MIT EECS
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

MIT STATISTICS + DATA SCIENCE CENTER

LABORATORY FOR INFORMATION & DECISION SYSTEMS
LIDS

An Automatic Finite-Data Robustness Check for Bayes and Beyond: Can Dropping a Little Data Change Conclusions?

Tamara Broderick

Associate Professor,
MIT

With Ryan Giordano, Rachael Meager

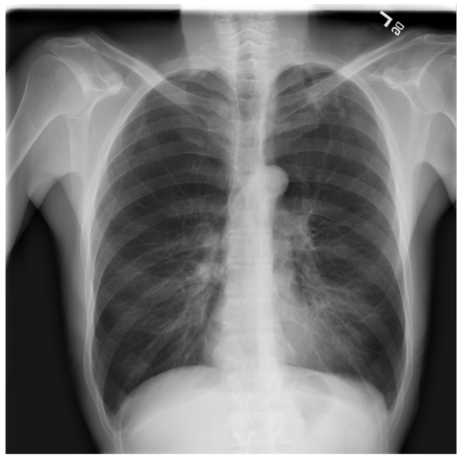
When can I trust my data analysis?

When can I trust my data analysis?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions

When can I trust my data analysis?

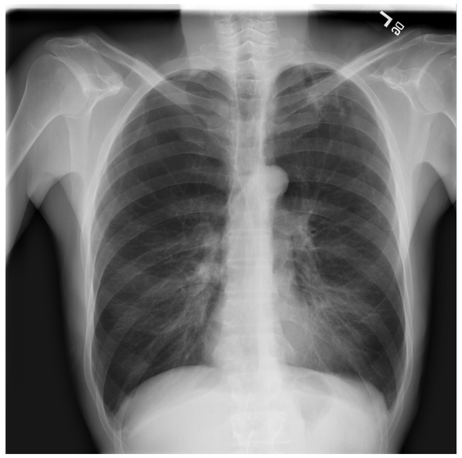
- More data & better computation ➔ data analyses increasingly drive life-changing decisions



[https://stanfordmlgroup.github.io/projects/chexnet/, https://www.ibmbigdatahub.com/blog/cyber-security-powered-ai-and-machine-learning, https://www.fairwinds.org/inside-fairwinds/fraud-protection-center/report-fraud.html]

When can I trust my data analysis?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions

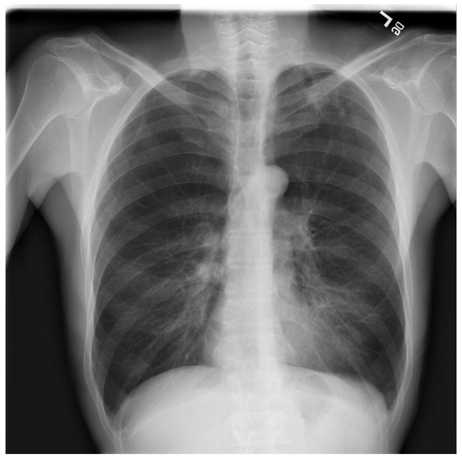


- Typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data

[https://stanfordmlgroup.github.io/projects/chexnet/, https://www.ibmbigdatahub.com/blog/cyber-security-powered-ai-and-machine-learning, https://www.fairwinds.org/inside-fairwinds/fraud-protection-center/report-fraud.html]

When can I trust my data analysis?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions

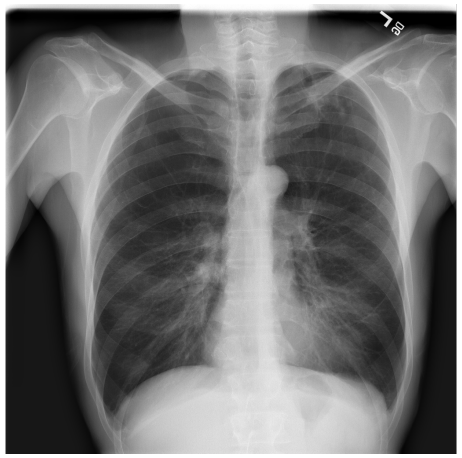


- Typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
- Might worry about generalizing if very small subset of data was instrumental to original analysis

[https://stanfordmlgroup.github.io/projects/chexnet/, https://www.ibmbigdatahub.com/blog/cyber-security-powered-ai-and-machine-learning, https://www.fairwinds.org/inside-fairwinds/fraud-protection-center/report-fraud.html]

When can I trust my data analysis?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions

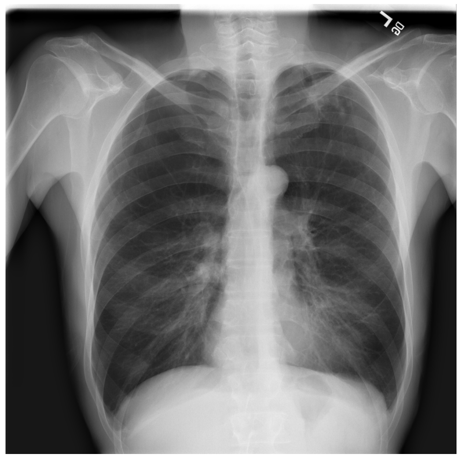


- Typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
- Might worry about generalizing if very small subset of data was instrumental to original analysis
  - E.g. in a study of microcredit with ~16,500 data points, we find just one data point drives the sign of the effect

[https://stanfordmlgroup.github.io/projects/chexnet/, https://www.ibmbigdatahub.com/blog/cyber-security-powered-ai-and-machine-learning, https://www.fairwinds.org/inside-fairwinds/fraud-protection-center/report-fraud.html]

When can I trust my data analysis?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions

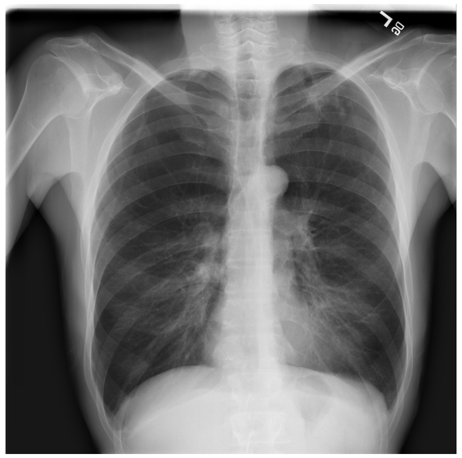


- Typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
- Might worry about generalizing if very small subset of data was instrumental to original analysis
  - E.g. in a study of microcredit with ~16,500 data points, we find just one data point drives the sign of the effect
- **Challenge**: Impossibly costly to check every data subset

[https://stanfordmlgroup.github.io/projects/chexnet/, https://www.ibmbigdatahub.com/blog/cyber-security-powered-ai-and-machine-learning, https://www.fairwinds.org/inside-fairwinds/fraud-protection-center/report-fraud.html]

When can I trust my data analysis?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions



- Typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
- Might worry about generalizing if very small subset of data was instrumental to original analysis
  - E.g. in a study of microcredit with ~16,500 data points, we find just one data point drives the sign of the effect
- **Challenge**: Impossibly costly to check every data subset
- **Our Solution**: a fast, automated, accurate *approximation*

[https://stanfordmlgroup.github.io/projects/chexnet/, https://www.ibmbigdatahub.com/blog/cyber-security-powered-ai-and-machine-learning, https://www.fairwinds.org/inside-fairwinds/fraud-protection-center/report-fraud.html]

Roadmap

Roadmap

- When do we care about dropping data subsets?

