SANDIA REPORT

SAND2017-6776
Unlimited Release
Printed June, 2017

Uncertainty Quantification for Machine Learning

David J. Stracuzzi, Maximillian G. Chen, Michael C. Darling, Matthew G. Peterson, Charles Vollmer

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology and Engineering Solutions of Sandia, LLC.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online ordering:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2017-6776
Unlimited Release
Printed June, 2017

Uncertainty Quantification for Machine Learning

David J. Stracuzzi
Data-Driven and Neural Computing Department
Sandia National Laboratories
djstrac@sandia.gov

Maximillian G. Chen
Mission Analytics Solutions 2 Department
Sandia National Laboratories
mgchen@sandia.gov

Michael C. Darling
Data-Driven and Neural Computing Department
Sandia National Laboratories
mcdarli@sandia.gov

Matthew G. Peterson
Scalable Analytics and Visualization Department
Sandia National Laboratories
mgpeter@sandia.gov

Charles Vollmer
Data-Driven and Neural Computing Department
Sandia National Laboratories
cvollme@sandia.gov

Abstract

In this paper, we assert the importance of uncertainty quantification for machine learning and sketch an initial research agenda. We define uncertainty in the context of machine learning, identify its sources, and motivate the importance and impact of its quantification. We then illustrate these issues with an image analysis example. The paper concludes by identifying several specific research issues and by discussing the potential long-term implications of uncertainty quantification for data analytics in general.

Acknowledgements

This work was funded by the Sandia National Laboratories Laboratory Directed Research and Development (LDRD) program. The authors thank Kristina Czuchlewski, Stephen Dauphin, Rich Field, and Jim Stewart for helpful comments and discussion.

Contents

List of Figures

Chapter 1

Introduction

Machine learning establishes statistical relationships between observable variables and some unobservable state of the world. Such mappings are important for domains in which theoretical mappings are incomplete or missing, ranging from image analysis to cyber security to climate modeling. However, statistical models are fundamentally stochastic: predicted and inferred values are random variables and inherently uncertain. A rigorous evaluation of uncertainty can improve the precision and quality of model predictions, the model sensitivity to rare or weakly indicated phenomena, and end-user decision making processes.

Uncertainty quantification provides a measure of sufficiency of the available data and the selected modeling approach for answering a question of interest. Currently, such determinations depend heavily on expert opinion, using a mix of domain and modeling expertise, and accuracy-based validation metrics, such as precision-recall or ROC curves. Model induction is an ill-posed problem (Hadamard, 1923) however, meaning that the resulting model parameterizations may not be unique and may be highly sensitive to the input data. Accuracy-based validation metrics do not address questions of uniqueness or stability, so these issues often go unexamined.

In this paper, we propose uncertainty quantification as a research area that requires attention by the machine learning community. Although many with a background in statistics are familiar with uncertainty, most machine learning research focuses on model induction and performance evaluation as opposed to the rigorous evaluation of the suitability of a learned model to a particular task. Our goal is to demonstrate that uncertainty analysis of learned models provides insight and information that is not otherwise available, and to provide an overview of research needed to bring uncertainty quantification into mainstream practice.

Chapter 2

Motivating Uncertainty Quantification

If machine learning establishes relationships between observable and unobservable variables, then uncertainty quantification characterizes the variability in those relationships. Variability arises from multiple points within the machine learning process, but generally reduces to a single interpretation: *What is the range of possible responses that a model might make given the available data?* Explicitly quantifying variability in a model and its outputs has a number of implications for the machine learning community, which we highlight throughout the paper. In this section, we discuss what uncertainty means in a machine learning context, where it comes from, and why analyzing uncertainty is important.

2.1 Why Uncertainty Matters

Consider spell checking with automatic replacement as implemented in many smartphones, tablets, and other environments. Most approaches combine limited context and the potentially incorrect or incomplete spelling of the word in question. Spell checkers often do a mediocre job of providing corrections and completions. The results can turn a simple typo into completely nonsensical statements, or worse, alter the author's intent. We argue that such issues do not stem from poor language modeling — performance is limited given the available information and computational resources — but from a neglect of uncertainty.

For a given misspelling and context, the number of possible corrections can be large. More importantly, many of the corrections may be similarly appropriate given the available data. This case corresponds to high uncertainty — many equally plausible possible solutions — and automatically selecting a replacement amounts to random guessing among alternatives. The available information is not sufficient to strongly indicate one word over another. This is a key point of our discussion. *The output of a learned model is a random variable, and the strength of evidence that supports selection of one value over another is important.*

The issue becomes more complex for problems in which data comes from multiple sources. For example, if we observe a single geospatial location using optical, lidar, and infrared sensors, we then have a multimodal analysis problem that relies on color, height, and heat information to distinguish objects in the scene. Ideally, two or more sources may provide complementary information, which improves performance and reduces variability in detection

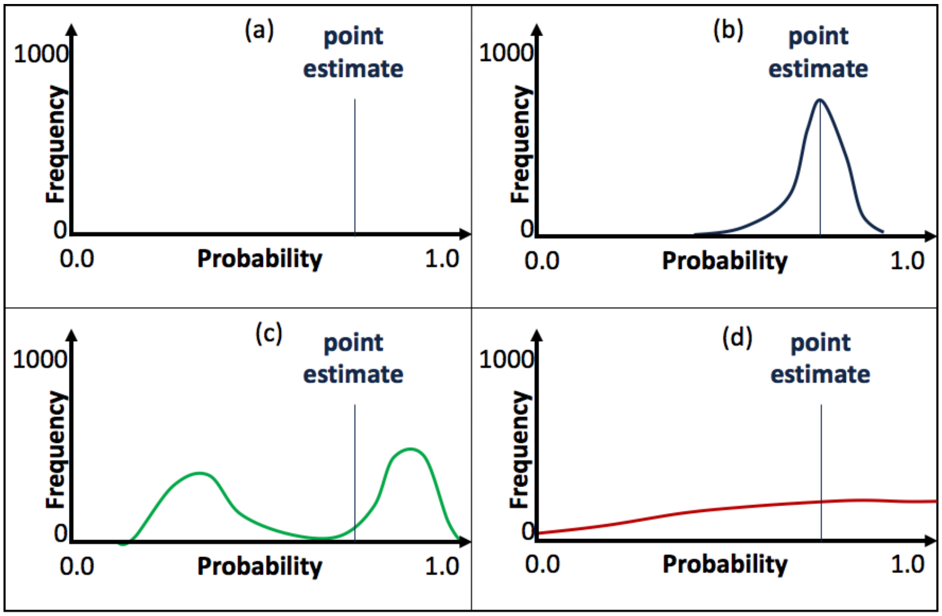


Figure 2.1. Illustration of the difference between probability and uncertainty. Each panel indicates a different uncertainty distributions that yield similar probability point estimates.

or segmentation analyses. A second possibility is that some sources contribute very little to the final result. This occurs when a distinguishing feature, such as heat, is not captured by a given modality, such as optical imagery. In some cases, data sources may disagree on identification, which, at a minimum, increases the variability in the results. Importantly, the cases outlined above can only be distinguished by evaluating uncertainty.

2.2 Probability Versus Uncertainty

Quantifying variability and characterizing a range of possible outputs naturally evokes the notion of probability. However, probability and uncertainty are two separate concepts. Probability measures the likelihood or belief in an inferred outcome, while uncertainty characterizes the range of possible outcomes or beliefs due to imperfect information. Many statistical models are probabilistic, but evaluating uncertainty requires additional analysis.

