-values above. In other words, we considered the city-specific confusion matrices as fixed. This is to some extent plausible since the sample size is rather large and hence the estimation variance of Θ is small. To confirm this, we run a few outer bootstrap loops, which are described and reported in the online supplementary material.

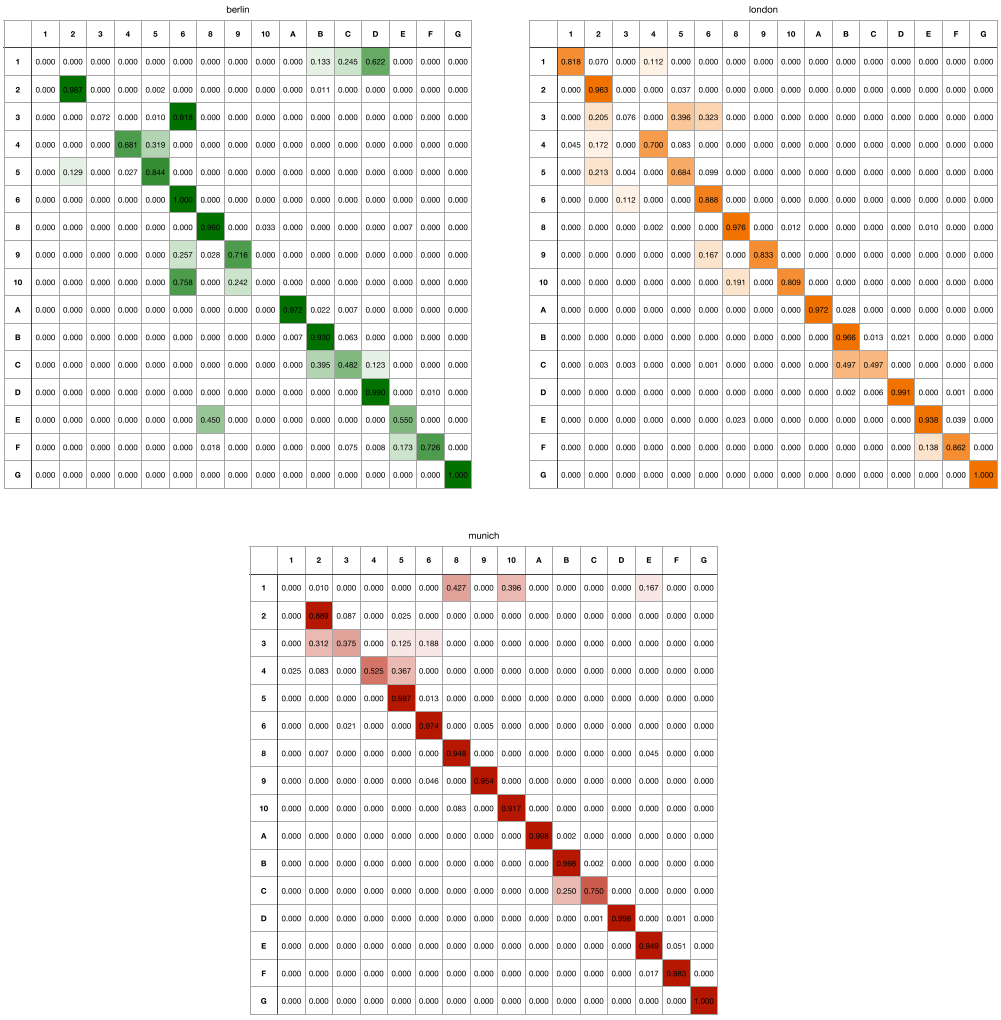


Figure 8. The matrices shows the true confusion of the voted local climate zones (columns) with the true classes (rows), for Berlin (top left matrix), London (top right matrix), and Munich (bottom matrix).

Table 2. *p*-Values of the pairwise tests for differences between cities, with (*) denoting significance on a level of 0.05

	Cologne	London	Madrid	Milan	Munich	Paris	Rome	Zurich
Berlin	0.760	0.011 (*)	0.429	0.136	0.677	0.809	0.390	0.003 (*)
Cologne		0.006 (*)	0.945	0.237	0.000 (*)	0.837	0.346	0.001 (*)
London			0.001 (*)	0.103	0.506	0.015 (*)	0.208	0.397
Madrid				0.065	0.106	0.938	0.470	0.004 (*)
Milan					0.415	0.021 (*)	0.414	0.008 (*)
Munich						0.112	0.902	0.740
Paris							0.213	0.001 (*)
Rome								0.116

5 Discussion

The paper demonstrates that the labelling of images is subject to error, misclassification, and heterogeneity of labellers. The results are relevant for all machine learning applications where image classification is pursued on multiply labelled data. Ambiguity is inevitable and the current paper aims to quantify this. It is important to note that error and uncertainty in the labelling process might stem from different sources and is multi-dimensional, as we showed in Section 4. In the context of classifying satellite images into climate zones, we were able to detect three primary sources of label uncertainty. These can be analysed based on the assumption that a latent ground truth label exists, on which the labellers condition their assessment.

First, the distinguishability of classes is not equal. This aspect is crucial not only for the earth observation domain but also relevant for most applications of image classification, be it medical image or face recognition. On the one hand, our analysis confirms that urban classes are much more challenging to identify than non-urban classes. On the other hand, the identifiability of classes also depends strongly on the database. Therefore, balanced classes are desirable and could improve and stabilise the labelling procedure.

Second, a fundamental aspect is labellers' heterogeneity and voting behaviour. Here, special training of the labellers was required as the classification of satellite images is a non-trivial task, even for earth observation experts. As the experts received the same training, one would expect homogeneity. However, we demonstrated an approach to assess homogeneity and found differences between the labellers. These can play a huge role, particularly if the panel of human labellers differ between the images, a problem not occurring in our data by design of the labelling process. Generally, labeller heterogeneity should not be neglected in the analysis of uncertainty.

Third, external properties of the instances to be classified can impact the labelling accuracy and therefore increase label uncertainty. In our case, the origin city for each image impacted the classification probabilities. While we only analysed this aspect by using separately estimated parameters, one could also include the variable in the model and inspect its impact. We have dispensed with this here, as the resulting model would have required estimating a huge number of parameters. Nevertheless, we note that incorporating external knowledge about the images could presumably lead to improved results. As already indicated in Section 4.3, the cause of the observed differences can not only be the voting behaviour of the experts but also arise due to the pursued strategy of selecting images and polygons for labelling. Imbalanced classes might induce a bias and increase the variance in the results which should be taken into account. This is an important aspect and leads to a number of new questions, related to the topics of survey methodology, sampling theory, and active learning, see, e.g. Settles (2009) or more recently Budd et al. (2021).

As a next step, it would be helpful to include the uncertainty in the machine learning process as well. The labelling process only serves as a preprocessing step for the data at hand and produces labelled training data. This data has been used for building elaborate models and networks to classify satellite images into LCZ automatically. Therefore, if the training data is flawed or suffers from high label uncertainty already, the same holds for the subsequent models. The results obtained by analysing the sources of labelling uncertainty and being able to quantify them could create possibilities to improve and stabilise machine learning processes in terms of overall uncertainty, a topic apparently beyond the scope of this paper.

Acknowledgments

K.H. was supported by the Helmholtz Association under the joint research school 'Munich School for Data Science - MUDS' (Award Number HIDSS-0006).

Conflict of interest: None declared.

Data availability

The data that support the findings of this study are available from TU Munich, Department of Data Science in Earth Observation. The exact data set is a version of the openly available data

set So2Sat LCZ42 (doi:10.14459/2018MP1454690) and will be made available publically as part of an accepted and soon-to-be-published paper by TU Munich.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

Appendix A. EM algorithm

We look at the (artificial) complete likelihood, resulting when $Z^{(i)}$ is known, i.e. the true image class is given. In this case, the complete log-likelihood results by assuming independence among the images and voters as

$$\begin{aligned}\log p(Y, Z; \boldsymbol{\pi}, \Theta) &= \sum_{i=1}^{n} \log p(\boldsymbol{Y}^{(i)}, \boldsymbol{Z}^{(i)}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{Z}^{(i)}, \boldsymbol{\pi}) + \sum_{i=1}^{n} \log p(\boldsymbol{Y}^{(i)} \mid \boldsymbol{Z}^{(i)}; \boldsymbol{\theta}_{Z^{(i)}}) \\ &= \sum_{i=1}^{n} \log \pi_{Z^{(i)}} + \sum_{i=1}^{n} \sum_{k=1}^{K} Y_k^{(i)} \log \theta_{Z^{(i)}k} + \text{parameter-free terms} \\ &= \sum_{i=1}^{n} \sum_{l=1}^{K} \mathbb{1}_{\{Z^{(i)}=l\}} \log \pi_l + \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \mathbb{1}_{\{Z^{(i)}=l\}} Y_k^{(i)} \log \theta_{lk} + \text{parameter-free terms}.\end{aligned}$$

This is a fairly simple model and could easily be estimated by maximum likelihood leading to the estimates

$$\hat{\pi}_l = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Z^{(i)} = l\}; \quad \hat{\theta}_{lk} = \frac{\sum_{i=1}^{n} \mathbb{1}\{Z^{(i)} = l\} \mathbb{1}\{Y^{(i)} = k\}}{\sum_{i=1}^{n} \mathbb{1}\{Z^{(i)} = l\}}.$$

Apparently, the likelihood above assumes that the true image class is given in the data. This is not the case, which brings us to the popular estimation strategy of the EM algorithm, described below.

Appendix B. Label switching in SEM

Put into an algorithmic format, we get the following procedure:

1. For $L = \{L_1, \ldots, L_K\}$ and $C = \{C_1, \ldots, C_K\}$:
   (a) Sort L in descending order according to the relative frequency of the labellers' votes.
   (b) Apply the permutation to all $l = 1, \ldots, K$, s.t.: $\sigma^{-1}(k) = \arg\max_l(P(Z^{(i)} = l \mid V^{(i)} = k))$.
2. If a relation $k \to l$ found in step (b) is unique, delete C_k and L_l from C and L.
3. Repeat until a unique allocation $C \to L$ is found.

Appendix C. Location of images and polygons

As mentioned in Section 4.3, the observed differences between the cities might not be caused solely by different voting behaviours. The pursued sampling strategy and selection of images and polygons possibly lead to imbalanced labels and therefore bias in the observed results. Looking at the coordinates of the images and their distribution across the cities, as shown in Figure C1 supports this hypothesis.

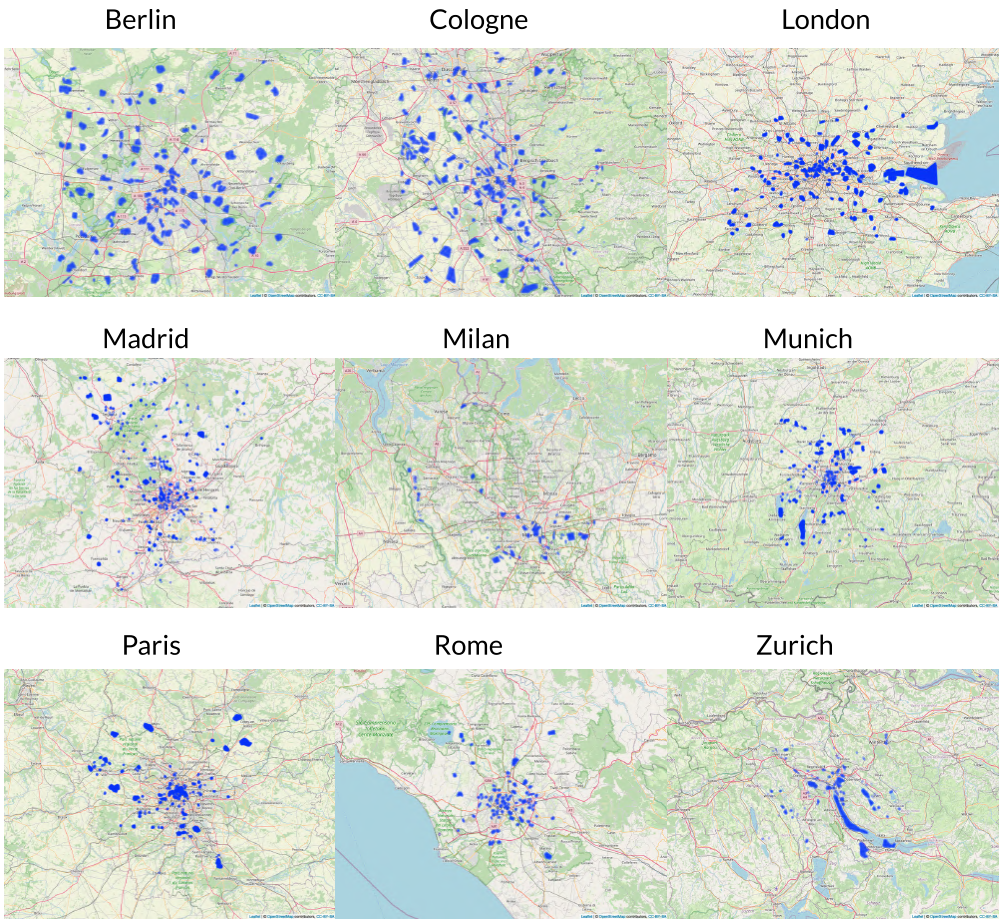


Figure C1. Maps of all cities with the coordinates of the images indicated as small dots forming clusters.

References

Budd S., Robinson E. C., & Kainz B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, *71*, 102062. https://doi.org/10.1016/j.media.2021.102062

Cadez I. V., Smyth P., Ip E., & Mannila H. (2001). *Predictive profiles for transaction data using finite mixture models* (Technical Report No. 01-67). Information and Computer Science Department, University of California, Irvine, CA.

Celeux G., Chauveau D., & Diebolt J. (1996). Stochastic versions of the EM algorithm: An experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, *55*(4), 287–314. https://doi.org/10.1080/00949659608811772

Chang J. C., Amershi S., & Kamar E. (2017). Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 2334–2346). https://doi.org/10.1145/3025453.3026044

Dawid A. P., & Skene A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28*(1), 20–28. https://dx.doi.org/10.2307/2346806

Dempster A. P., Laird N. M., & Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dgani Y., Greenspan H., & Goldberger J. (2018). Training a neural network based on unreliable human annotation of medical images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (pp. 39–42). IEEE.

Estellés-Arolas E., & González-Ladrón-de Guevara F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, *38*(2), 189–200. https://doi.org/10.1177/0165551512437638
Fraley C., & Raftery A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611–631. https://doi.org/10.1198/016214502760047131
Frenay B., & Verleysen M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 845–869. https://doi.org/10.1109/TNNLS.2013.2292894
Friedman J., Hastie T., & Tibshirani R. (2001). *The elements of statistical learning*. Series in Statistics. Springer.
Gawlikowski J., Tassi C. R. N., Ali M., Lee J., Humt M., Feng J., Kruspe A., Triebel R., Jung P., Roscher R., Shahzad M., Yang W., Bamler R., & Zhu X. X. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 1–77.
Geng X. (2016). Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, *28*(7), 1734–1748. https://doi.org/10.1109/TKDE.2016.2545658
Goodman L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231. https://doi.org/10.1093/biomet/61.2.215
Hüllermeier E., & Waegeman W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3
Ju L., Wang X., Wang L., Mahapatra D., Zhao X., Harandi M., Drummond T., Liu T., & Ge Z. (2021). 'Improving medical image classification with label noise using dual-uncertainty estimation', arXiv, arXiv:2103.00528, preprint: not peer reviewed.
Kamar E., Hacker S., & Horvitz E. (2012). Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1* (AAMAS '12) (pp. 467–474). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC.
Karger D. R., Oh S., & Shah D. (2013). Efficient crowdsourcing for multi-class labeling. In *Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems* (pp. 81–92). Association for Computing Machinery (ACM).
Lazarsfeld P. F. (1950). The logical and mathematical foundation of latent structure analysis. In *Measurement and prediction*. [Studies in social psychology in World War II. Vol.4 (pp. 362–412)]. Princeton University Press.
Little R. J., & Rubin D. B. (2002). *Statistical analysis with missing data* (Vol. *793*). John Wiley & Sons.
Luo J., Wang Y., Ou Y., He B., & Li B. (2021). Neighbor-based label distribution learning to model label ambiguity for aerial scene classification. *Remote Sensing*, *13*(4), 755. https://doi.org/10.3390/rs13040755
Magidson J., Vermunt J. K., & Madura J. P. (2020). *Latent class analysis*. SAGE Publications Limited Thousand Oaks.
McLachlan G., & Peel D. (2000). *Finite mixture models*. Wiley.
McLachlan G. J., Lee S. X., & Rathnayake S. I. (2019). Finite mixture models. *Annual Review of Statistics and its Application*, *6*(1), 355–378. https://doi.org/10.1146/statistics.2019.6.issue-1
Northcutt C., Jiang L., & Chuang I. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, *70*, 1373–1411. https://doi.org/10.1613/jair.1.12125
Peterson J. C., Battleday R. M., Griffiths T. L., & Russakovsky O. (2019). Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9617–9626). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00971
Phillips P. J., Yates A. N., Hu Y., Hahn C. A., Noyes E., Jackson K., Cavazos J. G., Jeckeln G., Ranjan R., Sankaranarayanan S., Chen J.-C., Castillo C. D., Chellappa R., White D., & O'Toole A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, *115*(24), 6171–6176. https://doi.org/10.1073/pnas.1721355115
Qiu C., Mou L., Schmitt M., & Zhu X. X. (2019). Local climate zone-based urban land cover classification from multi-seasonal sentinel-2 images with a recurrent residual network. *ISPRS Journal of Photogrammetry and Remote Sensing*, *154*, 151–162. https://doi.org/10.1016/j.isprsjprs.2019.05.004
Qiu C., Schmitt M., Mou L., Ghamisi P., & Zhu X. (2018). Feature importance analysis for local climate zone classification using a residual convolutional neural network with multi-source datasets. *Remote Sensing*, *10*(10), 1572. https://doi.org/10.3390/rs10101572
Raykar V. C., & Yu S. (2011). Ranking annotators for crowdsourced labeling tasks. *Advances in Neural Information Processing Systems*, *24*, 1809–1817. https://doi.org/10.5555/2986459.2986661
Robbins H. E. (1992). An empirical Bayes approach to statistics. In: S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics* (pp. 388–394). Springer Series in Statistics. Springer. https://doi.org/10.1007/978-1-4612-0919-5_26
Rubin D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Russwurm M., Ali M., Zhu X. X., Gal Y., & Körner M. (2020). Model and data uncertainty for satellite time series forecasting with deep recurrent models. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium* (pp. 7025–7028). IEEE.

