Analysis of Coarsened Likelihood using Large Deviation Asymptotics

Application to Outlier Detection and Robust Model Estimation

Miheer Dewaskar

Department of Statistical Science, Duke University

IIT Bombay, 4th October, 2023

My Research Areas

Training[1] in Systems Science:

- Applied Probability with *Amarjit Budhiraja* and *Shankar Bhamidi* at UNC – asymptotics of queuing models with the *power-of-choice* routing. Tools: Stochastic Calculus.
- Design and Verification of Controllers with *Nathalie Bertrand* and *Blaise Genest* at INRIA Rennes (France) and *PS Duggirala at UNC*. Tools: Automata Theory and Games.

[1] PhD in Statistics and Operations Research from University of North Carolina (UNC) at Chapel Hill. MSc in Computer Science from CMI.

My Research Areas

Training[1] in Systems Science:

- Applied Probability with *Amarjit Budhiraja* and *Shankar Bhamidi* at UNC – asymptotics of queuing models with the *power-of-choice* routing. Tools: Stochastic Calculus.
- Design and Verification of Controllers with *Nathalie Bertrand* and *Blaise Genest* at INRIA Rennes (France) and *PS Duggirala at UNC*. Tools: Automata Theory and Games.

Current interest in Data Science:

- Iterative Testing with *Andrew Nobel* at UNC – adapt *combinatorial algorithms* to noisy data by introducing hypothesis testing at each step. Theory using a dynamical systems perspective.
- Coarsened Inference with *David Dunson* at Duke (this talk).

[1] PhD in Statistics and Operations Research from University of North Carolina (UNC) at Chapel Hill. MSc in Computer Science from CMI.

Today's talk

Fit Interpretable Models to Big Data
- Motivation and Challenges
- Coarsened Inference Framework
- Asymptotics of the Coarsened Likelihood

Application to Outlier Detection and Robust Model Estimation
- Population setup and assumptions
- Estimator for Optimistic Kullback Leibler (OKL)
- Optimistically Weighted Likelihoods (OWL)

Application Examples and Summary
- Micro Credit study
- Clustering of scRNA-Seq data

Contents

Big Data and Statistical Challenges

Special issue of Statistics & Probability letters, Vol. 136

Some examples of Big Data:

1. Retail: Walmart generates 1 million customer transactions/hr.
2. Health: A billion Electronic Health Records are collected in the US/year.
3. Science: Sloan Digital Sky Survey (200 GB/night) and Large Hadron Collider experiments (25 petabytes/year)

Does "big" data mean that there is no need for statistics anymore?

Big Data and Statistical Challenges

Special issue of Statistics & Probability letters, Vol. 136

Some examples of Big Data:

1. Retail: Walmart generates 1 million customer transactions/hr.
2. Health: A billion Electronic Health Records are collected in the US/year.
3. Science: Sloan Digital Sky Survey (200 GB/night) and Large Hadron Collider experiments (25 petabytes/year)

Does "big" data mean that there is no need for statistics anymore?

No. The data:

- may have sampling or selection bias
- may not be very reliable
- may have unknown data collection artifacts

Big Data and Statistical Challenges

Special issue of Statistics & Probability letters, Vol. 136

Some examples of Big Data:

1. Retail: Walmart generates 1 million customer transactions/hr.
2. Health: A billion Electronic Health Records are collected in the US/year.
3. Science: Sloan Digital Sky Survey (200 GB/night) and Large Hadron Collider experiments (25 petabytes/year)

Does “big” data mean that there is no need for statistics anymore?

No. The data:

- may have sampling or selection bias
- may not be very reliable
- may have unknown data collection artifacts

Need a new framework for statistical modeling of big data.

- Classical theory only assumes sampling uncertainty, leading to order $n^{-1/2}$ estimation errors.
- For big data (large n) these error are wrongly overconfident.

Fit Interpretable Models to Big Data

- Focus on inference using interpretable models with finitely many parameters <u>and not</u> black boxes for prediction.

Fit Interpretable Models to Big Data

- Focus on inference using interpretable models with finitely many parameters and not black boxes for prediction.
- Inevitable misspecification due to: outliers, data contamination, and assumptions like Gaussianity.

Fit Interpretable Models to Big Data

- Focus on inference using interpretable models with finitely many parameters and not black boxes for prediction.
- Inevitable misspecification due to: outliers, data contamination, and assumptions like Gaussianity.
- But how to account for this? Usual method does not account for additional uncertainty due to misspecification.

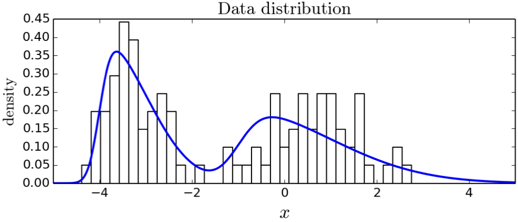
Fit Interpretable Models to Big Data

- Focus on inference using interpretable models with finitely many parameters and not black boxes for prediction.
- Inevitable misspecification due to: outliers, data contamination, and assumptions like Gaussianity.
- But how to account for this? Usual method does not account for additional uncertainty due to misspecification.
- Concern with brittleness: sometimes even slight misspecification can have substantial impact on inference, especially for large sample sizes (big-data settings).

Example I: Brittleness of Mixture models

Example from Miller & Dunson (2015) that has minor misspecification in the kernel

Data is generated from a mixture of two skew Gaussians:

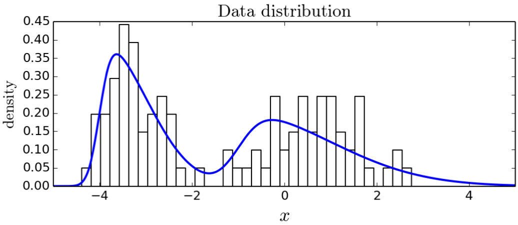


Fit a Gaussian mixture model with prior on the # of components:

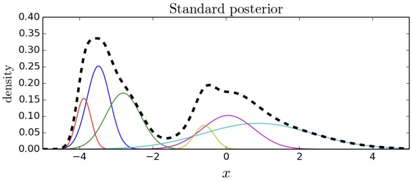
Example I: Brittleness of Mixture models

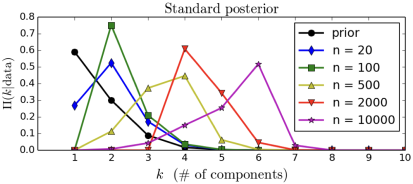
Example from Miller & Dunson (2015) that has minor misspecification in the kernel

Data is generated from a mixture of two skew Gaussians:



Fit a Gaussian mixture model with prior on the # of components:

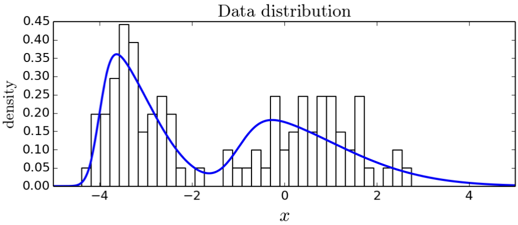




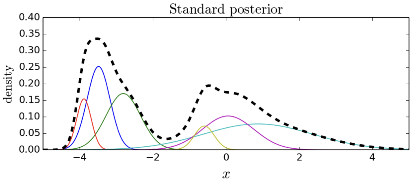
Example I: Brittleness of Mixture models

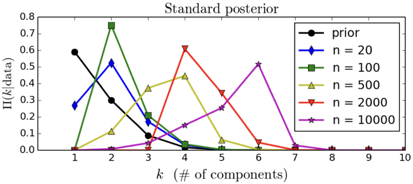
Example from Miller & Dunson (2015) that has minor misspecification in the kernel

Data is generated from a mixture of two skew Gaussians:



Fit a Gaussian mixture model with prior on the # of components:





Brittleness: as $n \to \infty$, the posterior favors large # of components. Theory by Cai, Campbell, Broderick (2021). Miller & Dunson (2015-19) introduced the coarsened posterior to fix this problem.