Roadmap

- When do we care about dropping data subsets?
- How should we drop data subsets?

Roadmap

- When do we care about dropping data subsets?
- How should we drop data subsets?
- Why is dropping data subsets computationally expensive?

Roadmap

- When do we care about dropping data subsets?
- How should we drop data subsets?
- Why is dropping data subsets computationally expensive?
- We provide a fast & automatic approximation

Roadmap

- When do we care about dropping data subsets?
- How should we drop data subsets?
- Why is dropping data subsets computationally expensive?
- We provide a fast & automatic approximation
- Many analyses are robust but some aren't

Roadmap

- When do we care about dropping data subsets?
- How should we drop data subsets?
- Why is dropping data subsets computationally expensive?
- We provide a fast & automatic approximation
- Many analyses are robust but some aren't
  - Being Bayesian doesn't guarantee robustness

Roadmap

- When do we care about dropping data subsets?
- How should we drop data subsets?
- Why is dropping data subsets computationally expensive?
- We provide a fast & automatic approximation
- Many analyses are robust but some aren't
  - Being Bayesian doesn't guarantee robustness
  - Here non-robustness isn't just a product of gross outliers, large p-values, heavy tails, etc.

Roadmap

- When do we care about dropping data subsets?
- How should we drop data subsets?
- Why is dropping data subsets computationally expensive?
- We provide a fast & automatic approximation
- Many analyses are robust but some aren't
  - Being Bayesian doesn't guarantee robustness
  - Here non-robustness isn't just a product of gross outliers, large p-values, heavy tails, etc.
  - Non-robustness is a product of low signal-to-noise

Roadmap

- When do we care about dropping data subsets?
- How should we drop data subsets?
- Why is dropping data subsets computationally expensive?
- We provide a fast & automatic approximation
- Many analyses are robust but some aren't
  - Being Bayesian doesn't guarantee robustness
  - Here non-robustness isn't just a product of gross outliers, large p-values, heavy tails, etc.
  - Non-robustness is a product of low signal-to-noise

Why care about dropping data subsets?

Why care about dropping data subsets?

- Any useful data analysis should be sensitive to *some* change in the data

Why care about dropping data subsets?

- Any useful data analysis should be sensitive to *some* change in the data
- What types of sensitivity concern us? Varies by problem

Why care about dropping data subsets?

- Any useful data analysis should be sensitive to *some* change in the data
- What types of sensitivity concern us? Varies by problem
- Let’s look at economics (e.g. because of the important applications and wonderful reproducibility)

Why care about dropping data subsets?

- Any useful data analysis should be sensitive to *some* change in the data
- What types of sensitivity concern us? Varies by problem
- Let’s look at economics (e.g. because of the important applications and wonderful reproducibility)
  - Report a convenient proxy (e.g. mean)

Why care about dropping data subsets?

- Any useful data analysis should be sensitive to *some* change in the data
- What types of sensitivity concern us? Varies by problem
- Let's look at economics (e.g. because of the important applications and wonderful reproducibility)
  - Report a convenient proxy (e.g. mean)
  - Policy population different from analyzed population

Why care about dropping data subsets?

- Any useful data analysis should be sensitive to *some* change in the data
- What types of sensitivity concern us? Varies by problem
- Let’s look at economics (e.g. because of the important applications and wonderful reproducibility)
  - Report a convenient proxy (e.g. mean)
  - Policy population different from analyzed population
  - Small fractions of data often missing not-at-random

Why care about dropping data subsets?

- Any useful data analysis should be sensitive to *some* change in the data
- What types of sensitivity concern us? Varies by problem
- Let’s look at economics (e.g. because of the important applications and wonderful reproducibility)
  - Report a convenient proxy (e.g. mean)
  - Policy population different from analyzed population
  - Small fractions of data often missing not-at-random
  - Models are necessarily misspecified

Why care about dropping data subsets?

- Any useful data analysis should be sensitive to *some* change in the data
- What types of sensitivity concern us? Varies by problem
- Let’s look at economics (e.g. because of the important applications and wonderful reproducibility)
  - Report a convenient proxy (e.g. mean)
  - Policy population different from analyzed population
  - Small fractions of data often missing not-at-random
  - Models are necessarily misspecified
- In all these cases, we’d be concerned if dropping a very small fraction of data changed our conclusions

Why care about dropping data subsets?

- Any useful data analysis should be sensitive to *some* change in the data
- What types of sensitivity concern us? Varies by problem
- Let's look at economics (e.g. because of the important applications and wonderful reproducibility)
  - Report a convenient proxy (e.g. mean)
  - Policy population different from analyzed population
  - Small fractions of data often missing not-at-random
  - Models are necessarily misspecified
- In all these cases, we'd be concerned if dropping a very small fraction of data changed our conclusions
- Concerns not specific to economics

Why care about dropping data subsets?

- Any useful data analysis should be sensitive to *some* change in the data
- What types of sensitivity concern us? Varies by problem
- Let’s look at economics (e.g. because of the important applications and wonderful reproducibility)
    - Report a convenient proxy (e.g. mean)
    - Policy population different from analyzed population
    - Small fractions of data often missing not-at-random
    - Models are necessarily misspecified
- In all these cases, we’d be concerned if dropping a very small fraction of data changed our conclusions
- Concerns not specific to economics
- Even if doesn’t bother you, should be up front about it

Dropping data & computational cost

Dropping data & computational cost

- Might worry if removing very small fraction $\alpha \in (0, 1)$ of data changes your decision.

Dropping data & computational cost

- Might worry if removing very small fraction $\alpha \in (0, 1)$ of data changes your decision. E.g.
  - Changes sign of estimated effect (e.g. posterior mean)

Dropping data & computational cost

- Might worry if removing very small fraction $\alpha \in (0, 1)$ of data changes your decision. E.g.
  - Changes sign of estimated effect (e.g. posterior mean)
  - Changes whether Bayesian credible interval includes 0

Dropping data & computational cost

- Might worry if removing very small fraction $\alpha \in (0, 1)$ of data changes your decision. E.g.
  - Changes sign of estimated effect (e.g. posterior mean)
  - Changes whether Bayesian credible interval includes 0
  - Changes whether confidence interval includes 0

Dropping data & computational cost

- Might worry if removing very small fraction $\alpha \in (0, 1)$ of data changes your decision. E.g.
  - Changes sign of estimated effect (e.g. posterior mean)
  - Changes whether Bayesian credible interval includes 0
  - Changes whether confidence interval includes 0
- Brute force approach: re-run data analysis with every very-small subset dropped

Dropping data & computational cost

- Might worry if removing very small fraction $\alpha \in (0, 1)$ of data changes your decision. E.g.
  - Changes sign of estimated effect (e.g. posterior mean)
  - Changes whether Bayesian credible interval includes 0
  - Changes whether confidence interval includes 0
- Brute force approach: re-run data analysis with every very-small subset dropped
- Angelucci et al 2015 microcredit study: over 16,000 points

Dropping data & computational cost

- Might worry if removing very small fraction $\alpha \in (0, 1)$ of data changes your decision. E.g.
  - Changes sign of estimated effect (e.g. posterior mean)
  - Changes whether Bayesian credible interval includes 0
  - Changes whether confidence interval includes 0
- Brute force approach: re-run data analysis with every very-small subset dropped
- Angelucci et al 2015 microcredit study: over 16,000 points
  - Take $\alpha = 0.001$; 0.1% of 16,000 is 16

