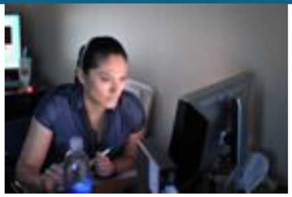
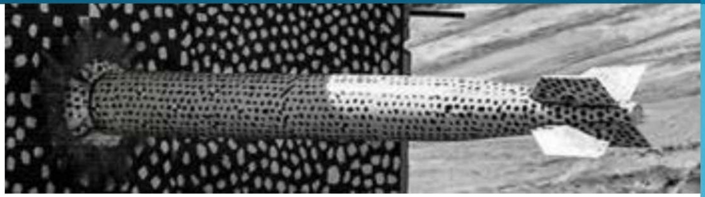
SAND2019-9591PE

UNCERTAINTY QUANTIFICATION IN MULTIMODAL IMAGE ANALYSIS





PRESENTED BY

Maximillian Chen

Johns Hopkins Applied Physics Laboratory

Acknowledgements

- Joint work with David Stracuzzi, Michael Darling, Matthew Peterson, and Charles Vollmer
- This work has been funded by the Laboratory Directed Research and Development (LDRD) Program at Sandia National Laboratories.
- 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.

Decision-Driven Uncertainty Quantification (UQ) at Sandia National Laboratories (SNL)

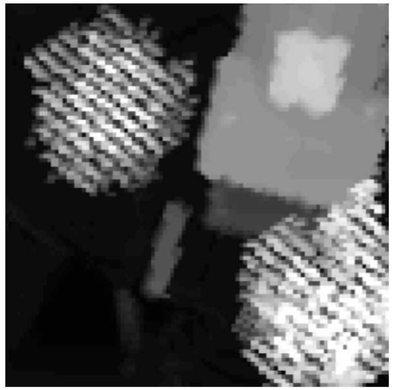
"Exceptional service in the national interest"

- Contributing novel decision-driven UQ research methods that lead to better informed and improved decision-making
- Interested in ***mission-oriented*** and ***high-risk*** national security applications

Motivating Example: Multimodal Image Analysis

- **Multimodal Imagery**: images from different sensors (e.g. optical and lidar) covering the same scene

Optical



Lidar

Core Questions:

- How do we know which images we need in order to have a complete and informative analysis?
  - Does each image help our understanding of what's going on in the scene? Or could one even hurt our understanding? *That* would be good to know.
- How do we know when to trust our analytic results?

Outline

1. Uncertainty in Statistics and Machine Learning
   - Lessons Learned from Uncertainty
2. Bootstrapping Methods
3. Uncertainty Analysis Using Gaussian Mixture Model (GMM)
4. Uncertainty Analysis Using Bayesian Consensus Clustering (BCC)
5. Conclusions

Uncertainty in Statistics and Machine Learning (ML)

Background

- Current ML algorithms and statistical models usually provide a point estimate to answer an analysis task (e.g. estimate of mean or standard deviation, one classification of an image pixel).
- Potentially many sources of varying quality
- Many choices of how to model
- Targets may be rare or hard to detect

Core Questions

- *How do we know that our data analyses are complete, accurate, and informative?*
- *How do we know when to trust our analytic results? Under what conditions?*

Rethinking Data Analysis

Much data science research focuses on

- Volume – handling very large data sets
- Velocity – data arrives and must be processed quickly
- Variety – (recent) unified handling of data from many sources

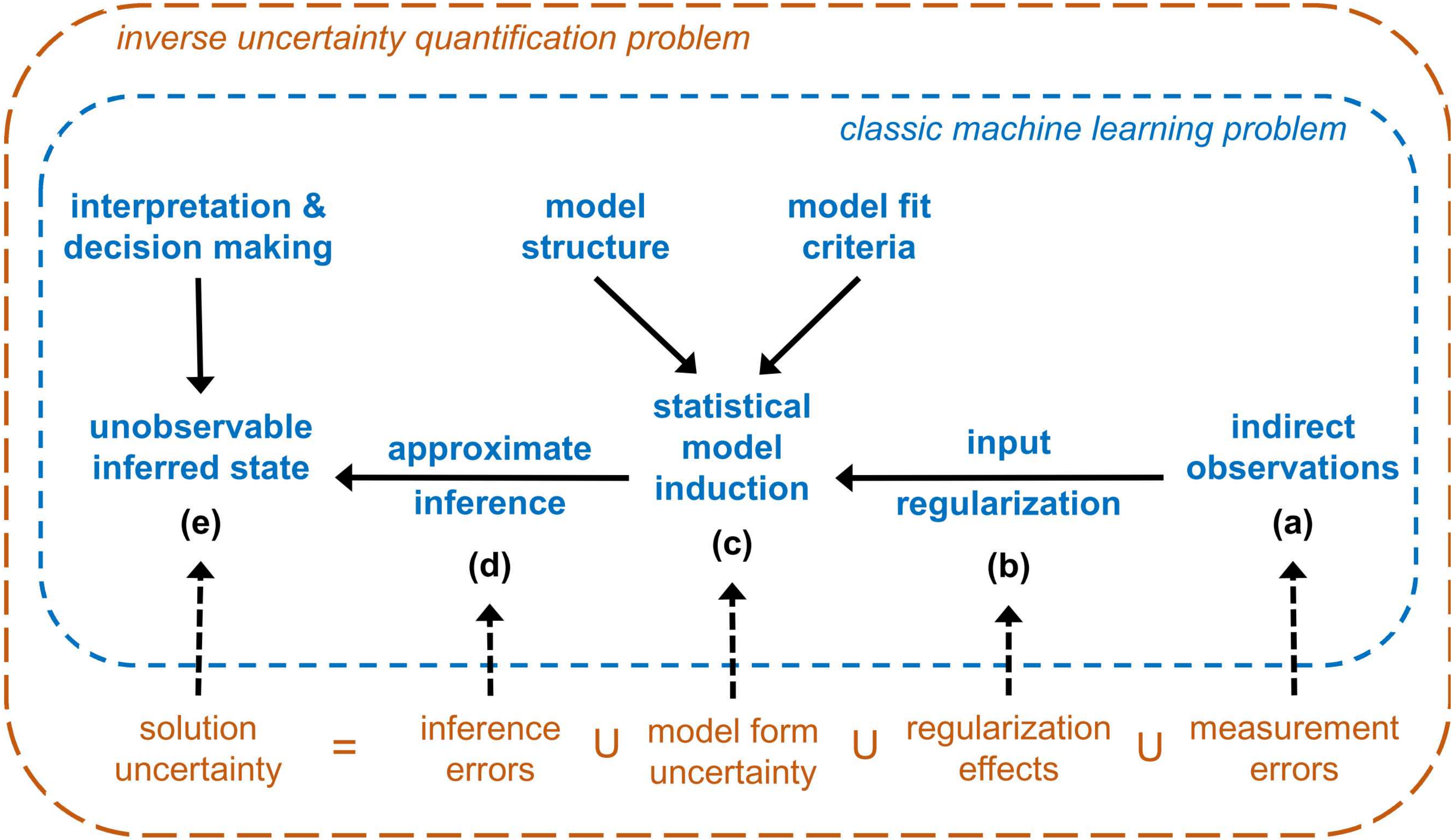
✧ Philosophy: Data-driven processing

This work focuses on

- Veracity – assessing the reliability of analytic results
- Value – assessing the contribution of data to results
- Variety – integrating sources to improve veracity and value
- Communication – make information usable by decision makers

✧ Philosophy: Decision-driven processing

Big Picture: Sources of Uncertainty



What Can We Learn from Uncertainty?