Example II: Brittleness of MLE to outliers

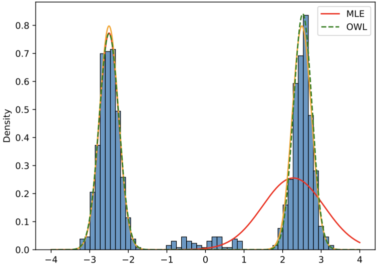
Outliers/data contamination corresponds to misspecification in Total Variation (TV)

95% of data points are drawn from an equal mixture of true Gaussians while 5% are contaminated in some way..
Can we fit our model in a way that is resistant to the 5% contaminated data?

Example II: Brittleness of MLE to outliers

Outliers/data contamination corresponds to misspecification in Total Variation (TV)

95% of data points are drawn from an equal mixture of true Gaussians while 5% are contaminated in some way..
Can we fit our model in a way that is resistant to the 5% contaminated data?

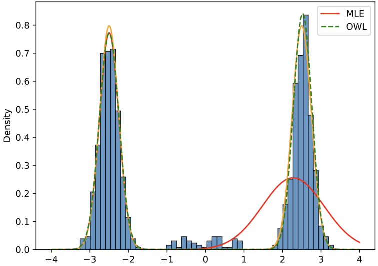


- Maximum Likelihood Estimates (MLE) is known to be brittle to data contamination. This has lead to the field of robust statistics (see Maronna, Martin, Yohai, 2019).

Example II: Brittleness of MLE to outliers

Outliers/data contamination corresponds to misspecification in Total Variation (TV)

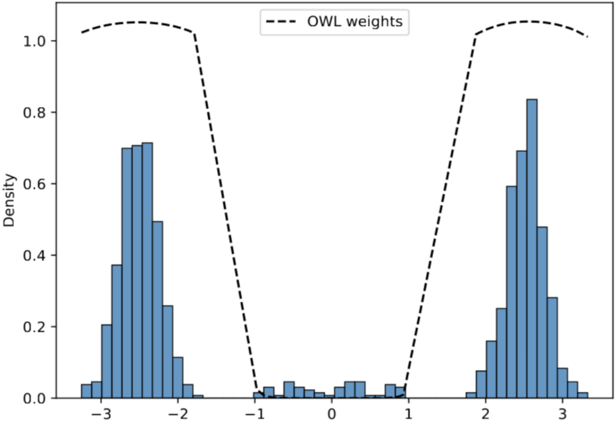
95% of data points are drawn from an equal mixture of true Gaussians while 5% are contaminated in some way.. Can we fit our model in a way that is resistant to the 5% contaminated data?



- Maximum Likelihood Estimates (MLE) is known to be brittle to data contamination. This has lead to the field of robust statistics (see Maronna, Martin, Yohai, 2019).
- This is small misspecification in the total-variation distance. Optimistically Weighted Likelihood (OWL) re-weights the data points to correct for this misspecification.

Optimism: re-weight the data to look like the model

Actively "correct" for the misspecification



Best-case data perturbation, rather than worst-case used in Distributionally Robust Optimization (e.g. Namkoong & Duchi, 2016).

Formalizing what optimism means

Suppose data $x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$ and a model $\{P_\theta\}_{\theta \in \Theta}$ is given.

We find weights $w_1, \ldots, w_n \geq 0$ and $\sum_{i=1}^n w_i = n$ such that

$$\frac{1}{n}\sum_{i=1}^{n}|w_i - 1| \leq \epsilon \qquad [\epsilon\text{-total variation (TV) perturbation}]$$

Formalizing what optimism means

Suppose data $x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$ and a model $\{P_\theta\}_{\theta \in \Theta}$ is given.

We find weights $w_1, \ldots, w_n \geq 0$ and $\sum_{i=1}^n w_i = n$ such that

$$\frac{1}{n}\sum_{i=1}^{n} |w_i - 1| \leq \epsilon \qquad [\epsilon\text{-total variation (TV) perturbation}]$$

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(x_i)^{w_i} \qquad [\text{Weighted Likelihood}]$$

that satisfy

$$P_{\hat{\theta}} \approx \frac{1}{n}\sum_{i=1}^{n} w_i \delta_{x_i} \qquad [\text{Optimism}].$$

- In the well-specified case, optimism holds for $\epsilon = 0$ (i.e. MLE)

Formalizing what optimism means

Suppose data $x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$ and a model $\{P_\theta\}_{\theta \in \Theta}$ is given.

We find weights $w_1, \ldots, w_n \geq 0$ and $\sum_{i=1}^n w_i = n$ such that

$$\frac{1}{n}\sum_{i=1}^{n} |w_i - 1| \leq \epsilon \qquad [\epsilon\text{-total variation (TV) perturbation}]$$

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(x_i)^{w_i} \qquad [\text{Weighted Likelihood}]$$

that satisfy

$$P_{\hat{\theta}} \approx \frac{1}{n}\sum_{i=1}^{n} w_i \delta_{x_i} \qquad [\text{Optimism}].$$

- In the well-specified case, optimism holds for $\epsilon = 0$ (i.e. MLE)
- In the misspecified case, optimistic weights exists $\iff$ $\mathrm{d}_{\mathrm{TV}}(P_o, P_{\theta^*}) \leq \epsilon$ for some $\theta^* \in \Theta$ (for contaminated data).

Contents

Handle misspecification by "coarsening" posterior

From Miller and Dunson (2019). Trust the data less.

We observe data $\boldsymbol{x} = x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$ from unknown $P_o \in \mathcal{P}(\mathcal{X})$.

Bayesian model: $\boldsymbol{X} = X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\vartheta$ and $\vartheta \sim \pi_0$
where $\{P_\theta\}_{\theta\in\Theta}$ is a parametric family, π_0 is a prior on Θ.

Handle misspecification by "coarsening" posterior

From Miller and Dunson (2019). Trust the data less.

We observe data $\boldsymbol{x} = x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$ from unknown $P_o \in \mathcal{P}(\mathcal{X})$.

Bayesian model: $\boldsymbol{X} = X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\vartheta$ and $\vartheta \sim \pi_0$
where $\{P_\theta\}_{\theta \in \Theta}$ is a parametric family, π_0 is a prior on Θ.

Empirical measure: $\hat{P}_{\boldsymbol{x}} \doteq \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ is a sufficient statistic.

Standard Posterior:

$$p(d\theta|\boldsymbol{x}) \doteq Pr\left(\vartheta \in d\theta \middle| \hat{P}_{\boldsymbol{X}} = \hat{P}_{\boldsymbol{x}}\right)$$

Handle misspecification by "coarsening" posterior

From Miller and Dunson (2019). Trust the data less.

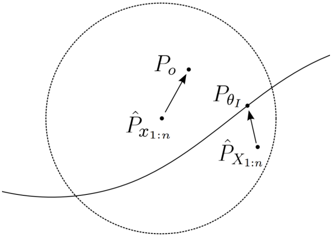
We observe data $\boldsymbol{x} = x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$ from unknown $P_o \in \mathcal{P}(\mathcal{X})$.

Bayesian model: $\boldsymbol{X} = X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\vartheta$ and $\vartheta \sim \pi_0$
where $\{P_\theta\}_{\theta\in\Theta}$ is a parametric family, π_0 is a prior on Θ.

Empirical measure: $\hat{P}_{\boldsymbol{x}} \doteq \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is a sufficient statistic.

Standard Posterior:

$$p(d\theta|\boldsymbol{x}) \doteq Pr\left(\vartheta \in d\theta \middle| \hat{P}_{\boldsymbol{X}} = \hat{P}_{\boldsymbol{x}}\right)$$



Coarsened (C-) posterior:

$$p_\epsilon(d\theta|\boldsymbol{x}) \doteq Pr\left(\vartheta \in d\theta \middle| \boldsymbol{d}(\hat{P}_{\boldsymbol{X}}, \hat{P}_{\boldsymbol{x}}) \le \epsilon\right)$$

Handle misspecification by "coarsening" posterior

From Miller and Dunson (2019). Trust the data less.

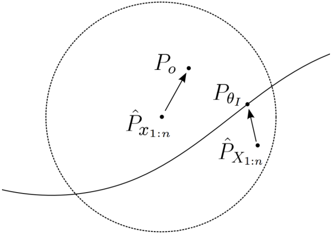
We observe data $\boldsymbol{x} = x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$ from unknown $P_o \in \mathcal{P}(\mathcal{X})$.