Dropping data & computational cost

- Might worry if removing very small fraction $\alpha \in (0, 1)$ of data changes your decision. E.g.
  - Changes sign of estimated effect (e.g. posterior mean)
  - Changes whether Bayesian credible interval includes 0
  - Changes whether confidence interval includes 0
- Brute force approach: re-run data analysis with every very-small subset dropped
- Angelucci et al 2015 microcredit study: over 16,000 points
  - Take $\alpha = 0.001$; 0.1% of 16,000 is 16
  - A dataset of size 16,000 has ~10^{53} subsets of size 16

Dropping data & computational cost

- Might worry if removing very small fraction $\alpha \in (0, 1)$ of data changes your decision. E.g.
  - Changes sign of estimated effect (e.g. posterior mean)
  - Changes whether Bayesian credible interval includes 0
  - Changes whether confidence interval includes 0
- Brute force approach: re-run data analysis with every very-small subset dropped
- Angelucci et al 2015 microcredit study: over 16,000 points
  - Take $\alpha = 0.001$; 0.1% of 16,000 is 16
  - A dataset of size 16,000 has ~10^{53} subsets of size 16
  - If analysis takes 1 second, check takes >10^{46} years

Dropping data & computational cost

- Might worry if removing very small fraction $\alpha \in (0, 1)$ of data changes your decision. E.g.
  - Changes sign of estimated effect (e.g. posterior mean)
  - Changes whether Bayesian credible interval includes 0
  - Changes whether confidence interval includes 0
- Brute force approach: re-run data analysis with every very-small subset dropped
- Angelucci et al 2015 microcredit study: over 16,000 points
  - Take $\alpha = 0.001$; 0.1% of 16,000 is 16
  - A dataset of size 16,000 has ~10^{53} subsets of size 16
  - If analysis takes 1 second, check takes >10^{46} years
  - Parallel computing can't save you here!

Dropping data & computational cost

- Might worry if removing very small fraction $\alpha \in (0, 1)$ of data changes your decision. E.g.
  - Changes sign of estimated effect (e.g. posterior mean)
  - Changes whether Bayesian credible interval includes 0
  - Changes whether confidence interval includes 0
- Brute force approach: re-run data analysis with every very-small subset dropped
- Angelucci et al 2015 microcredit study: over 16,000 points
  - Take $\alpha = 0.001$; 0.1% of 16,000 is 16
  - A dataset of size 16,000 has ~10^{53} subsets of size 16
  - If analysis takes 1 second, check takes >10^{46} years
  - Parallel computing can't save you here!
- We provide a fast, automated, accurate approximation

A Motivating Example: Microcredit

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
  - *Fantastic reproducibility and data sharing!*

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
  - *Fantastic reproducibility and data sharing!*

  - Original model:

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
    - *Fantastic reproducibility and data sharing!*

    - Original model: $y_n = \theta_0 + \theta_1 x_n + \epsilon_n, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
  - *Fantastic reproducibility and data sharing!*

microcredit indicator

  - Original model: $y_n = \theta_0 + \theta_1 x_n + \epsilon_n, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
    - *Fantastic reproducibility and data sharing!*

profit microcredit indicator

- Original model: $y_n = \theta_0 + \theta_1 x_n + \epsilon_n, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
    - *Fantastic reproducibility and data sharing!*

profit parameters microcredit indicator

- Original model: $y_n = \theta_0 + \theta_1 x_n + \epsilon_n, \epsilon_n \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
  - *Fantastic reproducibility and data sharing!*

profit parameters microcredit indicator

  - Original model: $y_n = \theta_0 + \theta_1 x_n + \epsilon_n, \epsilon_n \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$
  - Result: $\hat{\theta}_1 =$ -4.55 USD PPP/2 weeks, std error 5.88

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
    - *Fantastic reproducibility and data sharing!*

profit parameters microcredit indicator

    - Original model: $y_n = \theta_0 + \theta_1 x_n + \epsilon_n, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
    - Result: $\hat{\theta}_1 =$ -4.55 USD PPP/2 weeks, std error 5.88
- Our approximation:

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
    - *Fantastic reproducibility and data sharing!*

profit parameters microcredit indicator

- Original model: $y_n = \theta_0 + \theta_1 x_n + \epsilon_n, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
    - Result: $\hat{\theta}_1 =$ -4.55 USD PPP/2 weeks, std error 5.88
- Our approximation:
    - Takes 2 seconds to run (not 10^{46} years)

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
  - *Fantastic reproducibility and data sharing!*

profit parameters microcredit indicator

  - Original model: $y_n = \theta_0 + \theta_1 x_n + \epsilon_n, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
  - Result: $\hat{\theta}_1 =$ -4.55 USD PPP/2 weeks, std error 5.88
- Our approximation:
  - Takes 2 seconds to run (not 10^{46} years)
  - Can remove 1 household & change sign: neg → pos

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
  - *Fantastic reproducibility and data sharing!*

profit parameters microcredit indicator

  - Original model: $y_n = \theta_0 + \theta_1 x_n + \epsilon_n, \epsilon_n \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$
  - Result: $\hat{\theta}_1 =$ -4.55 USD PPP/2 weeks, std error 5.88
- Our approximation:
  - Takes 2 seconds to run (not 10^{46} years)
  - Can remove 1 household & change sign: neg → pos
  - Can remove 15 points to get $\hat{\theta}_1 =$ 7.03, std err 2.55

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
  - *Fantastic reproducibility and data sharing!*

profit parameters microcredit indicator

  - Original model: $y_n = \theta_0 + \theta_1 x_n + \epsilon_n, \epsilon_n \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$
  - Result: $\hat{\theta}_1 =$ -4.55 USD PPP/2 weeks, std error 5.88
- Our approximation:
  - Takes 2 seconds to run (not 10^{46} years)
  - Can remove 1 household & change sign: neg → pos
  - Can remove 15 points to get $\hat{\theta}_1 =$ 7.03, std err 2.55
  - Can re-run regression to check directly
    - We provide theoretical support, but no need to trust it

A Motivating Example: Microcredit

- Angelucci et al 2015: Largest (16,561 households) of 7 randomized controlled trials examining effect of microcredit
  - *Fantastic reproducibility and data sharing!*

profit parameters microcredit indicator

  - Original model: $y_n = \theta_0 + \theta_1 x_n + \epsilon_n, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
  - Result: $\hat{\theta}_1 =$ -4.55 USD PPP/2 weeks, std error 5.88
- Our approximation:
  - Takes 2 seconds to run (not 10^{46} years)
  - Can remove 1 household & change sign: neg → pos
  - Can remove 15 points to get $\hat{\theta}_1 =$ 7.03, std err 2.55
  - Can re-run regression to check directly
    - We provide theoretical support, but no need to trust it
- It's not just non-significance, gross outliers, heavy tails, reporting means, or not using Bayes; issue is signal-to-noise

Being Bayesian Isn't a Panacea

Being Bayesian Isn’t a Panacea

- Meager 2020 provides a hierarchical Bayesian analysis of 7 randomized controlled trials examining effect of microcredit

Being Bayesian Isn't a Panacea

- Meager 2020 provides a hierarchical Bayesian analysis of 7 randomized controlled trials examining effect of microcredit
    - +*Original studies: fantastic reproducibility & data sharing!*

Being Bayesian Isn't a Panacea

- Meager 2020 provides a hierarchical Bayesian analysis of 7 randomized controlled trials examining effect of microcredit
  - *+Original studies: fantastic reproducibility & data sharing!*
  - Meager 2020: Carefully chosen likelihoods and priors

