Robustifying Likelihoods by Optimistically Re-Weighting Data

Miheer Dewaskar

Department of Statistical Science, Duke University

February 15, 2024

Robustify likelihoods by Optimistic Data Re-weighting

Introduction

Why? Slightly wrong models + Big-data = Brittleness
What? Fuzzily fit "slightly wrong" models
How? Optimism: perturb data to improve model fit

OWL Methodology

Theoretical foundations of Optimism
Optimistically Weighted Likelihoods (OWL)

Applications

Micro Credit study
Clustering of scRNA-Seq data
Concluding remarks

Why? Slightly wrong models + Big-data = Brittleness

Big Data and Statistical Challenges

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)

Isn't big data all that is necessary? Do we still need Statistics?

Big Data and Statistical Challenges

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)

Isn't big data all that is necessary? Do we still need Statistics?

Yes. The data:

- may have sampling or selection bias
- may have unknown data collection artifacts

Reference: Special issue of Statistics & Probability letters, Vol. 136

Big Data and Statistical Challenges

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)

Isn't big data all that is necessary? Do we still need Statistics?

Yes. The data:

- may have sampling or selection bias
- may have unknown data collection artifacts

Reference: Special issue of Statistics & Probability letters, Vol. 136

Need a new framework for statistical modeling of large datasets.

- For example, classical theory only assumes sampling uncertainty, leading to order $n^{-1/2}$ estimation errors (CLT).
- For large n, these errors are often 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 model 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 model misspecification due to: outliers, data contamination, and assumptions like Gaussianity.
- Concern with brittleness: sometimes even slight misspecification can have substantial impact on inference, especially for large sample sizes.

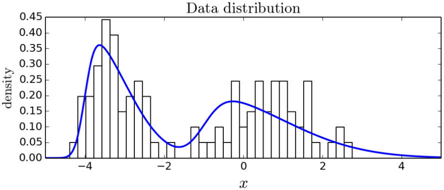
Fit Interpretable Models to Big Data

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

Example I: Brittleness in mixture model selection

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

Data is generated from a mixture of two skew Gaussians:

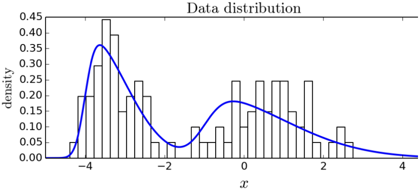


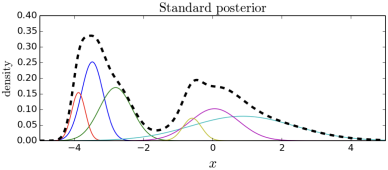
Fit a Gaussian mixture model with prior on the # of components:

Example I: Brittleness in mixture model selection

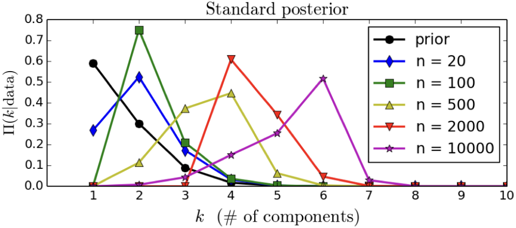
Example from Miller & Dunson (2015) that has minor misspecification in the kernel

Data is generated from a mixture of two skew Gaussians:





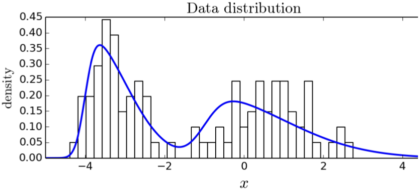
Fit a Gaussian mixture model with prior on the # of components:

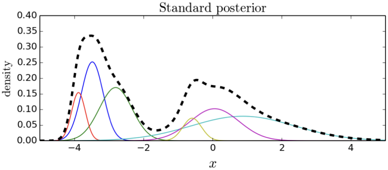


Example I: Brittleness in mixture model selection

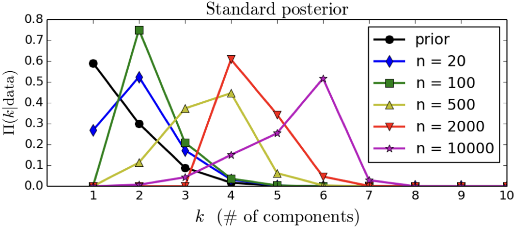
Example from Miller & Dunson (2015) that has minor misspecification in the kernel

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.

References: Miller & Dunson (2019). Theory by Cai, Campbell, Broderick (2021).

What? Fuzzily fit “slightly wrong” models

Example II: Brittleness of MLE to outliers

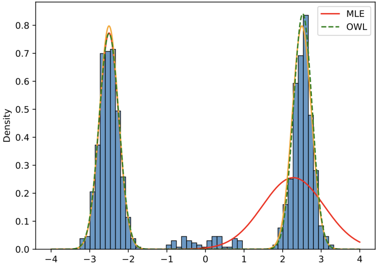
Outliers/data contamination corresponds to misspecification in Total Variation (TV)

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 robust 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 robust to the 5% contaminated data?

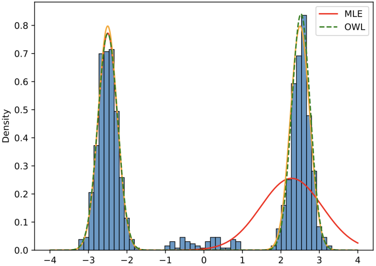


- Maximum Likelihood Estimates (MLE) is known to be brittle to data contamination. This has led to the field of robust statistics (e.g. influence functions). Reference: Maronna, Martin, Yohai (2019).