- **Data sufficiency** – Is there enough information in the data to identify stable decision boundaries?
- **Multi-source data contributions** – Which data should be included in an analysis, and how important is it relative to other data sources?
- **Trustworthiness** – How do we distinguish robust models from those that are sensitive to small changes in the data?
- **Performance diagnosis** – Under what conditions will a model perform well?
- **Alternate data interpretations** – Less likely, yet plausible, data interpretations

Following examples focus on one source of uncertainty: model

Lessons are general: should apply to any ML application

Bootstrapping: Generating Distributions to Quantify Uncertainty

- Proposed by Efron (1979), this is a method for estimating the properties of an estimator (mean, standard error, confident interval, etc.) for a parameter.

- Given a random sample $\mathbf{X} = (\mathbf{X}_1,\ldots,\mathbf{X}_n)$, which are *n* independent and identically distributed (i.i.d.) *k*-dimensional ($k \geq 1$) data points from a distribution function *F*, we wish to estimate the sampling distribution of a random variable $R(\mathbf{X},F)$.

Procedure:

1. Observed data: $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$
2. Compute the point estimate of $R(\mathbf{X},F)$ using the observed data.
3. Generate *B* bootstrap samples by resampling the observed data *with replacement*. Each bootstrap sample contains *n* observations. *B* should be a sufficiently large number to approximate the sampling distribution of the parameter estimate.
   - Bootstrap sample: $(\mathbf{x}_1^*,\ldots,\mathbf{x}_n^*)$
4. Compute the point estimate of $R(\mathbf{X},F)$ using the bootstrap samples data.

Unsupervised Learning Example: Clustering, Image Segmentation, and Multimodal Analysis (Stracuzzi et al, 2018)

- Multimodal Imagery Data: images from different sensors (e.g. optical and lidar) covering the same scene
- We wish to evaluate the relative contribution of each image.
  - Which data do we really need?
  - Which data do we wish we had?
- To do this, we will quantify the uncertainty surrounding the clustering results.
- Data: one optical (RBG values) and one lidar (height of objects) image of the same scene in Philadelphia.
- For each image, we model each pixel as a Gaussian mixture model (GMM).
  - In a GMM, each pixel is clustered into an unlabeled cluster.
  - Each cluster follows a multivariate normal distribution with its own mean vector and covariance matrix.
  - Each image will have its own GMM, so they can have different numbers of clusters and the clusters can have different meanings.

Unsupervised Learning Example: Clustering, Image Segmentation, and Multimodal Analysis (Stracuzzi et al, 2018)

- GMM density function:

Density for pixel x:

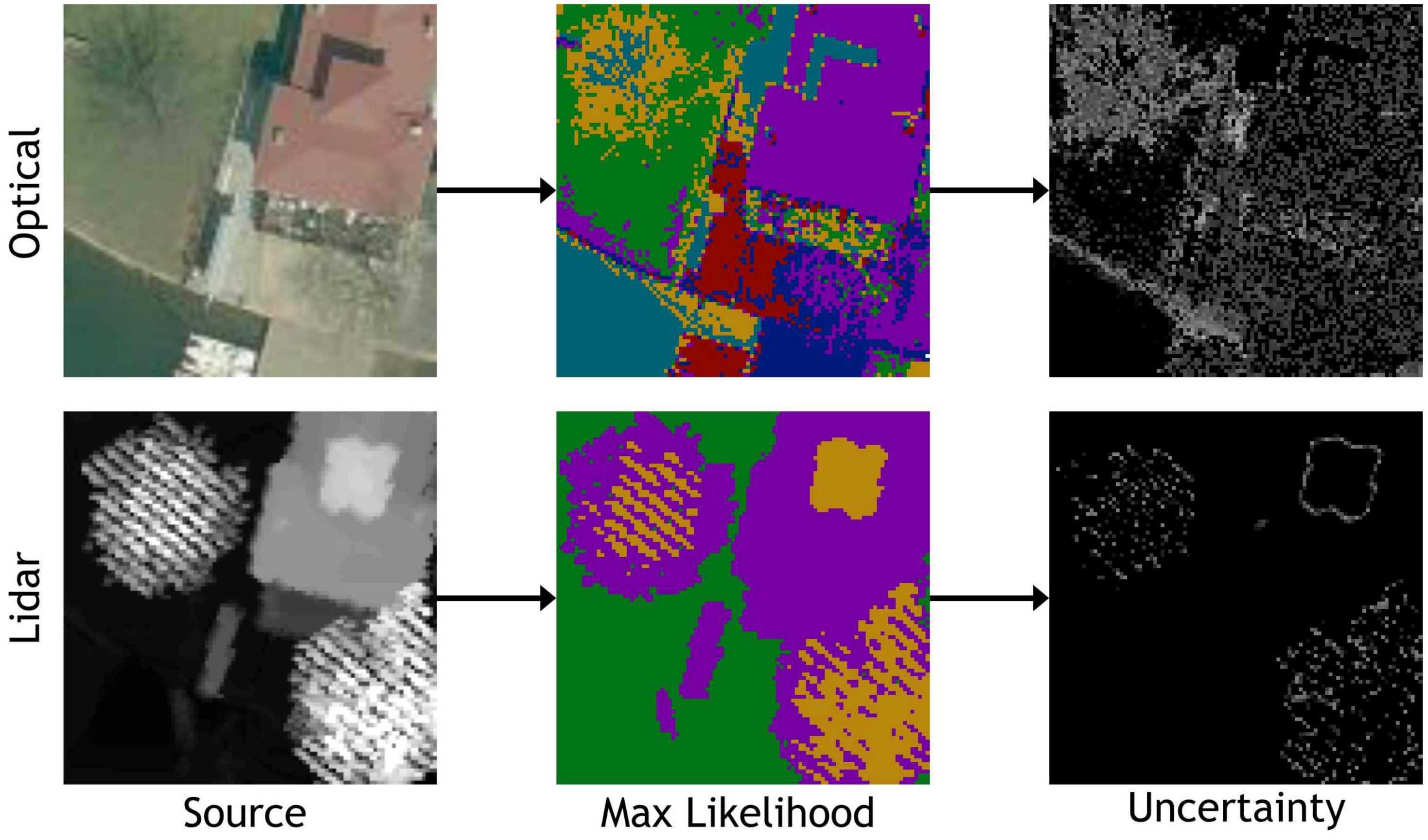
$$f(\mathbf{x}|\vartheta) = \sum_{k=1}^{K} \pi_k \frac{\exp\{-\frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma_g^{-1} (\mathbf{y} - \mu_k)\}}{\sqrt{\det(2\pi\Sigma_k)}},$$

where μ_k is the mean vector and Σ_k is the covariance matrix of component k.