Being Bayesian Isn't a Panacea

- Meager 2020 provides a hierarchical Bayesian analysis of 7 randomized controlled trials examining effect of microcredit
    - *+Original studies: fantastic reproducibility & data sharing!*
    - Meager 2020: Carefully chosen likelihoods and priors
- Meager approximates posterior with MCMC (Stan)

Being Bayesian Isn't a Panacea

- Meager 2020 provides a hierarchical Bayesian analysis of 7 randomized controlled trials examining effect of microcredit
  - *+Original studies: fantastic reproducibility & data sharing!*
  - Meager 2020: Carefully chosen likelihoods and priors
- Meager approximates posterior with MCMC (Stan)
  - We use variational Bayes (VB)

Being Bayesian Isn't a Panacea

- Meager 2020 provides a hierarchical Bayesian analysis of 7 randomized controlled trials examining effect of microcredit
  - *+Original studies: fantastic reproducibility & data sharing!*
  - Meager 2020: Carefully chosen likelihoods and priors
- Meager approximates posterior with MCMC (Stan)
  - We use variational Bayes (VB)
  - We check that VB matches Stan MCMC output

Being Bayesian Isn’t a Panacea

- Meager 2020 provides a hierarchical Bayesian analysis of 7 randomized controlled trials examining effect of microcredit
  - *+Original studies: fantastic reproducibility & data sharing!*
  - Meager 2020: Carefully chosen likelihoods and priors
- Meager approximates posterior with MCMC (Stan)
  - We use variational Bayes (VB)
  - We check that VB matches Stan MCMC output
  - Note: MCMC & VB match only when we use linear response covariance correction [Giordano, Broderick, Jordan 2018, 2015]

Being Bayesian Isn’t a Panacea

- Meager 2020 provides a hierarchical Bayesian analysis of 7 randomized controlled trials examining effect of microcredit
    - *+Original studies: fantastic reproducibility & data sharing!*
    - Meager 2020: Carefully chosen likelihoods and priors
- Meager approximates posterior with MCMC (Stan)
    - We use variational Bayes (VB)
    - We check that VB matches Stan MCMC output
    - Note: MCMC & VB match only when we use linear response covariance correction [Giordano, Broderick, Jordan 2018, 2015]
- We find that dropping < 0.1% of data changes the sign of the posterior expected average effect of microcredit

Being Bayesian Isn't a Panacea

- Meager 2020 provides a hierarchical Bayesian analysis of 7 randomized controlled trials examining effect of microcredit
  - +*Original studies: fantastic reproducibility & data sharing!*
  - Meager 2020: Carefully chosen likelihoods and priors
- Meager approximates posterior with MCMC (Stan)
  - We use variational Bayes (VB)
  - We check that VB matches Stan MCMC output
  - Note: MCMC & VB match only when we use linear response covariance correction [Giordano, Broderick, Jordan 2018, 2015]
- We find that dropping < 0.1% of data changes the sign of the posterior expected average effect of microcredit
  - Still sensitive like the ordinary least squares analyses

Setup & the Approximation

- A data analysis:

Setup & the Approximation

- A data analysis:

d_n

datum; e.g. (x_n, y_n)

Setup & the Approximation

parameters

- A data analysis: θ, d_n

datum; e.g. (x_n, y_n)

Setup & the Approximation

- A data analysis:

$$\underbrace{f}_{\text{loss}}(\underbrace{\theta}_{\text{parameters}}, \underbrace{d_n}_{\text{datum; e.g. } (x_n,\, y_n)})$$

Setup & the Approximation

- A data analysis: $\sum_{n=1}^{N} f(\theta, d_n)$

loss

parameters

datum; e.g. (x_n, y_n)

Setup & the Approximation

- A data analysis: $$\underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} f(\theta, d_n)$$

loss parameters

datum; e.g. (x_n, y_n)

Setup & the Approximation

loss parameters

- A data analysis: $\hat{\theta} := \underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} f(\theta, d_n)$

estimator

datum; e.g. (x_n, y_n)

Setup & the Approximation

- A data analysis: $\hat{\theta} := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} f(\theta, d_n) \boxed{+ R(\theta)}$

estimator
loss
parameters
penalty
datum; e.g. (x_n, y_n)

Setup & the Approximation

loss parameters penalty

- A data analysis: $\hat{\theta} := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} f(\theta, d_n) + R(\theta)$

estimator

datum; e.g. (x_n, y_n)

Setup & the Approximation

loss parameters penalty

- A data analysis: $$\hat{\theta} := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} f(\theta, d_n) + R(\theta)$$

estimator

datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)

Setup & the Approximation

loss parameters penalty

- A data analysis: $\hat{\theta} := \operatorname{argmin}_{\theta} \sum_{n=1}^{N} f(\theta, d_n) + R(\theta)$

estimator

datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest ϕ

Setup & the Approximation

loss parameters penalty

- A data analysis: $\hat{\theta} := \underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} f(\theta, d_n) + R(\theta)$

estimator

datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest ϕ : E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint

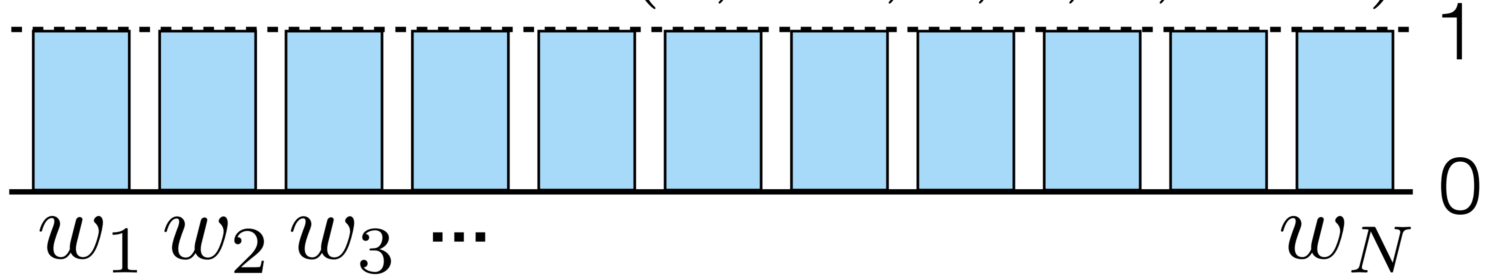
Setup & the Approximation

loss parameters penalty

- A data analysis: $\hat{\theta} := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} f(\theta, d_n) + R(\theta)$

estimator datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest ϕ : E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