Bayesian model: $\boldsymbol{X} = X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\vartheta$ and $\vartheta \sim \pi_0$
where $\{P_\theta\}_{\theta \in \Theta}$ is a parametric family, π_0 is a prior on Θ.

Empirical measure: $\hat{P}_{\boldsymbol{x}} \doteq \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ is a sufficient statistic.

Standard Posterior:

$$p(d\theta|\boldsymbol{x}) \doteq Pr\left(\vartheta \in d\theta \middle| \hat{P}_{\boldsymbol{X}} = \hat{P}_{\boldsymbol{x}}\right)$$



Coarsened (C-) posterior:

$$p_\epsilon(d\theta|\boldsymbol{x}) \doteq Pr\left(\vartheta \in d\theta \middle| \boldsymbol{d}(\hat{P}_{\boldsymbol{X}}, \hat{P}_{\boldsymbol{x}}) \leq \epsilon\right)$$

- Allows misspecification: $\hat{P}_{\boldsymbol{X}}$ is ϵ-close in the discrepancy $\boldsymbol{d}$ on $\mathcal{P}(\mathcal{X})$ (but not necessarily equal) to the observed data $\hat{P}_{\boldsymbol{x}}$.

Handle misspecification by "coarsening" posterior

From Miller and Dunson (2019). Trust the data less.

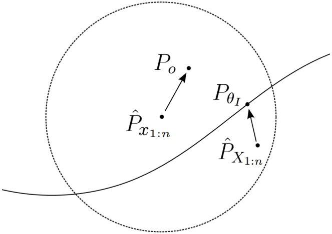
We observe data $\boldsymbol{x} = x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$ from unknown $P_o \in \mathcal{P}(\mathcal{X})$.

Bayesian model: $\boldsymbol{X} = X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\vartheta$ and $\vartheta \sim \pi_0$
where $\{P_\theta\}_{\theta\in\Theta}$ is a parametric family, π_0 is a prior on Θ.

Empirical measure: $\hat{P}_{\boldsymbol{x}} \doteq \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is a sufficient statistic.

Standard Posterior:

$$p(d\theta|\boldsymbol{x}) \doteq Pr\left(\vartheta \in d\theta \middle| \hat{P}_{\boldsymbol{X}} = \hat{P}_{\boldsymbol{x}}\right)$$



Coarsened (C-) posterior:

$$p_\epsilon(d\theta|\boldsymbol{x}) \doteq Pr\left(\vartheta \in d\theta \middle| \boldsymbol{d}(\hat{P}_{\boldsymbol{X}}, \hat{P}_{\boldsymbol{x}}) \leq \epsilon\right)$$

- Allows misspecification: $\hat{P}_{\boldsymbol{X}}$ is ϵ-close in the discrepancy $\boldsymbol{d}$ on $\mathcal{P}(\mathcal{X})$ (but not necessarily equal) to the observed data $\hat{P}_{\boldsymbol{x}}$.
- $p_\epsilon(d\theta|\boldsymbol{x}) \to p(d\theta|\boldsymbol{x})$ as $\epsilon \to 0$ under suitable conditions.

Computation of coarsened posterior

Bayes rule shows: $p_\epsilon(d\theta|\boldsymbol{x}) \propto L_\epsilon(\theta|\boldsymbol{x})\pi_0(d\theta)$ where

$$L_\epsilon(\theta|\boldsymbol{x}) \doteq Pr\left(\boldsymbol{d}(\hat{P}_{\boldsymbol{X}}, \hat{P}_{\boldsymbol{x}}) \leq \epsilon \middle| \vartheta = \theta\right)$$

is the coarsened likelihood. But difficult to use MCMC, as even evaluating $L_\epsilon(\theta|\boldsymbol{x})$ involves estimating a high dimensional integral.

Computation of coarsened posterior

Bayes rule shows: $p_\epsilon(d\theta|\boldsymbol{x}) \propto L_\epsilon(\theta|\boldsymbol{x})\pi_0(d\theta)$ where

$$L_\epsilon(\theta|\boldsymbol{x}) \doteq Pr\left(\boldsymbol{d}(\hat{P}_{\boldsymbol{X}}, \hat{P}_{\boldsymbol{x}}) \leq \epsilon \middle| \vartheta = \theta\right)$$

is the coarsened likelihood. But difficult to use MCMC, as even evaluating $L_\epsilon(\theta|\boldsymbol{x})$ involves estimating a high dimensional integral.

Coarsened posterior is an average of standard posteriors:

$$p_\epsilon(d\theta|\boldsymbol{x}) = \mathbb{E}\left[p(d\theta|\boldsymbol{X}) \middle| \boldsymbol{d}(\hat{P}_{\boldsymbol{X}}, \hat{P}_{\boldsymbol{x}}) \leq \epsilon\right].$$

Computation of coarsened posterior

Bayes rule shows: $p_\epsilon(d\theta|\boldsymbol{x}) \propto L_\epsilon(\theta|\boldsymbol{x})\pi_0(d\theta)$ where

$$L_\epsilon(\theta|\boldsymbol{x}) \doteq Pr\left(\boldsymbol{d}(\hat{P}_{\boldsymbol{X}}, \hat{P}_{\boldsymbol{x}}) \leq \epsilon \middle| \vartheta = \theta\right)$$

is the coarsened likelihood. But difficult to use MCMC, as even evaluating $L_\epsilon(\theta|\boldsymbol{x})$ involves estimating a high dimensional integral.

Coarsened posterior is an average of standard posteriors:

$$p_\epsilon(d\theta|\boldsymbol{x}) = \mathbb{E}\left[p(d\theta|\boldsymbol{X}) \middle| \boldsymbol{d}(\hat{P}_{\boldsymbol{X}}, \hat{P}_{\boldsymbol{x}}) \leq \epsilon\right].$$

Rejection sampling based approach leads to Approximate Bayesian Computation (ABC), which is very slow because the conditioning event is "rare".

Computation of coarsened posterior

Bayes rule shows: $p_\epsilon(d\theta|\boldsymbol{x}) \propto L_\epsilon(\theta|\boldsymbol{x})\pi_0(d\theta)$ where

$$L_\epsilon(\theta|\boldsymbol{x}) \doteq Pr\left(\boldsymbol{d}(\hat{P}_{\boldsymbol{X}}, \hat{P}_{\boldsymbol{x}}) \leq \epsilon \middle| \vartheta = \theta\right)$$

is the coarsened likelihood. But difficult to use MCMC, as even evaluating $L_\epsilon(\theta|\boldsymbol{x})$ involves estimating a high dimensional integral.

Coarsened posterior is an average of standard posteriors:

$$p_\epsilon(d\theta|\boldsymbol{x}) = \mathbb{E}\left[p(d\theta|\boldsymbol{X}) \middle| \boldsymbol{d}(\hat{P}_{\boldsymbol{X}}, \hat{P}_{\boldsymbol{x}}) \leq \epsilon\right].$$

Rejection sampling based approach leads to Approximate Bayesian Computation (ABC), which is very slow because the conditioning event is "rare".

Asymptotic approximation: When $\boldsymbol{d} = \mathrm{KL}$ and $\epsilon \sim \mathrm{Exp}(\alpha)$, Miller & Dunson (2019) develop the power-likelihood approximation:

$$\int L_\epsilon(\theta|\boldsymbol{x})\alpha e^{-\alpha\epsilon} d\epsilon \;\widetilde{\propto}\; \prod_{i=1}^{n} p_\theta(x_i)^{\frac{\alpha}{n+\alpha}} = L(\theta|\boldsymbol{x})^{\frac{\alpha}{n+\alpha}}$$

Computation of coarsened posterior

Bayes rule shows: $p_\epsilon(d\theta|\boldsymbol{x}) \propto L_\epsilon(\theta|\boldsymbol{x})\pi_0(d\theta)$ where

$$L_\epsilon(\theta|\boldsymbol{x}) \doteq Pr\left(\boldsymbol{d}(\hat{P}_{\boldsymbol{X}}, \hat{P}_{\boldsymbol{x}}) \leq \epsilon \middle| \vartheta = \theta\right)$$

is the coarsened likelihood. But difficult to use MCMC, as even evaluating $L_\epsilon(\theta|\boldsymbol{x})$ involves estimating a high dimensional integral.