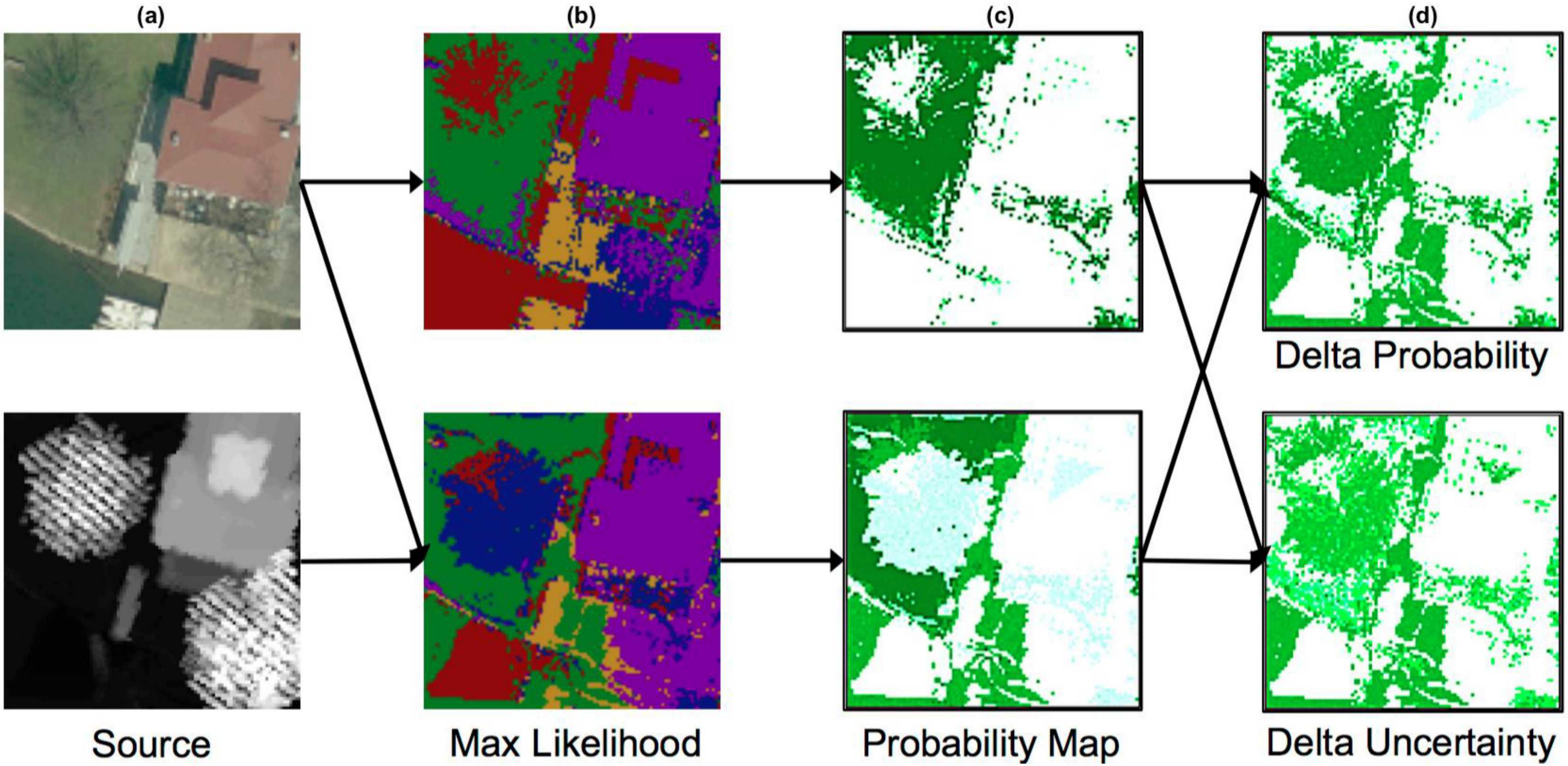
- Parameter of interest: cluster probabilities for each pixel (p_{ij}).
- Bootstrap procedure for each image:
  - Observed data: ($\mathbf{x}_1,\ldots,\mathbf{x}_n$) for n pixels in the image
  - Fit a GMM to the observed data. All parameters in the model are estimated via an EM algorithm.
  - Obtain B bootstrap samples by sampling with replacement from the n observed pixels.
  - Fit a GMM to each of the B bootstrap samples.
  - We now have a sample of estimates of the cluster probabilities for each pixel being classified to each cluster.

Unsupervised Learning Example: Clustering, Image Segmentation, and Multimodal Analysis (Stracuzzi et al, 2018)

- High uncertainty implies data are insufficient to reliably categorize pixels
- Example: pixel color not predictive around trees



Unsupervised Learning Example: Clustering, Image Segmentation, and Multimodal Analysis (Stracuzzi et al, 2018)



Bayesian Consensus Clustering (BCC) [Lock et al, 2013]

- **Consensus clustering (ensemble clustering)**: determination of an overall partition of the observations in a dataset that agrees the most with the source-specific clusterings
- BCC presents a general Bayesian framework for estimating the integrative clustering model relative to the source-specific clusterings
  - Both the source-specific clusterings and the consensus clustering are modeled in a statistical way that allows for uncertainty in all parameters.
  - The source-specific clusterings and the consensus clustering are estimated simultaneously.
- Key assumption of BCC: source-specific clusterings adhere loosely to the consensus clustering

BCC: Assumptions

Integrative model:

- Source-specific and overall clusterings all have K clusters
- Data from M sources: $X_1, \ldots, X_M$ (each data source may have disparate structure)
- Each data source is available for a common set of N objects
- X_{mn}: data m for object n
- Probability model for each data source: $f_m(X_n|\theta_m)$
- Each X_{mn}, $n = 1, \ldots, N$, is drawn independently from a K-component mixture distribution specified by the parameters $\theta_{m1}, \ldots, \theta_{mK}$.
- $L_{mn} \in \{1, \ldots, K\}$: component corresponding to X_{mn}
- $C_n \in \{1, \ldots, K\}$: overall mixture component for object n
- The source-specific clusterings $\mathbb{L}_m = (L_{m1}, \ldots, L_{mN})$ are dependent on the overall clustering $\mathbb{C} = (C_1, \ldots, C_N)$:

  $$P(L_{mn} = k|C_n) = \nu(k, C_n, \alpha_m)$$

  where α_m adjusts the dependence function ν.
- Data $\mathbb{X}_m$ are independent of $\mathbb{C}$ conditional on the source-specific clustering $\mathbb{L}_m$.
- Conditional model:

  $$P(L_{mn} = k|X_{mn}, C_n, \theta_{mk}) \propto \nu(k, C_{n,m})f_m(X_{mn|\theta_{mk}})$$

BCC: Assumptions

- We assume ν has the simple form

$$\nu(L_{mn}, C_n, \alpha_m) = \begin{cases} \alpha_m, & \text{if } C_n = L_{mn} \\ \frac{1-\alpha_m}{K-1}, & \text{otherwise} \end{cases}$$

where $\alpha_m = P(L_{mn} = C_n)$.

- Assume a Dirichlet(β) prior for $\Pi = (\pi_1, ..., \pi_K)$, where $\pi_k = P(C_n = k)$
- Probability that an object belongs to a given source-specific cluster:

$$P(L_{mn} = k|\Pi) = \pi_k \alpha_m + (1 - \pi_k)\frac{1 - \alpha_m}{K - 1}$$

- Conditional distribution of $\mathbb{C}$:

$$P(C_n = k|\mathbb{L}, \Pi, \alpha) \propto \pi_k \prod_{m=1}^{M} \nu(L_{mn}, k, \alpha_m)$$

- Joint marginal distribution of $\mathbb{L}_1, ..., \mathbb{L}_m$:

$$P(\{L_{mn} = k_m\}_{m=1}^{M}|\Pi, \alpha) \propto \sum_{k=1}^{K} \pi_k \prod_{m=1}^{M} \nu(k_m, k, \alpha_m)$$

BCC: Estimation

Data:

- $\mathbb{X}_i$ has a normal-gamma mixture with cluster-specific mean and variance

$$X_{mn}|L_{mn} = k \sim N(\mu_{mk}, \Sigma_{mk})$$

- μ_{mk} is a D_m dimensional mean vector, where D_m is the dimension of the data source m
- Σ_{mk} is a $D_m \times D_m$ diagonal covariance matrix, $\Sigma_{mk} = \text{Diag}(\sigma_{mk1}, ..., \sigma_{mkD_m})$
- Prior distribution for θ_{mk}: D_m dimensional normal-inverse-gamma distribution

$$\theta_{mk} = N\Gamma^{-1}(\eta_{m0}, \lambda_0, A_{m0}, B_{m0})$$

  where η_{m0}, λ_0, A_{m0}, and B_{m0} are hyperparameters.
- It follows that
  - $\frac{1}{\sigma^2_{mkd}} \sim \text{Gamma}(A_{m0d}, B_{m0d})$
  - $\mu_{mkd} \sim N(\eta_{m0}, \frac{\sigma^2_{mkd}}{\lambda_0})$ for $d = 1, ..., D_m$

BCC: Estimation

Conjugate prior distributions:

- $\alpha_m \sim \text{TBeta}(a_m = 1, b_m = 1, \frac{1}{K})$ (prior for α_m is uniformly distributed between $\frac{1}{K}$ and 1)
- $\pi \sim \text{Dirichlet}(\beta_0 = (1, 1, ..., 1))$ (prior for Π is uniformly distributed on the standard $(M-1)$-simplex
- $\theta_{mk} \sim N\Gamma^{-1}(\eta_{m0}, \lambda_0, A_{m0}, B_{m0})$

Conditional posterior distributions (iteratively sampled via MCMC):

- $\Theta_m | \mathbb{X}_m, \mathbb{L}_m \sim p_m(\theta_{mk} | \mathbb{X}_m, \mathbb{L}_m)$ for $k = 1, ..., K$

$$\theta_{mk} \sim N\Gamma^{-1}(\eta_{mk}, \lambda_k, A_{m0}, B_{m0})$$

- $\mathbb{L}_m | \mathbb{X}_m, \Theta_m, \alpha_m, \mathbb{C} \sim P(k | X_{mn}, C_n, \theta_{mk}, \alpha_m)$ for $n = 1, ..., N$, where

$$P(k | X_{mn}, C_n, \theta_{mk}, \alpha_m) \propto \nu(k, C_n, \alpha_m) f_m(X_{mn} | \theta_{mk}).$$

- $\alpha_m | \mathbb{C}, \mathbb{L}_m \sim \text{TBeta}(a_m + \tau_m, b_m + N - \tau_m, \frac{1}{K})$, where τ_m is the number of samples n satisfying $L_{mn} = C_n$.
- $\mathbb{C} | \mathbb{L}_m, \Pi, \alpha \sim P(k | \Pi, \{L_{mn}, \alpha_m\}_{m=1}^{M})$ for $n = 1, ..., N$, where

$$P(k | \Pi, \{L_{mn}, \alpha_m\}_{m=1}^{M}) \propto \pi_k \prod_{m=1}^{M} \nu(k, L_{mn}, \alpha_m)$$

- $\Pi | \mathbb{C} \sim \text{Dirichlet}(\beta_0 + \rho)$, where ρ_k is the number of samples allocated to cluster k in $\mathbb{C}$

Multi-modal Image Segmentation Uncertainty Analysis

- Use the *variance* as our measure of uncertainty
- Uncertainty of the overall consensus clustering: $\text{Var}[P(C_n = k)]$
- Uncertainty of the source-specific clusterings: $\text{Var}[P(L_{mn} = k_m)]$
- Question: Can we relate the uncertainty of the overall consensus clustering to the uncertainty of the source-specific clusterings?

$$\text{Var}[P(C_n = k)] = f(\text{Var}[P(L_{mn} = k_m)]_{m=1}^{M})?$$

Multi-modal Image Segmentation Uncertainty Analysis

$$P(C_n = k|\mathbb{L}, \Pi, \alpha) \propto \pi_k \prod_{m=1}^{M} \nu(L_{mn}, k, \alpha_m)$$

$$\mathrm{Var}[P(C_n = k|\mathbb{L}, \Pi, \alpha)] \propto \mathrm{Var}[\pi_k \prod_{m=1}^{M} \nu(L_{mn}, k, \alpha_m)]$$