For example, given a probabilistic classifier M, uncertainty characterizes the variability in the probability estimates. Figure 2.1 illustrates the difference. The horizontal axes correspond to the probability that a particular example x belongs to class c. The point estimate shown in panel (a) then corresponds to the probability $P(x = c|M)$ assigned by M that x is a member of c.

Now suppose that we repeatedly sample our training data and construct a new classifier, M_i, for each sample. The vertical axis then indicates the frequency with which the target example is assigned a given probability (a histogram) across all of the classifiers. Panel (b) shows one possible distribution $P(x = c|M_i)$ clustered tightly around the point estimate, suggesting that the point estimate is a stable approximation of the true probability.

Figure 2.1(c) shows a second distribution over the probability values, this time bimodal. Some of the M_i produced a high value for $P(x = c|M_i)$, while others set it low. Importantly, an ensemble that votes or averages the individual probabilities produces a point estimate similar to that shown in panel (b). The bimodal curve indicates far more uncertainty and variability in the response, suggesting at least two plausible interpretations of the example.

Figure 2.1(d) shows a more uniform distribution over probabilities, which indicates a strong dependence between classifier output and the specific training sample. Again, an average over $P(x = c|M_i)$ creates a point estimate similar to the other panels. The voted probability reasonably estimates the degree to which our target example fits into class c, but it does not characterize the variability in that estimate. For many practical machine learning applications, the variability matters as we illustrate throughout the paper.

2.3 Performance Versus Uncertainty

Traditional performance evaluation metrics also differ from uncertainty quantification. Confusion matrices, ROC curves, and F-Scores all use the true and false positive and negative rates to quantify a classifier's performance. These methods provide a global measure of a classifier's ability to discriminate among examples of different classes. However, none of these methods accounts for the variability in a classifier's output.

Returning to Figure 2.1, the models in panels (b-d) all predict the same class assignment and get the same credit for a correct response. Given a slightly different example or a slightly different training set, the models in (b) would tend to remain stable in their predictions while those in (c) or (d) might vary substantially. The outputs in the latter cases resemble (biased) random guessing. Traditional performance evaluation does not distinguish between a lucky guess and a consistently correct response. Thus, a model that exhibits high discrimination and high uncertainty is far less reliable than one that exhibits low uncertainty. Minor changes in the training or test data can lead to large and unpredictable shifts in performance for an uncertain model.

2.4 Uncertainty Quantification

Quantifying the uncertainty in a learned model relative to a given input example requires first identifying the different sources of error and then combining those sources into an overall uncertainty in the predicted quantity. Figure 2.2 shows a breakdown of the problem

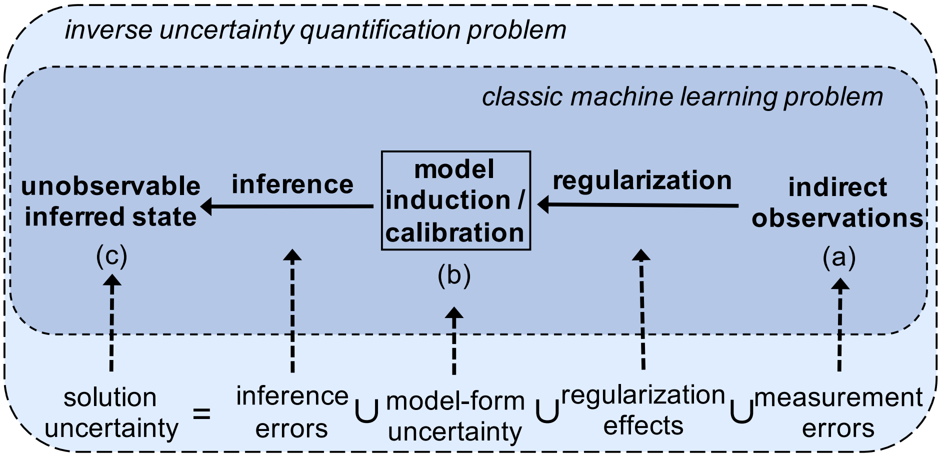


Figure 2.2. The relationship between the classic machine learning problem and inverse uncertainty quantification. The steps of the standard machine learning task and their associated sources of uncertainty. Uncertainty in model predictions combines the uncertainties associated with each machine learning processing step.

space. To illustrate the issues involved, we first review the basic steps of a machine learning application, and then highlight the sources of uncertainty in the context of an example.

Machine learning maps observed data (a) to unobservable properties of interest (c). For example, seismic event detection maps noisy waveforms into signal onset times by learning a model (Figure 2.2b) of the data from which the onset time can be estimated. In this case, one model captures the noise preceding the signal while a second includes both signal and noise. A learning algorithm optimizes the fit of each model such that the time point at which the two models meet represents the onset.

Many applications preprocess data, such as by a low-pass filter in the seismic example, and bias the model, such as by a prior, to push the resulting parameterization toward specific areas of the solution space (called regularization) to improve model fit. Likewise, inference procedures map examples through the model into output values. Depending on the specific model used, inference may only be approximate. Taken together, these steps comprise the major design decisions of a machine learning problem (inner box of Figure 2.2), and they mirror the *inverse problem* frequently solved in natural science and engineering design domains.

The difference between classical inverse problems and machine learning is that the model in Figure 2.2b would traditionally be composed of theoretical equations that define the mapping from observations to unobservables instead of an induced statistical model. Uncertainty arises from many sources, as shown in the bottom row of Figure 2.2, and must be quantified and aggregated to fully characterize the modeled system (outer box), known as the *inverse uncertainty quantification problem.* The problem is well-studied for cases in which an equation-based model is known (see Smith, 2014, for example). The switch to stochastic models, however, raises new questions about uncertainty quantification methods.

Returning to the seismic example, measurement errors arise at the seismograph, which requires calibration, receives vibrations from non-seismic sources such as wind, and has limited sensitivity. Regularization effects arise from data preprocessing, which removes irrelevant signals, but may also alter the seismic information. Model-form uncertainty arises from the learning process: many plausible model parameterizations exist and each provides a slightly different output. The issue is exacerbated if we consider different classes of models, such as different types of autoregressive models in this case. All of these sources, plus any inference errors as noted above, combine to produce an uncertainty in the output value. To fully characterize a seismic onset, we must extract a probability density function over possible onset times, which reveals alternative plausible hypotheses and their relative likelihood, as opposed to just the most likely onset.

Finally, note that confidence intervals provide only a weak approximation of uncertainty. For example, confidence intervals placed around the curves in Figure 2.1 might show that one distribution has greater variance than another, but still hides the underlying shape. While sometimes sufficient, we argue that the shape is important in general.

Chapter 3

A Preliminary Data-Driven Uncertainty Analysis Approach

To illustrate the importance of uncertainty quantification in machine learning applications, we consider an image analysis case study. In this section, we outline a simple analysis framework, demonstrating its use on a specific example in Section 4. Our preliminary work focuses almost entirely on quantifying model-form uncertainty. In section 5, we return to the general problem of uncertainty analysis to outline research questions associated with other sources of uncertainty and longer-term implications for the work.

To quantify uncertainty in unsupervised image classification, we use the Gaussian Mixture Model (GMM) with bootstrap sampling. Our approach builds on the Bayesian pixel classification methods surveyed by Falk et al. (2015). Importantly, given that the goal is to illustrate uncertainty quantification, we set aside discussion of which model is most appropriate to the task. We choose to use the GMM and bootstrap since they are well known and allow our discussion to focus on uncertainty analysis rather than the model itself. A variety of Bayesian methods are capable of producing the types of posterior distributions that we discuss below. However, this is rarely done in practice, and often requires either additional processing or even modification of the underlying algorithms.