Settles B. (2009). *Active learning literature survey* (Technical Report). University of Wisconsin-Madison Department of Computer Sciences.

Stewart I., & Oke, T.R. (2012). Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, *93*(12), 1879–1900.

Zhang L., Tanno R., Xu M.-C., Jin C., Jacob J., Ciccarelli O., Barkhof F., & Alexander D. C. (2020). 'Disentangling human error from the ground truth in segmentation of medical images', arXiv, arXiv:2007.15963, preprint: not peer reviewed.

Zhu X. X. (2021). So2sat lcz42. *Dataset*. TU Munich. https://doi.org/10.1109/MGRS.2020.2964708

Zhu X. X., Hu J., Qiu C., Shi Y., Kang J., Mou L., Bagheri H., Häberle M., Hua Y., Huang R., Hughes L., Li H., Sun Y., Zhang G., Han S., Schmitt M., & Wang Y. (2020). So2sat lcz42: A benchmark dataset for global local climate zones classification. *IEEE Geoscience and Remote Sensing Magazine*, *8*(3), 76–89. https://doi.org/10.1109/MGRS.2020.2964708

Zhu X. X., Qiu C., Hu J., Shi Y., Wang Y., Schmitt M., & Taubenböck H. (2022). The urban morphology on our planet–global perspectives from space. *Remote Sensing of Environment*, *269*, 112794. https://doi.org/10.1016/j.rse.2021.112794

Zhu X. X., Tuia D., Mou L., Xia G.-S., Zhang L., Xu F., & Fraundorfer F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

6. More Labels or Cases? Assessing Label Variation in Natural Language Inference

Contributing article:

Gruber*, C., Hechinger*, K., Aßenmacher, M., Kauermann, G. and Plank, B. (2024). More Labels or Cases? Assessing Label Variation in Natural Language Inference. *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language* https://aclanthology.org/2024.unimplicit-1.2.pdf

Copyright information:

Code and data:

The supplementary code is publicly available on GitHub: https://github.com/corneliagru/label-variation-nli.git.

Author contributions:

The initial idea of modeling annotation uncertainty in the context of Natural Language Inference using a multinomial mixture model stems from Cornelia Gruber, Göran Kauermann and Barbara Plank. The final model was developed by Göran Kauermann and Katharina Hechinger following the procedure outlined in the first contributing article. Katharina Hechinger devised the bootstrap approach to evaluate the model's stability and implemented the initial version of the model in `Python`. Cornelia Gruber developed the visual representation of the results, cleaned the code, and set up the final GitHub repository. The initial manuscript draft was collaboratively composed by Katharina Hechinger and Cornelia Gruber, with Katharina Hechinger primarily responsible for Sections 4, 5.2, and 6. Matthias Aßenmacher and Barbara Plank contributed valuable insights and expertise from the NLP domain. All authors were involved in improving and proofreading the manuscript.

More Labels or Cases?
Assessing Label Variation in Natural Language Inference

Cornelia Gruber[*1♠] **Katharina Hechinger**[*1♠] **Matthias Aßenmacher**[1,2♠]
Göran Kauermann[1♠] **Barbara Plank**[2,3♣]
[1] Department of Statistics, LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML), Munich, Germany
[3] Center for Information and Language Processing (CIS), LMU Munich, Germany
♠{cornelia.gruber, katharina.hechinger, matthias, goeran.kauermann}@stat.uni-muenchen.de
♣bplank@cis.uni-muenchen.de

Abstract

In this work, we analyze the uncertainty that is inherently present in the labels used for supervised machine learning in natural language inference (NLI). In cases where multiple annotations per instance are available, neither the majority vote nor the frequency of individual class votes is a trustworthy representation of the labeling uncertainty. We propose modeling the votes via a Bayesian mixture model to recover the data-generating process, i.e., the posterior distribution of the "true" latent classes, and thus gain insight into the class variations. This will enable a better understanding of the confusion happening during the annotation process. We also assess the stability of the proposed estimation procedure by systematically varying the numbers of i) instances and ii) labels. Thereby, we observe that few instances with many labels can predict the latent class borders reasonably well, while the estimation fails for many instances with only a few labels. This leads us to conclude that multiple labels are a crucial building block for properly analyzing label uncertainty.

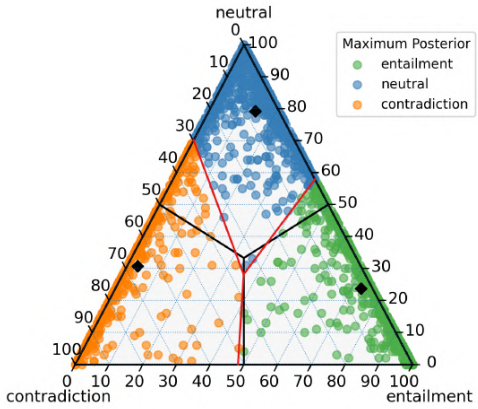


Figure 1: Scatter plot of the vote distribution of ChaosSNLI. Each point represents one instance. Its location is determined by the vote distribution. Corner points represent 100 votes for the respective class, i.e., *entailment*, *neutral*, *contradiction* for the bottom right, top, and bottom left, respectively. Solid black lines represent the border of class membership by majority vote. The color of the points is determined by the estimated latent class given by our model. Black diamonds describe the center points of the latent classes. Solid red lines represent the borders of latent class membership.

1 Introduction

Commonly, binary or multi-class classification settings in machine learning assume that a single gold label—representing the "true" class of an instance—can easily be acquired via human annotation. However, there are numerous examples where remarkable variations between different annotators exist, challenging the validity of this assumption (Uma et al., 2021). This issue is especially prevalent in datasets relating to the difficult task of perceiving human language, such as natural language inference (NLI). In NLI, the textual entailment of two sentences is to be determined. There exists an increasing body of work documenting inherent disagreement in labeling for NLI (Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Zhang and de Marneffe, 2021; Jiang et al., 2023). Such human label variation can be caused by context dependency and subjectivity, amongst others, and is ubiquitous (Plank, 2022). Moreover, human label variation is different from annotation errors, as plausible, linguistic reasons for such variation exist (Jiang and de Marneffe, 2022).

To provide new grounds to study human variation in labeling, Nie et al. (2020) collected the ChaosNLI (Collective HumAn OpinionS on Natural Language Inference) dataset. ChaosNLI comprises 100 labels per instance from quality-controlled annotators for each of the ambiguous

instances from multiple NLI-related datasets. In this paper, we analyze ChaosSNLI, a sub-dataset of ChaosNLI based on the Stanford Natural Language Inference (SNLI) data (Bowman et al., 2015). Several works on NLI (Pavlick and Kwiatkowski, 2019; Nie et al., 2020) show that many instances exhibit high human disagreement or uncertainty, i.e., human labelers do not agree on a single class, resulting in a high spread of the annotators' votes among multiple classes. Less work has looked at label variation and stability from a data-generating process viewpoint in light of uncertainty.

Uncertainty in machine learning and NLP is, however, gaining increased attention recently (Hüllermeier and Waegeman, 2021; Gruber et al., 2023; Baan et al., 2023). Different lines of research study sources of uncertainty in various parts of machine learning, such as the data itself, the model choice, the estimation procedure, and model deployment (Gruber et al., 2023). Early works characterize uncertainty in terms of reducible and irreducible randomness (Hüllermeier and Waegeman, 2021), while some works argue that this line is fuzzy (Gruber et al., 2023; Baan et al., 2023).

Variation in labels is part of the uncertainty in the data and is ubiquitous given the inherent ambiguity of language (Zhang et al., 2021). Yet, understanding the uncertainty in labels enables us to not only empirically investigate human confusion in annotated data, but also to gain insights on the classification task itself. For example, the complexity of detecting certain classes or the composition of class structures can be derived from voting patterns—this information can provide useful insights into task characteristics.

Therefore, in order to analyze the uncertainty in the label vote distribution of ChaosSNLI, we model the data-generating process and analyze the stability of the resulting estimation. To do so, we employ a Bayesian mixture model and recover the latent "true" class label, see also Hechinger et al. (2024). More precisely, we obtain the posterior probability for each of the classes and can thus assess the certainty for the class labels given the votes.

Our results could further be incorporated into a machine learning pipeline, e.g., by fitting a model on our latent classes instead of majority vote classes or class frequencies. This is, however, beyond the scope of this paper. In this work, we focus on the fundamental step of quantifying labeling uncertainty instead. We propose an estimation procedure and analyze its *stability* for different amounts of i) instances and ii) labels. Our work shows that more labels are more beneficial for stable estimation of uncertainty, while only a few instances already suffice. We also suggest new tools for *visual assessment* of the uncertainty in labels for three-way classification tasks (see Fig. 1).

Contributions With this paper, we contribute to a better understanding of label variation via a deep assessment of trustworthiness by 1) quantifying labeling uncertainty with Bayesian mixture models, 2) providing a novel visual tool for a better assessment of labeling uncertainty, and 3) deriving practical guidance for labeling tasks. We identify the benefit of using fewer cases with many labels rather than the other way around.[1]

2 Related Work

The need to analyze diverse human opinions in natural language inference is discussed by works including Pavlick and Kwiatkowski (2019) and Nie et al. (2020). Nie et al. (2020) show that some state-of-the-art models (including BERT, RoBERTa, XLNET, AL- BERT, DistilBERT, and BART) are neither designed nor able to capture human variation in labels and are therefore not appropriate. Their work also states that predicting the majority vote and predicting the human label distribution are distinct and seemingly conflicting objectives. In their benchmark study, all considered models performed consistently worse on examples with low human agreement. This indicates that analyzing label variation is of significant relevance for a more complete understanding of natural language inference.

Hovy et al. (2013) already advocated that majority voting might be the simplest but not most appropriate strategy for finding the correct label and, that modeling the votes leads to improved predicted label accuracy. The authors propose a method to separately model annotations from spamming and non-spamming annotators. Our methods differ in the way variation in labels is modeled. Hovy et al. (2013) explicitly model the behavior of annotators and assumes non-spamming annotators always provide the correct label, while votes by spamming annotators are drawn from a multinomial distribution. In contrast, our approach models human confusion in the annotation process, assuming equal levels of annotation skills. This is a reasonable assumption

[1]Code and data available at: https://github.com/corneliagru/label-variation-nli

for ChaosSNLI as all annotators undergo strict quality control, see Nie et al. (2020) for details. Nevertheless, both methods share the goal of estimating the distribution of the data-generating process and its parameters via an expectation-maximization (EM) algorithm.

Paun et al. (2018) compare various Bayesian approaches for modeling annotation. Based on their taxonomy, we employ a pooled model, i.e., assuming equal quality of the annotators. They conclude that such pooled models underperform, as the assumption that all annotators share the same ability is inappropriate in typical crowdsourcing settings. However, when information on individual annotators is unavailable, as is the case for the investigated ChaosSNLI dataset, pooling is inevitable.

The benefits of harnessing multiple labels are presented in Zhang et al. (2021). They demonstrate that improvements in accuracy can be achieved by varying the number of annotations for some examples within a given annotation budget. Our findings show a more nuanced picture supporting their claims, as we show the necessity of multiple annotations but a flattening value curve (see section 5).

3 Dataset and Problem Setting

We examine label uncertainty in NLI, a task for which textual entailment of two sentences is typically classified as either *entailment*, *neutral*, or *contradiction*. In ChaosSNLI (Nie et al., 2020), *multiple* annotations for *each* instance are provided. Example sentences of ChaosSNLI with their respective votes are shown in Table 1. Since those annotators do not necessarily agree with each other, we face a high degree of (human) label uncertainty. We chose this dataset as it provides a unique ground to explore label variation. Having access to a high amount of labels per instance is particularly valuable, but unfortunately not a common setting.

Our analysis is based on $N = 1,514$ instances with $J = 100$ labels, each, that originate from the development set of the SNLI dataset (Bowman et al., 2015). The original SNLI development set was generated by a multistep procedure, where first an initial annotator provides a text description of an image, i.e., generating the *premise*. Second, a different annotator constructs three *hypotheses* as an entailing, neutral, and contradicting description of the premise. Third, four more annotators, independent of the first two steps, provide labels for the premise-hypothesis pairs, i.e., classify the pairs into *entailment, neutral* or *contradiction*. This procedure yields five annotations per instance in total. In ChaosSNLI, examples, where only three out of those five annotators agree, are then relabeled by 100 quality-controlled annotators. For details on the quality control procedure, we refer to Nie et al. (2020). This relabeling procedure leads to a dataset, where instances with a high degree of uncertainty are overrepresented. Such a biased sample is valuable, as our main interest lies in understanding exactly those uncertain and hard-to-classify cases.

In the dataset, we observe that the most common class according to majority voting is *neutral*, with 53.7% of all examples, while *entailment* and *contradiction* amount to 27.8% and 18.5%, respectively. This already suggests that identifying *neutral* seems to be more challenging than discerning the other classes, as human annotators do not agree on those especially challenging examples that were collected for ChaosSNLI.

To gain a better understanding of label uncertainty in NLI, we analyze the annotations for the premise-hypothesis pairs available in ChaosSNLI. In order to detect hidden structures and comprehend label variation, we follow a statistical approach for modeling the label distribution. It is thus distinct from classical machine learning, where models are optimized for predictive power. However, our approach can ultimately be incorporated as a preprocessing step for predictive models. A precise description of our methodology can be found in section 4.

4 Modeling Approach

The main goal of this work is to explore the uncertainty inherent in the (multiple) labels of the sentence pairs in ChaosSNLI which is expressed by the distribution of the annotations. In order to formally describe the dataset with its multiple annotations and to assess label uncertainty, we use tools from statistical modeling. The multinomial mixture model provides the possibility to put multiple annotations into a distributional framework and subsequently estimate the associated parameters. Based on these parameters, a latent ground truth label can be derived for each instance, incorporating the uncertainty expressed by the distributions of the annotations over all instances. We follow the methodology proposed in Hechinger et al. (2024) for modeling multiple annotations via a Bayesian

Context/Premise	Statement/Hypothesis	[E, N, C]
A boy in an orange shirt sells fruit from a street cart.	A boy is a street vendor.	[90, 10, 0]
A woman wearing a red hat and black coat.	The woman is asleep.	[0, 87, 13]
People walk amongst a traffic jam in a crowded city.	The cars are zooming past the people.	[3, 15, 82]
A woman holding a child in a purple shirt.	The woman is asleep at home.	[1, 53, 46]

Table 1: Examples of ChaosSNLI. Annotators answered the question: "Given a context, a statement can be either: definitely correct (Entailment); or definitely incorrect (Contradiction); or neither (Neutral). Your goal is to choose the correct category for a given pair of context and statement."

mixture model.

First, let us introduce a formal description of the data. Each instance is a pair of $(X^{(i)}, \boldsymbol{Y}^{(i)}), i = 1, \ldots, N$, where $X^{(i)}$ denotes the sentence pair of premise and hypothesis and $\boldsymbol{Y}^{(i)}$ denotes the corresponding vote distribution. For this work, our focus lies on the latter exclusively, i.e., we only consider the vector of annotations for each instance. To explicitly represent votes for K possible classes by J different annotators, $\boldsymbol{Y}^{(i)}$ is set to $\boldsymbol{Y}^{(i)} = (Y_1^{(i)}, \ldots, Y_K^{(i)})$ with $Y_k^{(i)} = \sum_{j=1}^{J} \mathbb{1}(V_j^{(i)} = k)$. Here, $V_j^{(i)}$ denotes the individual vote for instance i by annotator j. In ChaosSNLI we do not have access to individual annotator-specific votes, but observe $\boldsymbol{Y}^{(i)}$ directly. As mentioned above, we model the uncertainty inherent in the labels, so we omit $X^{(i)}$ and only analyze $\boldsymbol{Y}^{(i)}$. It is worth mentioning that incorporating the actual text is still possible for downstream tasks, but is out of the scope of this work.

In order to make use of the multinomial mixture model, we assume that each instance is associated with one true label, i.e., there exists an unambiguous ground truth. However, due to the inherent uncertainty in the perception of language, annotators are not easily capable of recovering the ground truth and they might vote for different classes. We denote the latent ground truth of each instance $X^{(i)}$ with $Z^{(i)} \in \{1, \ldots, K\}$. Again, to match our notation with the definition of a multivariate variable, we define $\boldsymbol{Z}^{(i)}$ as a one-hot encoded vector indicating the latent class, i.e., $\boldsymbol{Z}^{(i)} = (\mathbb{1}\{Z^{(i)} = 1\}, \ldots, \mathbb{1}\{Z^{(i)} = K\})$.