For $M = 2$,

$$\mathrm{Var}[P(C_n = k|\mathbb{L}, \Pi, \alpha)] \propto \mathrm{Var}\{\pi_k[\nu(L_{1n}, k, \alpha_1)][\nu(L_{2n}, k, \alpha_2)]\}$$

$$\propto \begin{cases} \mathrm{Var}[\pi_k \alpha_1 \alpha_2], & \text{if } L_{1n} = k, L_{2n} = k \\ \mathrm{Var}[\pi_k \alpha_1 \frac{1-\alpha_2}{K-1}], & \text{if } L_{1n} = k, L_{2n} \neq k \\ \mathrm{Var}[\pi_k \frac{1-\alpha_1}{K-1} \alpha_2], & \text{if } L_{1n} \neq k, L_{2n} = k \\ \mathrm{Var}[\pi_k \frac{1-\alpha_1}{K-1} \frac{1-\alpha_2}{K-1}], & \text{if } L_{1n} \neq k, L_{2n} \neq k \end{cases}$$

Multi-modal Image Segmentation Uncertainty Analysis

Since π_k, α_1, and α_2 are all dependent,

$$\begin{aligned}\mathsf{Var}(\pi_k\alpha_1\alpha_2) &= \mathsf{Cov}(\pi_k^2,\alpha_1^2\alpha_2^2) + (\mathsf{Var}(\pi_k) + [E(\pi_k)]^2)(\mathsf{Var}(\alpha_1\alpha_2) + [E(\alpha_1\alpha_2)]^2) \\ &\quad - [\mathsf{Cov}(\pi_k,\alpha_1\alpha_2) + E(\pi_k)E(\alpha_1\alpha_2)]^2 \\ \mathsf{Var}(\alpha_1\alpha_2) &= \mathsf{Cov}(\alpha_1^2,\alpha_2^2) + (\mathsf{Var}(\alpha_1) + [E(\alpha_1)]^2)(\mathsf{Var}(\alpha_2) + [E(\alpha_2)]^2) \\ &\quad - [\mathsf{Cov}(\alpha_1,\alpha_2) + E(\alpha_1)E(\alpha_2)]^2\end{aligned}$$

$$\begin{aligned}&\mathsf{Var}(\pi_k\alpha_1\alpha_2) = \mathsf{Cov}(\pi_k^2,\alpha_1^2\alpha_2^2) + (\mathsf{Var}(\pi_k) + [E(\pi_k)]^2)(\mathsf{Cov}(\alpha_1^2,\alpha_2^2) \\ &+ (\mathsf{Var}(\alpha_1) + [E(\alpha_1)]^2)(\mathsf{Var}(\alpha_2) + [E(\alpha_2)]^2) - \underbrace{[\mathsf{Cov}(\alpha_1,\alpha_2) + E(\alpha_1)E(\alpha_2)]^2}_{[E(\alpha_1\alpha_2)]^2} \\ &+ [E(\alpha_1\alpha_2)]^2) - [\mathsf{Cov}(\pi_k,\alpha_1\alpha_2) + E(\pi_k)E(\alpha_1\alpha_2)]^2 \\ &= \mathsf{Cov}(\pi_k^2,\alpha_1^2\alpha_2^2) + (\mathsf{Var}(\pi_\mathbf{k}) + [E(\pi_k)]^2)(\mathsf{Cov}(\alpha_1^2,\alpha_2^2) \\ &+ (\mathsf{Var}(\alpha_1) + [E(\alpha_1)]^2)(\mathsf{Var}(\alpha_2) + [E(\alpha_2)]^2)) - [\mathsf{Cov}(\pi_k,\alpha_1\alpha_2) + E(\pi_k)E(\alpha_1\alpha_2)]^2\end{aligned}$$

- We conclude that the **uncertainty for the overall clustering is directly proportional to the uncertainties for the adherence of the source-specific clusterings to the overall clustering,** which affect the results of the source-specific clusterings.

Multi-modal Image Segmentation Uncertainty Analysis

We have that

$$\Pi|\mathbb{C} \sim \text{Dirichlet}(\beta_0 + \rho), \beta_0 = (1, 1, ..., 1)$$

$$\Rightarrow \pi_k|\mathbb{C} \sim \text{Beta}(1 + \rho_k, \sum_{i=1}^{K}(1 + \rho_i) - (1 + \rho_k)) = \text{Beta}(1 + \rho_k, K - 1 + \sum_{i \neq k} \rho_i)$$

$$E(\pi_k|\mathbb{C}) = \frac{1 + \rho_k}{K - 1 + \sum_{i \neq k} \rho_i}$$

$$\text{Var}(\pi_k|\mathbb{C}) = \frac{(1 + \rho_k)(K - 1 + \sum_{i \neq k} \rho_i)}{(K + \sum_{i=1}^{K} \rho_i)^2(1 + K + \sum_{i=1}^{K} \rho_i)}$$

Multi-modal Image Segmentation Uncertainty Analysis

We have that

$$\alpha_m|\mathbb{C},\mathbb{L}_m \sim \text{TBeta}(a_m+\tau_m, b_m+N-\tau_m, \frac{1}{K}) = \text{TBeta}(1+\tau_m, 1+N-\tau_m, \frac{1}{K})$$

Let $B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$ be the incomplete Beta function. Then

$$E[\alpha_m|\mathbb{C},\mathbb{L}_m] = \frac{B(\frac{1}{K};2+\tau_m,1+N-\tau_m) - B(1;2+\tau_m,1+N-\tau_m)}{B(\frac{1}{K};1+\tau_m,1+N-\tau_m) - B(1;1+\tau_m,1+N-\tau_m)}$$

$$E[\alpha_m^2|\mathbb{C},\mathbb{L}_m] = \frac{B(\frac{1}{K};3+\tau_m,1+N-\tau_m) - B(1;3+\tau_m,1+N-\tau_m)}{B(\frac{1}{K};1+\tau_m,1+N-\tau_m) - B(1;1+\tau_m,1+N-\tau_m)}$$

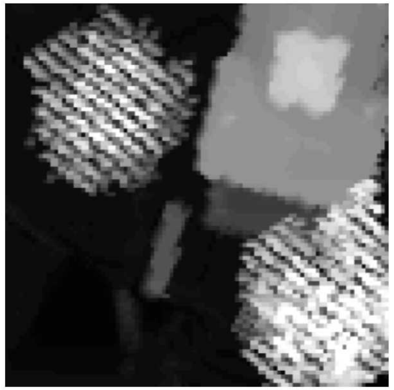
$$\text{Var}(\alpha_m|\mathbb{C},\mathbb{L}_m) = \frac{B(\frac{1}{K};3+\tau_m,1+N-\tau_m) - B(1;3+\tau_m,1+N-\tau_m)}{B(\frac{1}{K};1+\tau_m,1+N-\tau_m) - B(1;1+\tau_m,1+N-\tau_m)} - \{\frac{B(\frac{1}{K};2+\tau_m,1+N-\tau_m) - B(1;2+\tau_m,1+N-\tau_m)}{B(\frac{1}{K};1+\tau_m,1+N-\tau_m) - B(1;1+\tau_m,1+N-\tau_m)}\}^2$$