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 robust to the 5% contaminated data?

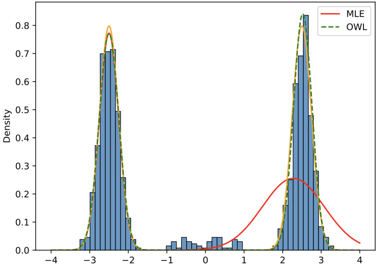


- Maximum Likelihood Estimates (MLE) is known to be brittle to data contamination. This has led to the field of robust statistics (e.g. influence functions). Reference: Maronna, Martin, Yohai (2019).
- Problem persists even if you try to fit $k \geq 2$ mixture components. Small contamination can badly affect the MLE.

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 robust to the 5% contaminated data?



- Maximum Likelihood Estimates (MLE) is known to be brittle to data contamination. This has led to the field of robust statistics (e.g. influence functions). Reference: Maronna, Martin, Yohai (2019).
- Problem persists even if you try to fit $k \geq 2$ mixture components. Small contamination can badly affect the MLE.
- This is small misspecification in the total-variation distance. Optimistically Weighted Likelihood (OWL) automatically corrects for this problem.

Problem summary

Interpretable models will tend to be slightly misspecified.

Problem summary

Interpretable models will tend to be slightly misspecified.

For large n, standard inference (Bayes/MLE) can be problematic even under minor misspecification.

Problem summary

Interpretable models will tend to be slightly misspecified.

For large n, standard inference (Bayes/MLE) can be problematic even under minor misspecification.

Problem: Find a way to fuzzily/inexactly fit models that may be slightly misspecified.

Problem summary

Interpretable models will tend to be slightly misspecified.

For large n, standard inference (Bayes/MLE) can be problematic even under minor misspecification.

Problem: Find a way to fuzzily/inexactly fit models that may be slightly misspecified.

Formalism

Suppose $\{P_\theta\}_{\theta \in \Theta}$ is our model family, and P_o is the true distribution of the observed data. Assume misspecification: $P_o \notin \{P_\theta\}_{\theta \in \Theta}$.

Problem summary

Interpretable models will tend to be slightly misspecified.

For large n, standard inference (Bayes/MLE) can be problematic even under minor misspecification.

Problem: Find a way to fuzzily/inexactly fit models that may be slightly misspecified.

Formalism

Suppose $\{P_\theta\}_{\theta \in \Theta}$ is our model family, and P_o is the true distribution of the observed data. Assume misspecification: $P_o \notin \{P_\theta\}_{\theta \in \Theta}$.

The Bayesian posterior and MLE target [Kleijn and van der Vaart (2012/2006)]

$$\theta_1 = \arg\min_{\theta \in \Theta} \mathrm{KL}(P_o | P_\theta).$$

which may be brittle to the tails and support of P_o.

Problem summary

Interpretable models will tend to be slightly misspecified.

For large n, standard inference (Bayes/MLE) can be problematic even under minor misspecification.

Problem: Find a way to fuzzily/inexactly fit models that may be slightly misspecified.

Formalism

Suppose $\{P_\theta\}_{\theta \in \Theta}$ is our model family, and P_o is the true distribution of the observed data. Assume misspecification: $P_o \notin \{P_\theta\}_{\theta \in \Theta}$.

The Bayesian posterior and MLE target [Kleijn and van der Vaart (2012/2006)]

$$\theta_1 = \arg\min_{\theta \in \Theta} \mathrm{KL}(P_o | P_\theta).$$

which may be brittle to the tails and support of P_o.

We want to find $\theta_0 \in \Theta$ such that $P_{\theta_0} \approx P_o$ (in Wasserstein, TVD, etc.).

How? Optimism: perturb data to improve model fit

Key idea: fit models robustly by trusting data less

1. Assume that observed data are a biased, unreliable, or corrupted version of the "ideal" data drawn from the model P_{θ_0}.
2. We should be skeptical of the observed data and trust it less.
3. Data points that do not conform with the model P_{θ_0} should not be allowed to unduly influence the parameter estimates.

This can address the examples of brittleness we saw earlier. But we don't know the true model P_{θ_0} (we want to estimate it!)

Key idea: fit models robustly by trusting data less

1. Assume that observed data are a biased, unreliable, or corrupted version of the "ideal" data drawn from the model P_{θ_0}.
2. We should be skeptical of the observed data and trust it less.
3. Data points that do not conform with the model P_{θ_0} should not be allowed to unduly influence the parameter estimates.

This can address the examples of brittleness we saw earlier. But we don't know the true model P_{θ_0} (we want to estimate it!)

This idea has appeared in the literature on learning from imprecise data.

> *Roughly, the idea is to [...] fit the model to the data and the data to the model [simultaneously]. – Eyke Hüllermeier*

[Hüllermeier, 14], [Hüllermeier & Cheng, 15], [Hüllermeier, Destercke and Couso, 19], [Lienen, HÏlerme, 21a,b]

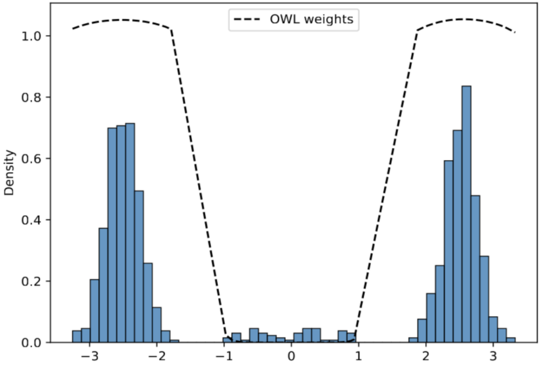
Optimistically re-interpret data

Compute MLE based on a best-case dataset "near" the observed dataset.