In the context of this particular dataset, as described in section 3, there exists a clearly defined ground truth that annotators should recover. This is due to the fact, that the annotator had one specific class in mind while inventing the hypothesis. Thus, the assumption of exactly one underlying "true" label is justified. However, this methodology can be applied beyond scenarios with known ground truth. In cases where no such information is available, the distributions of votes can serve as a valuable tool for deducing the latent labels.

Model Framework Let us now proceed to the analysis of the voting distribution $\boldsymbol{Y}^{(i)}$, which carries information about the latent true labels. We employ the following Bayesian modeling framework. First, considering the ground truth labels to be unobserved (or unobservable), they are assumed to follow a multinomial distribution

$$\boldsymbol{Z}^{(i)} \sim Multi(\boldsymbol{\pi}, 1) \text{ i.i.d.},$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ denote the prior probabilities for all classes. This distribution is also called the *prior* distribution. Given the true classes, the annotations are also assumed to be distributed multinomially, i.e.,

$$\boldsymbol{Y}^{(i)} | Z^{(i)} \sim Multi(\boldsymbol{\theta}_p, J). \quad (1)$$

This multinomial distribution describes the data *likelihood* conditional on Z. Here, the parameter vector $\boldsymbol{\theta}_p$ depends on the latent true class $Z^{(i)}$, i.e., the multinomial probabilities vary based on what we consider to be the true label. Hence, this parameter describes the probability of voting for a class given the true label. We can summarize the multinomial probability vectors of each latent, true class into a matrix $\boldsymbol{\Theta} = (\theta_{pk}, p, k = 1, \ldots, K)$, which can be interpreted as a confusion matrix. Formally, θ_{pk} describes the probability of an annotator voting for class k given the instance has the true class p, i.e., using the notation in Eq. (1) we have $\boldsymbol{\theta}_p = (\theta_{p1}, \theta_{p2}, \ldots, \theta_{pK})$.

The key component of the model is the *posterior* distribution, i.e., the probabilities for an instance to truly belong to each of the classes given the observed annotations. These probabilities are cal-

culated as

$$\begin{aligned}\tau_p^{(i)} &= P(Z^{(i)} = p|\boldsymbol{Y}^{(i)}; \boldsymbol{\pi}, \boldsymbol{\Theta}) \\ &= \frac{P(Z^{(i)} = p; \boldsymbol{\pi}) P(\boldsymbol{Y}^{(i)}|Z^{(i)} = p; \boldsymbol{\Theta})}{P(\boldsymbol{Y}^{(i)}; \boldsymbol{\pi}, \boldsymbol{\Theta})} \\ &= \frac{\pi_p P(\boldsymbol{Y}^{(i)}; \boldsymbol{\theta}_p)}{\sum_{p'=1}^{K} \pi_{p'} P(\boldsymbol{Y}^{(i)}; \boldsymbol{\theta}_{p'})}.\end{aligned}$$

The class with the maximal posterior serves as an estimate for the latent ground truth, it is however also possible to use $\boldsymbol{\tau}$ in downstream tasks directly, i.e., for training a classifier on the probabilities instead of discrete class labels and thus directly incorporate the label uncertainty.

It is important to note that the prior modeling assumption of a single ground truth does not dictate the reality to be discrete, much more it enables us to compute the posterior distribution and quantify the evidence for each class, given the vote distribution. It thus allows us to model settings with ambiguous labels.

Estimation Procedure The model above includes unknown parameters, which we suggest estimating through maximum likelihood. As we are in the latent variable framework, straightforward estimation of the model parameters via maximum likelihood is, however, not possible. Instead, we apply an iterative estimation procedure to obtain parameter estimates. With the help of the expectation-maximization (EM) algorithm as introduced by Dempster et al. (1977), we can replace the latent class label $\boldsymbol{Z}^{(i)}$ with its expectation for each voting distribution. The expected latent class is thereby calculated given the data and the current parameter estimates and can be used afterward to update the estimates, leading to an iterative procedure that is performed until convergence. The algorithm can be outlined as follows, with additional details available in Hechinger et al. (2024) and in Appendix A. For the current parameter values at estimation iteration (t), $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{(t)}$ and $\boldsymbol{\pi} = \boldsymbol{\pi}_{(t)}$, one iterates over the two steps:

1. **E-Step:** Calculate the expectation of the full data likelihood given the data and the current estimates. Applying Bayes' rule, this simplifies to the computation of the expected latent class, given by posterior probabilities $\tau_p^{(i)}, i = 1, ..., N$ and $p = 1, ..., K$.

2. **M-Step:** Update the parameters $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{(t+1)}$ and $\boldsymbol{\pi} = \boldsymbol{\pi}_{(t+1)}$ based on the posterior $\boldsymbol{\tau}$.

The final estimates are denoted as $\hat{\boldsymbol{\Theta}}$ and $\hat{\boldsymbol{\pi}}$. Our modeling approach harnesses the information retrieved from the annotations from all instances, as in every EM-step all instances are used for recalculating the estimates. This enables our method to incorporate knowledge about all annotation uncertainties and provide a comprehensive and holistic view of label variation.

Label Switching The classes obtained through mixture models are subject to label switching, i.e., their numbering is arbitrary and does not correspond to the original order anymore. This is a common issue in mixture models and can be resolved in various ways depending on the specific application at hand, as outlined by Stephens (2000). In this case, we apply a simple heuristic permutation to the latent classes. The original classes *entailment, contradiction, neutral*, denoted with index $k = 1, 2, 3$, are assigned to the respective latent classes $p = 1, 2, 3$ based on the diagonal entries of the estimated confusion matrix $\boldsymbol{\Theta}$. E.g., the class *entailment* is assigned to the mixture component, where the highest voting probability is *entailment*. This corresponds to the permutation $\sigma^{-1}(p) = \arg\max_k(\hat{\boldsymbol{\theta}}_p)$ and the latent classes are re-ordered accordingly.

To summarize, by allowing for human uncertainty, i.e., human confusion while labeling a certain instance, we can recover information on a latent class Z. The posterior distribution of the latent class is then a more trustworthy representation of the "true" class an instance belongs to, since all information contained in the full dataset is used for estimation, and not only the specific label distribution.

5 Results

5.1 Introspection by Visualization

As described earlier, the dataset ChaosSNLI (Nie et al., 2020) consists of $J = 100$ annotations for $K = 3$ classes. We propose to analyze human label variation in NLI with a novel visualization tool, to help gain insights into labeling. Figure 1 illustrates the distribution of votes present in ChaosSNLI, which we then contrast to the majority vote and our model's estimated class membership votes.

Each point in Figure 1 represents one instance, where its location is determined by the empirical distribution of votes. It is clearly visible by the density of dots that most instances cluster around the top of the plot, i.e., with many votes for *neutral*.

This is consistent with the distribution of majority votes (with *neutral* being observed 53.7% of times, as discussed in section 3). Furthermore, we observe that there is little confusion between *contradiction* and *entailment*, as almost no points lie close to the lower horizontal line or the vertical line starting in the center. This observation is intuitively plausible, due to the contrasting nature of the two labels of *entailment* vs. *contradiction*. Interestingly, this visualization tool helps us to quickly identify that there are cases in the datasets where many labels for both *entailment* and *contradiction* were observed.

In order to analyze our modeling result in relation to majority voting, we examine the borders between the three classes. Figure 1 shows the borders of the majority voting as solid black lines, which connect the center points of the axes, i.e., 50:50 votes for two of the classes, to the center, i.e., 33.33 votes for all three classes.

The borders between the latent classes are shown as red lines. To calculate these borders between two latent classes, we determine the vote combinations that lead to equal posterior probabilities. That is, we calculate the specific vote distribution $\boldsymbol{Y}^{(i)}$ such that $\tau_k = \tau_j$ for two classes $k, j \in \{1, 2, 3\}, k \neq j$, while there are no votes for the third class. This gives us the critical points lying on the axis connecting classes k and j. For the middle point, i.e., the connection between all three classes, the equation $\tau_1 = \tau_2 = \tau_3$ is solved for the corresponding vote distribution. This results in four critical points. By connecting the points on the axes to the center, we obtain the new borders of the latent classes, which are now based on posterior probability estimates and not just on the empirical distribution of the votes for one instance. In other words, they are estimated by taking all data into account. The exact border points are described in Appendix A.

In Figure 1, for all instances that lie between the black and red borders, the latent class label does not agree with the majority vote. It is especially evident that the latent class *neutral* comprises a smaller fraction of vote distributions than it would have by majority voting (black line). More precisely, considering all cases with a majority vote for *neutral*, our model agrees for 83.3%, however, *entailment* is estimated for 6.9% of cases and *contradiction* for the remaining 9.8%, i.e., 16.7% of the majority vote *neutral* are assigned a different label by our model. This is however desirable, as many votes for one of the more informative classes (*entailment* or *contradiction*) strongly speak for exactly those classes, even if there is no majority. For example, having 40 votes for *contradiction*, 60 for *neutral*, and none for *entailment*, indicates that *entailment* is unlikely. Likewise, if *neutral* would be the "true" latent class, at least some votes for entailment are expected. Thus, in this setting, a latent *contradiction* is most probable. Analogous reasoning can be applied for instances with many votes for *entailment*, without *entailment* as the majority. Further, we argue that negative votes by the annotators can be regarded as a stronger signal for the instance actually being *contradiction* as fewer of them are required for our model to assign the label *contradiction*, compared to *entailment*. This becomes evident from Figure 1 as the red border between *neutral* and *contradiction* is much closer to the *neutral* corner compared to its counterpart between *neutral* and *entailment*.

To summarize, the model especially refines the class *neutral* and alleviates the issue that the majority class *neutral* does not only contain true neutral statements, but might also be conflated with examples where the annotators were indecisive or had conflicting interpretations (Nighojkar et al., 2023).

5.2 Stability Analysis

Having provided a visualization tool that allows valuable insights into the dataset, we are now interested in the *stability* of the modeling procedure. One common approach to assess the estimation uncertainty and stability of the resulting parameter estimates is to employ a resampling method, like bootstrapping (Efron, 1979). We therefore analyze the stability of the estimation procedure in relation to three aspects:

1. overall stability,
2. stability in the number of instances, N,
3. stability in the number of labels, J.

Overall stability In order to assess the uncertainty of the estimation procedure itself, we employ a classical bootstrap. That is, we sample from the data with replacement[2] and subsequently estimate the model parameters. Repeating this multiple times allows us to assess how the estimation would change if we had different datasets coming from the same distribution as the initial one.

[2] i.e., the same instance can be present multiple times, while other instances might not be included at all.

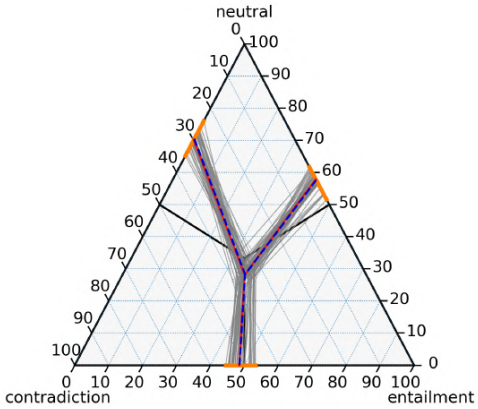


Figure 2: The ternary plot contains the decision borders between the three classes calculated based on $B = 50$ bootstrapped estimates as gray lines. The range of the gray lines is outlined in orange. The blue dashed line indicates the mean of the bootstrapped versions and the red line shows the original borders for comparison.

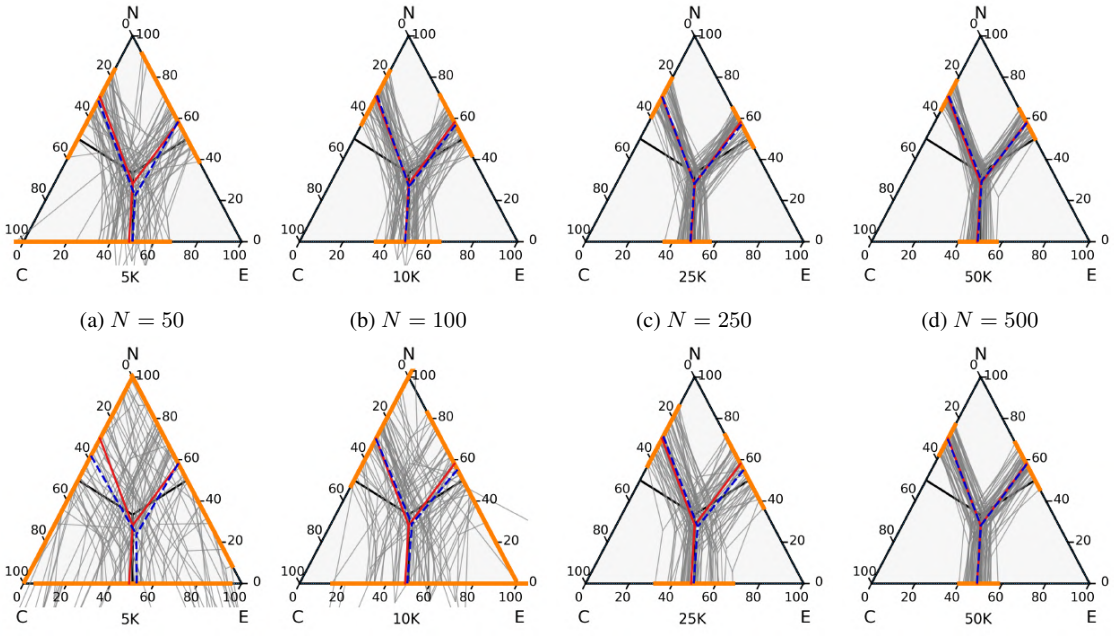
We run B bootstrap iterations, producing bootstrapped versions of the parameter estimates $\boldsymbol{\pi}$ and $\boldsymbol{\Theta}$. Based on these values, the borders of the latent classes can be recalculated B times. Figure 2 shows the estimated borders for $B = 50$ bootstrap replicates in gray alongside the borders computed based on the full dataset in red (cf. Fig. 1). This leads us to conclude that the estimation of the parameters and, therefore, the latent classes is stable for the full dataset. Due to the high number of instances in the dataset, this result is not surprising. However, the question arises whether stable estimation is also possible with a smaller dataset. Reducing N, the number of multiple annotated instances on the one hand, and reducing J, the number of annotations for the instances on the other hand, could lead to substantially reduced labeling effort. Hence, these aspects will be analyzed in the following.

Stability of Number of Instances In many real-world applications, the number of instances that can be annotated multiple times is often limited to a couple of hundred instances (as an example, the earlier multi-annotated NLI dataset from Pavlick and Kwiatkowski (2019) contained five annotations for less than 500 instances as available in ChaosNLI). Therefore, it is worthwhile to examine the stability of the estimation procedure and the resulting estimates for a smaller dataset in terms of sample size (less than 1.5k instances). Specifically, we are interested in the location of the decision borders regarding the latent classes and their stability for fewer instances.

Therefore, we employ a bootstrap again but this time randomly sample smaller datasets, i.e., $N < 1,514$ with replacement to artificially reduce the sample size. Figure 3 shows $B = 50$ bootstrapped borders of the latent classes for various numbers of samples N with fixed $J = 100$. While the bootstrapped borders still show quite some variation for very small sample sizes (e.g., $N = 50$), the average of all bootstrapped borders already aligns quite well with the original borders. For a sample size of $N = 100$, the variation has already decreased noticeably, and for even larger samples, like $N = 500$, which is only one-third of the original sample size, almost no differences to the original results are visible. Hence, we conclude that reducing the sample size leads to reasonably good and stable estimation results if a certain minimum of instances is kept.

Stability of Number of Labels While this work focuses on the ChaosSNLI dataset with $J = 100$ annotations, the original SNLI development dataset only contains five labels per instance. In practice, annotating instances many times is costly and might seem inefficient. Hence, we are also interested in the stability of the estimation procedure in terms of the number of labels as well as the *minimal* number of labels needed per instance for stable parameter estimates.

Again, we draw bootstrap samples from the original dataset. This time, the sample size is kept constant at $N = 1000$ but the number of annotations per sample is reduced. Therefore, we randomly choose $J < 100$ annotations from the original ones. The resulting bootstrapped borders are shown in Figure 3. As expected, only using $J = 5$ annotations leads to large variations and unstable results. For $J = 25$ annotations, the procedure is already quite stable. For more than $J = 50$ annotations, the results show diminishing returns: they depict similar behavior to the original ones with the double amount of J, i.e., $J = 100$ (see Fig. 2). Therefore, we note that acquiring a smaller number of labels for each instance is possible, but a sufficient amount of annotations is needed for stable estimation. Particularly, the number of annotations seems to be more crucial for the stability of the results than the sample size. Additional results for simultaneously varying the amount of N and J that further support this finding can be found in Figure 4, Appendix A.



(a) $N = 50$ (b) $N = 100$ (c) $N = 250$ (d) $N = 500$

(e) $J = 5$ (f) $J = 10$ (g) $J = 25$ (h) $J = 50$

Figure 3: The ternary plots show the bootstrapped latent class borders as gray lines, the range of the gray lines in orange, the mean of the bootstrapped as blue dashed lines, and for comparison, the original borders as red lines for various sample sizes and annotations. In the top row J is set to $J = 100$ and $N \in \{50, 100, 250, 500\}$. In the bottom row, we set $N = 1000$ and $J \in \{5, 10, 25, 50\}$. The total number of annotations, i.e., $N \cdot J$, is below each plot.