3.1 Setting

Consider N data sources, D^i for $i = 1, \ldots, N$, such that each source consists of a regularly spaced lattice indexed by $j = 1, \ldots, n$. Each index refers to a specific pixel which may belong to one of K categories. We use *category* to refer to unsupervised clusters and *class* to refer to the semantic tags used to label the categories. Classes can be assigned either manually or through a supervised labeling process, which we defer to future work.

For simplicity and without loss of generality, we assume that the N data sources are combined into observations $D = \mathbf{y}_1, \ldots, \mathbf{y}_n$, such that each $\mathbf{y}_j$ represents the concatenated data vectors from all sources. We revisit this assumption in the discussion of future work. Each $\mathbf{y}_j$ is a noisy observation of pixel j. The true category for each pixel, z_j, is not directly observable, but is a random variable conditioned on the observations and assumed model.

Our goal is the posterior estimation of the model parameters, and therefore the true category, conditioned on the observations. To estimate the desired posteriors, we take a two-step approach. First, we use a bootstrap procedure to resample the original image data, $D = \mathbf{y}_1, \ldots, \mathbf{y}_n$, and obtain multiple samples that allow us to estimate the distribution of the probability, $P(z_j = k|D)$, that k is the true category of pixel j. Second, we fit a GMM to each bootstrapped sample, obtaining point estimates for $P(z_j = k|D)$ from each sample. The aggregation of point estimates provides the posterior distribution over category probabilities.

3.2 Gaussian Mixture Model

Given data D with independent observations $\mathbf{y}_1, ..., \mathbf{y}_n$, the likelihood for a mixture model with K components is

$$\mathcal{L}_{MIX}(\theta_1, ..., \theta_K; \tau_1, ..., \tau_K|\mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \tau_k f_k(\mathbf{y}_i|\theta_k)$$

where f_k and θ_k are the respective density and parameters of the k^{th} components in the mixture and τ_k is the probability that an observation belongs to the k^{th} component ($\tau_k \geq 0; \sum_{k=1}^{K} \tau_k = 1$). Most commonly, f_k is the multivariate normal (Gaussian) density ϕ_k, which is parameterized by its mean μ_k and covariance Σ_k, with probability distribution function (pdf):

$$\phi_k(\mathbf{y}_i|\mu_k, \Sigma_k) \equiv \frac{\exp\{-\frac{1}{2}(\mathbf{y}_i - \mu_k)^T \Sigma_k^{-1}(\mathbf{y}_i - \mu_k)\}}{\sqrt{\det(2\pi\Sigma_k)}}. \tag{3.1}$$

The Expectation-Maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997) finds the maximum-likelihood estimate for unknown parameters in a data model. The data, which consists of n multivariate observations and is denoted by $\mathbf{X} = (\mathbf{x}_i, i = 1, ..., n)$, is recoverable from the parameter space $(\mathbf{Y}, \mathbf{Z})$, in which $\mathbf{Y} = (\mathbf{y}_i, i = 1, ..., n)$ is observed and $\mathbf{Z} = (\mathbf{z}_i, i = 1, ..., n)$ is unobserved.

If the $\mathbf{x}_i$ are independent and identically distributed (iid) according to a probability distribution f with parameters θ, then the *complete-data likelihood* is

$$\mathcal{L}_C(\mathbf{x}_i|\theta) = \prod_{i=1}^{n} f(\mathbf{x}_i|\theta).$$

Further, if the probability that a particular variable is unobserved depends only on the observed data $\mathbf{y}$ and not on $\mathbf{z}$, then the *observed-data likelihood*, $\mathcal{L}_O(\mathbf{y}|\theta)$, can be obtained by integrating $\mathbf{z}$ out of the complete-data likelihood,

$$\mathcal{L}_O(\mathbf{y}|\theta) = \int \mathcal{L}_C(\mathbf{x}|\theta) d\mathbf{z}. \tag{3.2}$$

The MLE for θ based on the observed data maximizes $\mathcal{L}_O(\mathbf{y}|\theta)$.

The EM algorithm alternates between two steps, an "E step" and an "M step." In the "E step," the conditional expectation of the complete data log-likelihood given the observed data and the current parameter estimates are computed, i.e. compute

$$Q(\theta|\theta^{(t)}) = E_{\mathbf{Z}|\mathbf{X},\theta^{(t)}}[\log L(\theta; \mathbf{X}, \mathbf{Z})]. \tag{3.3}$$

In the "M step," parameters that maximize the expected log-likelihood from the E step are determined, i.e. compute

$$Q^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}). \tag{3.4}$$

The unobserved portion of the data may involve values that are missing due to nonresponse and/or quantities that are introduced to reformulate the problem for EM. Under fairly mild regularity conditions, EM can be shown to converge to a local maximum of the observed-data likelihood (e.g. (Dempster et al., 1977; Boyles, 1983; Wu, 1983; McLachlan and Krishnan, 1997)). Although these conditions do not always hold in practice, the EM algorithm has been widely used for maximum likelihood estimation for mixture models with good results.

In EM for GMMs, the data is a set of n iid multivariate Gaussian distributions, which means $\mathbf{x}_i, i = 1, ..., n$, follow the pdf given by (3.1). The "complete" data are considered to be $\mathbf{x}_i = (\mathbf{y}_i, \mathbf{z}_i)$, where $\mathbf{z}_i = (z_{i1}, ..., z_{iK})$ is the unobserved portion of the data. , with

$$z_{ik} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ belongs to category } k \text{ ,} \\ 0 & \text{otherwise.} \end{cases}$$

Assuming that each $\mathbf{z}_i$ is iid according to a multinomial distribution of one draw from K categories with probabilities $\tau_1, ..., \tau_K$, and that the density of an observation $\mathbf{y}_i$ given $\mathbf{z}_i$ is given by $\prod_{k=1}^{K} \phi_k(\mathbf{y}_i|\theta_k)^{z_{ik}}$, the resulting complete-data log-likelihood that we want to maximize with the MLEs of θ_k is

$$l(\theta_k, \tau_k, z_{ik}|\mathbf{x}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log[\tau_k \phi_k(\mathbf{y}_i|\theta_k)].$$

We wish to estimate θ_k that maximizes $l(\theta_k, \tau_k, z_{ik}|\mathbf{x})$, i.e.,

$$\begin{aligned} &\max_{\theta_k \in \Theta} \quad l(\theta_k, \tau_k, z_{ik}|\mathbf{x}), \\ \Theta = \{&(\mu_k, \Sigma_k, \tau_k) : \mu_k \in \mathbb{R}^d, \Sigma_k = \Sigma_k^K > 0, \Sigma_k \in \mathbb{R}^{d \times d}, \\ &\tau_k \geq 0, \sum_{k=1}^{K} \tau_k = 1\}. \end{aligned}$$

In EM for GMMs we can estimate z_{ik}, the conditional probability that observation i belongs to category k in the E-step with the estimate:

$$\hat{z}_{ik} = \frac{\hat{\tau}_k \phi(\mathbf{y}_i; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \hat{\tau}_j \phi(\mathbf{y}_i; \mu_k, \Sigma_k)}.$$