We use re-weightings to represent nearby datasets

We use re-weightings to represent nearby datasets

Optimistic re-weighting in Example 2 perturbs the data to look like it was drawn from a mixture of two Gaussians.



This is best-case data perturbation in contrast to the worst-case perturbation used in DRO (Namkoong & Duchi, 2016).

Why use data re-weightings?

Suppose $x_1, \ldots, x_n \stackrel{i.i.d.}{\sim} P_o$. We can consider the re-weighted distribution:

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

for weights $w_1, \ldots, w_n \geq 0$ and $\sum_{i=1}^{n} w_i = n$.

Why use data re-weightings?

Suppose $x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$. We can consider the re-weighted distribution:

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

for weights $w_1, \ldots, w_n \geq 0$ and $\sum_{i=1}^{n} w_i = n$.

Re-weightings:

- **are powerful**. Re-weightings can transform samples from P_o to those of any absolutely continuous distribution Q using Radon-Nikodym derivatives. (E.g. Gaussian to a t or Gamma distribution.)

Why use data re-weightings?

Suppose $x_1, \dots, x_n \overset{i.i.d.}{\sim} P_o$. We can consider the re-weighted distribution:

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

for weights $w_1, \dots, w_n \geq 0$ and $\sum_{i=1}^n w_i = n$.

Re-weightings:

- **are powerful**. Re-weightings can transform samples from P_o to those of any absolutely continuous distribution Q using Radon-Nikodym derivatives. (E.g. Gaussian to a t or Gamma distribution.)
- **require simple modifications to existing algorithms**. It turns out that one only needs to adopt existing algorithms for MLE or Bayes posteriors to work with weighted likelihoods $\prod_{i=1}^n p_\theta(x_i)^{w_i}$.

Why use data re-weightings?

Suppose $x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$. We can consider the re-weighted distribution:

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

for weights $w_1, \ldots, w_n \geq 0$ and $\sum_{i=1}^n w_i = n$.

Re-weightings:

- **are powerful**. Re-weightings can transform samples from P_o to those of any absolutely continuous distribution Q using Radon-Nikodym derivatives. (E.g. Gaussian to a t or Gamma distribution.)
- **require simple modifications to existing algorithms**. It turns out that one only needs to adopt existing algorithms for MLE or Bayes posteriors to work with weighted likelihoods $\prod_{i=1}^n p_\theta(x_i)^{w_i}$.
- **are interpretable**. Weights provide a summary of how much each observation is trusted by the estimated model.

Optimistic re-weighting: an operational definition

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

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

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

Optimistic re-weighting: an operational definition

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

Optimistic weights $w_1, \ldots, w_n \geq 0$ and $\sum_{i=1}^n w_i = n$ are 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

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

- There is no need for optimism in the well-specified case. That is, [OWL] holds with $\epsilon = 0$ and $w_i = 1$ (MLE).

Optimistic re-weighting: an operational definition

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

Optimistic weights $w_1, \ldots, w_n \geq 0$ and $\sum_{i=1}^n w_i = n$ are 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

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

- There is no need for optimism in the well-specified case. That is, [OWL] holds with $\epsilon = 0$ and $w_i = 1$ (MLE).
- Otherwise the degree of optimism is related to the degree of misspecification. Weights satisfying [OWL] exists $\iff$ $\mathrm{d}_{\mathtt{TV}}(P_o, P_{\theta^*}) \leq \epsilon$ for some $\theta^* \in \Theta$.

Robustify likelihoods by Optimistic Data Re-weighting

Introduction

Why? Slightly wrong models + Big-data = Brittleness
What? Fuzzily fit "slightly wrong" models
How? Optimism: perturb data to improve model fit

OWL Methodology

Theoretical foundations of Optimism
Optimistically Weighted Likelihoods (OWL)

Applications

Micro Credit study
Clustering of scRNA-Seq data
Concluding remarks

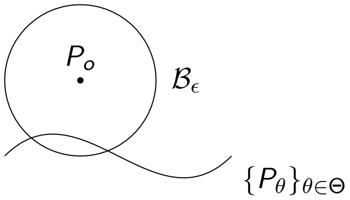
Theoretical foundations of Optimism

Robust model estimation: setup and assumptions

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

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

Assumption: $\Theta_I \neq \emptyset$.



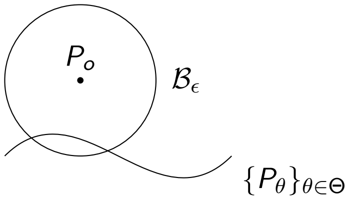
$$\mathcal{B}_\epsilon = \{Q : \mathrm{d}_{\mathtt{TV}}(Q, P_o) \leq \epsilon\}$$

Robust model estimation: setup and assumptions

Setup: 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: $\Theta_I \neq \emptyset$.



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

Example (Huber's contamination model)

If $P_o = (1-\epsilon)P_{\theta^*} + \epsilon C$, then $\theta^* \in \Theta_I$.

- Identifiability only upto Θ_I. This is not a practical problem as Θ_I is small when ϵ is small [Huber, 1964].
- OWL aims to estimate some element from Θ_I.