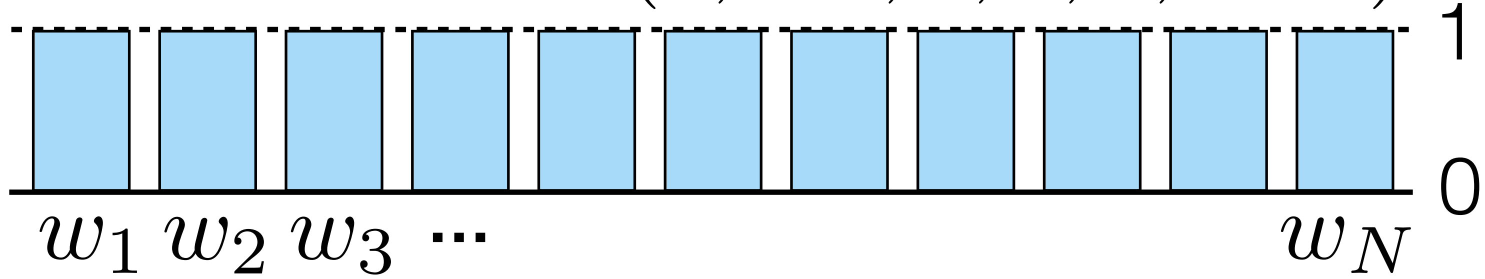
Setup & the Approximation

loss parameters penalty

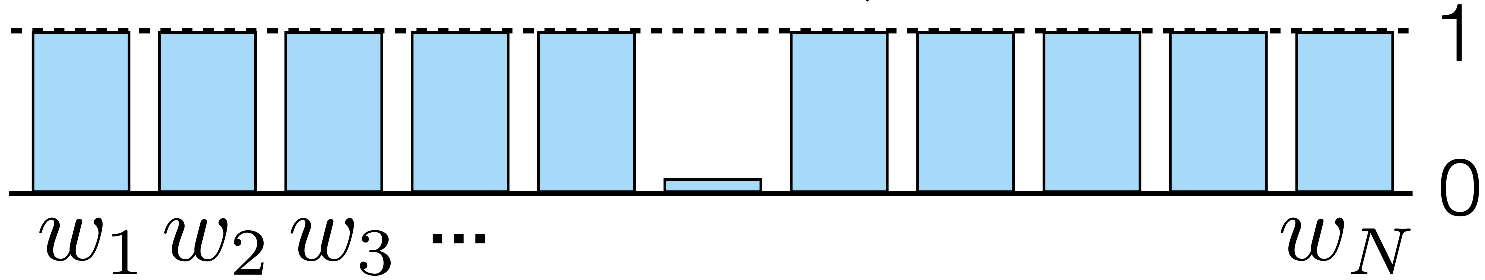
- A data analysis: $\hat{\theta} := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} f(\theta, d_n) + R(\theta)$

estimator datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest ϕ : E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



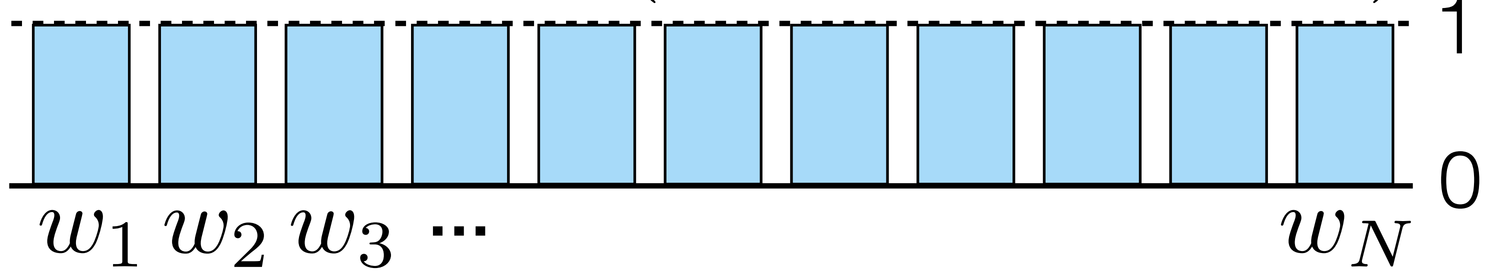
Setup & the Approximation

loss parameters penalty

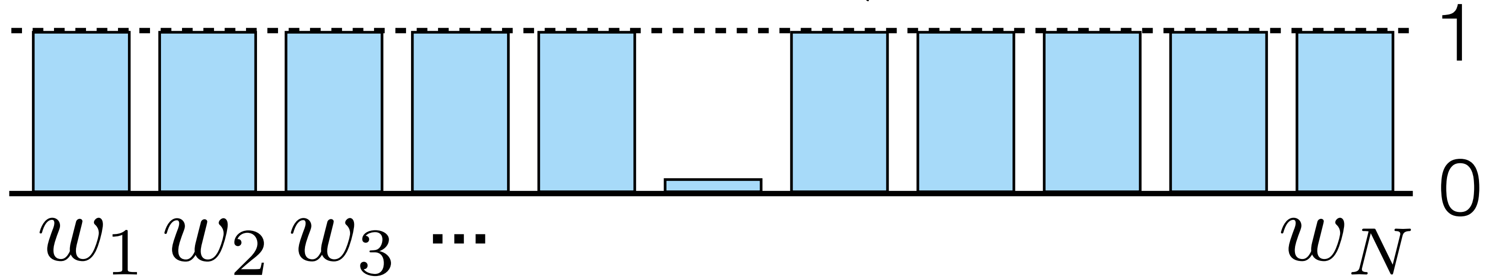
- A data analysis: $$\hat{\theta} := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} f(\theta, d_n) + R(\theta)$$

estimator datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest ϕ : E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



- Each dropped data subset corresponds to a different w

Setup & the Approximation

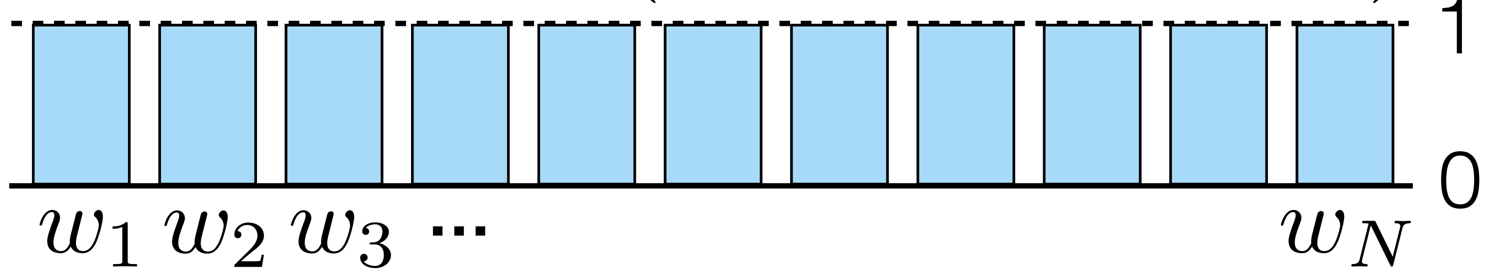
loss parameters penalty

- A data analysis: $\hat{\theta} := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} f(\theta, d_n) + R(\theta)$

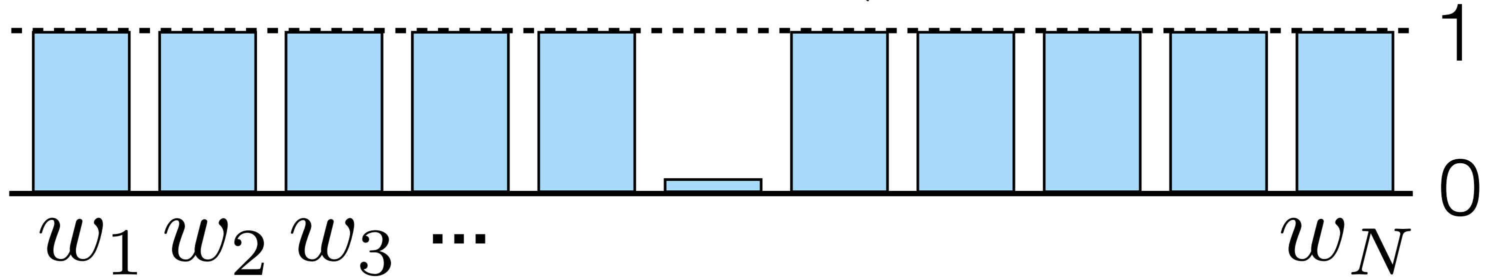
  estimator

  datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest ϕ : E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