In the M-step, (3.5) is maximized in terms of τ_k and θ_k with z_{ik} fixed at the values computed in the E step, $\hat{z}_{ik}$. The resulting estimates for the category probabilities, means, and covariances are

$$\hat{\tau}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ik}, \quad \hat{\mu}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ik} \mathbf{y}_i}{\sum_{i=1}^{n} \hat{z}_{ik}}, \tag{3.5}$$

$$S_k = \frac{\sum_{i=1}^{n} \hat{z}_{ik} (\mathbf{y}_i - \mu_k)(\mathbf{y}_i - \mu_k)^T}{\sum_{i=1}^{n} \hat{z}_{ik}}. \tag{3.6}$$

(Fraley and Raftery, 2002)

With the above estimated values, we can compute any of the following parameters that estimate probability and uncertainty:

1. z^*_{ik}: estimated conditional probability that observation i belongs to category k
2. $1 - z^*_{ik}$: measure of the uncertainty that observation i belongs to category k
3. $\{j | z^*_{ij} = \max_k z^*_{ik}\}$: classification of observation i to the category with the maximum estimated conditional probability
4. $1 - \max_k z^*_{ik}$: measure of the uncertainty in the classification

3.3 Bootstrap Method

The bootstrap method considers the following problem, initially studied by Efron (1979). Given a random sample $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ from an unknown probability distribution F, estimate the sampling distribution of random variable $R(\mathbf{Y}, F)$ on the basis of the observed data, $\mathbf{y}$. To do so, we re-sample $\mathbf{y}$ with replacement to get a new sample $\mathbf{y}^*$ from the pool of observed data.

In the image analysis problem, we have only one image from each data source. To compute the uncertainty of the class probabilities, we can generate many similar images from which we can estimate a distribution over class probabilities. After using the bootstrap method to generate the sample of images, we can obtain sampling distributions and statistics for the model parameters, τ_k, μ_k, Σ_k, and the pixel categories $\hat{z}_i k$. Note that this approach only captures the model-form uncertainty.

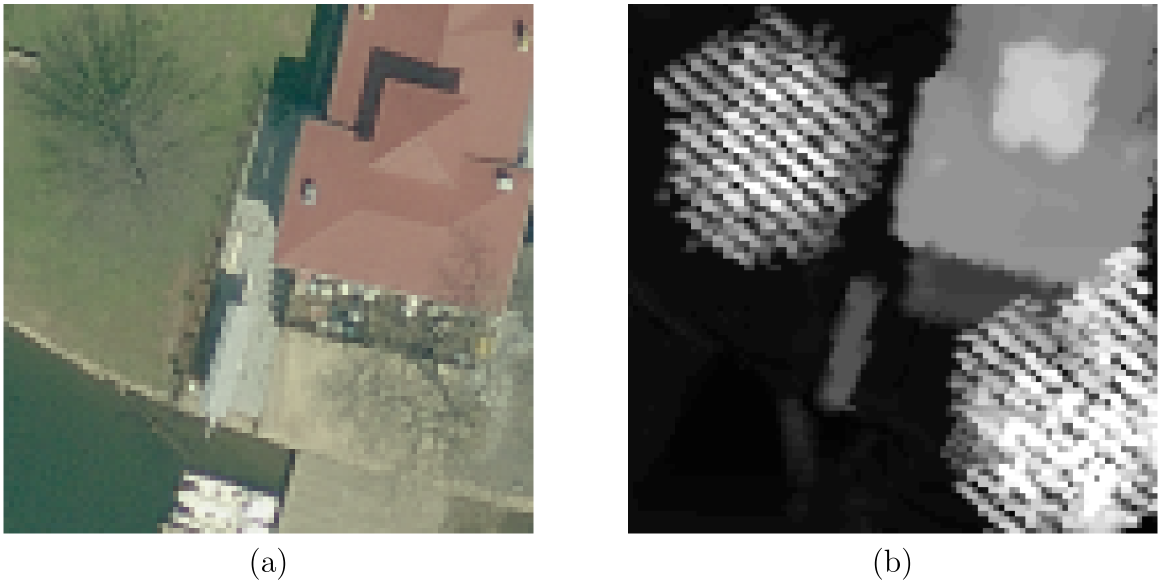
The procedure is as follows. First, obtain S bootstrap samples by sampling the observed data $\mathbf{y} = \mathbf{y}_1, ..., \mathbf{y}_n$ with replacement and obtaining n pixels. Denote the bootstrap samples $\mathbf{y}^{*1}, \mathbf{y}^{*2}, ..., \mathbf{y}^{*S}$. Then, for each of the S bootstrap samples, fit a GMM and obtain estimates for the parameter z_{ik} using the EM algorithm. The S values of $\hat{z}_{ik}$ provide a sampling distribution for $\hat{z}_{ik}$.

Chapter 4

Demonstration on Optical and Lidar Data

We demonstrate uncertainty quantification on a multimodal imagery analysis task. The images cover a small region in Philadelphia that contains trees, grass, water, pavement, a building, and a variety of small, undetermined objects. Figure 4.1a shows the 100x100 pixel optical image of the target area. Each pixel contains red, green, and blue values scaled from 0 to 1. Figure 4.1b shows the same region imaged with lidar (Light Radar) which has been preprocessed into a height map (lighter colors indicate taller data points). We use the data as prepared by O'Neil-Dunne et al. (2013) and merge the two sources into a single, four-dimensional vector containing the R, G, B, and height-above-ground values.

We apply the GMM with $K = 5$ categories and $S = 1,000$ bootstrap iterations to the data shown in Figure 4.1. A common representation of the most likely category for each


(a)

(b)

Figure 4.1. 100x100 pixel images of target area captured by optical (a) and lidar (b) sensors.

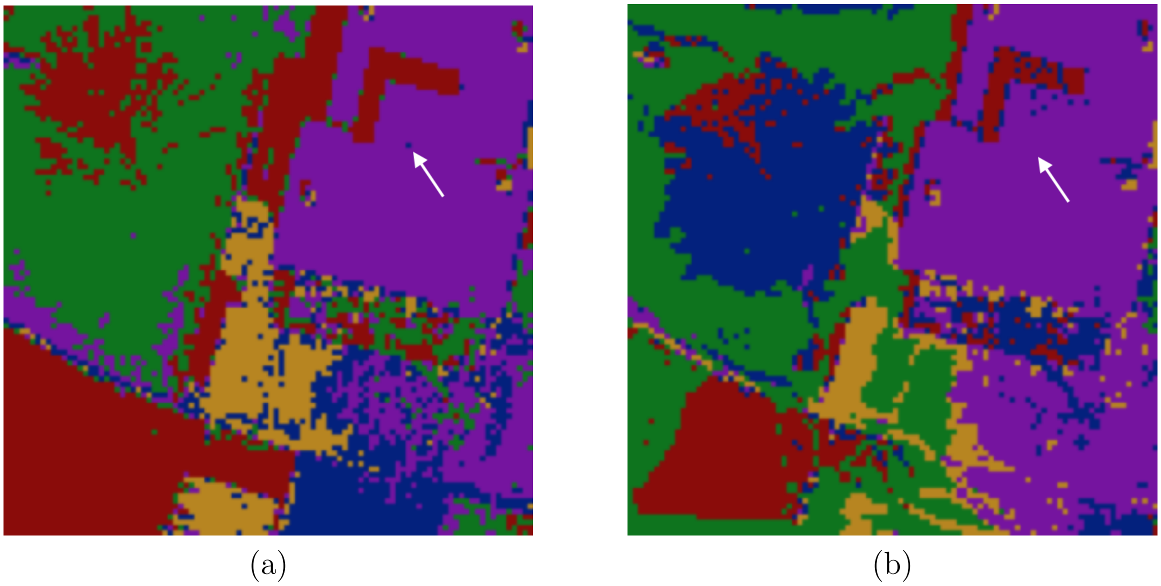


Figure 4.2. Most likely categories assigned to each pixel for optical only (a) and combined optical with lidar (b). The arrow identifies a roof pixel that changes in both category and uncertainty with the addition of lidar data.

pixel assigns a unique color to each category, as illustrated in Figure 4.2. Panel (a) shows the category estimates when using only optical data while panel (b) shows the categorization when using both the optical and lidar sources. Note that the two trees in the scene become clearer while the boats and the water become less defined. The improved tree recognition follows from lidar's ability to detect the branches, even without foliage, while the loss of definition in the water follows from the typically poor performance of lidar over water.