Robustly identified parameters are minimizers of OKL

Suppose P_o and $\epsilon \geq 0$ are known.
We define the following population level objective function:

Optimistic Kullback Leibler (OKL) from Large Deviations

$$I_\epsilon(\theta) = \min_{Q : \mathrm{d}_{\mathtt{TV}}(Q, P_o) \leq \epsilon} \mathrm{KL}(Q | P_\theta),$$

A unique minimizer Q^θ exists [Csiszar, 1975] and is called the information projection of P_θ on $\mathcal{B}_\epsilon = \{Q : \mathrm{d}_{\mathtt{TV}}(Q, P_o) \leq \epsilon\}$.

Robustly identified parameters are minimizers of OKL

Suppose P_o and $\epsilon \geq 0$ are known.
We define the following population level objective function:

Optimistic Kullback Leibler (OKL) from Large Deviations

$$I_\epsilon(\theta) = \min_{Q : \mathrm{d}_{\mathtt{TV}}(Q, P_o) \leq \epsilon} \mathrm{KL}(Q|P_\theta),$$

A unique minimizer Q^θ exists [Csiszar, 1975] and is called the information projection of P_θ on $\mathcal{B}_\epsilon = \{Q : \mathrm{d}_{\mathtt{TV}}(Q, P_o) \leq \epsilon\}$.

- When $\epsilon = 0$, $I_0(\theta) = \mathrm{KL}(P_o|P_\theta)$.
- For $\epsilon > 0$, one finds an Optimistic data re-interpretation $Q^\theta \in \mathcal{B}_\epsilon$ that minimizes the KL divergence to P_θ.

The robust parameters Θ_I can be seen as the minimizers of OKL:

$$\arg\min_{\theta \in \Theta} I_\epsilon(\theta) = \Theta_I$$

Optimistically Weighted Likelihoods (OWL)

Minimize the OKL function using alternating optimization

Minimizing the OKL function corresponds to solving the double minimization:

$$\min_{\theta \in \Theta} \min_{Q : \mathrm{d}_{\mathtt{TV}}(Q, P_o) \leq \epsilon} \mathsf{KL}\left(Q | P_\theta\right).$$

Following EM/MM algorithms, this can be done using the following alternating minimization scheme.

Minimize the OKL function using alternating optimization

Minimizing the OKL function corresponds to solving the double minimization:

$$\min_{\theta \in \Theta} \min_{Q : \mathrm{d}_{\mathtt{TV}}(Q, P_o) \leq \epsilon} \mathrm{KL}\left(Q | P_\theta\right).$$

Following EM/MM algorithms, this can be done using the following alternating minimization scheme.

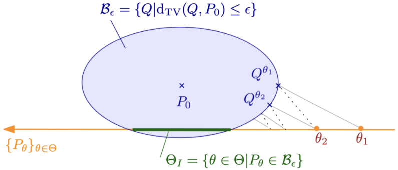
Start from some θ_1 and iterate (for $t = 1, \ldots$) until convergence:

Information-projection:

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

Maximize log-likelihood:

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



The Optimistically Weighted Likelihood Algorithm

Emulate the previous algorithm based on a consistent estimator of the OKL function using samples $x_1, \ldots, x_n \sim P_o$. We want to solve

$$\min_{\theta \in \Theta} \min_{Q_w : \mathrm{d}_{\mathrm{TV}}(Q_w, P_o) \leq \epsilon} \hat{\mathrm{KL}}\left(Q_w | P_\theta\right).$$

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

The Optimistically Weighted Likelihood Algorithm

Emulate the previous algorithm based on a consistent estimator of the OKL function using samples $x_1, \dots, x_n \sim P_o$. We want to solve

$$\min_{\theta \in \Theta} \min_{Q_w : \mathrm{d}_{\mathrm{TV}}(Q_w, P_o) \leq \epsilon} \hat{\mathrm{KL}}\left(Q_w | P_\theta\right).$$

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

This leads following alternating optimization steps (for $t = 1, \dots$)

Approx I-projection:

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

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 Likelihood (OWL) summary

- Theoretically motivated from OKL minimization.
- We jointly 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 are defined by the I-projection of $\{p_\theta(x_i)\}_{i=1}^n$ onto the intersection of ℓ_1 ball $\mathcal{B}_\epsilon = \{w : \|w - \mathbf{1}\|_1 \leq n\epsilon\}$ and the simplex of weights.
- $\epsilon \in (0, 1)$ denotes amount of model misspecification, which can automatically be tuned from data.

Optimistically Weighted Likelihood (OWL) summary

- Theoretically motivated from OKL minimization.
- We jointly 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 are defined by the I-projection of $\{p_\theta(x_i)\}_{i=1}^n$ onto the intersection of ℓ_1 ball $\mathcal{B}_\epsilon = \{w : \|w - \mathbf{1}\|_1 \leq n\epsilon\}$ and the simplex of weights.
- $\epsilon \in (0, 1)$ 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`

Robustify likelihoods by Optimistic Data Re-weighting

Introduction

Why? Slightly wrong models + Big-data = Brittleness

What? Fuzzily fit "slightly wrong" models

How? Optimism: perturb data to improve model fit

OWL Methodology

Theoretical foundations of Optimism

Optimistically Weighted Likelihoods (OWL)

Applications

Micro Credit study

Clustering of scRNA-Seq data

Concluding remarks

Micro Credit study

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:

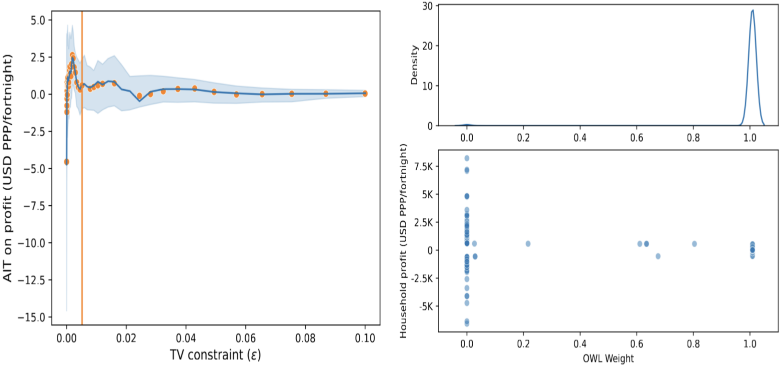
$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i \qquad i = 1, \ldots, n$$

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 $\epsilon \ll \epsilon_0$.

Clustering of scRNA-Seq data

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]).

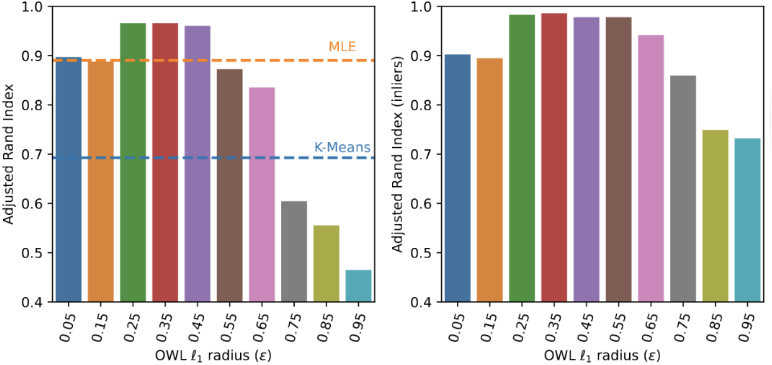
Ground truth cell-lines available:

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]

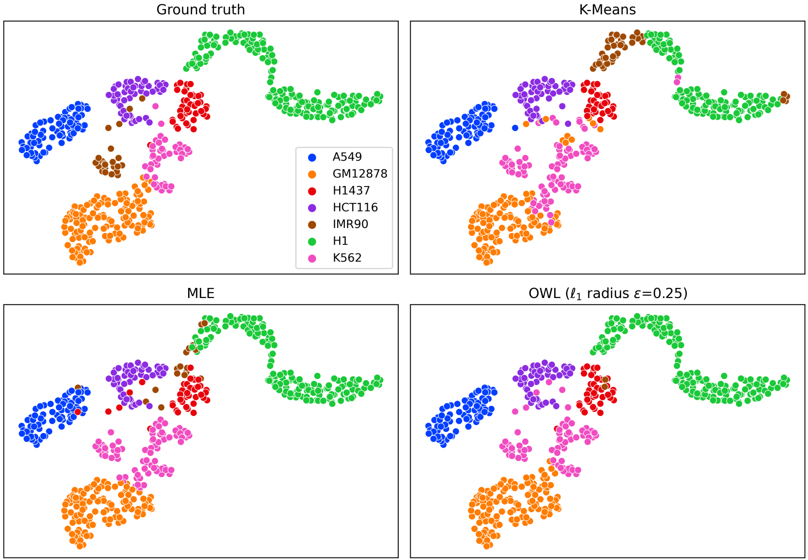
OWL improves clustering, especially on inliers



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.



Concluding remarks

Summary

Our primary objective has been to develop methods to fuzzily/inexactly fit likelihood-based models to complex data.

- Our key idea is to let the model inform which data points we trust, and would like to allow to influence our model fit.

Summary

Our primary objective has been to develop methods to fuzzily/inexactly fit likelihood-based models to complex data.

- Our key idea is to let the model inform which data points we trust, and would like to allow to influence our model fit.
- OWL (Optimistically weighted likelihood) implements a practical version of this scheme by looking for weighted perturbations in a small TV-neighborhood of the observed data that can improve the model fit.

Summary

Our primary objective has been to develop methods to fuzzily/inexactly fit likelihood-based models to complex data.

- Our key idea is to let the model inform which data points we trust, and would like to allow to influence our model fit.
- OWL (Optimistically weighted likelihood) implements a practical version of this scheme by looking for weighted perturbations in a small TV-neighborhood of the observed data that can improve the model fit.
- OWL is implemented as an alternating minimization that jointly estimates the model (via weighted MLE) and the optimistic weights (via I-projection).

Summary

Our primary objective has been to develop methods to fuzzily/inexactly fit likelihood-based models to complex data.

- Our key idea is to let the model inform which data points we trust, and would like to allow to influence our model fit.
- OWL (Optimistically weighted likelihood) implements a practical version of this scheme by looking for weighted perturbations in a small TV-neighborhood of the observed data that can improve the model fit.
- OWL is implemented as an alternating minimization that jointly estimates the model (via weighted MLE) and the optimistic weights (via I-projection).
- OWL weights down-weighted outliers in Micro credit study and improved clustering on inliers in scRNASeq data.

Future directions

Tons of exciting areas to work on!