- Each dropped data subset corresponds to a different w

Setup & the Approximation

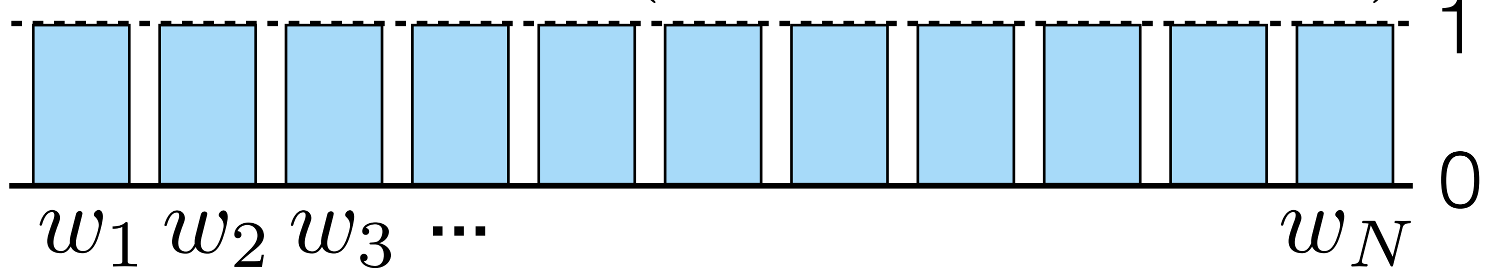
loss parameters penalty

- A data analysis: $\hat{\theta} := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} f(\theta, d_n) + R(\theta)$

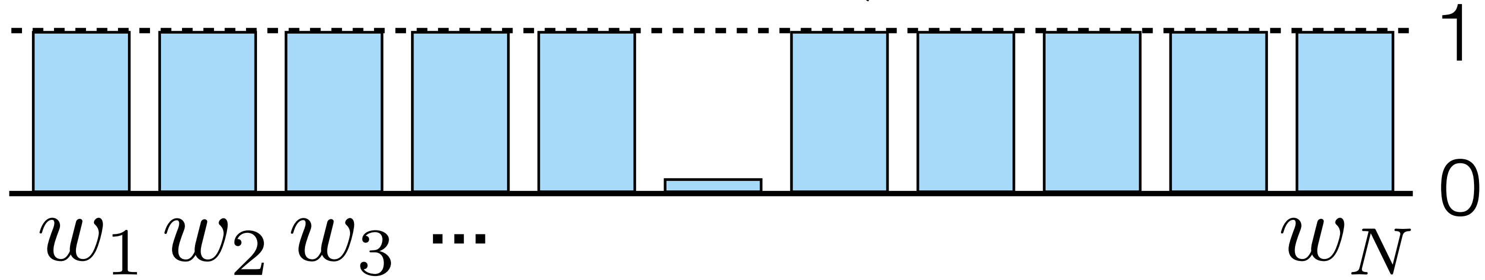
estimator

datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest ϕ : E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



- Each dropped data subset corresponds to a different w

Setup & the Approximation

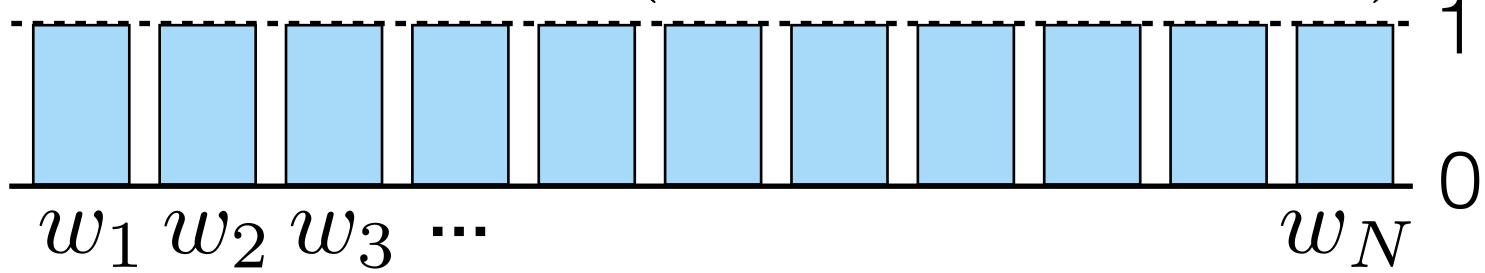
loss parameters penalty

- A data analysis: $\hat{\theta} := \underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} w_n f(\theta, d_n) + R(\theta)$

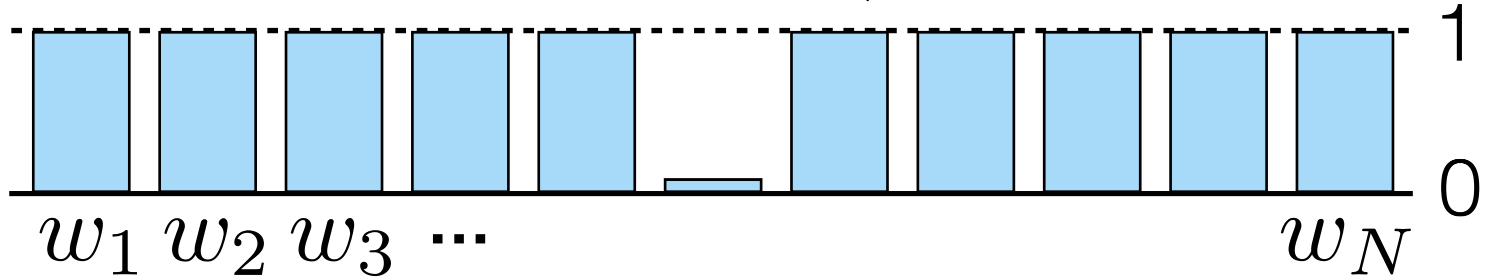
estimator

datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest ϕ : E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



- Each dropped data subset corresponds to a different w

Setup & the Approximation

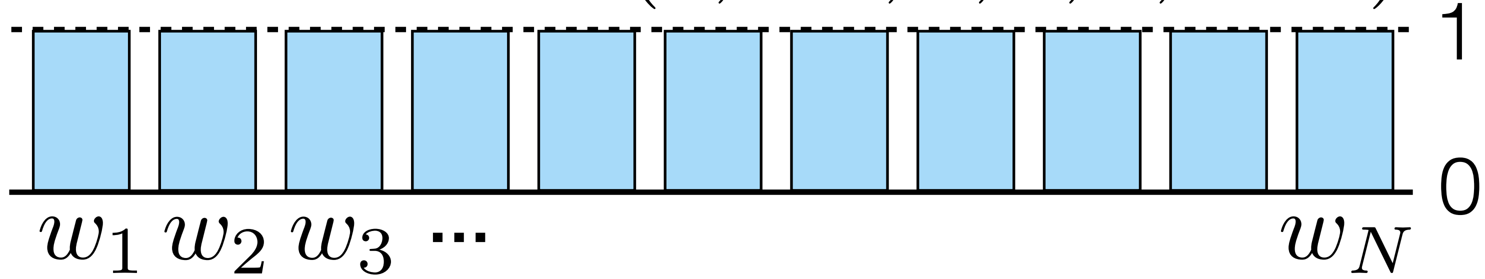
loss parameters penalty

- A data analysis: $\hat{\theta} := \underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} w_n f(\theta, d_n) + R(\theta)$