Multi-modal Image Segmentation Uncertainty Analysis

$$
\begin{aligned}
&\mathrm{Var}(\pi_k\alpha_1\alpha_2) \\
&= \mathrm{Cov}(\pi_k^2, \alpha_1^2\alpha_2^2) + \{\frac{(1+\rho_k)(K-1+\sum_{i\neq k}\rho_i)}{(K+\sum_{i=1}^{K}\rho_i)^2(1+K+\sum_{i=1}^{K}\rho_i)} + [\frac{1+\rho_k}{K-1+\sum_{i\neq k}\rho_i}]^2\}\times \\
&\{\mathrm{Cov}(\alpha_1^2,\alpha_2^2) + (\frac{B(\frac{1}{K};3+\tau_1,1+N-\tau_1)-B(1;3+\tau_1,1+N-\tau_1)}{B(\frac{1}{K};1+\tau_1,1+N-\tau_1)-B(1;1+\tau_1,1+N-\tau_1)} \\
&-\{\frac{B(\frac{1}{K};2+\tau_1,1+N-\tau_1)-B(1;2+\tau_1,1+N-\tau_1)}{B(\frac{1}{K};1+\tau_1,1+N-\tau_1)-B(1;1+\tau_1,1+N-\tau_1)}\}^2+ \\
&[\frac{B(\frac{1}{K};2+\tau_1,1+N-\tau_1)-B(1;2+\tau_1,1+N-\tau_1)}{B(\frac{1}{K};1+\tau_1,1+N-\tau_1)-B(1;1+\tau_1,1+N-\tau_1)}]^2)\times \\
&(\frac{B(\frac{1}{K};3+\tau_2,1+N-\tau_2)-B(1;3+\tau_2,1+N-\tau_2)}{B(\frac{1}{K};1+\tau_2,1+N-\tau_2)-B(1;1+\tau_2,1+N-\tau_2)} \\
&-\{\frac{B(\frac{1}{K};2+\tau_2,1+N-\tau_2)-B(1;2+\tau_2,1+N-\tau_2)}{B(\frac{1}{K};1+\tau_2,1+N-\tau_2)-B(1;1+\tau_2,1+N-\tau_2)}\}^2)+ \\
&[\frac{B(\frac{1}{K};2+\tau_2,1+N-\tau_2)-B(1;2+\tau_2,1+N-\tau_2)}{B(\frac{1}{K};1+\tau_2,1+N-\tau_2)-B(1;1+\tau_2,1+N-\tau_2)}]^2\} \\
&-[\mathrm{Cov}(\pi_k,\alpha_1\alpha_2) + \frac{1+\rho_k}{K-1+\sum_{i\neq k}\rho_i}\times E[\alpha_1\alpha_2]]^2
\end{aligned}
$$

Philadelphia Multi-modal Imagery

Optical

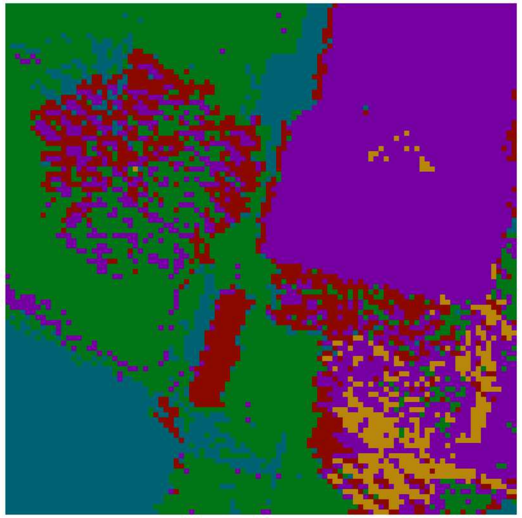


Lidar

- Data: one optical (RBG values) and one lidar (height of objects) image of the same scene in Philadelphia. (M=2)
- Each image contains 10,000 pixels (100 pixels x 100 pixels). (N=10,000)
- Run BCC to obtain the following clusterings, each with six clusters: (K=6)
  - Source-specific clustering for optical image
  - Source-specific clustering for lidar image
  - Overall clustering considering data from both images
- 1,000 MCMC iterations using R with the bayesCC package
- Point estimates of clustering probabilities: MAP estimates

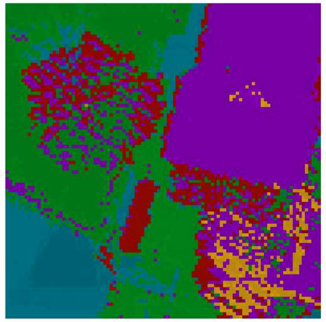
Philadelphia Multi-modal Imagery

Overall Clustering



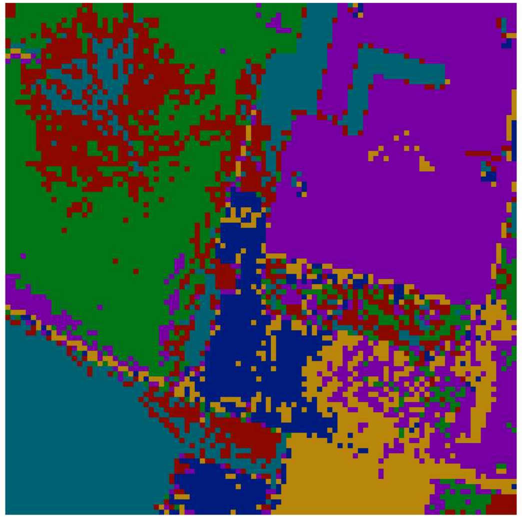
Cluster Assignments

Variance



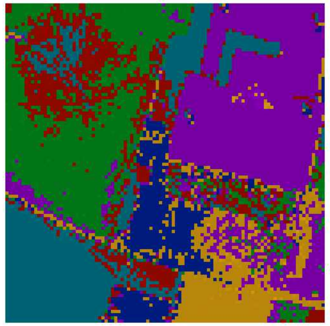
Variance Overlayed with Cluster Assignments

Optical Clustering ($\alpha = 0.646$) ($Var(\alpha) = 0.0001248$)