Figure 4.2 provides no information about the probabilities associated with each category or the uncertainties. We can capture the former by applying Shannon entropy (Shannon, 2001) to the probability distribution over the five categories. For a discrete random variable X with possible values $\{x_1, ..., x_n\}$ and probability mass function $P(X)$, entropy is: $H(X) = \sum_{i=1}^{n} -P(x_i) \ln P(x_i)$. Entropy is maximized if the probability mass is equally distributed across all of the available categories and minimized when all of the mass is assigned to a single category.

We can use per-pixel entropy scores to produce a grayscale image in which the brightness indicates the entropy value, as illustrated in the top row of Figure 4.3. The same idea applies to the category-colored images by adjusting the colors toward white based on the entropy as shown in the bottom row. Lighter shades indicate greater entropy. The resulting images identify (approximately) the degree to which pixels strongly identify with a particular category.

Based on the optical image alone (panel a), the largest areas of category confusion (high

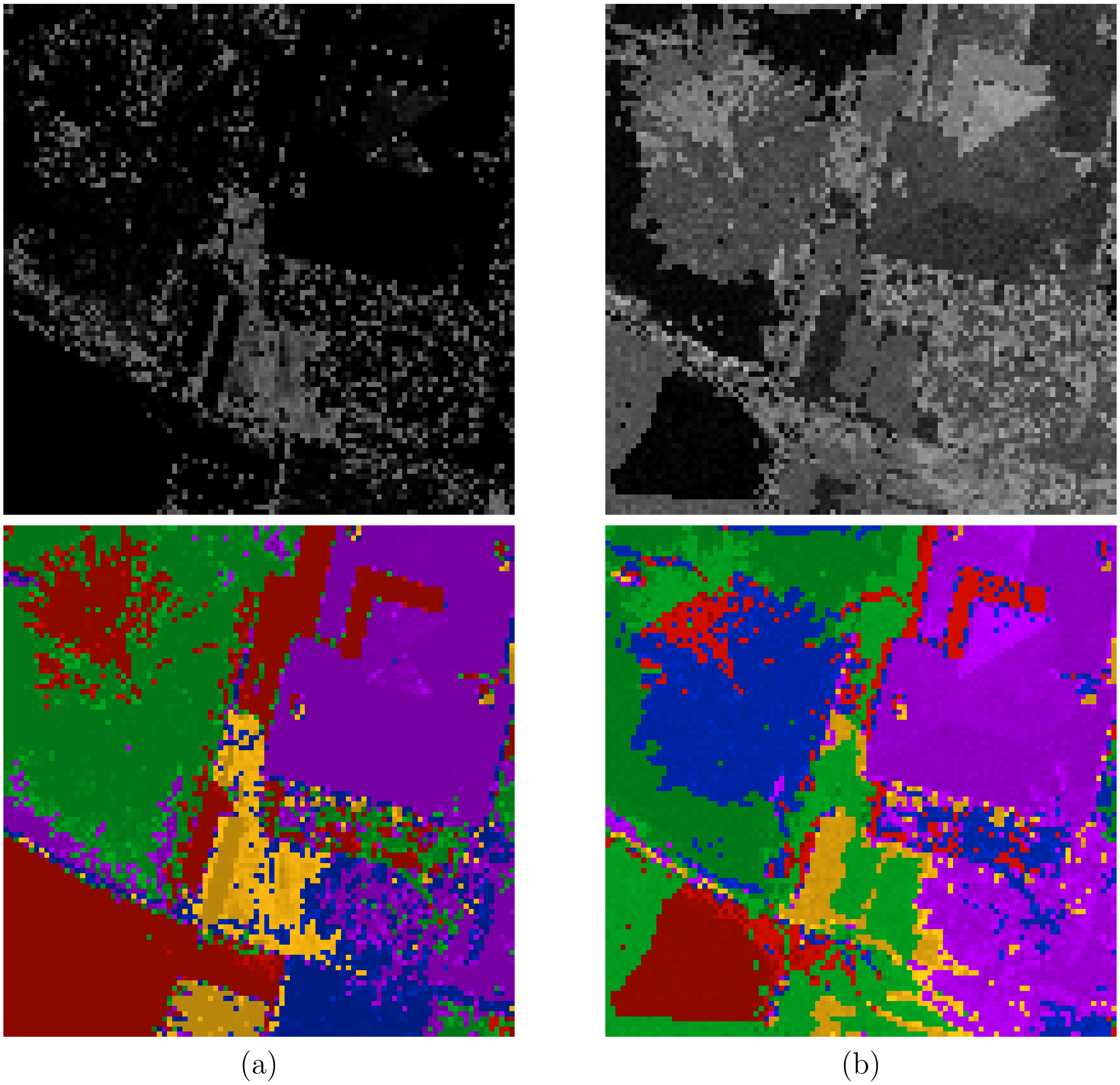


Figure 4.3. Category confusion plots showing the entropy only (top row) and the entropy overlaid onto the most probable category (bottom row) for optical (a) and combined (b) imagery.

entropy) center on the concrete near the bottom of the image, while the trees indicate very little confusion. However, the addition of lidar data in panel b increases the confusion substantially across the entire image, even while the results in many areas represent an improvement over the optical data alone. Note in particular the confusion over the roof and shadow areas. The analysis is struggling with the relatively steep changes in height associated with these areas without a corresponding change in color, which makes assigning a category difficult.