Coarsened posterior is an average of standard posteriors:

$$p_\epsilon(d\theta|\boldsymbol{x}) = \mathbb{E}\left[p(d\theta|\boldsymbol{X})\middle|\boldsymbol{d}(\hat{P}_{\boldsymbol{X}}, \hat{P}_{\boldsymbol{x}}) \leq \epsilon\right].$$

Rejection sampling based approach leads to Approximate Bayesian Computation (ABC), which is very slow because the conditioning event is "rare".

Asymptotic approximation: When $\boldsymbol{d} = \mathrm{KL}$ and $\epsilon \sim \mathrm{Exp}(\alpha)$, Miller & Dunson (2019) develop the power-likelihood approximation:

$$\int L_\epsilon(\theta|\boldsymbol{x})\alpha e^{-\alpha\epsilon} d\epsilon \;\widetilde{\propto}\; \prod_{i=1}^{n} p_\theta(x_i)^{\frac{\alpha}{n+\alpha}} = L(\theta|\boldsymbol{x})^{\frac{\alpha}{n+\alpha}}$$

Usual likelihood with finite effective sample size $n_0 = \frac{n\alpha}{\alpha+n} < \infty$.

Contents

Sanov's theorem from large deviations

Setup: $\boldsymbol{X} = X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$ and $\hat{P}_{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \in \mathcal{P}(\mathcal{X})$.

Sanov's theorem: As $n \to \infty$, the random measures $\hat{P}_{\boldsymbol{X}}$ satisfy a Large Deviations Principle on $\mathcal{P}(\mathcal{X})$ with rate $\cdot \mapsto \mathrm{KL}(\cdot|P_\theta)$, the Kullback Leiber divergence.

Intuitively:

$$\Pr[\hat{P}_{\boldsymbol{X}} \approx Q|\theta] = e^{-nKL(Q|P_\theta)+o(n)},$$

and that for nice subsets $B \subseteq \mathcal{P}(\mathcal{X})$:

$$Pr\left(\hat{P}_{\boldsymbol{X}} \in B\middle|\theta\right) = e^{-n\inf_{Q\in B}\mathrm{KL}(Q|P_\theta)+o(n)}.$$

Sanov's theorem from large deviations

Setup: $\boldsymbol{X} = X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$ and $\hat{P}_{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \delta_{X_i} \in \mathcal{P}(\mathcal{X})$.

Sanov's theorem: As $n \to \infty$, the random measures $\hat{P}_{\boldsymbol{X}}$ satisfy a Large Deviations Principle on $\mathcal{P}(\mathcal{X})$ with rate $\cdot \mapsto \mathrm{KL}(\cdot|P_\theta)$, the Kullback Leiber divergence.

Intuitively:

$$\Pr[\hat{P}_{\boldsymbol{X}} \approx Q|\theta] = e^{-nKL(Q|P_\theta)+o(n)},$$

and that for nice subsets $B \subseteq \mathcal{P}(\mathcal{X})$:

$$Pr\left(\hat{P}_{\boldsymbol{X}} \in B\middle|\theta\right) = e^{-n\inf_{Q\in B}\mathrm{KL}(Q|P_\theta)+o(n)}.$$

- Really only useful when $\inf_{Q\in B}\mathrm{KL}(Q|P_\theta) > 0 \implies P_\theta \notin B$.
- Recall Gliveco Cantelli: $\hat{P}_{\boldsymbol{X}} \to P_\theta$ as $n \to \infty$. Thus a statement about the tail distribution of $\hat{P}_{\boldsymbol{X}}$.

Sanov's theorem from large deviations

Setup: $\boldsymbol{X} = X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$ and $\hat{P}_{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \in \mathcal{P}(\mathcal{X})$.

Sanov's theorem: As $n \to \infty$, the random measures $\hat{P}_{\boldsymbol{X}}$ satisfy a Large Deviations Principle on $\mathcal{P}(\mathcal{X})$ with rate $\cdot \mapsto \mathrm{KL}(\cdot|P_\theta)$, the Kullback Leiber divergence.

Intuitively:

$$\Pr[\hat{P}_{\boldsymbol{X}} \approx Q|\theta] = e^{-nKL(Q|P_\theta)+o(n)},$$

and that for nice subsets $B \subseteq \mathcal{P}(\mathcal{X})$:

$$Pr\left(\hat{P}_{\boldsymbol{X}} \in B \middle| \theta\right) = e^{-n \inf_{Q \in B} \mathrm{KL}(Q|P_\theta)+o(n)}.$$

- Really only useful when $\inf_{Q\in B} \mathrm{KL}(Q|P_\theta) > 0 \implies P_\theta \notin B$.
- Recall Gliveco Cantelli: $\hat{P}_{\boldsymbol{X}} \to P_\theta$ as $n \to \infty$. Thus a statement about the tail distribution of $\hat{P}_{\boldsymbol{X}}$.
- KL divergence: related to information theory and likelihoods!

Application of large deviations to coarsened inference

Recall: Coarsened inference conditions on the event $E = \{\hat{P}_{\boldsymbol{X}} \in B_\epsilon(\hat{P}_{\boldsymbol{x}})\}$ when $\boldsymbol{X} = X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$.

Here $B_\epsilon(\hat{P}_{\boldsymbol{x}}) = \{Q : \boldsymbol{d}(Q, \hat{P}_{\boldsymbol{x}}) \leq \epsilon\}$ is the ϵ neighborhood around the observed empirical distribution $\hat{P}_{\boldsymbol{x}}$.

Application of large deviations to coarsened inference

Recall: Coarsened inference conditions on the event $E = \{\hat{P}_{\boldsymbol{X}} \in B_\epsilon(\hat{P}_{\boldsymbol{x}})\}$ when $\boldsymbol{X} = X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$.

Here $B_\epsilon(\hat{P}_{\boldsymbol{x}}) = \{Q : \boldsymbol{d}(Q, \hat{P}_{\boldsymbol{x}}) \leq \epsilon\}$ is the ϵ neighborhood around the observed empirical distribution $\hat{P}_{\boldsymbol{x}}$.

Sanov's theorem:

$$L_\epsilon(\theta|\boldsymbol{x}) \doteq Pr\left(E|\vartheta = \theta\right) = e^{-nI_\epsilon(\theta) + o_P(n)},$$

where

$$I_\epsilon(\theta) \doteq \inf_{\substack{Q \in \mathcal{P}(\mathcal{X}) \\ \boldsymbol{d}(Q, P_o) \leq \epsilon}} \mathrm{KL}(Q|P_\theta).$$

is called the Optimistic Kullback Leibler (OKL).

Application of large deviations to coarsened inference

Recall: Coarsened inference conditions on the event $E = \{\hat{P}_{\boldsymbol{X}} \in B_\epsilon(\hat{P}_{\boldsymbol{x}})\}$ when $\boldsymbol{X} = X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$.

Here $B_\epsilon(\hat{P}_{\boldsymbol{x}}) = \{Q : \boldsymbol{d}(Q, \hat{P}_{\boldsymbol{x}}) \leq \epsilon\}$ is the ϵ neighborhood around the observed empirical distribution $\hat{P}_{\boldsymbol{x}}$.

Sanov's theorem:

$$L_\epsilon(\theta|\boldsymbol{x}) \doteq Pr\left(E|\vartheta = \theta\right) = e^{-nI_\epsilon(\theta) + o_P(n)},$$

where

$$I_\epsilon(\theta) \doteq \inf_{\substack{Q \in \mathcal{P}(\mathcal{X}) \\ \boldsymbol{d}(Q, P_o) \leq \epsilon}} \mathrm{KL}(Q|P_\theta).$$

is called the Optimistic Kullback Leibler (OKL).

- $\boldsymbol{d}$ must be a nice, e.g. Maximum Mean Discrepancy, or Wasserstein, or smoothed TV distance.
- Search over "optimistic" data Q_θ in the $(\boldsymbol{d}, \epsilon)$ ball around P_o.