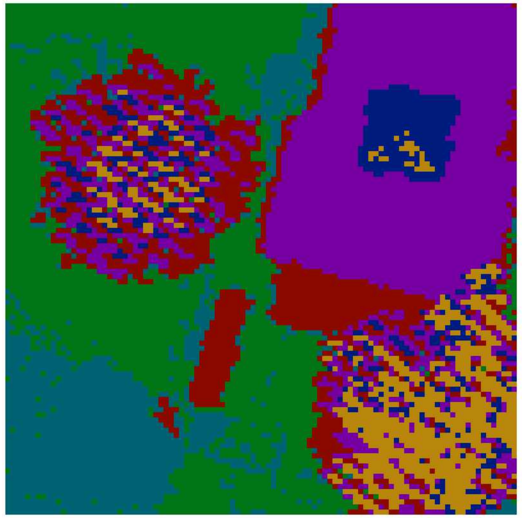
Cluster Assignments

Variance



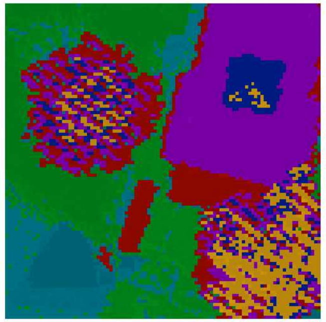
Variance Overlayed with Cluster Assignments

Lidar Clustering ($\alpha = 0.721$) ($Var(\alpha) = 0.0001689$)



Cluster Assignments

Variance



Variance Overlayed with Cluster Assignments

Discussion

- Recent advances have made both data and the patterns they contain more accessible.
- However, *simply finding patterns of interest is not sufficient to support high consequence decision making*.
- We must also assess confidence in results, which includes consideration of uncertainty.

Lessons Learned:

- Measure the credibility of each prediction that the model makes on unseen samples (Performance Diagnostics, Trustworthiness).
- Evaluate data sufficiency that considers not only the number of examples, but their usefulness in identifying decision boundaries. (Multi-source data contributions)

Discussion

- Presented preliminary work on mathematically relating the uncertainty in the results of source-specific clusterings (clusterings of optical and lidar images separately) to the uncertainty of the results of an overall consensus clustering
- Visually, we inspect the results of the clusterings and see how the uncertainty results for the source-specific clusterings and the uncertainty of the estimated adherences factor into the uncertainty in the overall consensus clustering.
- **Main takeaway: overall consensus clustering uncertainty is directly proportional to the uncertainties in the adherences of each source-specific clustering to the overall clustering**

Future Work: BCC

- Deriving expressions in closed form, particularly with the covariances
- Extending to the case of any number of M data sources
- Related to BCC implementation:
  - Nonparametric distributions
  - Different number of clusters and semantic meanings of clusters for each data source and overall clustering
  - Computational scalability

Future Work: Frequentist Approach

Probabilistic Feature Fusion (PFF) [Simonson et al, 2017]

- Combines evidence arising from multiple features and classifiers expressed in the form of (generally dependent) hypothesis tests
- Used in one-class classification (finding one target class)
- Sum of transformed p-values approximately follows a Gamma distribution

Problem:

- PFF only works when you are testing for the same target class (i.e. same hypothesis)
- When we fit Gaussian mixture models (GMMs) to each image (optical and lidar), even with the same number of clusters, the semantic meanings of the clusters in each image likely will be different.
- Questions:
  1. How do we formulate a frequentist framework for fusing p-values together when they are testing different hypotheses, but they are related because they are describing the same scene of interest?
  2. Can we do the same for uncertainty measures, such as variance or entropy?
  3. Is there a frequentist method for doing consensus clustering where information is shared between the overall clustering and the source-specific clusterings?

Future Work: Inverse Uncertainty Quantification

- Rigorous investigation of all sources of uncertainty
- Statistical properties of uncertainty propagation through the inverse UQ analysis pipeline, as well as between multiple analysis tasks
- Evaluating the quality of benchmarks used for ML algorithm evaluation and UQ

Thank you for your kind attention!

Questions?

References

Darling, M., Heileman, G., Gressel, G., Ashok, A., and Poornachandran, P. (2015). A lexical approach for classifying malicious urls. In *High Performance Computing & Simulation (HPCS), 2015 International Conference on*, pages 195–202. IEEE.

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

Lock, E.F., Dunson, D.B. Bayesian consensus clustering, *Bioinformatics*, Volume 29, Issue 20, 15 October 2013, Pages 2610–2616, https://doi.org/10.1093/bioinformatics/btt425

Lock, E.F., Dunson, D.B. Supplement to "Bayesian Consensus Clustering".

Simonson, K.M., West, R.D., Hansen, R.L., LaBruyere, T.E., Van Benthem, M.H. 2017. A statistical approach to combining multisource information in one-class classifiers. Stat. Anal. Data Min. 10, 4 (August 2017), 199-210. DOI: https://doi.org/10.1002/sam.11342

Stracuzzi, D. J., Darling, M. C., Chen, M. G., and Peterson, M. G. (2018). Data-driven uncertainty quantification for multi-sensor analytics. In *SPIE Defense + Security, Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR IX*, 10635. Orlando, FL. SPIE. Also available as Sandia Report 2018-3541C.

Vollmer, C., Peterson, M., Stracuzzi, D. J., and Chen, M. G. (2017). Using data-driven uncertainty quantification to support decision making. In *Statistical Data Science Workshop*.

BACKUP SLIDES

Future Work: Frequentist Approach

Probabilistic Feature Fusion (PFF) [Simonson et al, 2017]

- Combines evidence arising from multiple features and classifiers expressed in the form of (generally dependent) hypothesis tests
- Used in one-class classification (finding one target class)

Steps:

- Data consists of N feature vectors each of length K. Let X_i, $i = 1, ..., K$, denote the ith feature.
- The marginal probability distribution function of X_i for the target class is denoted F_i. For X_i drawn from the target distribution, the quantity $F_i(X_i)$ will be distributed $U[0, 1]$.
- For each feature, compute the p-value $1 - F_i(X_i)$.
- Transform the p-values using the transformation $Y_i = -\log(1 - F_i(X_i))$, which has a standard exponential distribution under the null hypothesis.

Future Work: Frequentist Approach

- Sum the transformed values

$$S_{\text{fused}} = \sum_{i=1}^{K} Y_i,$$

which follows the gamma distribution with shape parameter $\alpha = K$ and $\beta = 1$ when the individual tests are independent. However, since the tests are likely dependent, we have an approximate gamma distribution. Let $\hat{r}_{ij}$ be the sample correlation coefficient between exponential random variables Y_i and Y_j. We estimate the mean and variance of the sum S_{fused} with the quantities $\hat{E}_k$ and $\hat{V}_k$, respectively:

$$\hat{E}_K = K$$

$$\hat{V}_K = K + 2\sum_{i=1}^{K}\sum_{j>i} \hat{r}_{ij}$$

The estimates of the shape and rate parameters are

$$\hat{\alpha} = \frac{\hat{E}_K^2}{\hat{V}_K}$$

$$\hat{\beta} = \frac{\hat{E}_K}{\hat{V}_K}.$$

- Compute the fused p-value as $P_{\text{fused}}(s_{\text{fused}}) = 1 - F_{\text{fused}}(s_{\text{fused}})$.