6 Discussion

Reliable and correct labels are crucial for classification models. While it is common practice to gather multiple annotations to ensure high-quality labels, these are often summarized into one single final label via a majority vote (Paun et al., 2018). However, this strategy leads to a major loss of information and uncertain ground truth labels in applications where a high degree of label variation is present. The statistical approach pursued in this work offers the possibility to condense information, given in multiple labels through the whole dataset, into a single ground truth label. To evaluate the results, we compared the borders between the classes, i.e., we examined the voting combinations where the ground truth label changes for an instance. By choosing the estimated latent ground truth instead of the majority vote, these borders shifted reasonably, from a semantic perspective.

Additionally, we showed that the parameters of the model and, hence, the borders can be estimated reliably based on the available instances and annotations. However, in many realistic applications, the data basis might be smaller in terms of both aspects. Hence, we also conducted a stability analysis for random subsets of the number of instances (N) and the number of votes per instance (J) of the dataset. The results show that stable estimation is already possible for a smaller dataset and that human labeling effort can be decreased, without loss of information. The quantity of accessible labels proves to be more important for ensuring a stable model performance than the sample size. We assume that this is because the annotations bear the majority of the inherent uncertainty. Therefore, acquiring multiple labels, particularly for uncertain instances, i.e., instances where label variation is expected, is advisable.

While the results and decision borders obtained via the proposed model in this work showcase the problem of label uncertainty, future directions of research could include the incorporation of this information into the ML pipeline or the development of a quantitative measure for label uncertainty. This could then lead to a detailed strategy for acquiring labels efficiently. Though these questions are highly relevant and should be tackled in the future, they are beyond the scope of the current work.

7 Conclusion

In conclusion, by analyzing ChaosSNLI we showcase the suitability of Bayesian mixture models to recover the true data-generating process of annotation tasks with access to multiple labels. Our work provides a framework to deal with multi-annotation settings in classification and is applicable regardless of the underlying task, i.e., NLI. Furthermore, our results suggest that in the annotation process, the focus should lie on increasing the number of labels per instance, instead of more instances in total, as this promotes capturing the labeling uncertainty.

Limitations

Our proposed method analyzes uncertainty in labels for a three-way classification task. However, since the concept of *uncertainty* is by definition vague and fuzzy, it is important to determine which aspects of uncertainty *should be* or *can be* specified. In our work, we focus on modeling the annotation process. If other aspects of uncertainty are of relevance, our method might not be the most appropriate anymore. This points to the individuality of dealing with uncertainty and that no one-fits-all approach exists.

Further limitations might arise upon the application of the model to other datasets. 1) Multiple annotations per instance are needed. 2) Visual assessment of class memberships (c.f. Fig 1) or the stability of class borders (c.f. Fig 3) works reasonably well for up to three classes. Analyzing datasets with labels of higher dimensions is straightforward, as shown by Hechinger et al. (2024) for the classification of ambiguous images. However, assessing the stability of class borders needs to be done quantitatively, e.g., by computing confidence intervals of the bootstrapped borders. 3) In case annotator IDs are available, we recommend extending our approach in order to incorporate all available information. This could be done by determining the impact of individual annotators or a general annotator effect on the results, e.g., by discarding votes by certain annotators and re-estimating the model, see Hechinger et al. (2024).

Our work contributes to the understanding of NLI tasks and provides guidance for the early stage of data collection. Therefore, analyzing the impact on the full machine learning pipeline, i.e., improvements on the predictive power of classifiers is beyond the scope of this paper, but is open for future work.

Acknowledgements

CG is supported by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. KH is supported by the Helmholtz Association under the joint research school HIDSS-006 - Munich School for Data Science@Helmholtz, TUM&LMU. MA has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of BERD@NFDI - grant number 460037581. BP is supported by European Research Council (ERC) grant agreement No. 101043235.

References

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

B. Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. 2023. Sources of Uncertainty in Machine Learning – A Statisticians' View. ArXiv:2305.16703 [cs, stat].

Katharina Hechinger, Xiao Xiang Zhu, and Göran Kauermann. 2024. Categorising the world into local climate zones: towards quantifying labelling uncertainty for machine learning models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73:143–161.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Ecologically valid explanations for label variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What Can We Learn from Collective Human Opinions on Natural Language Inference Data? ArXiv:2010.03532 [cs].

Animesh Nighojkar, Antonio Laverghetta Jr., and John Licato. 2023. No strong feelings one way or another: Re-operationalizing neutrality in natural language inference. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 199–210, Toronto, Canada. Association for Computational Linguistics.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian Models of Annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. Place: Cambridge, MA Publisher: MIT Press.

Barbara Plank. 2022. The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. ArXiv:2211.02570 [cs].

Matthew Stephens. 2000. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with Different Amounts of Annotation: From Zero to Many Labels. ArXiv:2109.04408 [cs].

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

A Appendix

Details on Model and Estimation

The EM algorithm is initialized with $\boldsymbol{\pi}_{(0)} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, and $\boldsymbol{\Theta}_{(0)}$ is drawn from a Dirichlet distribution where $\boldsymbol{\alpha}$ is set to be a vector with K entries, where each value is $2 \cdot K$. In this case $\boldsymbol{\alpha} = (6, 6, 6)$.

The model estimated on the full dataset (i.e., $N = 1514, J = 100$), which is also depicted in Figure 1, results the following final parameter estimates:

$$\hat{\boldsymbol{\pi}} = (0.314, 0.448, 0.238)$$

$$\hat{\boldsymbol{\Theta}} = \begin{pmatrix} \hat{\boldsymbol{\theta}}_{entailment} \\ \hat{\boldsymbol{\theta}}_{neutral} \\ \hat{\boldsymbol{\theta}}_{contradiction} \end{pmatrix} = \begin{pmatrix} 0.73 & 0.24 & 0.03 \\ 0.14 & 0.79 & 0.07 \\ 0.03 & 0.31 & 0.66 \end{pmatrix}$$

In both parameters, the order of entries/columns is *entailment*, *neutral*, *contradiction*.

Based on the estimated parameters obtained via the procedure described in section 4 the decision borders are defined by connecting the points ([E, N, C]):

- center point: $[35.98, 28.15, 35.86]$
- EC axis: $[48.46, 0.0, 51.54]$
- EN axis: $[42.03, 57.97, 0.0]$
- NC axis: $[0.0, 70.13, 29.87]$

Combined Stability Analysis

Figure 4 shows the estimation results and their bootstrapped stability for various sample sizes and numbers of annotations. Reducing N and J simultaneously leads to unstable results for very small datasets. However, this visualization supports the earlier finding that a sufficient number of annotations is more crucial than a large sample for stable and reliable estimation.

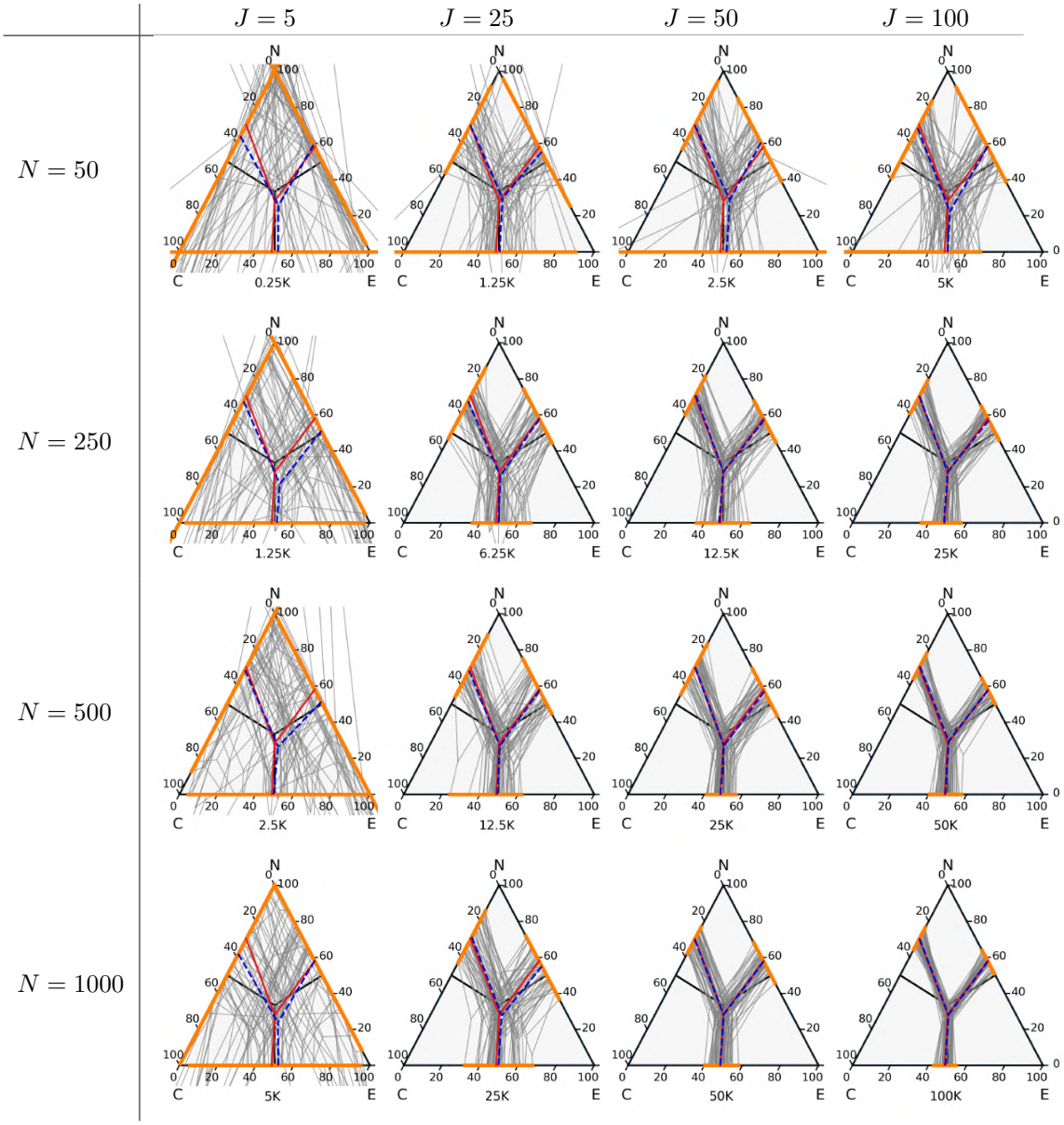


Figure 4: The Figure shows the bootstrapped latent class borders as gray lines, the range of the gray lines in orange, the mean of the bootstraps as blue dashed lines and the original borders as red lines for different values of N and J. The total number of annotations, i.e., $N \cdot J$, is below each plot.

Part III.

Multi-dimensional Ground Truth

7. Human-in-the-loop: Towards Label Embeddings for Measuring Classification Difficulty

Contributing article:

Hechinger, K., Koller, C., Zhu, X.X. and Kauermann, G. (2024). Human-in-the-loop: Towards Label Embeddings for Measuring Classification Difficulty. *arXiv preprint arXiv:2311.08874*. Under review in *Journal of Computational and Graphical Statistics.*

Code and data:

The supplementary code is publicly available on GitHub: `https://github.com/katharinahech/labelembeddings.git`. Further information on the sources and availability of the datasets is provided in the article.

Author contributions:

Katharina Hechinger and Göran Kauermann developed the idea of moving away from a single ground truth label. Göran Kauermann proposed the final method, which involves employing a Dirichlet multinomial model and estimating it through Expectation Maximization with Markov Chain Monte Carlo steps. The implementation of the model in `R`, including data preparation and visualization of the results, was carried out by Katharina Hechinger. Sections 1, 2, 3, and 5 of the manuscript were primarily written by Katharina Hechinger, with valuable input from Göran Kauermann. Christoph Koller authored the central part of Section 4 as an outlook. Xiao Xiang Zhu provided the dataset and offered helpful input as a domain expert. All authors were involved in proofreading the manuscript.

HUMAN-IN-THE-LOOP: TOWARDS LABEL EMBEDDINGS FOR MEASURING CLASSIFICATION DIFFICULTY

A PREPRINT

Katharina Hechinger[*]
Department of Statistics
Ludwig-Maximilians-University
Munich, Germany

Christoph Koller
Chair of Data Science in Earth Observation
Technical University of Munich
Munich, Germany
German Aerospace Center
Wessling, Germany

Xiao Xiang Zhu
Chair of Data Science in Earth Observation
Technical University of Munich
Munich, Germany
Munich Center for Machine Learning
Munich, Germany

Göran Kauermann
Department of Statistics
Ludwig-Maximilians-University
Munich, Germany

June 25, 2024

ABSTRACT

Uncertainty in machine learning models is a timely and vast field of research. In supervised learning, uncertainty can already occur in the first stage of the training process, the annotation phase. This scenario is particularly evident when some instances cannot be definitively classified. In other words, there is inevitable ambiguity in the annotation step and hence, not necessarily a single "ground truth" associated with each instance.

The main idea of this work is to drop the assumption of a ground truth label and instead embed the annotations into a multidimensional space. This embedding is derived from the empirical distribution of annotations in a Bayesian setup, modeled via a Dirichlet-Multinomial framework. We estimate the model parameters and posteriors using a stochastic Expectation Maximisation algorithm with Markov Chain Monte Carlo steps. The methods developed in this paper readily extend to various situations where multiple annotators independently label instances. To showcase the generality of the proposed approach, we apply our approach to three benchmark datasets for image classification and Natural Language Inference, where multiple annotations per instance are available. Besides the embeddings, we can investigate the resulting correlation matrices, which reflect the semantic similarities of the original classes very well for all three exemplary datasets.

Keywords Annotation Uncertainty, Multiple Labels, Label Variation, Stochastic EM Algorithm, Dirichlet-Multinomial Model, Classification and Clustering

[*]katharina.hechinger@stat.uni-muenchen.de

1 Introduction

Machine Learning models are increasingly used for a growing number of applications, one of which is supervised classification, for example in the form of images or texts. While such models have achieved impressive standards over the last years in terms of accuracy, the assessment of uncertainty remains an active field of open problems and research challenges. Recent survey articles discussing the field include Gawlikowski et al. [2023] or Hüllermeier and Waegeman [2021]. Uncertainty thereby has numerous and intertwined sources as discussed in Gruber et al. [2023] or Baan et al. [2023] and is heavily impacted at multiple stages of the common machine learning pipeline. Gruber et al. [2023] also explicitly emphasize the role of the data itself for appropriately assessing uncertainty entirely.

In the field of deep learning, multiple major streams of research related to the quantification of uncertainty exist. Besides ensemble methods and Bayesian approaches, evidential neural networks have been gaining attention as a deterministic method of uncertainty quantification (Sensoy et al., 2018). Specifically, these methods conceptualize learning as the acquisition of evidence, where each new training example adds support to a learned evidential distribution. A recent survey by Ulmer et al. [2023] provides an extensive overview of evidential deep learning and discusses its strengths and weaknesses in depth. However, some lines of work also advise caution when employing evidential networks for uncertainty quantification. Jürgens et al. [2024] state that generally, epistemic uncertainty is not reliably represented by those methods and Meinert et al. [2023] showcase the issue of overparameterization for evidential regression.

However, while some parts of the overall uncertainty are already heavily studied in research, less attention has been paid to one of the major prerequisites for training classification models, namely the (un-)availability of reliable ground truth labels for the training data and their uncertainty. In fact, uncertainty already starts in the labeling process for supervised machine learning, where human annotators label images or texts. Any supervised model will rely on these "ground truth" labels, inherently incorporating their associated uncertainty. We refer to this type of uncertainty as "label uncertainty". Commonly, such gold labels are acquired with human labeling effort, leading to multiple annotations per instance. Depending on the complexity of the problem at hand, it might suffice to aggregate the annotations into a single ground truth label, e.g. by majority voting. However, in many realistic application areas, such as the classification of complex images or the assessment of language and speech, this assumption does not hold true.
Of course, humans are naturally prone to errors, leading to unreliable annotations or mistakes, and therewith, label noise or label errors. The problem was already tackled and discussed early in the statistical literature, see for example Dawid and Skene [1979]. In recent years, more and more methods have been developed for handling data despite human errors, for example in the context of neural networks (Dgani et al., 2018). Peterson et al. [2019] argue that incorporating human ambiguity can improve classification models in terms of robustness. However, training supervised machine learning models based on noisy or deficient labels can lead to poor performance and high uncertainties, see e.g. Frénay et al. [2014] or more recently Frénay and Verleysen [2014] for an overview. Also, the labels might introduce some bias, as shown by Jiang and Nachum [2020], that needs to be identified and corrected if possible. Different algorithms have been introduced to tackle the problem of noisy labels, see Algan and Ulusoy [2021] for an extensive survey on various methods.

However, ambiguity in annotations cannot always be attributed to the fallibility of human annotators. Instead, label variation is also likely to arise if the assumption of a singular ground truth label for each instance is questionable. In the context of language, Plank [2022] discusses the sources of label variation. Particularly, the authors argue that the absence of a singular ground truth is often reasonable and should not be considered erroneous by default. In this line, the survey by Uma et al. [2021] discusses the disagreement of annotators. The authors conclude that suitable evaluation methods are required if a single gold label cannot be assigned. Various works in multiple domains show that disregarding label variation and leaving it untreated can indeed lead to quality issues and uncertainties. The common approach to simply summarise the annotations into a single label does not only discard valuable information, it is also an inappropriate representation of the truth and introduces remarkable amounts of uncertainty, in particular in the "gold" label (Davani et al., 2022, Uma et al., 2021 or Aroyo and Welty, 2013).