Application of large deviations to coarsened inference

Recall: Coarsened inference conditions on the event $E = \{\hat{P}_{\boldsymbol{X}} \in B_\epsilon(\hat{P}_{\boldsymbol{x}})\}$ when $\boldsymbol{X} = X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$.

Here $B_\epsilon(\hat{P}_{\boldsymbol{x}}) = \{Q : \boldsymbol{d}(Q, \hat{P}_{\boldsymbol{x}}) \leq \epsilon\}$ is the ϵ neighborhood around the observed empirical distribution $\hat{P}_{\boldsymbol{x}}$.

Sanov's theorem:

$$L_\epsilon(\theta|\boldsymbol{x}) \doteq Pr\left(E|\vartheta = \theta\right) = e^{-nI_\epsilon(\theta) + o_P(n)},$$

where

$$I_\epsilon(\theta) \doteq \inf_{\substack{Q \in \mathcal{P}(\mathcal{X}) \\ \boldsymbol{d}(Q, P_o) \leq \epsilon}} \mathrm{KL}(Q|P_\theta).$$

is called the Optimistic Kullback Leibler (OKL).

- $\boldsymbol{d}$ must be a nice, e.g. Maximum Mean Discrepancy, or Wasserstein, or smoothed TV distance.
- Search over "optimistic" data Q_θ in the $(\boldsymbol{d}, \epsilon)$ ball around P_o.
- Use: Finding $\theta \in \Theta$ that maximizes $\theta \mapsto L_\epsilon(\theta|\boldsymbol{x})$ corresponds to minimizing OKL: $\theta \mapsto I_\epsilon(\theta)$ (asymptotically).
- Case $\epsilon = 0$, θ^* is MLE $\iff$ $\theta^* \in \arg\min_{\theta \in \Theta} \mathrm{KL}(P_o|P_\theta)$.

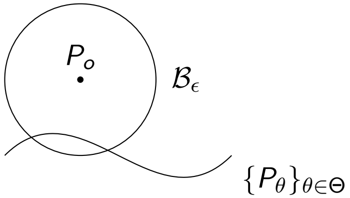
Contents

Robust model estimation: setup and assumptions

Setup: robustly fit model family $\{P_\theta\}_{\theta \in \Theta}$ based on data $x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$.

$\Theta_I = \{\theta \mid \mathrm{d}_{\mathtt{TV}}(P_o, P_\theta) \leq \epsilon\}$ are robustly identified parameters.

Assumption: $\Theta_I \neq \emptyset$.



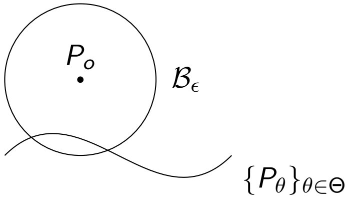
$$\mathcal{B}_\epsilon = \{Q : \mathrm{d}_{\mathtt{TV}}(Q, P_o) \leq \epsilon\}$$

Robust model estimation: setup and assumptions

Setup: robustly fit model family $\{P_\theta\}_{\theta\in\Theta}$ based on data $x_1,\ldots,x_n \overset{i.i.d.}{\sim} P_o$.

$\Theta_I = \{\theta \mid \mathrm{d_{TV}}(P_o, P_\theta) \le \epsilon\}$ are robustly identified parameters.

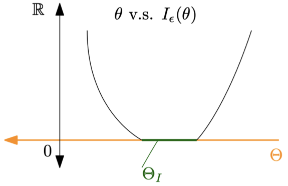
Assumption: $\Theta_I \neq \emptyset$.



$$\mathcal{B}_\epsilon = \{Q : \mathrm{d_{TV}}(Q, P_o) \le \epsilon\}$$

We find a point from Θ_I by minimizing an estimator for OKL (right) w.r.t. θ:

$$I_\epsilon(\theta) = \inf_{Q:\mathrm{d_{TV}}(Q,P_o)\le\epsilon} \mathrm{KL}(Q|P_\theta).$$



- Jointly minimize with respect to θ (model) and Q (pseudo data). Use alternate minimization in practice.

Contents

Estimation of the OKL using data re-weightings

Finite spaces

Given data $x_1, \dots, x_n \sim P_o \in \mathcal{P}(\mathcal{X})$, we use the estimator

$$\hat{I}_\epsilon(\theta) = \min_{\substack{w \in \Delta_n \\ \frac{1}{2}\|w - o\|_1 \leq \epsilon}} \sum_{i=1}^{n} w_i \log \frac{n w_i \hat{p}(x_i)}{p_\theta(x_i)}$$

for $I_\epsilon(\theta) = \min_{Q : \mathrm{d}_{\mathrm{TV}}(Q, P_o) \leq \epsilon} \mathrm{KL}(Q|P_\theta)$ and $o = (1/n, \dots, 1/n)$.

Theorem (D., Tosh, Knoblauch, Dunson, 2023)

If $\mathcal{X}$ is finite and $\mathrm{supp}(P_\theta) \subseteq \mathrm{supp}(P_o)$ for some $\theta \in \Theta$, then

$$\hat{I}_\epsilon(\theta) = \min_{w \in \Delta_n : \mathrm{d}_{TV}(Q_w, \hat{P}) \leq \epsilon} \mathrm{KL}(Q_w|P_\theta) \quad \text{and} \quad \left| I_\epsilon(\theta) - \hat{I}_\epsilon(\theta) \right| = O_p(n^{-1/2})$$

where $Q_w = \sum_{i=1}^{n} w_i \delta_{x_i}$.

Estimation of the OKL using data re-weightings

Continuous space $\mathcal{X} \subseteq \mathbb{R}^d$

Let κ_h be the Gaussian kernel on $\mathbb{R}^d$ with bandwidth $h > 0$, $q_w(x) = \sum_{i=1}^n w_i \kappa_h(x_i, x)$, and $A \in \mathbb{R}^{n \times n}$ with $A_{ij} = \frac{\kappa_h(x_i, x_j)}{n\hat{p}(x_i)}$.

$$\hat{I}_{h,\epsilon}(\theta) \doteq \min_{\substack{v \in A\Delta_n \\ \frac{1}{2}\|v - o\|_1 \leq \epsilon}} \sum_{i=1}^n v_i \log \frac{n v_i \hat{p}(x_i)}{p_\theta(x_i)}$$

$$= \min_{\substack{w \in \Delta_n \\ \mathrm{d}_{\mathtt{TV}}(q_w, \hat{p}) \leq \epsilon}} \frac{1}{n} \sum_{i=1}^n \frac{q_w(x_i)}{\hat{p}(x_i)} \log \frac{q_w(x_i)}{p_\theta(x_i)} \approx \min_{\substack{w \in \Delta_n \\ \mathrm{d}_{\mathtt{TV}}(q_w, p_o) \leq \epsilon}} \mathrm{KL}(q_w | p_\theta).$$

Theorem (D., Tosh, Knoblauch, Dunson, 2023)

If $\mathcal{X} \subseteq \mathbb{R}^d$ is compact and smooth densities p_o, p_θ are supported on $\mathcal{X}$:

$$\left| I_\epsilon(\theta) - \hat{I}_{h,\epsilon}(\theta) \right| = O_p(n^{-1/2} h^{-d} + \sqrt{h}).$$

Contents

Algorithm to estimate the OKL minimizer.

Population OKL minimization:

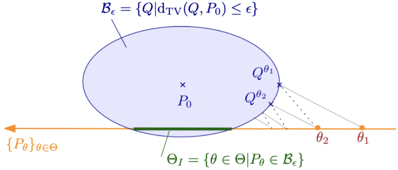
Alternatively update pseudo-data Q and model θ until convergence.

I-projection:

$$Q_t = \underset{Q:\mathrm{d}_{\mathrm{TV}}(Q,P_o)\le\epsilon}{\arg\min} \mathrm{KL}(Q|P_{\theta_t})$$

Maximize log-likelihood:

$$\theta_{t+1} = \underset{\theta\in\Theta}{\arg\max} \int \log p_\theta(x) Q_t(dx)$$



Algorithm to estimate the OKL minimizer.