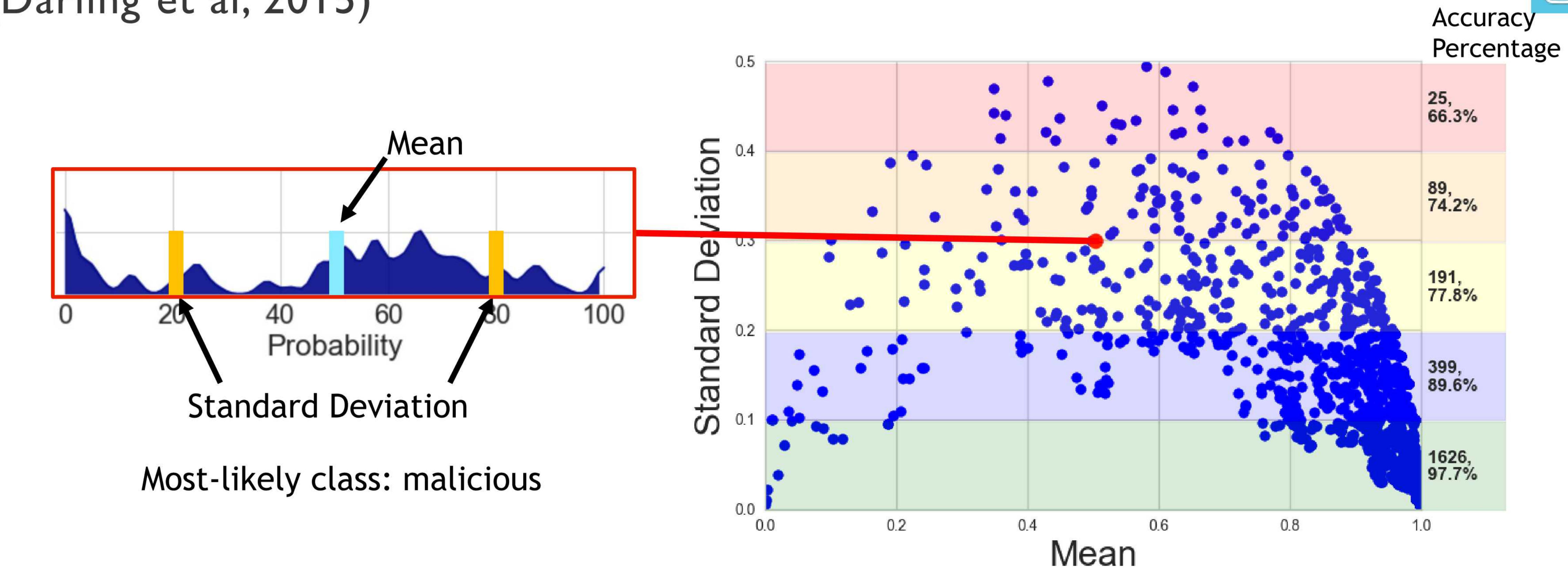
Supervised Learning Example: Classifying URLs of Malicious Websites (Darling et al, 2015)

- Observed data: n URLs
- Task: Classify the URLs as benign or malicious.
- UQ task: How confident should we be in a classifier's performance to accurately classify a URL as benign or malicious? We need to know how much variability is in the classifier we construct.

Approach:

- Obtain S bootstrap samples, each with n URLs, by sampling with replacement from the observed n URLs. Denote the bootstrap samples ($\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*S}$)
- For each of the S bootstrap samples, fit a CART decision tree and obtain probability p_{ik} estimates for k candidate labels y_{ik} for each $\mathbf{x}^*$.
- The S values of p_{ik} provide a probability distribution for each candidate label for each sample.

Supervised Learning Example: Classifying URLs of Malicious Websites (Darling et al, 2015)



Left panel: Probability distribution of predicting most-likely class for a particular URL
Right panel: statistics of prediction distributions. **Color bars annotate the accuracy for the samples they contain.**

- Lower left-hand corner of right panel: classifiers are consistently classifying URLs incorrectly
- Upper-middle of right panel: classifiers do not know how to classify the URLs

Event/Anomaly Detection Example: Time Series Analysis (Vollmer et al, 2017)

- $\mathcal{M}_1$ (noise model):

$$Y_t \sim N(0, \sigma_n^2)$$

where $\sigma_n^2 < \infty$ for $t = 1, ..., k-1$.

- $\mathcal{M}_2$ (signal model): auto-regressive, moving-average (ARMA) model with auto-regressive terms of order p and moving average terms of order q:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$$

where $\epsilon_t \sim N(0, \sigma_s^2)$ for $t = k, ..., T$, and σ_s^2 is the finite variance of the noise component of the signal model.

Event/Anomaly Detection Example: Time Series Analysis (Vollmer et al, 2017)

- Log-likelihood for $\mathcal{M}_1$:

$$l(\theta_1|y_1, ..., y_k) = -\frac{k}{2}\ln(2\pi) - \frac{k}{2}\ln(\sigma_n^2) - \frac{1}{2\sigma_n^2}\sum_{t=1}^{k} y_t^2$$

- Log-likelihood for $\mathcal{M}_2$:

$$l(\theta_2|y_{k+1}, ..., y_T) = -\frac{T-k-p}{2}\ln(2\pi) - \frac{(T-k-p)}{2}\ln(\sigma_s^2) - \frac{1}{2\sigma_s^2}\sum_{t=k+p+1}^{T} \varepsilon_t^2$$

where $\varepsilon_t = Y_t - c - \sum_{i=1}^{p} \phi_i Y_{t-i} - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$ for $t = p+1, p+2, \ldots, T$.

Event/Anomaly Detection Example: Time Series Analysis (Vollmer et al, 2017)

- Akaike Information Criterion (AIC) criterion to minimize for arrival time estimation:

$$
\begin{aligned}
I(\mathcal{M}) &= I(\theta_1|Y_1, ..., Y_k) + I(\theta_2|Y_{k+1}, ..., Y_T) \\
&= -\frac{k}{2}\ln(2\pi) - \frac{k}{2}\ln(\sigma_n^2) - \frac{1}{2\sigma_n^2}\sum_{t=1}^{k} y_t^2 - \frac{T-k-p}{2}\ln(2\pi) \\
&\quad - \frac{(T-k-p)}{2}\ln(\sigma_s^2) - \frac{1}{2\sigma_s^2}\sum_{t=k+p+1}^{T} \varepsilon_t^2 .
\end{aligned}
$$