- **Coarsened Inference**: Although we only considered robust point estimation and not inference, OWL is statistically motivated by coarsened inference framework of Miller & Dunson, 2019. Currently, we are working on a theory for robust Bayesian inference, allowing for robust uncertainty quantification and model selection.
- **Models beyond product likelihoods**: The coarsened inference philosophy allow us to move beyond models with product likelihoods. I am interested in extensions to hierarchical and spatio-temporal models, particularly in applications to climate modeling.
- **Interesting Applications** to differential private inferences, borrowing information across historical data in clinical trials, and data compression problem.

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.

Robustify likelihoods by Optimistic Data Re-weighting

Statistical Foundations

Coarsened Inference Framework

Computation of the coarsened posterior

Estimator for Optimistic Kullback Leibler (OKL)

Coarsened Inference Framework

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

Bayesian model: $\boldsymbol{X} = X_1, \ldots, X_n \stackrel{i.i.d.}{\sim} P_\vartheta$ 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.

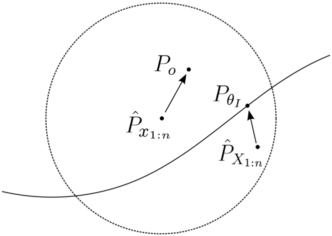
Empirical measure: $\hat{P}_{\boldsymbol{x}} \doteq \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ is a sufficient statistic.

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:

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



Handle misspecification by "coarsening" posterior

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

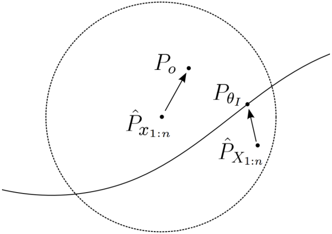
We observe data $\boldsymbol{x} = x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$ from unknown $P_o \in \mathcal{P}(\mathcal{X})$.

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.

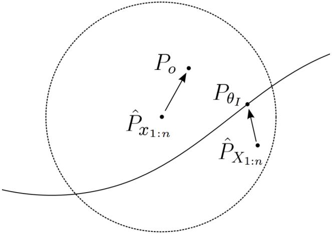
We observe data $\boldsymbol{x} = x_1, \ldots, x_n \overset{i.i.d.}{\sim} P_o$ from unknown $P_o \in \mathcal{P}(\mathcal{X})$.

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 the coarsened posterior

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). Slow convergence: 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). Slow convergence: 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). Slow convergence: 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 $n_0 = \frac{n\alpha}{\alpha+n} < \infty$.

General asymptotics of the coarsened likelihood

Sanov's theorem from Large Deviations shows:

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

$$-\frac{1}{n}\log L_\epsilon(\theta|\boldsymbol{x}) \xrightarrow{\mathrm{P}} I_\epsilon(\theta) \doteq \inf_{\substack{Q\in\mathcal{P}(\mathcal{X})\\ \boldsymbol{d}(Q,P_o)\leq\epsilon}} \mathrm{KL}(Q|P_\theta)$$

We call $I_\epsilon(\theta)$ 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 in the $(\boldsymbol{d}, \epsilon)$ ball around P_o.

General asymptotics of the coarsened likelihood

Sanov's theorem from Large Deviations shows:

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

$$-\frac{1}{n}\log L_\epsilon(\theta|\boldsymbol{x}) \xrightarrow{\text{P}} I_\epsilon(\theta) \doteq \inf_{\substack{Q\in\mathcal{P}(\mathcal{X})\\ \boldsymbol{d}(Q,P_o)\leq\epsilon}} \text{KL}(Q|P_\theta)$$

We call $I_\epsilon(\theta)$ 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 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 $\iff$ $\theta^*\in\arg\min_{\theta\in\Theta}\text{KL}(P_o|P_\theta)$.

Estimator for Optimistic Kullback Leibler (OKL)

Estimation of the OKL using data re-weightings

Finite spaces

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

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

with $o = (1/n, \ldots, 1/n)$ to target the OKL:

$$I_\epsilon(\theta) = \min_{Q : \mathrm{d}_{\mathtt{TV}}(Q, P_o) \leq \epsilon} \mathrm{KL}(Q | P_\theta).$$

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

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

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

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 $A_{ij} = \frac{\kappa_h(x_i, x_j)}{n\hat{p}(x_i)}$.

$$\begin{aligned} \hat{I}_{h,\epsilon}(\theta) &\doteq \min_{\substack{v \in A\Delta_n \\ \frac{1}{2}\|v - o\|_1 \le \epsilon}} \sum_{i=1}^n v_i \log \frac{n v_i \hat{p}(x_i)}{p_\theta(x_i)} \\ &= \min_{\substack{w \in \Delta_n \\ \mathrm{d}_{\mathrm{TV}}(q_w, \hat{p}) \le \epsilon}} \frac{1}{n} \sum_{i=1}^n \frac{q_w(x_i)}{\hat{p}(x_i)} \log \frac{q_w(x_i)}{p_\theta(x_i)} \approx \min_{\substack{w \in \Delta_n \\ \mathrm{d}_{\mathrm{TV}}(q_w, p_o) \le \epsilon}} \mathrm{KL}(q_w | p_\theta). \end{aligned}$$

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 $\mathcal{X}$:

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

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).
- Connections to differentially private inference and informative prior elicitation in clinical trials!