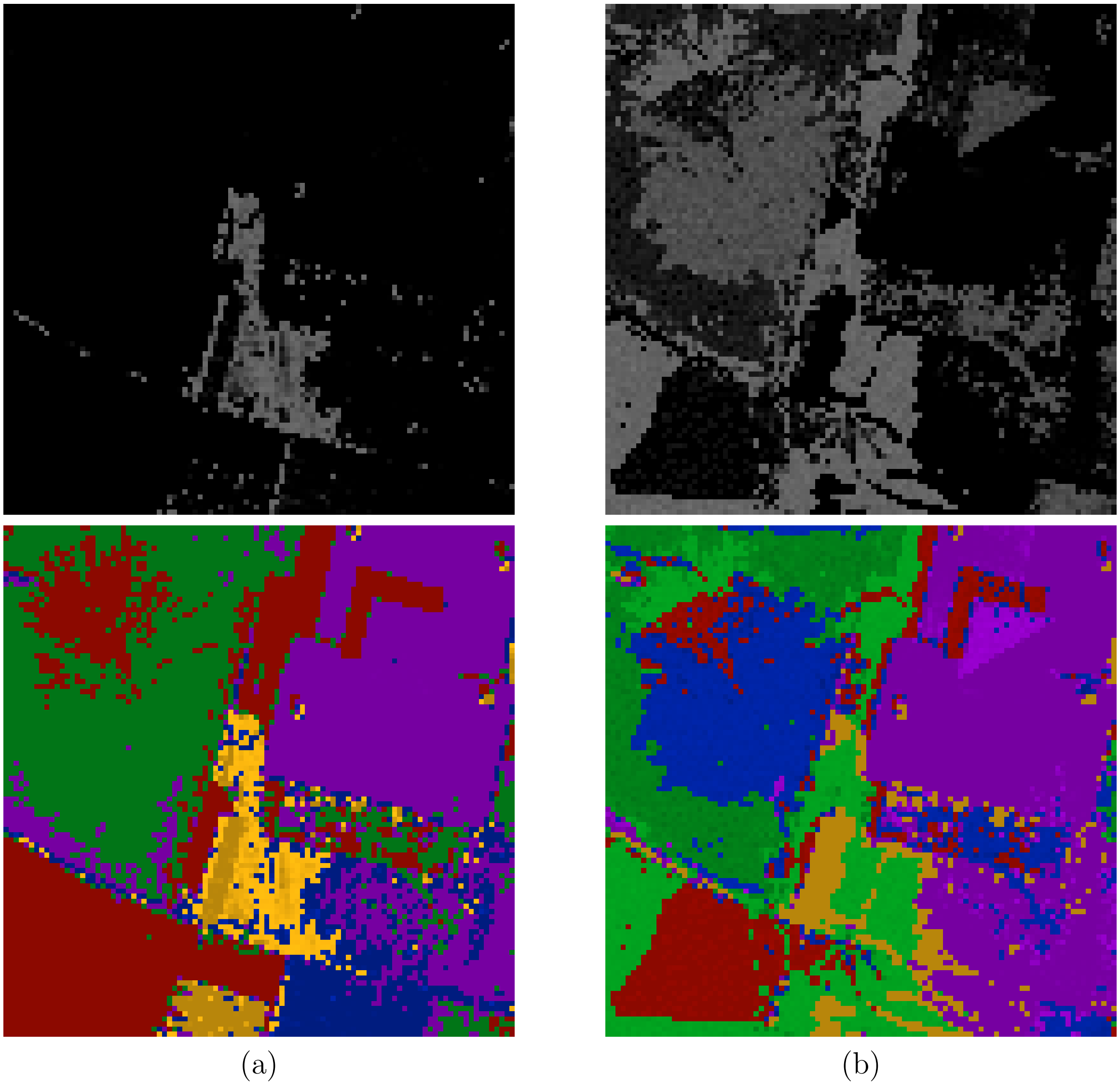


Figure 4.4. Category uncertainty plots showing the standard deviation of the posterior distribution alone (top row) and overlaid onto the most probable category (bottom row) for optical (a) and combined (b) imagery.

Importantly, the images in Figure 4.3 do not indicate variability in the probability values. The mean category probabilities do not provide information about the shape of the associated posterior distributions. We can approximate the width of the posteriors by plotting the standard deviations for each pixel's most probable category. Increasing variance results in lighter (grayscale) pixel values. As with the confusion plots, we can overlay the uncertainty information onto the colored category images as shown in Figure 4.4.

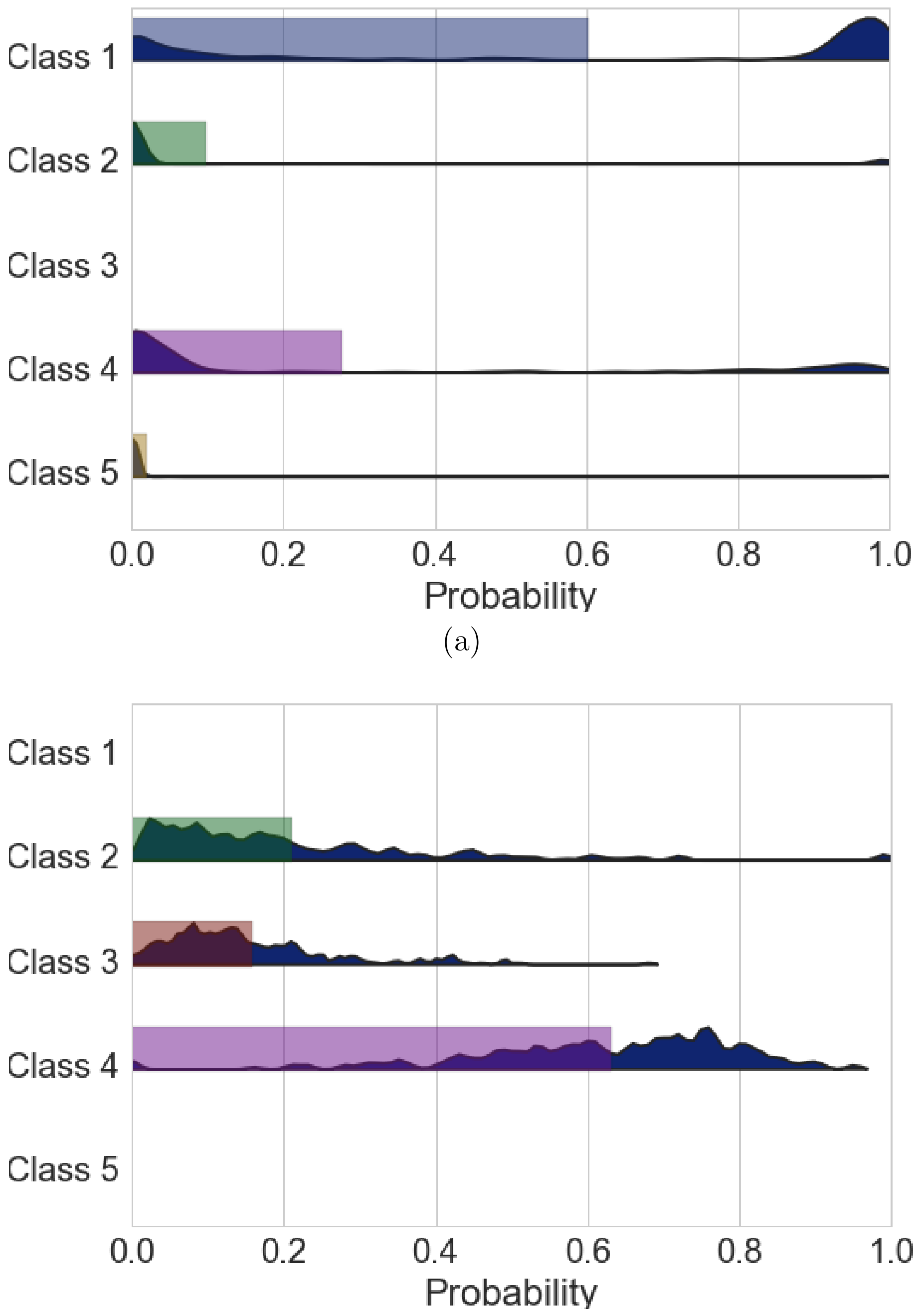
As with the entropy plots, uncertainty in the optical image is mostly confined to the concrete. However, comparing panel (b) of Figures 4.3 and 4.4 shows the difference between uncertainty and confusion. Although the roof shows high entropy in the category probabilities, the uncertainty is generally low. This implies that several categories received some probability mass, but the underlying posteriors are not broadly distributed. In other words, even though none of the categories provides a perfect match, the most likely category assignment should be reliable. This is not true of areas around the pavement, water, or the upper left tree, all of which show high uncertainty.

Each of the described visualizations provides a crude estimate of the confusion or uncertainty associated with pixel category posterior distributions. To fully examine the posteriors, we create modified violin plots which show kernel density estimations of the posteriors for each category individually, as illustrated in Figure 4.5. The shaded bars indicate the mean probability for each category. Though not practical in general, the violin plots provide a detailed view of important data points, such as those with high uncertainty or those near category boundaries.