Population OKL minimization:

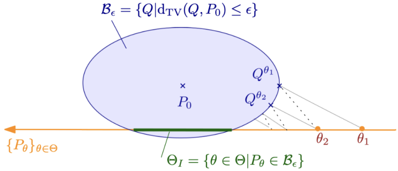
Alternatively update pseudo-data Q and model θ until convergence.

I-projection:

$$Q_t = \operatorname*{arg\,min}_{Q: \mathrm{d}_{\mathrm{TV}}(Q, P_o) \le \epsilon} \mathrm{KL}(Q|P_{\theta_t})$$

Maximize log-likelihood:

$$\theta_{t+1} = \operatorname*{arg\,max}_{\theta \in \Theta} \int \log p_\theta(x) Q_t(dx)$$



Estimating the OKL minimizer from samples $x_1, \ldots, x_n \sim P_o$.

Intuition: $Q_t \approx \sum_{i=1}^n w_i^t \delta_{x_i}$.

Approx I-projection:

$$w^{t+1} = \operatorname*{arg\,min}_{\substack{w \in \Delta_n \\ \frac{1}{2}\|w - o\|_1 \le \epsilon}} \sum_{i=1}^n w_i \log \frac{n w_i \hat{p}(x_i)}{p_{\theta_t}(x_i)}$$

Weighted-MLE:

$$\theta^{t+1} = \operatorname*{arg\,max}_{\theta \in \Theta} \sum_{i=1}^n w_i^{(t+1)} \log p_\theta(x_i)$$

- w-step is convex: Alternating Direction Method of Multipliers (ADMM) [Parikh & Boyd, 2014]
- θ-step: modification of algorithms for MLE.

Optimistically Weighted Likelihoods (OWL)

- Theoretically motivated by the coarsened likelihood framework of Miller & Dunson (2019)
- We estimate parameter and data-weights by repeated weighted likelihood maximization

$$\theta_{t+1} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(x_i)^{w_i(\theta_t)}$$

  where weights $\{w_i(\theta)\}_{i=1}^n$ sum to n and I-projection onto the ℓ_1 ball: $\|w(\theta) - \mathbf{1}\|_1 \leq n\epsilon$.
- $\epsilon \in (0, 1)$ denotes amount of model misspecification, which can automatically be tuned from data.

Optimistically Weighted Likelihoods (OWL)

- Theoretically motivated by the coarsened likelihood framework of Miller & Dunson (2019)
- We estimate parameter and data-weights by repeated weighted likelihood maximization

$$\theta_{t+1} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(x_i)^{w_i(\theta_t)}$$

where weights $\{w_i(\theta)\}_{i=1}^n$ sum to n and I-projection onto the ℓ_1 ball: $\|w(\theta) - \mathbf{1}\|_1 \le n\epsilon$.
- $\epsilon \in (0, 1)$ denotes amount of model misspecification, which can automatically be tuned from data.

Features

- Weights assign a confidence to each data point.
- Implemented for a variety of models with product likelihoods: Linear/Logistic Regression and Bernoulli/Gaussian Mixtures.
- Customizable code: `https://github.com/cjtosh/owl`

Contents

Micro-credit study by Angelucci et al. (2015)

Randomized credit rollout across 238 geographical regions in north-central Sonora state, Mexico; and 18-36 months after rollout, surveyed $n = 16,560$ households across the region to understand impact.

Consider the Average Treatment Effect (ATE) on household profits (i.e. the coefficient β_1) in the model:

$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i \qquad i = 1, \ldots, n$$

$Y_i =$ Profit of household i (outcome; units: USD PPP/2 weeks), $T_i \in \{0, 1\}$ indicates whether household i falls in a region where credit rollout happened (treatment).

Micro-credit study by Angelucci et al. (2015)

Randomized credit rollout across 238 geographical regions in north-central Sonora state, Mexico; and 18-36 months after rollout, surveyed $n = 16,560$ households across the region to understand impact.

Consider the Average Treatment Effect (ATE) on household profits (i.e. the coefficient β_1) in the model:

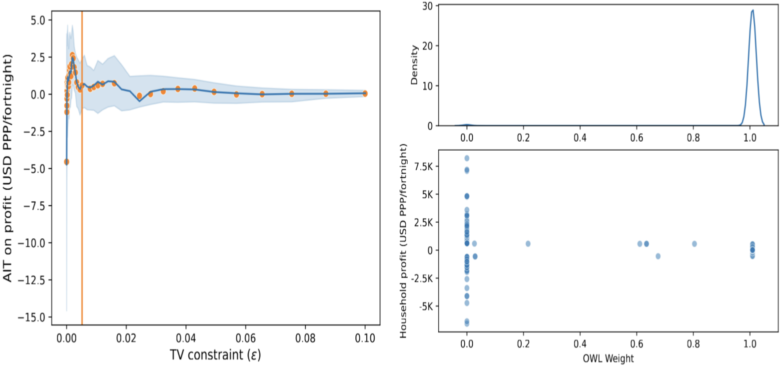
$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i \qquad i = 1, \ldots, n$$

$Y_i =$ Profit of household i (outcome; units: USD PPP/2 weeks), $T_i \in \{0, 1\}$ indicates whether household i falls in a region where credit rollout happened (treatment).

OLS estimate of β_1 is brittle [Broderick, Giordano & Meager, 2023]

Removing a single household changes β_1 from -4.55 (s.e. 5.88) to $\beta_1 = 0.4$ (s.e. 3.19); removing 15 households makes β_1 significant.

Estimating β_1 from the micro-credit study using OWL



- We estimate β_1 using OWL for 50 values of ϵ placed uniformly on $\log_{10}$-scale from -4 to -1.
- Tuning procedure selected $\epsilon_0 = 0.005$. OWL down-weighted 1% of the households with extreme profit values.
- Estimated ATE of $\beta_1 = 0.6$ USD PPP/2 weeks at $\epsilon = \epsilon_0$, is stable with respect to ϵ, and has relatively narrow bootstrap confidence bands than $\epsilon \ll \epsilon_0$.

Contents

Clustering single cell RNA-Seq using Gaussian mixtures

GSE81861 cell line dataset from Li et al. (2017)

Expression measurements for 7666 genes across 531 cells (after processing as in [Chandra et al., 2020]).

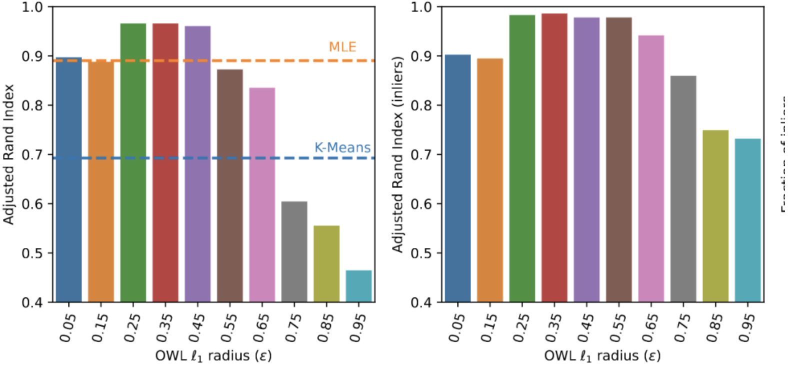
Ground truth cell-lines available:

Cell line	A549	GM12878	H1	H1437	HCT116	IMR90	K562
#	74	126	164	47	51	23	46

making this ideal to validate clustering methods.