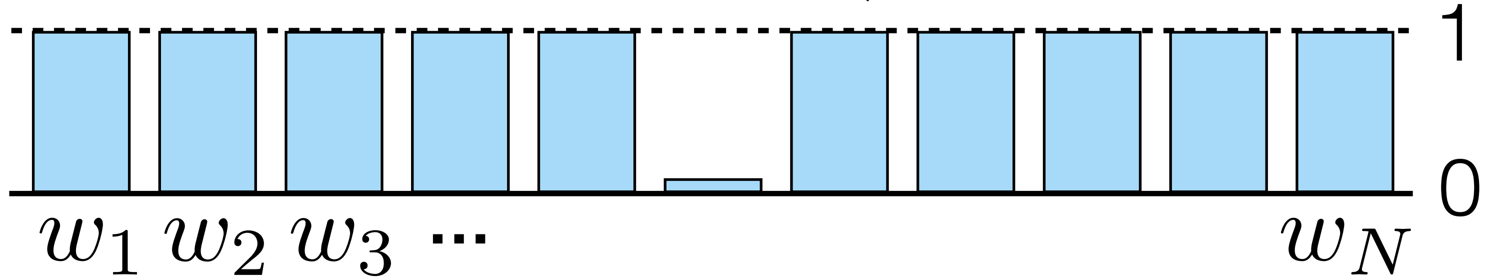
estimator

datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest ϕ : E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



- Each dropped data subset corresponds to a different w

Setup & the Approximation

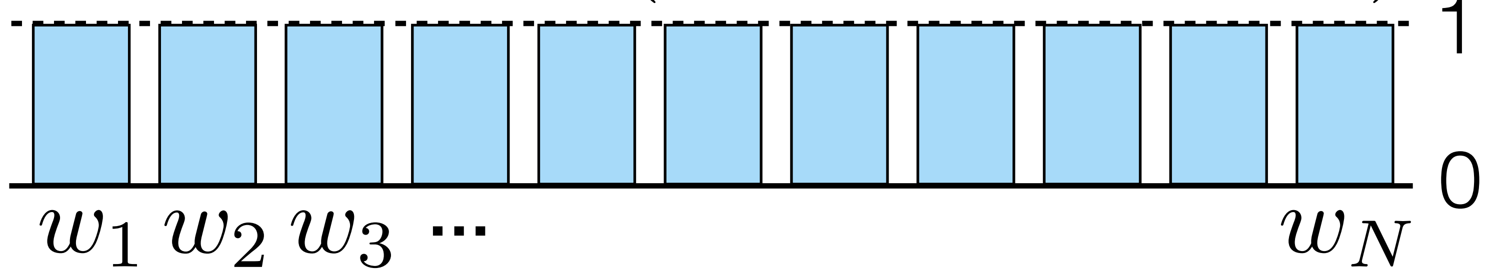
loss parameters penalty

- A data analysis: $\boxed{\hat{\theta}(w)} := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} w_n f(\theta, d_n) + R(\theta)$

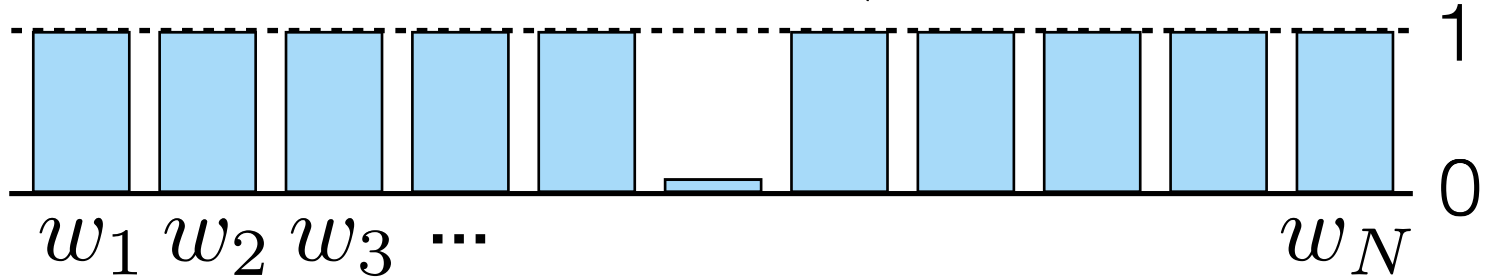
estimator

datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest ϕ : E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



- Each dropped data subset corresponds to a different w

Setup & the Approximation

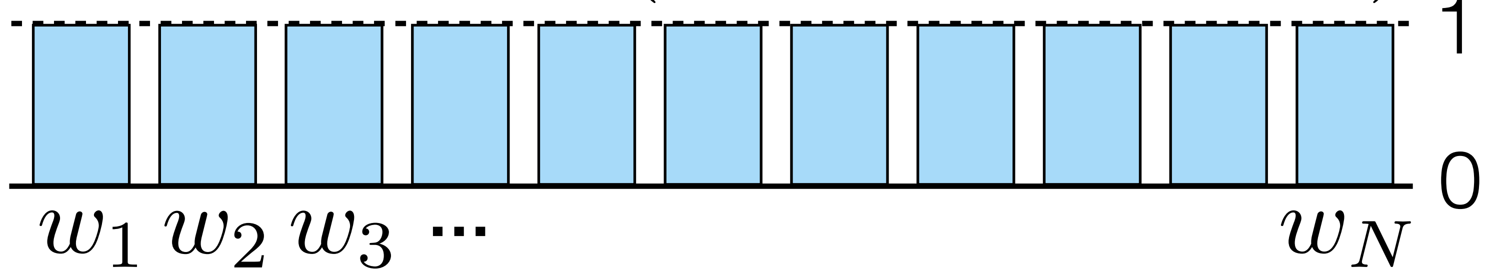
loss parameters penalty

- A data analysis: $\hat{\theta}(w) := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} w_n f(\theta, d_n) + R(\theta)$

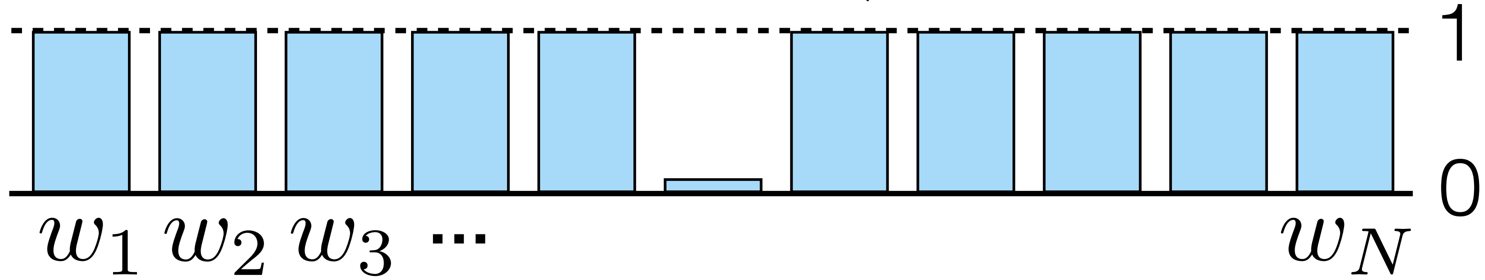
estimator

datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest ϕ : E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