Figure 4.5 shows the violin plots for the roof pixel identified by the arrow in Figure 4.2. Note the change in most likely category when lidar is included. Panel (a) shows that three of the posteriors for the optical analysis, categories 1, 2, and 4, have bimodal characteristics. Category 1, which differs from the other roof pixels, clearly dominates the other categories. However, category 4, which does agree with the other roof pixels, also includes a secondary mode that suggests that in some of the bootstrapped samples, category 4 was the best fit. For cases such as Figure 4.5a, in which some of the categories have a bimodal structure but the secondary modes are small, the class probabilities are sufficient.

Adding the lidar data in panel b changes the both the result and the utility of the uncertainty results substantially. Now category 4 is the most likely, and the category probability is higher than the best in panel a. However, the posteriors are distributed broadly, meaning that even though category 4 has a relatively high probability, the assignment is less certain than in the optical-only case. The variable posteriors associated with the categories in the lower panel suggest that the optical and lidar sources do not agree for this pixel.

For the case study, we can apply background knowledge about the physics of the two sensors and the scene to conclude that although the pixel may be off-color, it still belongs to the roof. Consider next a case in which we lack the domain knowledge to determine the correct interpretation manually. Then the role of uncertainty analysis becomes more important. We can use the uncertainty results to determine the stability and reliability of the learned model, even in the absence of ground truth. In the next section, we extend this argument to assessing the value or impact of each input sensor to the output.



(a)

(b)

Figure 4.5. Violin plots for the roof pixel identified in Figure 4.2 showing the change in posteriors before (a) and after (b) incorporating lidar data into the analysis.

Chapter 5

Discussion

The previous section provides a demonstration of the types of insights that uncertainty quantification can provide into a statistical model and its associated data. In this section, we look ahead to using the uncertainty information to identify how each data modality specifically contributes to the analytic results. We also identify several areas of future work required to realize these benefits.

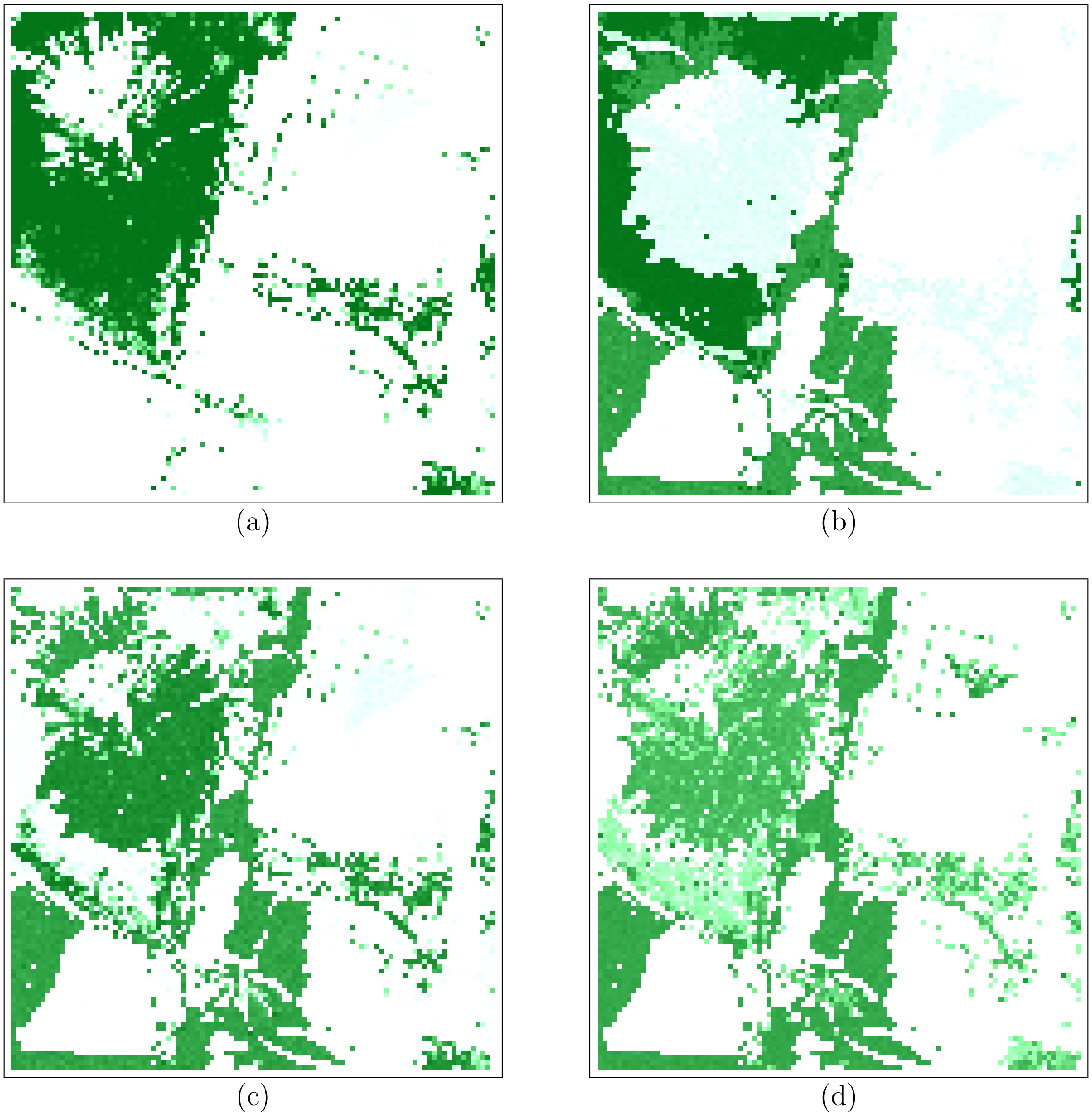
5.1 Evaluating Each Data Source

In our case study, we generate an initial analysis based on the optical data alone. When we add the lidar data and rerun the analysis, some of the pixels get assigned to different categories, some of them exhibit changes in category confusion, and some show changes in uncertainty. In financial domains, the value of adding information to an analysis is typically measured in dollars, such as expected loss or gain. In general however, methods for measuring the value of adding data to an analysis remain less well defined.

We define data from a given sensor as valuable to the extent that it changes the result of an analysis. Figure 5.1 shows a simple attempt at using the category probabilities and uncertainty information to define a notion of value for sensor data. Panel a shows a probability map for one of the five categories, corresponding roughly to grass, estimated from the optical data. The shade of each pixel in the image indicates the probability that the pixel belongs to the category, with dark green indicating high probability and white indicating zero probability. The grass is visible through the tree branches, so many of the pixels under the tree are recognized as grass in the absence of height information.

Panel b shows the probability map for the combined data. Notice that the tree has been mostly assigned to another category, but the region still gets assigned to the green category with low probability. Panel c shows the absolute value of the difference in probability between the first two panels, which corresponds to the impact on the green category of adding lidar to the optical data. Although it does not distinguish between increased and decreased probabilities, it does indicate which areas of the data are impacted by the new information.

Finally, panel d shows the Kullback-Leibler (K-L) divergence (Kullback and Leibler, 1951) between the posterior distributions for the green category at each pixel. Darker pixels


(a)

(b)

(c)

(d)

Figure 5.1. Probability maps showing green category pixel probabilities based on optical data (a), combined optical and lidar (b), and the difference between the two (c). Panel (d) shows the K-L divergence between the category posteriors associated with panels (a) and (b).

indicate greater divergence in the uncertainties due to the lidar. Although similar to the probability difference map in this case study, they are not equivalent. The degree to which these or related methods for evaluating the contribution of a data source to an analysis provide useful information remains to be seen. However, we note that both changes in probability and uncertainty may be useful for determining which data to collect, store, and analyze during future iterations of an analysis or alternate versions of an analysis (grass versus tree coverage, for example). This is particularly important in the content of remote sensing applications in which different sensors can be selected and scheduled to achieve a desired result, and background knowledge may not always be sufficient to determine the best sensor combination.