- We use PCA to project expressions to 10 dim and fit a mixture of 7 Gaussians using OWL for a grid of ϵ values.
- Compared the resulting clustering to the ground truth cluster labels using adjusted Rand Index [Hubert and Arabie, 1985]

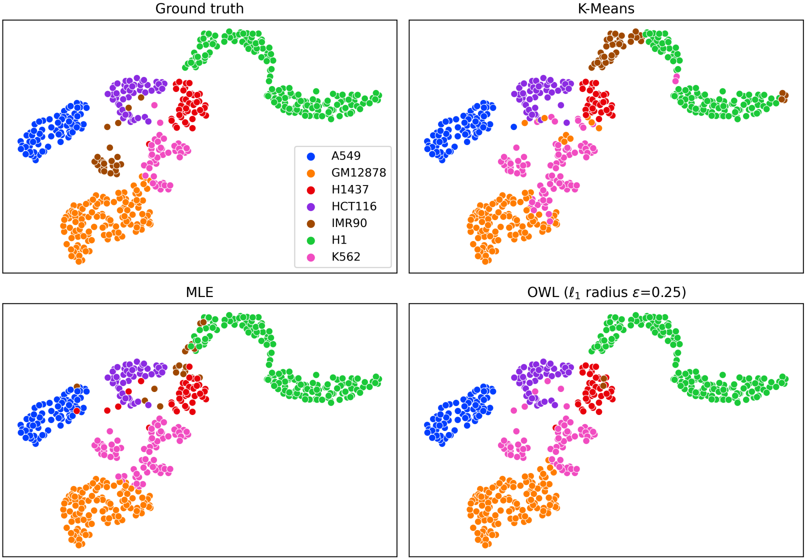
OWL improves clustering, especially on inliers



Left: Adjusted Rand index (ARI) over the entire dataset for OWL.
Right: ARI of inliers for the OWL methods.

Visualizing clusters using UMAP

Uniform Manifold Approximation and Projection. See GM12868 v.s. K562, and IMR90.



Summary

- Introduced the general coarsened likelihood framework from Miller & Dunson (2019) for inference under small misspecification in terms of a discrepancy $\boldsymbol{d}$ that define neighborhoods of empirical distribution of the observed data.

Summary

- Introduced the general coarsened likelihood framework from Miller & Dunson (2019) for inference under small misspecification in terms of a discrepancy $\boldsymbol{d}$ that define neighborhoods of empirical distribution of the observed data.
- Asymptotically approximated the coarsened likelihood using large deviation results. Used the large deviations formulas based on $\boldsymbol{d} = \mathrm{d}_{\mathtt{TV}}$ to describe a practical methodology (OWL) to robustly fit models and detect outliers.

Summary

- Introduced the general coarsened likelihood framework from Miller & Dunson (2019) for inference under small misspecification in terms of a discrepancy $\boldsymbol{d}$ that define neighborhoods of empirical distribution of the observed data.
- Asymptotically approximated the coarsened likelihood using large deviation results. Used the large deviations formulas based on $\boldsymbol{d} = \mathrm{d}_{\mathtt{TV}}$ to describe a practical methodology (OWL) to robustly fit models and detect outliers.
- OWL (Optimistically weighted likelihood) estimates the OKL minimizer by finding optimistic data re-weightings via alternating optimization. Weights down-weighted outliers in Micro credit study and improved clustering on inliers in scRNASeq data.

Thanks for your attention!

Code `https://github.com/cjtosh/owl`

Preprint `https://arxiv.org/abs/2303.10525`

Acknowledgement:

- 014-21-1-2510-P00001 from the ONR
- R01ES027498, U54 CA274492-01 and R37CA271186 from NIH
- Collaborators: Chris Tosh, Jeremias Knoblauch, and David Dunson.

Further research directions

- Use of Wasserstein neighborhoods to fit models with misspecified supports. For example, this allows us to fit models with discrete support to continuous data to perform data compression with uncertainty. Application: Brain Connectome.

Further research directions

- Use of Wasserstein neighborhoods to fit models with misspecified supports. For example, this allows us to fit models with discrete support to continuous data to perform data compression with uncertainty. Application: Brain Connectome.
- Coarsened inference for Hidden Markov Models. We can use LD formulas for HMMs (Hu and Wu, 2011) and divide & conquer ideas for fast posterior computation in long time series (Ou, Sen, Dunson, 2021).

Further research directions

- Use of Wasserstein neighborhoods to fit models with misspecified supports. For example, this allows us to fit models with discrete support to continuous data to perform data compression with uncertainty. Application: Brain Connectome.
- Coarsened inference for Hidden Markov Models. We can use LD formulas for HMMs (Hu and Wu, 2011) and divide & conquer ideas for fast posterior computation in long time series (Ou, Sen, Dunson, 2021).
- Connection to missing data problems and data privacy.

Simulation study overview

We adversarially corrupted between 0% to 25% of the observations with the largest likelihood values.

On the corrupted data we ran:

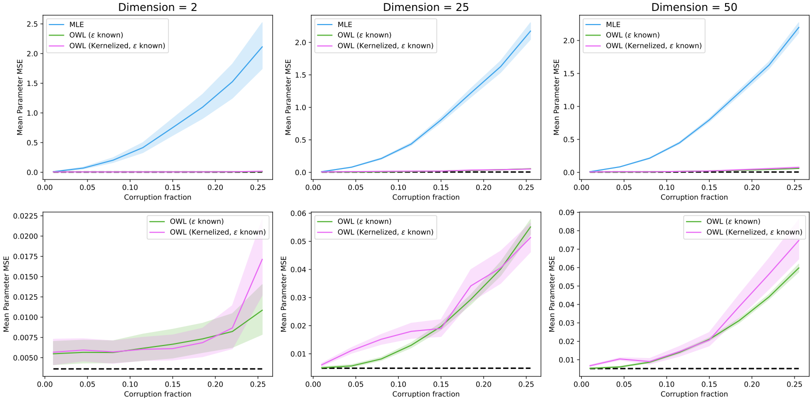
- MLE
- OWL with, both, known ϵ and tuned value of ϵ.
- Robust estimation methods when available: like Huber regression & RANSAC MLE.

We repeated the experiment 50 times to obtain error-bars. MLE on the uncorrupted sample was used as baseline.

OWL estimates with tuned ϵ are resistant to outliers, and have better (or comparable) performance than other methods.

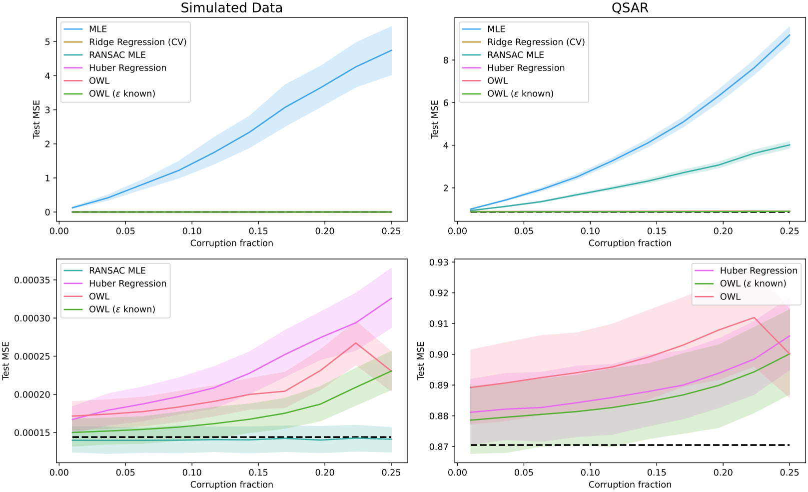
Gaussian Mean Estimation

OWL with and without the KDE have similar performance



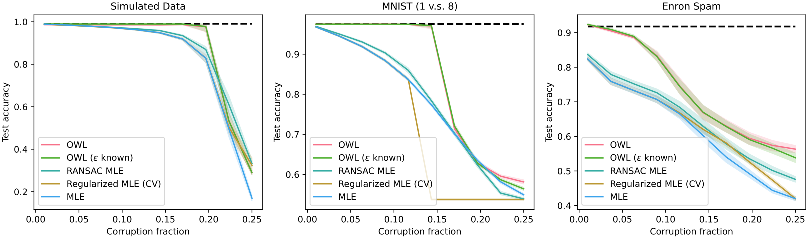
Linear Regression

OWL competitive with RANSAC MLE (left) and Huber Regression (right)



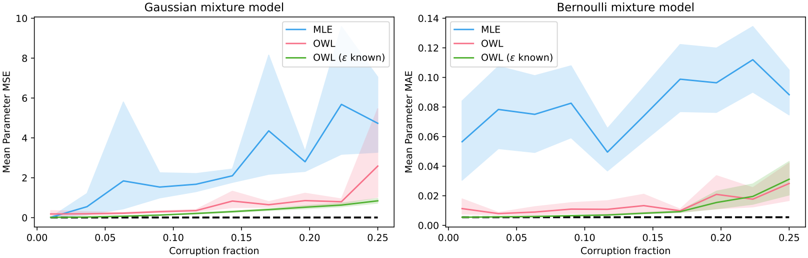
Logistic Regression

OWL most robust in terms of test-accuracy.