Simulation study overview

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

On the corrupted data we ran:

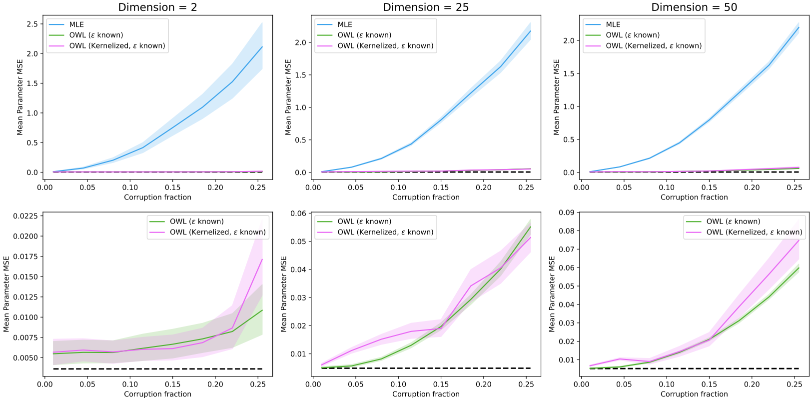
- MLE
- OWL with, both, known ϵ and tuned value of ϵ.
- Robust estimation methods when available: like Huber regression & RANSAC MLE.

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.

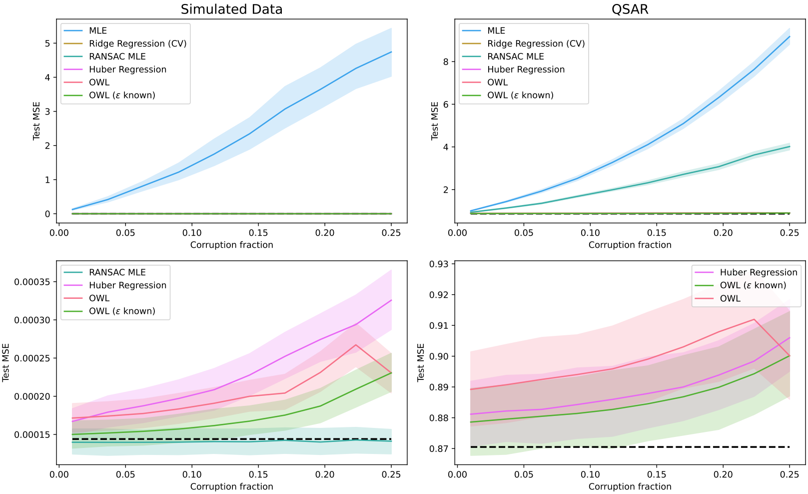
Gaussian Mean Estimation

OWL with and without the KDE have similar performance



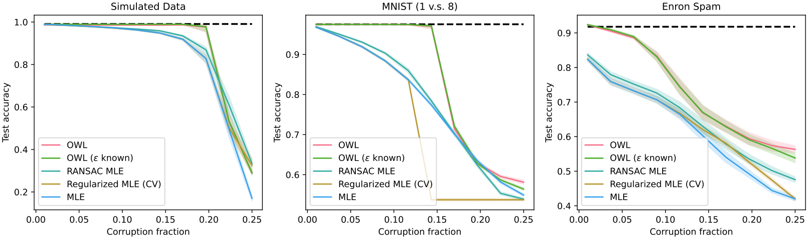
Linear Regression

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



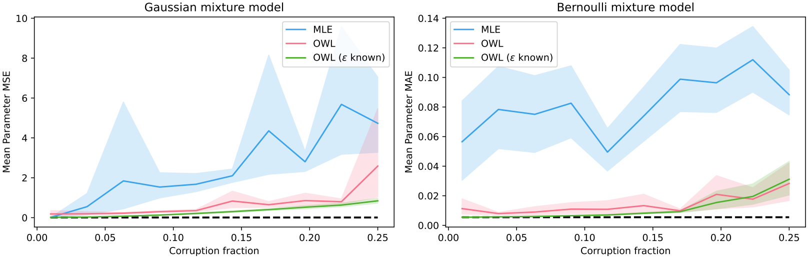
Logistic Regression

OWL most robust in terms of test-accuracy.

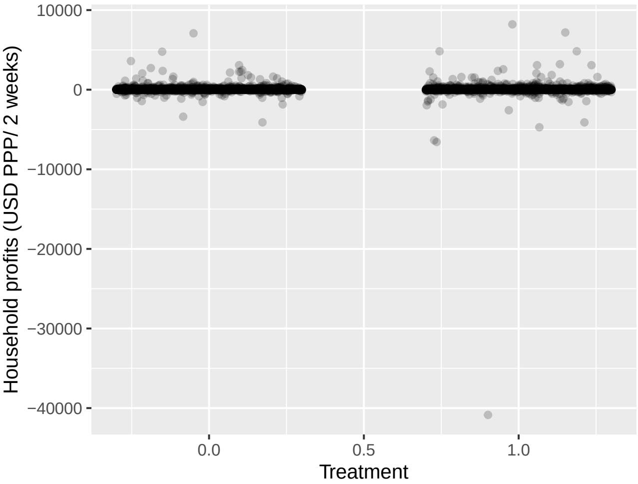


Mixture models

OWL does better than MLE for mixture models.