This problem is prevalent across various classification domains, specifically for application areas characterized by inherent ambiguity. To showcase this, let us first consider the domain of natural language processing (NLP) or more specifically natural language inference (NLI), where ambiguity is ubiquitous due to the subjective interpretation of language and speech. This issue has been already extensively discussed, see e.g. Plank [2022]. NLI corresponds to the task of discerning the logical relationship between two sentences, typically whether one entails the other, contradicts it, or is unrelated to it. Naturally, the perception of language differs for the human annotators causing high rates of disagreement (Nie et al., 2020, Pavlick and Kwiatkowski, 2019). Table 1 shows exemplary sentences, which are highly ambiguous. This ambiguity is clearly reflected by the annotations. Gruber et al. [2024] provide a statistical approach for modeling the data-generating process in order to gain a better understanding of the label uncertainty. However, their modeling approach assumes a latent ground truth label associated with each sentence pair. While this is only a modeling assumption, numerous works claim that the assumption of a single ground truth is not appropriate for NLI tasks and instead, a more realistic representation of the labels should be used, see e.g. Aroyo and Welty [2015], Uma et al. [2021] or Plank [2022].
Similar problems arise in the domain of image classification if either the categories or the images themselves are ambiguous. Depending on the nature of the problem, assigning a singular ground truth label is often simply impossible. Our second example comes from the field of remote sensing and satellite image classification. The ultimate goal is to categorize images into local climatic zones (LCZs), a classification scheme for satellite images developed by Stewart and Oke [2012]. A huge effort has been made to develop complex deep neural networks for this task (e.g., Zhu et al., 2020, Qiu et al., 2019 or Qiu et al., 2020). These models rely on large amounts of labeled data and, hence, it is necessary to manually annotate a vast amount of satellite images. Annotation of such images requires specialized expertise in earth observation and hence, the labeling is conducted by trained experts instead of laypersons. However, even with their domain knowledge, annotators frequently disagree due to the complexity of the task and the inherent ambiguity of both, the images and the categories. Figure 1 illustrates the climatic classes with accompanying low-resolution satellite images and high-resolution images from Google Earth, highlighting the challenges in classification and the potential for disagreement. Nevertheless, the annotations are commonly aggregated into a majority vote discarding valuable information about the initial disagreement (Zhu et al., 2020), and the uncertainty associated with the ground truth labels remains untreated. Hechinger et al. [2024] approach this problem from a statistical perspective and employ a Bayesian mixture model to estimate the latent posterior label distributions for the images based on the annotations. Based on the proposed modeling framework they assess factors contributing to the uncertainty and variation in the annotations. Their work especially showcases the heterogeneity of the individual annotators and the diversity of the images themselves. These factors lead to remarkable variation within the annotations and, hence, express the complexity of the labeling task. Their work is again based on the modeling assumption of a singular ground truth label for each image. However, the validity of this assumption in practice is questionable due to the ambiguous nature of the images themselves.
Lastly, situations without a distinct ground truth might also arise for presumably straightforward classification tasks, as shown in Figure 2. The exemplary images are part of the benchmark dataset Cifar-10H, introduced by Peterson et al. [2019]. The dataset is designed such that each instance is assigned to a single unambiguous class, at least in theory. Still, some images defy easy classification due to ambiguities caused by the size or quality of the picture, leading to high disagreement rates within the annotations. Consequently, relying solely on majority voting to assign a singular label in such cases does not accurately reflect the underlying truth. We take this data set as a third example to showcase our modeling approach.

In this work, we propose to move away from the premise of a sole ground truth. The three examples sketched above underpin that the assumption of a ground truth is not always appropriate. Therefore, we explicitly allow for ambiguity for each instance. This view extends to many real-world applications, where it is likely to encounter instances that can not be classified uniquely but are associated with a mixture or combination of different classes. In this paper, we aim to statistically model such situations in a distributional framework. Namely, we employ a Dirichlet Multinomial model, as discussed in Minka [2000] or Mosimann [1962]. Variants of this model class have been e.g. used for clustering of text documents (Yin and Wang, 2014) or genomics data (Holmes et al., 2012 or Harrison et al., 2020). Avetisyan and

Context/Premise	Statement/Hypothesis	Human Votes [C, N, E]
A man running a marathon talks to his friend.	There is a man running.	[0, 0, 100]
A black and white dog running through shallow water.	Two dogs running through water.	[42, 14, 44]
A woman holding a child in a purple shirt.	The woman is asleep at home.	[46, 53, 1]
An elderly woman crafts a design on a loom.	The woman is sewing.	[34, 31, 35]

Table 1: The table shows 4 examples of sentence pairs from ChaosSNLI, along with the annotations, see Gruber et al. [2024]. Each pair of context and statement is classified by 100 human annotators with the categories "contradiction" (C), "neutral" (N) and "entailment" (E).

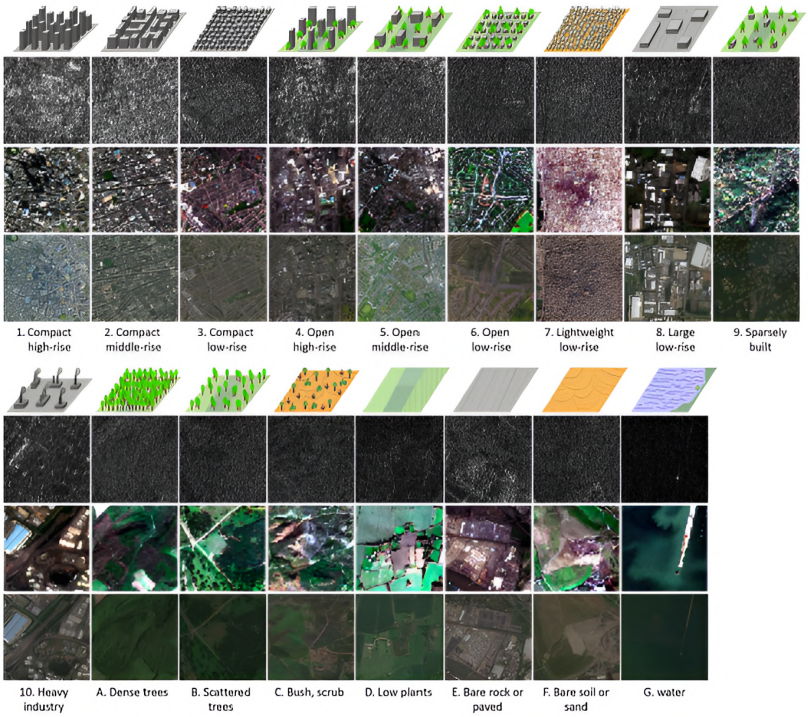


Figure 1: The figure shows the categories of the LCZ classification scheme (Stewart and Oke, 2012), along with exemplary images for each class from three different sources (top row: Sentinel 1, middle row: Sentinel 2, bottom row: Google Earth), images taken from Zhu et al. [2020].

Fox [2012] deploy a Dirichlet-Multinomial Mixture model to estimate survey response rates. Eswaran et al. [2017] also connected this model class to uncertainty quantification and modeled beliefs as Dirichlet distributions to capture uncertainty. In this work, we propose a Dirichlet Multinomial model to estimate **embedded ground truth values** that express the classification difficulty and uncertainty for the respective images based on human annotations. Specifically, we construct an embedding space so that each image is located in a K dimensional space, with K as the number of categories. To do so, we pursue an empirical Bayes approach in combination with Markov Chain Monte Carlo (MCMC) sampling and a stochastic version of the Expectation Maximization (EM) algorithm for estimation, as proposed by Celeux et al. [1996]. The presented strategy gives insights into the correlation (or confusion) patterns between different classes and simultaneously allows to express and quantify uncertainty.

Moving forward, the results can subsequently be integrated into the machine learning pipeline by directly training a model on the acquired embedded labels rather than relying on the majority-voted classes. Approaches in this direction


(a) Original label: Boat


(b) Original label: Frog


(c) Original label: Cat


(d) Original label: Deer

Figure 2: The figure shows exemplary images from Cifar-10H, where a high disagreement rate between the annotators could be observed hinting at the ambiguity of the images.

are based on the idea of label distribution learning, which incorporates the ambiguity of labels (e.g. Geng, 2016, Gao et al., 2017 or Xu et al., 2019). Koller et al. [2024] suggest integrating the label uncertainty into the training process by using distributional labels based on multiple votes for better generalization for unseen data and more stable performance in terms of uncertainty calibration. Extending their ideas by using a more sound representation of the labels, i.e. the label embeddings estimated by our approach, could allow us to improve the predictions even more in terms of uncertainty. We will sketch the idea but this paper emphasizes the statistical modelling step and not on the incorporation into a machine learning pipeline.

The paper is structured as follows. Section 2 describes the distributional framework and the algorithm used for estimation. First, we consider a binary case with two classes only for clarification purposes and then move on to the more general multiclass case. The results on three different datasets are reported in Section 3. We consider some possible further steps and applications in Section 4. Section 5 concludes the paper with a detailed discussion.

2 Model

2.1 Notations

Each image i, with $i = 1, ..., n$ is assessed by a set of annotators (labelers, voters) indexed with j, where $j = 1, ..., J_i$. We consider the images as independent and the same holds for the annotators. The labelers classify each image individually into the class k, where $k = 1, ..., K$. The corresponding vote of the expert is denoted by $V_{ij} \in \{1, \dots, K\}$. It is notationally helpful to rewrite this vote into the K dimensional indicator vector, which we denote in bold with $\boldsymbol{V}_{ij} = (\mathbf{1}\{V_{ij} = 1\}, ..., \mathbf{1}\{V_{ij} = K\})$, with $\mathbf{1}\{\cdot\}$ as indicator function. This allows to accumulate the annotators' votes into $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{iK})$ with $Y_{ik} = \sum_{j=1}^{J_i} \mathbf{1}(V_{ij} = k)$. This vector can be considered as the vote distribution for image i. To keep the notation simple we will drop index i from the number of voters per image and write J subsequently. We emphasize though, that images can be labeled by different numbers of voters, as our examples demonstrate.

2.2 Binary Case: $K = 2$

For a more straightforward presentation and interpretation of our modeling strategy, we start with the binary case $K = 2$. We assume a binary label representation, which is embedded into the two-dimensional space. That is, each instance (image or text) is allocated with

$$\boldsymbol{Z_i} = (Z_{i1}, Z_{i2}) \in \mathbb{R}^2.$$

The vector Z_i can be interpreted as **embedding** or **embedded ground truth values**, meaning that we represent the labeled instance i as a point in a two-dimensional space. To simplify the notation we will drop the index i in the following.

The embedding steers ambiguity as well as the uncertainty of the labeling process. This is achieved by relating Z to the coefficients of a Beta distribution. To be specific we define $\alpha_Z = \exp(Z_1)$ and $\beta_Z = \exp(Z_2)$ as parameters of a Beta distribution, from which we draw the binomial parameter π as

$$\pi \sim \text{Beta}(\alpha_Z, \beta_Z).$$

Given π, we obtain the image labels by drawing from the binomial distribution

$$Y|\pi \sim B(J, \pi),$$

where J is the number of votes or annotations of the respective image. Note that J can vary for different instances, which for simplicity of notation we ignore here. Apparently, if π is close to 0 or 1, the image has no or little ambiguity, i.e. it is easy to label. Note that π remains unobserved, so that given Z, we have

$$\begin{aligned} P(Y = y|\boldsymbol{Z}) &\propto \int_\pi \binom{J}{y} \pi^y (1-\pi)^{(1-y)} \pi^{\alpha_Z} (1-\pi)^{\beta_Z} d\pi \\ &\propto \binom{J}{y} \frac{B(\alpha_Z + y, \beta_Z + J - y)}{B(\alpha_Z, \beta_Z)}, \end{aligned} \tag{1}$$

where $B(.)$ denotes the univariate Beta function.
Within this model setup, we can derive a couple of interpretations. Interpreting $\boldsymbol{Z} \in R^2$ as ground truth, we obtain the Beta-Binomial model (1). This in turn allows us to derive the mean value of π through

$$E(\pi|\boldsymbol{Z}) = \frac{\exp(Z_1)}{\exp(Z_1) + \exp(Z_2)}.$$

Additionally, we can also quantify uncertainty by calculating the variance, which results as

$$Var(\pi|\boldsymbol{Z}) = \frac{\exp(Z_1)\exp(Z_2)}{\big(\exp(Z_1)+\exp(Z_2)\big)^2\big(\exp(Z_1)+\exp(Z_2)+1\big)}.$$

For different values of $\boldsymbol{Z}$, we plot the mean and the (log)-variance of the Beta-Binomial distribution in Figure 3. The variance expresses the uncertainty, which is how likely an image/text is misclassified given the data at hand. The larger Z_1, the more likely the instance is classified as "one". On the contrary, the larger Z_2, the more likely the image/text is classified as "zero". Moreover, the smaller the values of Z_1 and Z_2 and the smaller the difference between them, the larger the variance. Hence, the concrete location of Z_1 and Z_2 expresses how likely it is that we can quantify an image or text in one category and how certain we are with respect to the class.

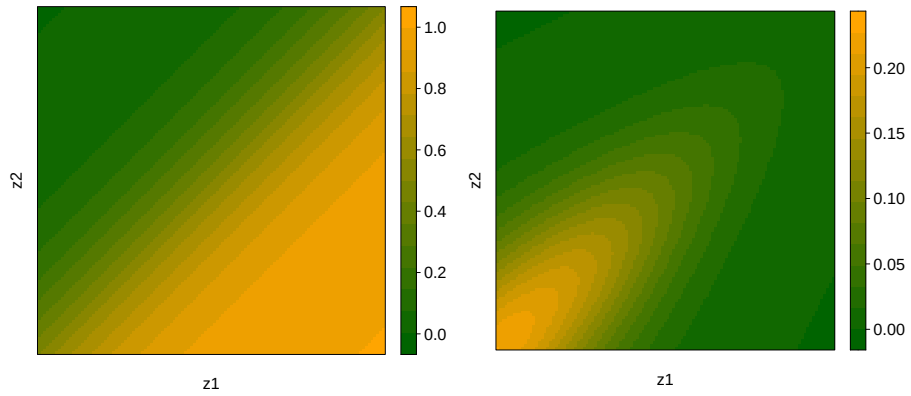


Figure 3: The figure shows the mean (left) and log-variance (right) of the Beta-Binomial distribution for different values of $\boldsymbol{Z}$, expressed through color.

Apparently, $\boldsymbol{Z}$ is latent, but we aim to draw information about Z given the votes Y. The two variables are connected via the Beta-Binomial model (1) and we can estimate the distribution of Z for given votes Y by drawing Markov Chain Monte Carlo (MCMC) samples. To do so, we sample from the posterior

$$\begin{aligned} f(\boldsymbol{Z}|Y=y) &\propto f(y|\boldsymbol{Z}) \cdot f_{prior}(\boldsymbol{Z}) \\ &\propto \binom{J}{y} \frac{B(\alpha_Z + y, \beta_Z + J - y)}{B(\alpha_Z, \beta_Z)} \cdot f_{prior}(\boldsymbol{Z}). \end{aligned}$$

As prior distribution for $\boldsymbol{Z}$, we use a bivariate normal with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ and estimate these parameters following an empirical Bayes approach. While this prior distribution is presumably too simple to completely capture the true hidden structure in the data and hence does not constitute the underlying data-generating process, it is a convenient modeling assumption. Postulating a multivariate Gaussian distribution for the latent embedded ground truth provides numerical stability during estimation and allows to express uncertainty, even if only one annotation is available.

With Y_i as votes on image i, J_i referring to the number of annotations on image i and $\boldsymbol{Z}_i \in \mathbb{R}^2$ as the embedded "true location" of image i within the two-dimensional space, we run the estimation algorithm laid out in Figure 4. We also refer to Appendix A for additional computational details. As a result, we obtain estimates $\hat{\boldsymbol{z}}_i$ for each image by averaging the MCMC samples for instance i of the last iteration. The point itself expresses the classification difficulty for the respective image and is thus far more informative than a singular ground truth label.

1. Initialize the prior parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and set them to $\boldsymbol{\mu}^{(m)}$ and $\boldsymbol{\Sigma}^{(m)}$, $m = 1$.
2. For $i = 1, \ldots, n$: Draw an MCMC sample from

$$f(\boldsymbol{Z}_i|Y_i = y_i) \propto \binom{J_i}{y_i} \frac{B(\exp(Z_{i1}) + y_i, \exp(Z_{i2}) + J_i - y_i)}{B(\exp(Z_{i1}), \exp(Z_{i2}))} \cdot \exp(-0.5(\boldsymbol{Z}_i - \boldsymbol{\mu}^{(m)})^T \cdot \boldsymbol{\Sigma}^{(m)-1} \cdot (\boldsymbol{Z}_i - \boldsymbol{\mu}^{(m)}))$$

   and take $\hat{\boldsymbol{z}}_i^{(m)} = E(\boldsymbol{Z}_i|Y_i)$ as the mean of the MCMC samples.
3. Update

$$\hat{\boldsymbol{\mu}}^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{z}}_i^{(m)}$$
$$\hat{\boldsymbol{\Sigma}}^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\boldsymbol{z}}_i^{(m)} - \hat{\boldsymbol{\mu}}^{(m+1)})(\hat{\boldsymbol{z}}_i^{(m)} - \hat{\boldsymbol{\mu}}^{(m+1)})^T$$

   and repeat from Step 2.