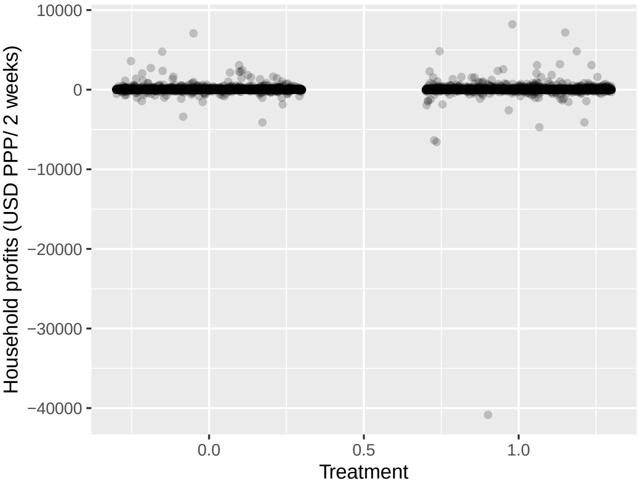


Mixture models

OWL does better than MLE for mixture models.

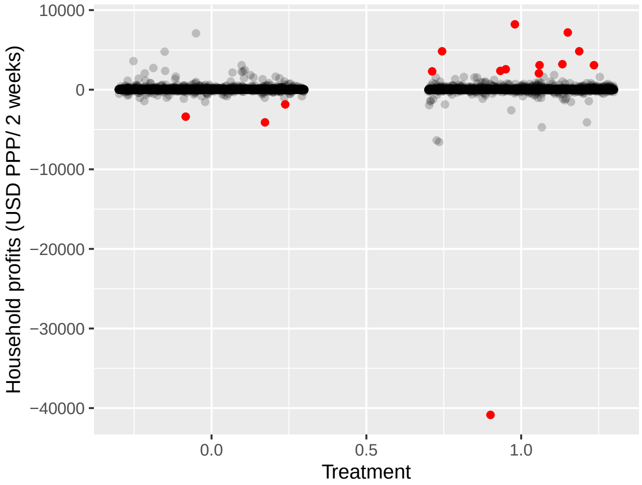


What is happening? Let's visualize the data



82% of the household profits are zero (after imputation).

What is happening? Let's visualize the data



82% of the household profits are zero (after imputation).
15 households removed by `zaminfluence` package [Broderick et al.]

OWL implementation details

Omitting KDE, extension to product likelihoods, and automatic tuning of ϵ

- Theory requires access to density estimator $\hat{p}$, but in practice we continue to get good empirical performance by omitting it.
- Thus we use the OKL estimator:

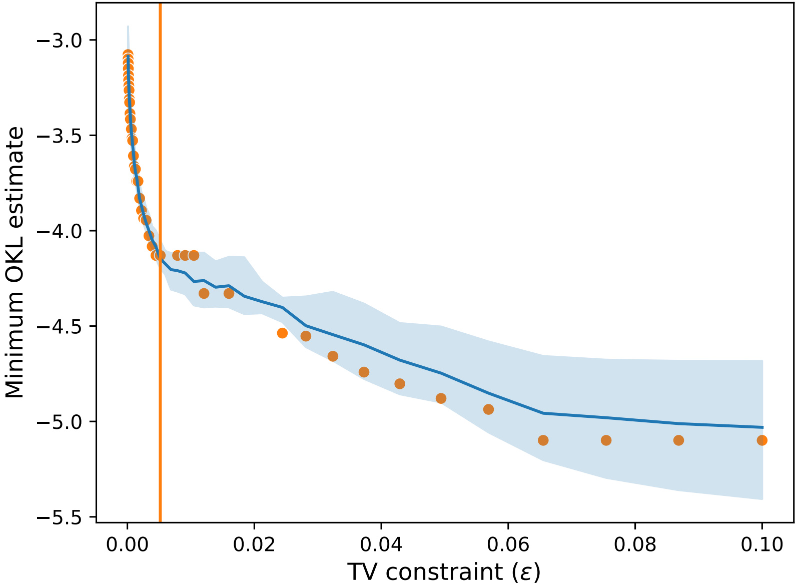
$$\hat{I}_\epsilon(\theta) = \min_{\substack{w \in \Delta_n \\ \frac{1}{2}\|w-o\|_1 \leq \epsilon}} \sum_{i=1}^{n} w_i \log w_i - \sum_{i=1}^{n} w_i \log p_\theta(x_i)$$

  which is easy to extend to likelihoods that take a conditionally product form, including regression and mixture models.

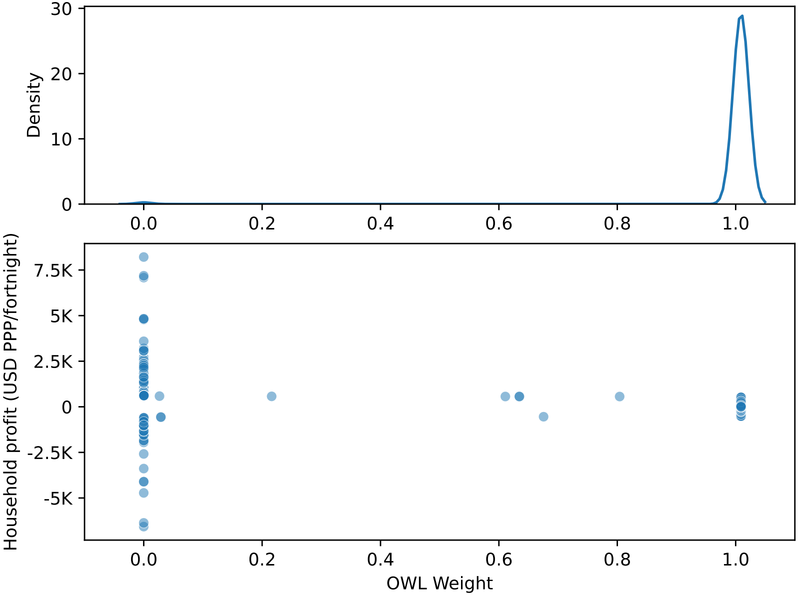
How to set parameter $\epsilon \in (0, 1)$?

- The non-increasing population function $R(\epsilon) = \min_{\theta \in \Theta} I_\epsilon(\theta)$ has a kink at $\varepsilon_0 = \min_{\theta \in \Theta} \mathrm{d}_{\mathtt{TV}}(p_0, p_\theta)$ after which it remains zero and A1 holds.
- We use an automatic procedure to find the best “kink” [Satopaa et al. 2011] in the $\hat{R}(\epsilon) = \min_{\theta \in \Theta} \hat{I}_\epsilon(\theta)$ v.s. ϵ plot.

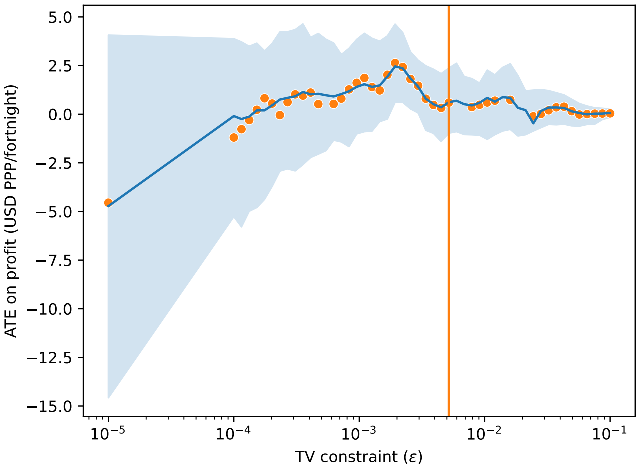
Choice of parameter $\epsilon_0 = 0.005$



OWL at ϵ_0 downweight 1% households with extreme profit.



OWL ATE estimates as function of ϵ



The leftmost point is the MLE. Confidence bands correspond to Outlier-Stratified Bootstrap.