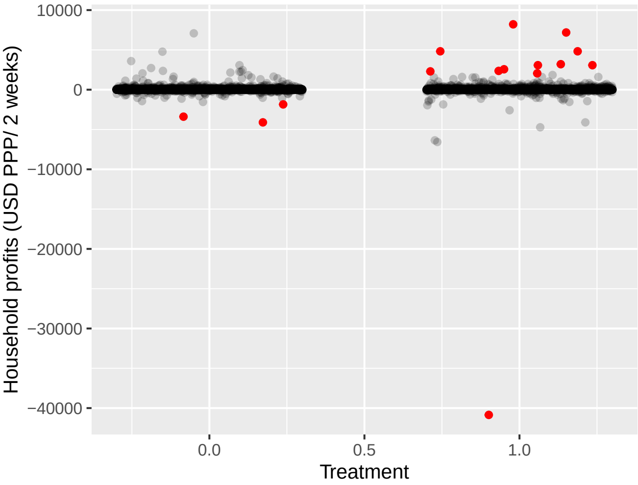


What is happening? Let's visualize the data



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:

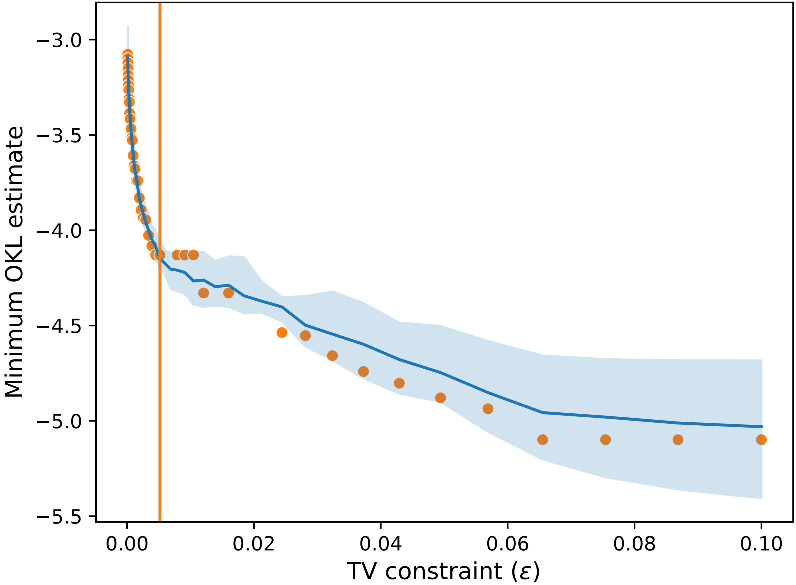
$$\hat{I}_\epsilon(\theta) = \min_{\substack{w \in \Delta_n \\ \frac{1}{2}\|w - o\|_1 \leq \epsilon}} \sum_{i=1}^{n} w_i \log w_i - \sum_{i=1}^{n} w_i \log p_\theta(x_i)$$

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

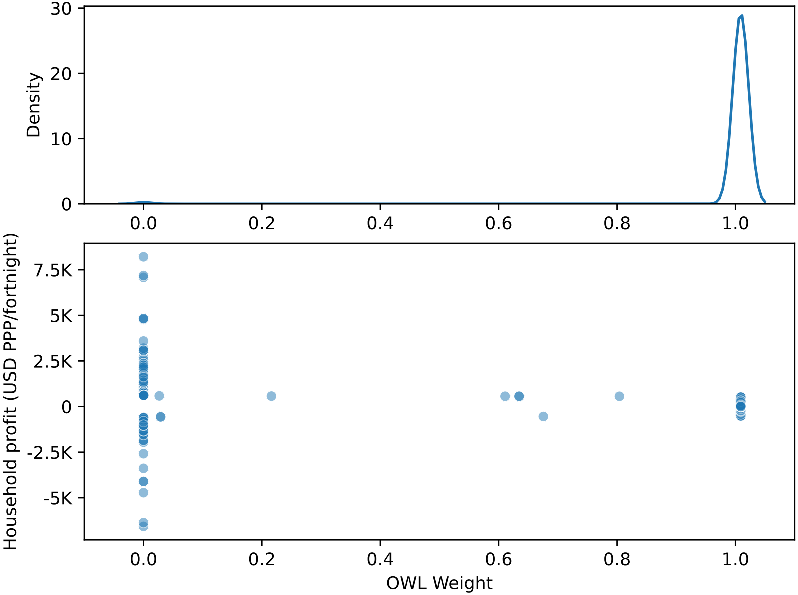
How to set parameter $\epsilon \in (0, 1)$?

- The non-increasing population function $R(\epsilon) = \min_{\theta \in \Theta} I_\epsilon(\theta)$ has a kink at $\varepsilon_0 = \min_{\theta \in \Theta} \mathrm{d}_{\mathrm{TV}}(p_0, p_\theta)$ 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.

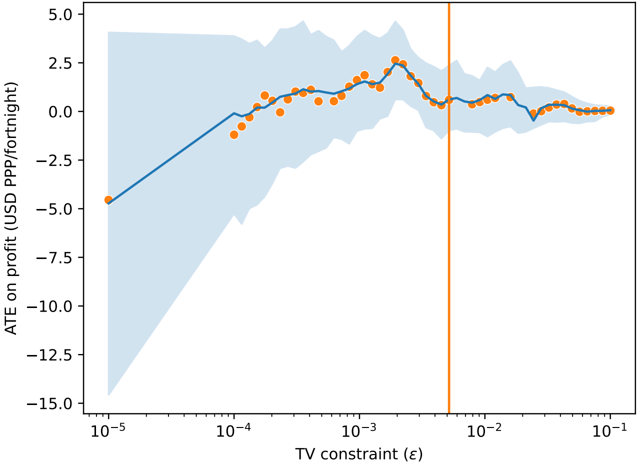
Choice of parameter $\epsilon_0 = 0.005$



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.