Figure 4: Estimating the embeddings.

2.3 Multiclass Case: $K > 2$

The binary model can now be easily extended to more than two classes by employing the Dirichlet distribution. Now, $\boldsymbol{Z} = (Z_1, ..., Z_K) \in \mathbb{R}^K$ is the embedded ground truth and we obtain the parameters $\alpha_k = \exp(Z_k), \forall k = 1, ..., K$. From these, we can draw

$$\boldsymbol{\pi} \sim Dir(\alpha_1, ..., \alpha_k).$$

This results in the multinomial parameter vector $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$, where K corresponds to the number of classes. Given $\boldsymbol{\pi}$, the votes are assumed to be drawn from a multinomial distribution:

$$\boldsymbol{Y}|\boldsymbol{\pi} \sim Mult(\boldsymbol{\pi}, J),$$

with J denoting the number of votes. This leads to the two probability functions

$$f(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\pi}) = \frac{J!}{y_1!...y_K!} \prod_k \pi_k^{y_k}$$
$$f(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_k \pi_k^{\alpha_k - 1}.$$

In this case, the function $B(.)$ denotes the multivariate version of the Beta function. The vector $\boldsymbol{\pi}$ remains unobserved and we can calculate the probability of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ by marginalizing over $\boldsymbol{\pi}$:

$$\begin{aligned} P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\alpha}) &= \int_\theta f(\boldsymbol{y}|\boldsymbol{\pi}) f(\boldsymbol{\pi}|\boldsymbol{\alpha}) d\boldsymbol{\pi} = \frac{J!}{y_1!...y_K!} \cdot \frac{1}{B(\boldsymbol{\alpha})} \cdot \int_{\boldsymbol{\pi}} \prod_k \pi_k^{y_k + \alpha_k - 1} d\boldsymbol{\pi} \\ &= \frac{J!}{y_1!...y_K!} \cdot \frac{\Gamma(\sum \alpha_k)}{\prod_k \Gamma(\alpha_k)} \cdot \frac{\prod_k (\Gamma(\alpha_k + y_k))}{\Gamma(\sum_k \alpha_k + y_k)} = \frac{J!}{y_1!...y_K!} \cdot \frac{\Gamma(\sum \alpha_k)}{\Gamma(\sum_k \alpha_k + y_k)} \cdot \prod_k \frac{\Gamma(\alpha_k + y_k)}{\Gamma(\alpha_k)} \\ &= \frac{J!}{y_1!...y_K!} \cdot \frac{B(\boldsymbol{\alpha} + \boldsymbol{y})}{B(\boldsymbol{\alpha})}. \end{aligned}$$

Again, the embedded ground truth values $\boldsymbol{Z}$ can be estimated given the votes $\boldsymbol{Y}$ using MCMC samples with the stochastic EM algorithm. Following the binary case and assuming a multivariate Gaussian prior for the embeddings $\boldsymbol{Z}$ leads to the posterior distribution

$$f(\boldsymbol{Z}|\boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{Z}) f_{\text{prior}}(\boldsymbol{Z}).$$

We obtain a Dirichlet-Multinomial model by assuming a K-dimensional embedded ground truth $\boldsymbol{Z}$ for each image. The parameter $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$ follows a Dirichlet distribution given $\boldsymbol{Z}$, and we can easily derive expectation and variance for all entries $Z_k \in \boldsymbol{Z}, k = 1, ..., K$:

$$\begin{aligned} E(\pi_k|\boldsymbol{Z}) &= \frac{\exp(Z_k)}{\sum_{k'=1}^{K} \exp(Z_{k'})} \\ Cov(\pi_k, \pi_{k'}|\boldsymbol{Z}) &= \frac{1}{1 + \sum_{k'=1}^{K} exp(Z_{k'})} \cdot \frac{\exp(Z_k)}{\sum_{k'=1}^{K} \exp(Z_{k'})} \cdot \left(1 - \frac{\exp(Z_k)}{\sum_{k'=1}^{K} \exp(Z_{k'})}\right). \end{aligned}$$

Each entry of $\boldsymbol{Z}$ corresponds to one of the K classes. The concrete values can again be interpreted in two ways. On the one hand, Z_k hints at how likely the image is classified into category k. On the other hand, the difference between the entries of $\boldsymbol{Z}$, i.e. the distance between classes k and k' for $k' \neq k$, corresponds to the certainty about the category k versus k'.

Following the estimation procedure described in Section 2.2 adapted to the multiclass case leads to values $\hat{\boldsymbol{z}}_i$ for all images $i = 1, ..., n$. Like above, these values form an **embedding** of the image in the K dimensional space. These embeddings express the classification (un-)certainty of the individual images in terms of distribution. The concrete algorithm is comparable to the case $K = 2$ discussed above and therefore not explicitly laid out here again.

Dataset	#Images N	#Classes K	#Distinct Annotation Patterns	#Annotations J
ChaosSNLI	1514	3	832	100
So2Sat LCZ42	159581	16	360	11
Cifar-10H	10000	10	3406	[50,63]

Table 2: Overview of the datasets.

3 Results

To showcase the generality and versatility of the proposed approach in various applications, the proposed model is applied to the three datasets described in the introduction. The datasets are typical examples in the field of multiple annotations and annotator disagreement and hence provide ground for analyzing the uncertainty associated with the labels. Table 2 contains general information about the three datasets discussed in this section.

3.1 ChaosSNLI

First, we explore the advantages of the proposed methodology in the context of the classification of language, i.e. the domain of NLI, as shortly introduced in Section 1. The multi-annotator dataset ChaosSNLI* is based on the development set of the Standford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) and was introduced in the context of label ambiguity by Nie et al. [2020]. It contains multiple annotations for sentence pairs, i.e. pairs of premise and hypothesis. For each premise, three hypotheses are originally generated by an annotator, as an entailing, neutral, and contradicting description of the premise. The resulting sentence pairs of premise and hypothesis can therefore be classified as *entailment, neutral* or *contradiction*. Note that a subjective ground truth, namely the original intention of the first annotator, is available for this specific dataset. However, it cannot be recovered by the annotators in many cases and the dataset ChaosSNLI especially showcases this problem as it contains sentence pairs exhibiting a high rate of disagreement. Particularly, $N = 1514$ sentence pairs are re-assessed by a large number of annotators, i.e. $J = 100$, and assigned to one of the three classes, as shown in Table 1. Due to the ambiguous nature of language and the individual perception, the disagreement rate in the annotations is high and the original true label cannot be recovered reliably. For the classification of language, the existence of a single "gold" label is especially doubtful and hence, the need for alternative and more appropriate representations of labels persists. Applying the methodology proposed in Section 2 allows us to estimate embedded ground truth vectors for the observations based on the provided annotations, which will be analyzed in this subsection.

First, let us return to the **exemplary sentence pairs** provided in Table 1 and inspect the respective embeddings. Figure 5 shows the estimated values for the exemplary sentence pairs, along with the observed annotations as well as the MCMC samples. While the estimated values for observation 34 (upper left plot) express clear affiliation to class *entailment*, all other embeddings reflect the ambiguity within the sentence pairs and also the associated annotations. Not only are the class-specific estimated values rather small and hence similar across the three categories, we also see a tendency towards class *neutral* for ambiguous instances, even though the majority vote might advocate otherwise. This expresses ambiguity with respect to classification and the "weaker" interpretation of class *neutral* compared to the other two categories. If all three classes received annotations, it is semantically unlikely that the respective instance can be uniquely classified into either *entailment* or *contradiction*. Hence, the model favours class *neutral* in such situations.

For this particular application, the classes themselves are by definition uncorrelated or negatively correlated. This property is also expressed by the resulting estimated embeddings. To visually inspect the results, we employ dimensionality reduction techniques for easier exploration. We specifically utilize principal component analysis (PCA) to extract the principal components from the estimated embeddings. PCA is a widely used technique for linear dimensionality reduction, commonly employed for exploratory analysis and visualization of high-dimensional data. For detailed

*Download available from `https://github.com/easonnie/ChaosNLI` (accessed on 04th of February 2024)

ID	Context/Premise	Statement/Hypothesis
34	A man running a marathon talks to his friend.	There is a man running.
1168	A black and white dog running through shallow water.	Two dogs running through water.
1177	A woman holding a child in a purple shirt.	The woman is asleep at home.
1371	An elderly woman crafts a design on a loom.	The woman is sewing.

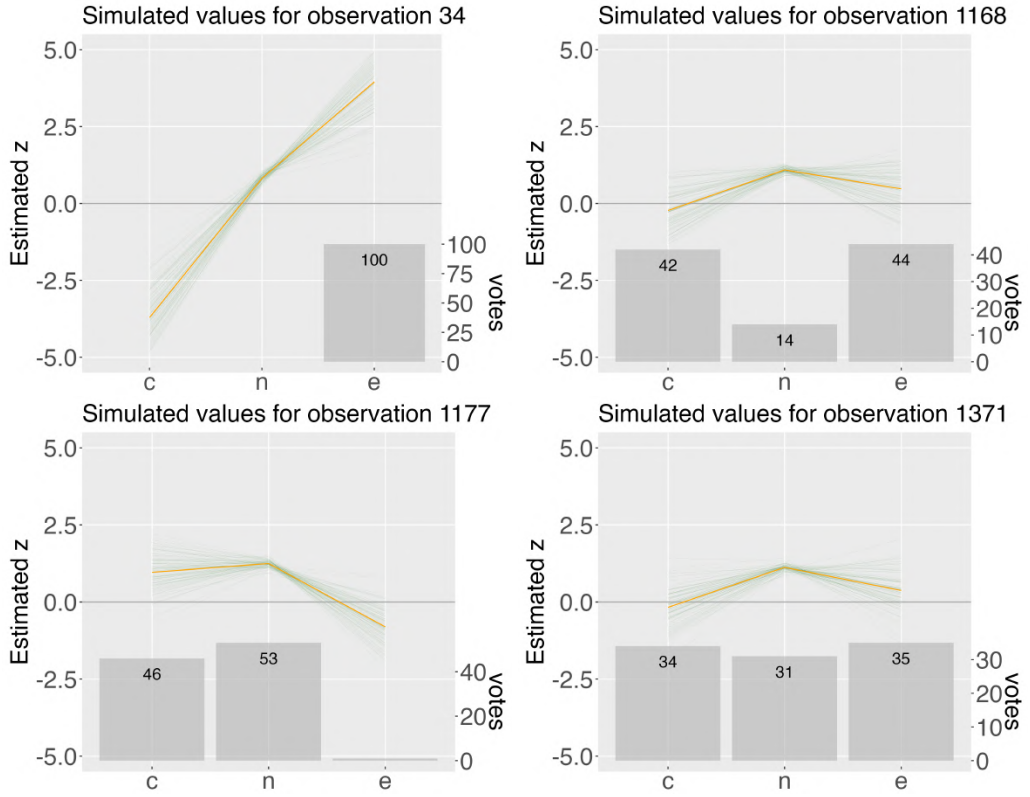


Figure 5: The plots show the estimated embedded ground truth vectors for exemplary sentence pairs from the dataset ChaosSNLI. The actual estimated vector is shown as orange line, the green lines represent the MCMC samples and the actual annotations are shown as grey bars.

information on the methodology, refer to e.g. Jolliffe [2002]. Here, we especially focus on the visualization benefits of the technique. Namely, it is possible to plot the observations in a so-called two-dimensional **biplot** after applying PCA. Figure 6 shows the respective plot for the dataset ChaosSNLI. The estimated embeddings are projected onto the two-dimensional space, spanned by the two first principal components. The embeddings of the instances are visualized as scattered points, where their overall similarity is expressed by proximity. In this case, no specific clustering is apparent. By coloring the scattered points according to the observed majority voting, we observe some overlap of the singular classes in the two-dimensional embedding space. Additionally, the vectors correspond to the original variables, i.e. the categories and dimensions of the embeddings. The angles between the vectors express the degree of correlation between the variables, i.e. small angles suggest a high positive correlation. However, in this case, the angles between variable *neutral* and the other two variables are roughly 90 degrees, indicating no correlation between the quantities. In contrast, the angle between variables *entailment* and *contradiction* is close to 180 degrees, hence expressing a negative correlation. Of course, this is reasonable from a semantic perspective and in line with the interpretation of the classes. Additional results for the ChaosSNLI dataset are reported in Appendices B and C.

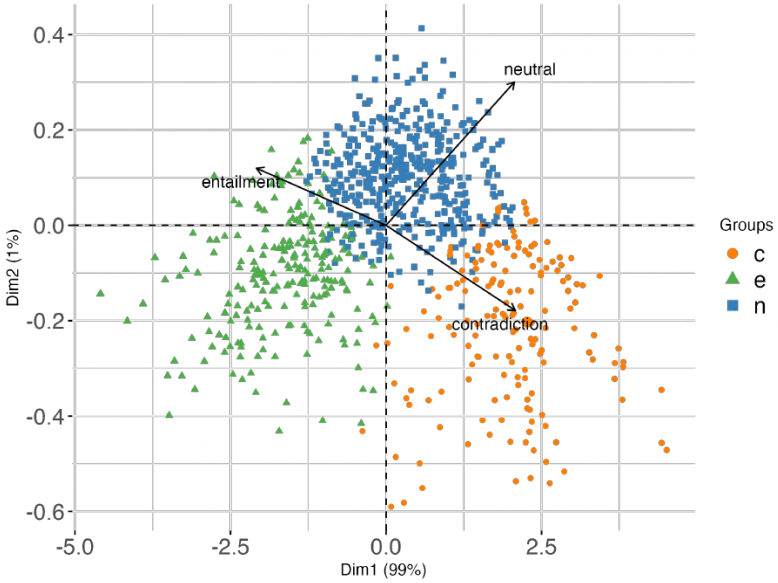


Figure 6: The biplot shows the estimated embeddings for ChaosSNLI, projected into two dimensions. The scatterpoints represent the instances, colored by majority vote. The original dimensions are represented as arrows.

3.2 So2Sat LCZ42

To showcase the generality of the proposed approach, we now move on to the analysis of a multi-annotator dataset for image classification. In the domain of image classification, the assumption of a singular ground truth label is especially doubtful if the images themselves are ambiguous. A prominent example is the classification of low-resolution satellite images. We analyze the earth observation benchmark dataset So2Sat LCZ42[*], see Table 2 for a brief overview. For a detailed description of the complete dataset, we refer to Zhu et al. [2020]. The task is to classify satellite images into one of 17 Local Climate Zones (LCZ). See also Figure 2 for a sketch of the LCZ. Note that zone 1 to 10 refer to urban areas while zones A to G are non-urban. The data comprises humanly classified satellite images from urban agglomerations in Europe, of which some images were assessed multiple times during the labeling phase to ensure label quality. Overall, we look at images from 9 European cities with multiple annotations. Figure 1 shows the categories for the classification of satellite images. The distribution of votes across the classes is quite imbalanced, and particularly LCZ 7 rarely occurs in the dataset. Due to its composition (lightweight low-rise building types, i.e. "slums"), it is doubtful to be observed in European cities. Therefore, we excluded LCZ 7 and restricted the model to $K = 16$ instead of $K = 17$, see also Hechinger et al. [2024]. The voting patterns suggest that most images seemed easy to classify. For 77.18 % of the images, the annotators agree on one single LCZ. Again, we apply the proposed framework and estimate an embedded ground truth value $\hat{z}^{(i)}$ for each instance represented as a vector of annotations.

First, it is again helpful to examine the explicit estimates for some **exemplary instances**. We get the sampled embeddings from Step 2 in the algorithm shown in Figure 4, where the parameters are set to the final estimated parameters. Figure 7 shows the explicit estimates, as well as the MCMC samples for four random images. The resulting posterior mean value (i.e. the estimated ground truth) is shown as an orange line and the observed votes as grey bars at the bottom of the plot. The first image with ID 18 (upper left plot) shows a non-confusion case, i.e. all experts agreed on class E. Looking at the ground truth estimate, we see a clear spike for the respective class, indicating that classification for this image is rather easy. Image 66 (upper right plot) gets nine votes for class C and two votes for class B, i.e. the voting pattern is

[*]Download available from: `https://mediatum.ub.tum.de/1659039` (accessed on 03rd of Febuary 2024)

restricted to non-urban classes. Here, we observe negative values for urban classes 1-10 and small positive values for all non-urban classes. Hence, the voting pattern is reflected in the estimated ground truth in terms of the magnitude of the values. Image 185 (lower left plot) shows confusion of two urban classes, namely classes 6 and 9. In this case, the estimated ground truth values clearly correspond to these votes with almost all other values being negative or very close to zero. The image might be ambiguous regarding the correct ground truth, as those two values are quite close. Lastly, confusion happened for urban classes only for the image with ID 349 (lower right plot). The majority of experts voted for class 4, three experts chose class 5, class 2 received 2 votes and one expert voted for class 1. This is also reflected in the estimated $\hat{z}_k$ values. We estimate similar positive values for classes 4, 5 and 2 and a slightly smaller value for class 1, reflecting the voting patterns but also incorporating additional information about the confusion risk of the classes.