5.2 Future Work

Our initial experiments with uncertainty quantification highlight several areas for future work, many of which are related to the inverse uncertainty quantification problem problem depicted in Figure 2.2. The first relates to the Gaussian assumption of the data model. For example, in a pilot experiment (Authors, 2016) that included synthetic aperture radar data, we found that GMMs do not perform well because the Gaussian assumption is inappropriate for the modality. An alternative is to eliminate the assumption by constructing a Bayesian nonparametric mixture model (see Orbanz and Buhmann, 2008, for example). Other alternatives include the distance-dependent Chinese restaurants process (ddCRP) (Blei and Frazier, 2011; Ghosh et al., 2011), which considers the distance between pixels, and probabilistic fusion (Simonson, 1998). Extracting the posteriors from these methods is not trivial, however. A related issue concerns *regularization* methods and their impact on modeling results.

A second line of research relates to the sampling process, which has substantial impact on computational efficiency. For example, biasing each run of the GMM with the result of a prior run may reduce computation, but may also yield overly optimistic uncertainty estimates. Another issue is the relationship between the number of data observations (pixels), the number of categories, and the number of bootstrap samples required. We can also consider alternatives to bootstrapping based on the *measurement errors*. If sensor performance is well-characterized, we can sample data points from within the known range of errors instead from the actual observations, which may produce more robust estimates of *model-form uncertainty*. Other bootstrap alternatives include blockwise updates (Hall et al., 1995; Barbu and Zhu, 2005). We can also investigate Markov Chain Monte Carlo (MCMC) methods for sampling the posterior distribution, such as parametric representations (Fan et al., 2007), min-marginal energies (Kohli and Torr, 2006), and random maximum a posterior (MAP) perturbation (Hazan et al., 2013; Papandreou and Yuille, 2011), and the differences between the uncertainty results obtained via methods from sampling the data and the posterior distribution.

A third area extends from the preceding discussion of source evaluation. Having estimated the category assignment posterior distributions, an alternate formulation of the multimodal data integration problem may be possible. Earlier, we merged the data sources (after co-registration) into a single vector for each location. A more efficient approach may entail analyzing each source independently using unsupervised methods and then merging the distributions at the supervised stage by marginalizing over data sources and categories. Generating the appropriate class posteriors requires substantial additional research, but entails an important benefit. The unsupervised step is task independent, so changes to the supervised task, or changes to the data sources included in the analysis, do not require restarting the expensive mixture modeling or sampling steps.

Finally, the simple visualization methods used above require further refinement. We have argued that uncertainty analysis uncovers information about the data and model not available elsewhere. However, using this information to improve human-machine interaction and trust requires visualizations that convey the implications of the analysis results succinctly and accurately. Very little research has been conducted on uncertainty visualization.

Chapter 6

Conclusions

Our goal in this work is to demonstrate the importance of uncertainty quantification in machine learning, illustrate the implications of uncertainty quantification for the larger problem of data analysis, and to identify several areas in which additional research is needed. Broadly speaking, uncertainty analysis provides detailed insights into the variability and reliability of predictions and results generated by machine learning and statistical models. To be clear, uncertainty does not speak to the *correctness* of an output. However, in the absence of ground truth and supervised data, uncertainty does *provide a alternative measure for model evaluation.* Greater variability in output values indicates that the model has very little (or conflicting) evidence to support it's result. Importantly, the uncertainty distributions refer to a particular example; they are not aggregated across many different examples.

Aside from the implications for determining the value of information and source selection discussed above, uncertainty can also speak to broader questions of resource allocation, algorithm selection, and so on. For example, given large quantities of data and limited processing capability, allocating extra cycles to examples that have high uncertainty based on an initial analysis may be productive. Similarly, we can start considering the trade-off between the accuracy and variability in results produced by learning algorithms. Finally, the results of many machine learning problems are used as the basis for decision making by human data analysts or systems operators. Providing uncertainty information may yield deeper insight into how a model produced it's result, and the trustworthiness of that result. Given these issues, we anticipate that uncertainty quantification will play an increasing role in data analysis applications as the diversity of data and information available continues to increase.

References

Barbu, A. and Zhu, S.-C. (2005). Generalizing swendsen-wang to sampling arbitrary posterior probabilities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1239–1253.

Blei, D. M. and Frazier, P. I. (2011). Distance dependent chinese restaurant processes. *J. Mach. Learn. Res.*, 12:2461–2488.

Boyles, R. A. (1983). On the convergence of the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(1):47–50.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26.

Falk, M., Alston, C., McGrory, C., Clifford, S., Henron, E., Leonte, D., Moores, M., Walksh, C., Pettitt, A., and Mengerson, K. (2015). Recent bayesian approaches for spatial analysis of 2-D images with application to environmental modeling. *Environmental and Ecological Statistics*, 22.

Fan, A. C., Fisher, J. W., Wells, W. M., Levitt, J. J., and Willsky, A. S. (2007). *MCMC Curve Sampling for Image Segmentation*, pages 477–485. Springer Berlin Heidelberg, Berlin, Heidelberg.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631.

Ghosh, S., Ungureanu, A. B., Sudderth, E. B., and Blei, D. M. (2011). Spatial distance dependent chinese restaurant processes for image segmentation. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 1476–1484. Curran Associates, Inc.

Hadamard, J. (1923). *Lectures on Cauchy's Problem in Linear Partial Differential Equations.* Yale University Press, New Haven, CT.

Hall, P., Horowitz, J. L., and Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, 82:561–574.

Hazan, T., Maji, S., and Jaakkola, T. (2013). On sampling from the gibbs distribution with random maximum a-posteriori perturbations. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 1268–1276. Curran Associates, Inc.

Kohli, P. and Torr, P. H. S. (2006). *Measuring Uncertainty in Graph Cut Solutions – Efficiently Computing Min-marginal Energies Using Dynamic Graph Cuts*, pages 30–43. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley, New York.

O'Neil-Dunne, J. P. M., MacFaden, S. W., Royar, A. R., and Pelletier, Keith, C. (2013). An object-based system for LiDAR data fusion and feature extraction. *Geocarto International*, 28(3):227–242.

Orbanz, P. and Buhmann, J. M. (2008). Nonparametric bayesian image segmentation. *International Journal of Computer Vision*, 77(1):25–45.

Papandreou, G. and Yuille, A. L. (2011). Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *2011 International Conference on Computer Vision*, pages 193–200. IEEE.

Shannon, C. (2001). A mathematical theory of communication. *SIGMOBILE Mobile Computing and Communications Review*, 5(1):355.

Simonson, K. M. (1998). Probabilistic fusion of atr results. Technical Report SAND98-1699, Sandia National Laboratories.

Smith, R. C. (2014). *Uncertainty Quantification: Theory, Implementation, and Applications*. Computational Science and Engineering, CS12. SIAM, Philadelphia, PA.

Wu, C. F. J. (1983). On the convergence properties of the em algorithm. *Ann. Statist.*, 11(1):95–103.

DISTRIBUTION:

1	MS 0899	Technical Library, 9536 (electronic copy)

Sandia National Laboratories