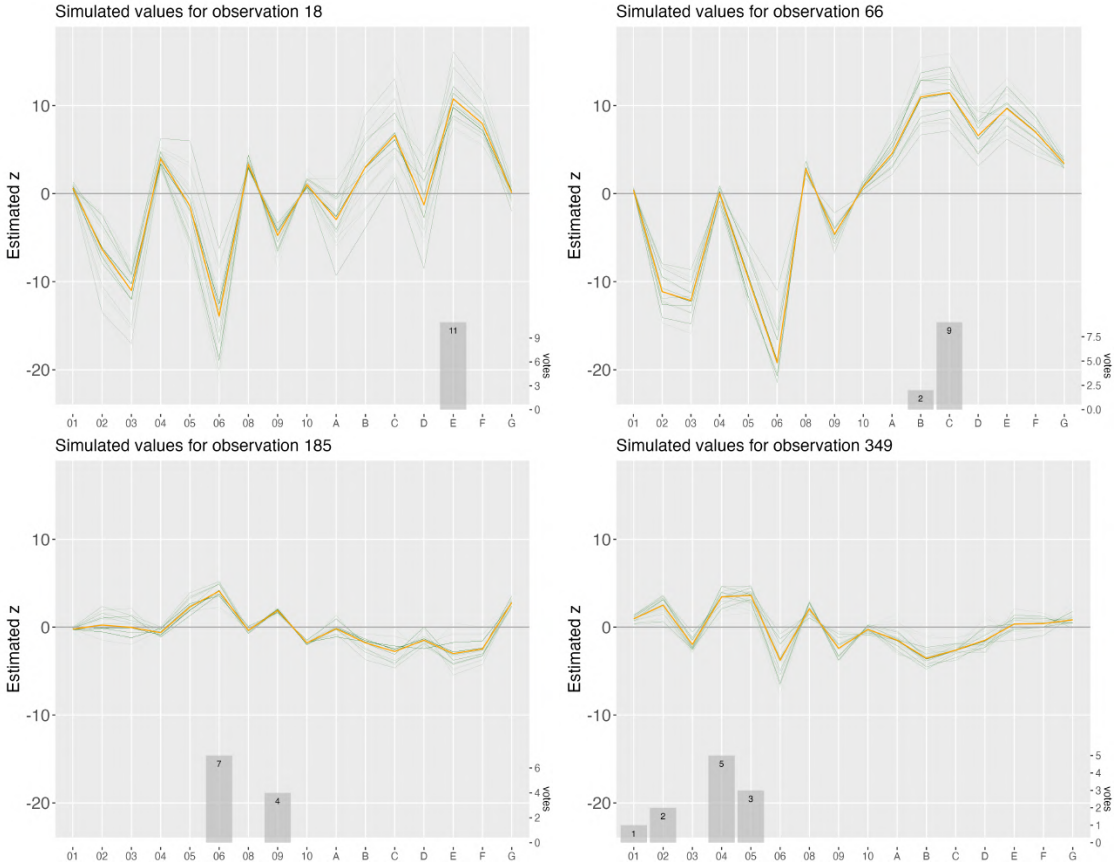
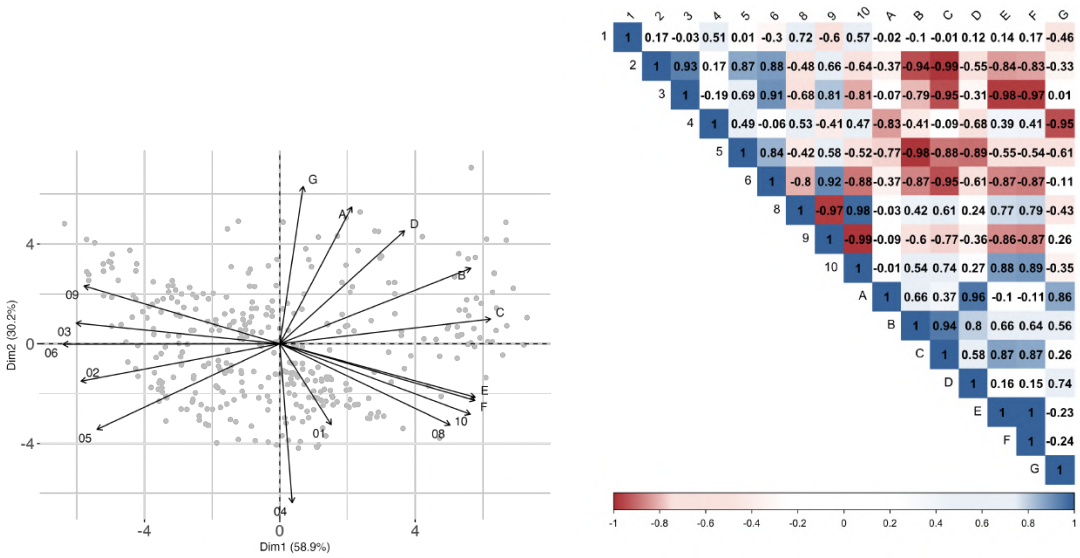


Figure 7: The plots show the MCMC samples as green lines and their mean value as an orange line for exemplary randomly chosen images. Additionally, the actual votings from the experts are shown as grey bars.

While low-resolution satellite images are of course prone to ambiguity, in this particular application, the classes themselves are also ambiguous and not easy to distinguish (Hechinger et al., 2024). Hence, the ambiguity of the classes contributes to the uncertainty in the annotations. This property is nicely expressed by the vectors shown in the **biplot** in Figure 8a, as well as the **estimated correlation matrix** of the ground truth embeddings in Figure 8b. This matrix can be interpreted as a generalized confusion matrix, in the case where a ground truth is not guaranteed. A high positive correlation indicates a high confusion risk. On the contrary, negative correlation values suggest that classes are well distinguishable. The correlation patterns between classes are visible in the correlation matrix, see Figure 8b. Exemplary, we investigate the correlation values for LCZ 2. Apparently, it is positively correlated with classes 3, 5 and 6, shows



(a) Biplot. (b) Correlation Matrix.

Figure 8: The subfigures show additional results for the dataset So2Sat-LCZ42. The biplot of the projected embeddings, as well as the correlation matrix of the estimates, show similarities of the categories leading to ambiguities and confusion.

weaker positive correlations to the other urban classes 1, 4 and 9 and is negatively correlated to all other classes. A semantic inspection of the description of the classes, given in Figure 1, makes this observation rather plausible. Class 2 refers to compact middle-rise areas and can therefore easily be confused with either other compact (LCZ 3), mid-rise areas (LCZ 5) or also open low-rise areas (LCZ 6), while distinction from classes 1, 4 and 10 as well as from nonurban classes A-G should be less difficult. We also see a strong correlation between class 8 and class 10, i.e. confusion of large low-rise areas (8) and heavy industry areas (10). This is again reasonable from a semantic perspective. Concerning nonurban classes A-G, we observe a higher correlation between classes A (dense trees) and D (low plants) as well as between B (scattered trees) and C (bush/scrub). The same holds for classes E and F, referring to bare rock/paved areas and bare soil/sand. Again, these classes are similar in their composition and confusion or overlap might occur. The matrix also shows that urban and nonurban classes (LCZ 1-10 vs. LCZ A-G) have mostly negative or close to zero correlation values, supporting the claim that confusion about these two types of LCZs rarely occurs. Overall, the entries of the estimated ground truth estimate are correlated in a way that has a meaningful interpretation and aligns with the initial hypotheses. The MCMC procedure additionally enables us to inspect the variation of the correlation matrix. The variances of the individual values are quite small and on average very close to zero, indicating a stable estimation and reliable results. The standard deviations of the entries in Figure 8b are displayed in Appendix D.

3.3 Cifar-10H

The third dataset is a version of Cifar-10, a popular benchmark dataset for image classification, as introduced by Krizhevsky et al. [2009]. The subset Cifar-10H* as introduced by Peterson et al. [2019] contains multiple annotations for images in the test set, reflecting the uncertainty stemming from differences in human perception. Here, the natural images are categorized into unambiguous classes, see Table 2. The original dataset has been extended with soft labels, i.e. multiple annotations, to achieve better generalization for classification models, specifically on out-of-sample datasets, see Peterson et al. [2019] and Battleday et al. [2020]. Therefore, $N = 10000$ images of $K = 10$ classes from

*Download available from `https://github.com/jcpeterson/cifar-10h` (accessed on 24th of August 2023)

the test set of Cifar-10 were annotated by 2571 Amazon Mechanical Turk workers. After an initial training phase, each worker annotated 200 images, 20 per category. To identify and remove low-performance annotators, attention checks were introduced after every 20 trials.
The current setting differs from the previously discussed dataset. Most images belong to one of the unambiguous categories and can be reliably classified by untrained annotators. Nevertheless, it is helpful to additionally inspect label embeddings reflecting the individual human perception, which can still be ambiguous. Due to the small size of the images, the pictured class is also not always identifiable, as shown in Figure 2. The dataset contains a high degree of human consensus due to its nature but also enough images where the annotation is still uncertain. This is also visible in the majority votes. While each class contains originally 1000 images, the number of images classified into the categories according to the majority vote varies slightly between 981 and 1015. Additionally, the images can be easily assessed and evaluated against the annotations, in contrast to the dataset presented previously. By applying the proposed model we generate embeddings of the images in the appropriate label space, which contain a notion of uncertainty and reflect the original annotations, without the loss of information by taking the majority voting.

Returning to the **exemplary images** from Figure 2, the estimated ground truth embeddings are shown in Figure 9. The upper plot in Figure 9 shows the label embedding for an image of a ship, which is clearly visible in the picture and therefore also identifiable by the annotators. This is reflected in the respective label embedding with a high positive value for class *ship*. For the second image, two annotators did not agree with the others and labeled the picture of a frog as *horse* or *deer*. Most of the labelers could correctly assign the label *frog*. The estimated label embedding reflects this by assigning the highest value to the class *frog* and small positive values to the other two classes. Nevertheless, the classification of the image is apparently easy, which is expressed by the embedding. This does not hold for the images, which correspond to the two lower plots. The correct label for the third image was *cat*, correctly identified by the majority vote. Nevertheless, the image is quite ambiguous, which is reflected in the annotations and therefore also in the label embedding. Only using the majority vote would therefore lead to a correct label while losing a large amount of information about the inherent uncertainty. For the last image, the annotators did not agree at all and could not recover the label *deer*. In fact, almost every class received votes. In this case, the label embedding clearly reflects this confusion by assigning similar values close to zero to all classes, reflecting the classification uncertainty.

Next, we repeat the analyses for the previous datasets, i.e. plotting the estimates of the embeddings onto a 2-dimensional **biplot** via PCA as well as calculating their **correlation matrix**. The biplot of the projected embeddings is displayed in Figure 10a and shows a clear separation between classes referring to animals and classes referring to objects. This is also expressed by the correlation matrix, shown in Figure 10b. Again, this reflects possible similarities between images from correlated classes, which occur due to individual human perception despite the clear separation of the classes by definition.

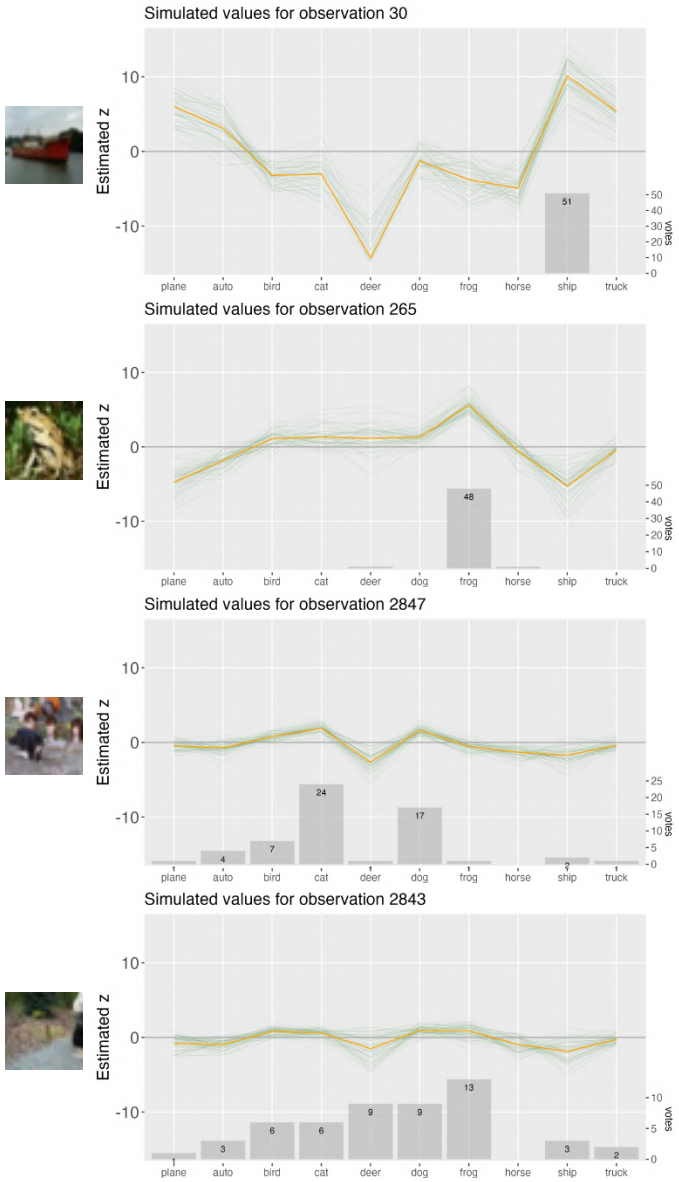
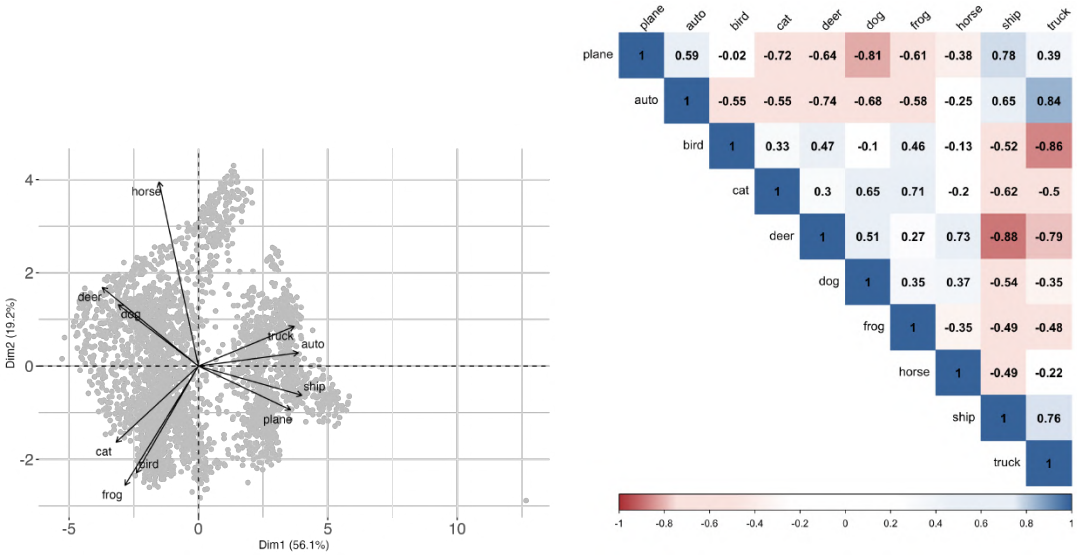


Figure 9: The plots show the estimated embeddings (orange) for exemplary images of the dataset Cifar-10H, along with the votes (bars) and the MCMC samples (green).



(a) Biplot. (b) Correlation Matrix.

Figure 10: The subfigures show additional results for the dataset Cifar-10H via the biplot and the correlation matrix of the estimated embeddings.

4 Outlook

Naturally, the question arises of how to use the information gained from embedding the ambiguous labels into a multidimensional space. The two main goals are to improve the supervised model assigned with the corresponding classification task and to possibly refine its uncertainty estimates. In many applications, training the classification model based on averaged labels or labels obtained via majority voting is still common practice. In the case of high annotator disagreement due to ambiguities, this can lead to major problems related to the associated uncertainties (Plank, 2022, Baan et al., 2023 and Davani et al., 2022). Koller et al. [2024] propose to instead integrate the annotation uncertainty via the empirical distribution of the annotations. Their work shows that incorporating this uncertainty leads to better generalization and calibration of the classification model. However, the benefit of the empirical distribution of course strongly depends on the number of annotations and is limited to the observed disagreement for one single instance only. The idea of estimating label embeddings via a distributional approach presented in this work offers a possibility to overcome said limitations. In particular, it is possible to train a classification model directly on the estimated embeddings $\hat{\boldsymbol{z}} = (\hat{\boldsymbol{z}}_1, ..., \hat{\boldsymbol{z}}_n)$, resulting from the estimation process as the mean of the MCMC samples in the last iteration. These embedded ground truth values retain information about all annotations for the respective observation and additionally incorporate knowledge about the annotations globally, across all instances. This leads to a more sound representation of the labels expressing uncertainties due to ambiguities of the instances themselves and also ambiguities due to the similarity of specific categories. Hence, this approach naturally handles images that cannot be directly classified into one class only. To integrate the embeddings into a deep learning framework, several strategies are available. While it is possible to directly learn the embedded ground truth vectors via a regression framework, reformulating the label embeddings into a Dirichlet function also allows us to stay within the world of classification. Either way, by incorporating the embeddings as labels in a machine learning framework, we expect the model to be better calibrated and yield more expressive predictive uncertainties. While this is beyond the scope of this paper, the results presented here serve as a valuable starting point for future work.

5 Discussion

For classification models, the dependence on labeled training data is a common practice, i.e. each instance is linked to an established ground truth label or "gold" label. Generating these ground truth labels requires substantial human effort and is prone to errors causing uncertainty. However, unreliable labels cannot always be attributed to human failure. In many applications, assigning a single label is unrealistic or even impossible due to the ambiguity of the instances themselves. A single ground truth label often cannot account for the complexity of e.g. images or sentences. This is often expressed through a high rate of disagreement in the annotations received from human labelers. Hence, the single-label approach results in a substantial loss of information and introduces additional uncertainty into the classification process. Therefore, moving beyond this limiting assumption is necessary in certain applications. This can be done by considering more flexible and adaptive strategies.
This paper focuses on classifying text or images addressing the specific case where we cannot assume that every observation can be uniquely classified into one class. Based on multiple annotations per observation, we propose to embed the images into a K-dimensional space instead of restricting them to a single label.

The proposed estimation procedure leads to interesting results, as reported in Section 3. We estimate label embeddings for three different datasets, emphasizing the generality of our approach and its usefulness in diverse settings. First, we apply the method to the dataset ChaosSNLI from the domain of language classification. The dataset contains especially ambiguous sentence pairs and a high number of annotations per instance. The assumption of a singular gold label is especially doubtful for the classification of language, due to its inherent ambiguity and subjectivity. Instead, multi-dimensional embeddings can serve as a more appropriate representation of the underlying truth. We show that the estimated vectors not only express the uncertainty associated with the instances due to the observed annotator disagreement but also the uncertainty due to a possibly insufficient number of annotations. Second, we move on to the domain of image classification and apply the proposed method to the earth observation dataset So2Sat LCZ42. Here, the satellite images themselves exhibit a high degree of ambiguity but also the categories are similar in terms of their composition, complicating the assignment of a singular label even more. Therefore, we inspect the correlation matrix based on the estimated embeddings. This matrix can be interpreted as a generalized confusion matrix, where a high correlation refers to a high confusion risk and vice versa. The correlation values directly reflect the semantic similarities of the LCZs. A small fraction of uncertainty remains as we only inspect a limited number of images. The chosen model framework allows estimating a distribution of the embeddings, namely a multivariate Gaussian, where the variance matrix expresses the uncertainty. Third, we move away from expert labels and inspect the performance of our model on a crowd-sourced dataset, namely the multiply annotated dataset Cifar-10H. The results show that even the classification of images into well-separated and naturally distinguishable categories could benefit from using label embeddings instead of hard-coded labels. The proposed model and the estimation framework are very flexible and hence, the presented work can be easily adapted to any classification problem with multiple annotations.

These insights can be valuable in multiple regards and pave the way for future research in various directions. While the presented results already deliver interesting insights into the annotation tasks, they of course rather serve as a preprocessing step for further work. The long-term goal is to use label embeddings within a complete machine-learning framework. In particular, we are interested in training classification models on multi-dimensional embeddings instead of single labels, i.e. incorporating information about label uncertainty directly into the model. This work can also serve as a basis for analyzing different design choices for label generation for image classification problems. The trade-off between the number of instances and the number of annotators is a well-known problem, related to experimental design. For problems with a high degree of ambiguity, determined by the proposed model, acquiring more annotations instead of more instances is beneficial. Vice versa, if classification is "easy", i.e. the embeddings reflect clear class affiliations, a smaller number of annotations might be sufficient and one should concentrate on generating more labeled instances instead. We believe that our modeling framework could be of great benefit for future steps towards better handling label uncertainty for machine learning models.

Acknowledgements

The present contribution is supported by the Helmholtz Association under the joint research school "HIDSS-006 - Munich School for Data Science@Helmholtz, TUM&LMU".

References

Görkem Algan and Ilkay Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771, 2021.

Lora Aroyo and Chris Welty. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013), 2013.

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1): 15–24, 2015.

Marianna Avetisyan and Jean-Paul Fox. The dirichlet-multinomial model for multivariate randomized response data and small samples. *Psicologica: International Journal of Methodology and Experimental Psychology*, 33(2):362–390, 2012.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*, 2023.

Ruairidh M Battleday, Joshua C Peterson, and Thomas L Griffiths. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, 11(1):5418, 2020.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi:10.18653/v1/D15-1075. URL `http://aclweb.org/anthology/D15-1075`.

Gilles Celeux, Didier Chauveau, and Jean Diebolt. Stochastic versions of the EM algorithm: An experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, 55(4):287–314, 1996. ISSN 00949655. doi:10.1080/00949659608811772.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110, 2022.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Yair Dgani, Hayit Greenspan, and Jacob Goldberger. Training a neural network based on unreliable human annotation of medical images. In *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*, pages 39–42. IEEE, 2018.

Dhivya Eswaran, Stephan Günnemann, and Christos Faloutsos. The power of certainty: A dirichlet-multinomial model for belief propagation. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2017.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. doi:10.1109/TNNLS.2013.2292894.

Benoît Frénay, Ata Kabán, et al. A comprehensive introduction to label noise. In *ESANN*. Citeseer, 2014.

Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, pages 1–77, 2023.

Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. Sources of uncertainty in machine learning–a statisticians' view. *arXiv preprint arXiv:2305.16703*, 2023.

Cornelia Gruber, Katharina Hechinger, Matthias Aßenmacher, Goeran Kauermann, and Barbara Plank. More labels or cases? assessing label variation in natural language inference. In *The Third Workshop on Understanding Implicit and Underspecified Language*, 2024. URL `https://openreview.net/forum?id=9vL3GBWt9w`.

Joshua G Harrison, W John Calder, Vivaswat Shastry, and C Alex Buerkle. Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Molecular ecology resources*, 20(2):481–497, 2020.

Katharina Hechinger, Xiao Xiang Zhu, and Göran Kauermann. Categorising the world into local climate zones: towards quantifying labelling uncertainty for machine learning models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(1):143–161, 2024.

Ian Holmes, Keith Harris, and Christopher Quince. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one*, 7(2):e30126, 2012.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.

Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

Mira Jürgens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? *arXiv preprint arXiv:2402.09056*, 2024.

Christoph Koller, Göran Kauermann, and Xiao Xiang Zhu. Going beyond one-hot encoding in classification: Can human uncertainty improve model performance in earth observation? *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11, 2024. doi:10.1109/TGRS.2023.3336357.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park. MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, 42(9):22, 2011. doi:10.18637/jss.v042.i09.

Nis Meinert, Jakob Gawlikowski, and Alexander Lavin. The unreasonable effectiveness of deep evidential regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9134–9142, 2023.

Thomas Minka. Estimating a dirichlet distribution, 2000.

James E Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1/2):65–82, 1962.

Yixin Nie, Xiang Zhou, and Mohit Bansal. What Can We Learn from Collective Human Opinions on Natural Language Inference Data?, 2020. URL `http://arxiv.org/abs/2010.03532`. arXiv:2010.03532 [cs].

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019.

Joshua Peterson, Ruairidh Battleday, Thomas Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. *Proceedings of the IEEE International Conference on Computer Vision*, pages 9616–9625, 2019. doi:10.1109/ICCV.2019.00971.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682. Association for Computational Linguistics, 2022. doi:10.18653/v1/2022.emnlp-main.731.

Chunping Qiu, Lichao Mou, Michael Schmitt, and Xiao Xiang Zhu. Local climate zone-based urban land cover classification from multi-seasonal sentinel-2 images with a recurrent residual network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 154:151–162, 2019.

Chunping Qiu, Xiaochong Tong, Michael Schmitt, Benjamin Bechtel, and Xiao Xiang Zhu. Multilevel feature fusion-based cnn for local climate zone classification from sentinel-2 images: Benchmark results on the so2sat lcz42 dataset. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:2793–2806, 2020.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Ian D Stewart and Tim R Oke. Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, 93(12):1879–1900, 2012.

Dennis Thomas Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Transactions on Machine Learning Research*, 2023.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.

Ning Xu, Yun-Peng Liu, and Xin Geng. Label enhancement for label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1632–1643, 2019.

Jianhua Yin and Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242, 2014.

Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Haberle, Yuansheng Hua, Rong Huang, Lloyd Hughes, Hao Li, Yao Sun, Guichen Zhang, Shiyao Han, Michael Schmitt, and Yuanyuan Wang. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8(3):76–89, 2020. doi:10.1109/MGRS.2020.2964708.

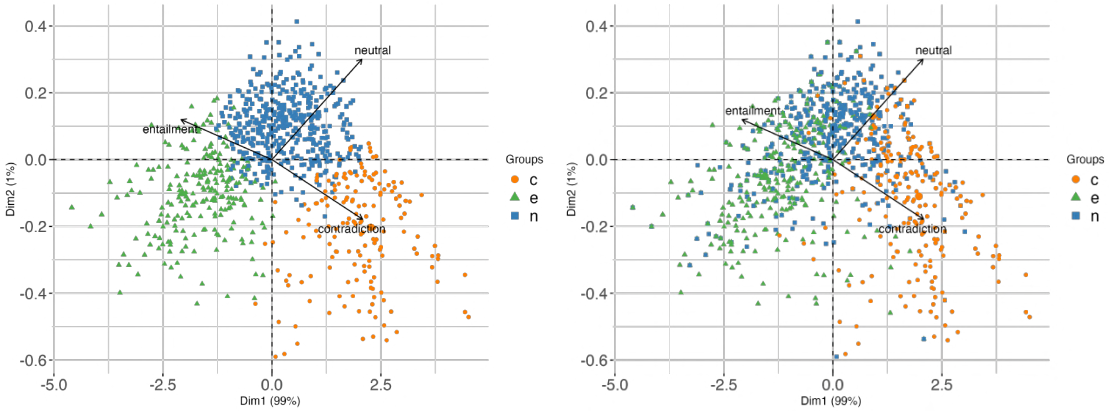


Figure 11: The biplots show the projected embeddings, colored by majority vote (left) and by original subjective ground truth label (right).

A Implementation Details

The algorithm presented in Figure 4 was initialized with

$$\mu^{(1)} = (0, ..., 0) \in R^K$$

$$\Sigma^{(1)} = \begin{pmatrix} 10 & 0 & ... & 0 & 0 \\ 0 & 10 & ... & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & ... & 10 & 0 \\ 0 & 0 & ... & 0 & 10 \end{pmatrix} \in R^{K \times K}.$$

Specifically, we employed the Metropolis Hasting algorithm implemented in the function `MCMCmetrop1R()` from the R package **MCMCpack** (Martin et al., 2011) for drawing MCMC samples (`mcmc` $= 1000$) from the logarithmic version of the posterior. Thereby, `burnin` $= 50$ samples were discarded during the burn-in phase and `thin` $= 20$ was used as thinning interval. The starting value `theta.init` was set to $\hat{z}_i^{(m-1)}$, i.e. the previously estimated embedding vector, or to $(0, ..., 0)$ in the first iteration. Additional details and the code can be found on github via `https://github.com/katharinahech/labelembeddings`

B ChaosSNLI: Majority Vote vs. Ground Truth

The dataset ChaosSNLI is especially interesting in the context of analyzing label variation and ambiguities. It does not only contain a huge number of annotations but also provides a subjective ground truth label for each pair of sentences due to its generation process, see Nie et al. [2020] for details. However, due to the general ambiguity of language, the subjective "true" label does not constitute a reliable ground truth in general. Nevertheless, it is interesting to examine the difference between the original label and the majority-voted label, as shown in Figure 11. While the left plot shows the biplot of the projected embeddings colored by the majority vote, the points in the right plot are colored according to the original "gold" label. For the latter, we see less separation of the three classes in the two-dimensional space. In particular, instances with the original label *neutral* are often classified into *contradiction* or *entailment* according to the majority vote. This observation suggests that the class "neutral" is harder to detect as the other classes are inherently more informative, see also Gruber et al. [2024].

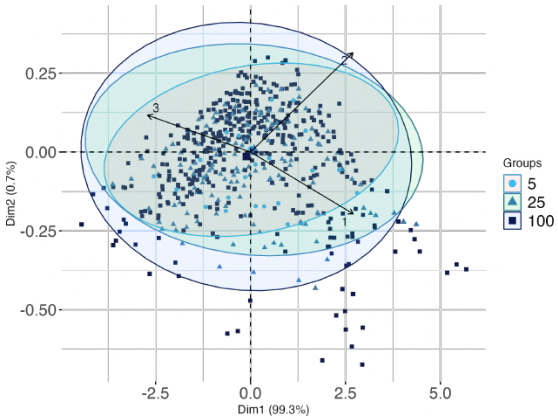


Figure 12: The biplot shows the projected estimated embeddings for a version of the dataset ChaosNLI, where for random instances 5, 25 or 100 annotations were sampled. The points are colored according to the number of annotations and concentration ellipses are shown for easier visual exploration.

C ChaosSNLI: Impact of the Number of Annotations

The dimensions of the dataset also allow us to investigate the dependence of the results on the number of annotations, J. Following the sampling strategy proposed by Gruber et al. [2024], we randomly sample annotations for a fraction of the original dataset and thereby create a new version, consisting of observations with a varying number of annotations. Specifically, $J_{100} = 100$ annotations remain for $N_{100} = 514$ observations, while for $N_5 = 500$ resp. $N_{25} = 500$ instances $J_5 = 5$ resp. $J_{25} = 25$ annotations were randomly sampled. Figure 12 shows the biplot created based on the estimated ground truth embeddings for the sampled dataset. The plot inherently expresses the variance of the instances, comparable to Figure 3. Points located close to the center of the graphic exhibit a high degree of variance. In contrast, instances that lie far from the origin have a smaller variance, as shown for $K = 2$ in Figure 3. Interestingly, Figure 12 clearly reveals that observations with fewer annotations lie closer to the center of the plot compared to similar instances with a higher number of annotations, indicating higher variance and higher uncertainty. This is of course a highly desirable property of the embedded ground truth vectors. The estimated values express both, the uncertainty due to the disagreement of the labelers as well as the uncertainty due to an insufficient number of annotations. Figure 12 additionally hints at the fact that a certain number of annotations is necessary to fully capture the associated ambiguity (Gruber et al., 2024). The concentration ellipses for J_{100} and J_{25} are almost similar, whereas the ellipse for J_5 is notably smaller, indicating a higher variance of instances with only five annotations.

D So2Sat LCZ42: Standard Deviation of the Correlation Matrix

As mentioned in Section 3, the MCMC samples enable us to additionally inspect the variance of the correlations of the estimated embeddings. Figure 13 explicitly shows the respective variances of entries of the correlation matrix given in Figure 8b.

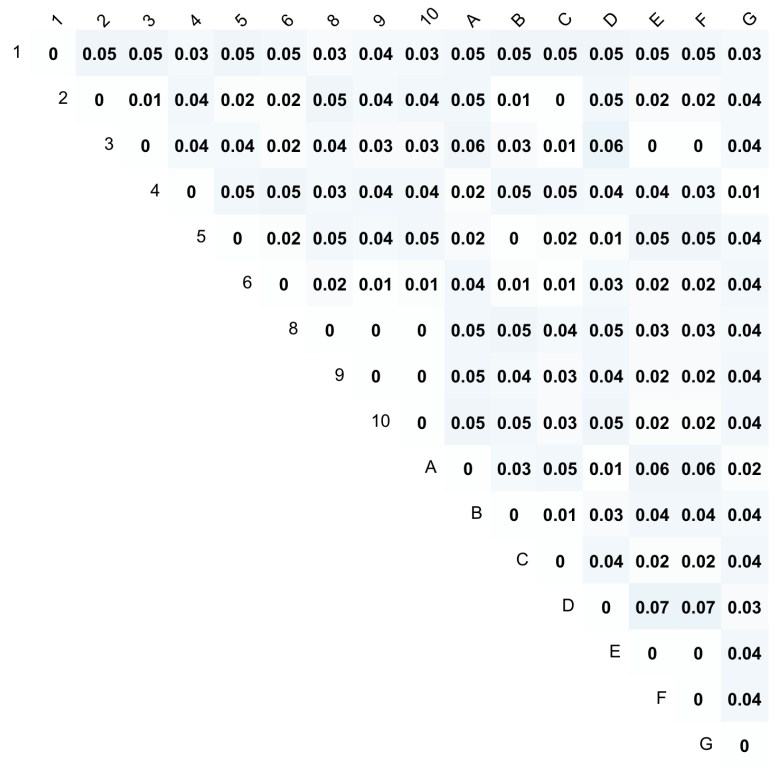


Figure 13: Standard deviation of the entries of the correlation matrix.

Contributing Publications

Hechinger, K., Zhu, X.X. and Kauermann, G. (2024). Categorising the World into Local Climate Zones: Towards Quantifying Labelling Uncertainty for Machine Learning Models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(1), 143-161. https://doi.org/10.1093/jrsssc/qlad089.

Gruber, C., Hechinger, K., Aßenmacher, M., Kauermann, G. and Plank, B. (2024). More Labels or Cases? Assessing Label Variation in Natural Language Inference. *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language* https://aclanthology.org/2024.unimplicit-1.2.pdf

Hechinger, K., Koller, C., Zhu, X.X. and Kauermann, G. (2024). Human-in-the-loop: Towards Label Embeddings for Measuring Classification Difficulty. *arXiv preprint arXiv:2311.08874*. Under review in *Journal of Computational and Graphical Statistics.*

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 22.07.2024 Katharina Hechinger