- Each dropped data subset corresponds to a different w

Setup & the Approximation

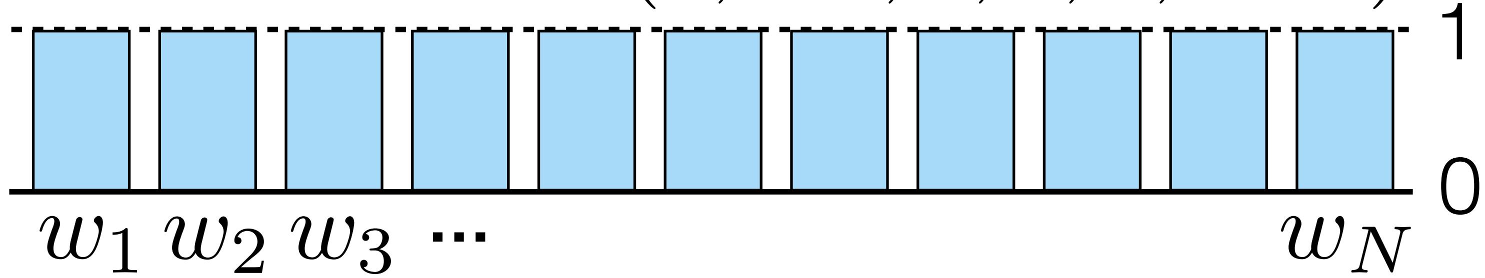
loss parameters penalty

- A data analysis: $\hat{\theta}(w) := \underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} w_n f(\theta, d_n) + R(\theta)$

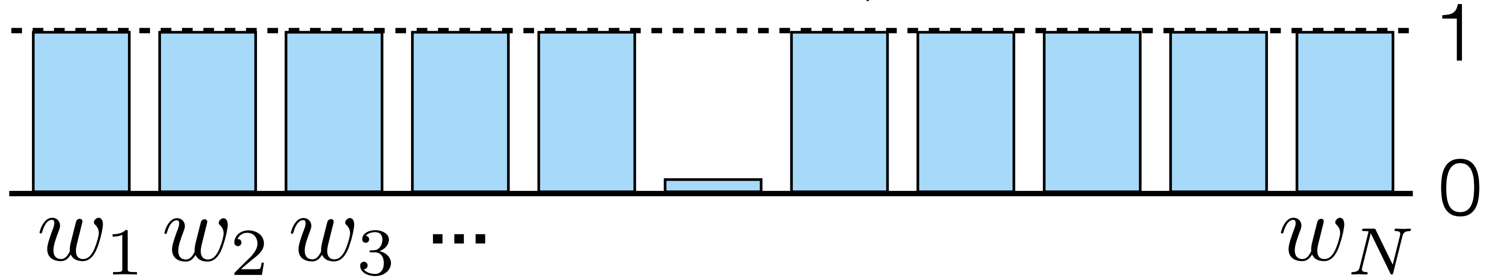
estimator

datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest $\phi(w)$: E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



- Each dropped data subset corresponds to a different w

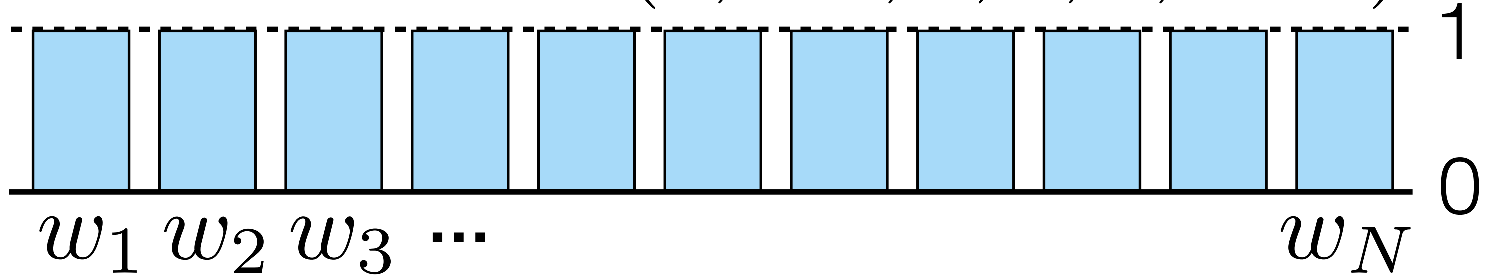
Setup & the Approximation

loss parameters penalty

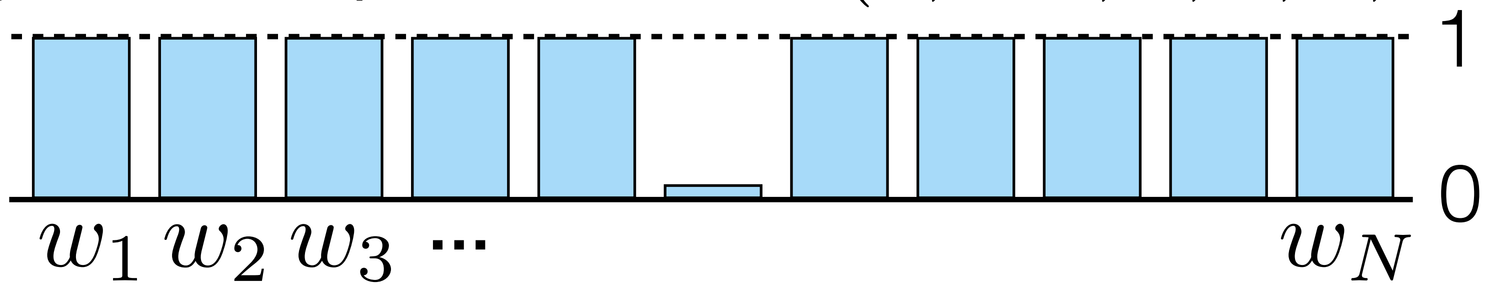
- A data analysis: $\hat{\theta}(w) := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} w_n f(\theta, d_n) + R(\theta)$

estimator datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest $\phi(w)$: E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



- Each dropped data subset corresponds to a different w
- $\phi(w) \approx \phi^{\text{lin}}(w)$

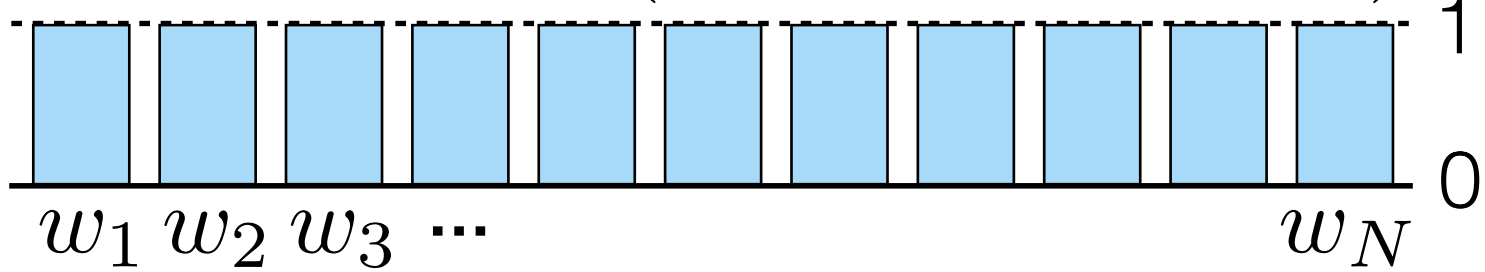
Setup & the Approximation

loss parameters penalty

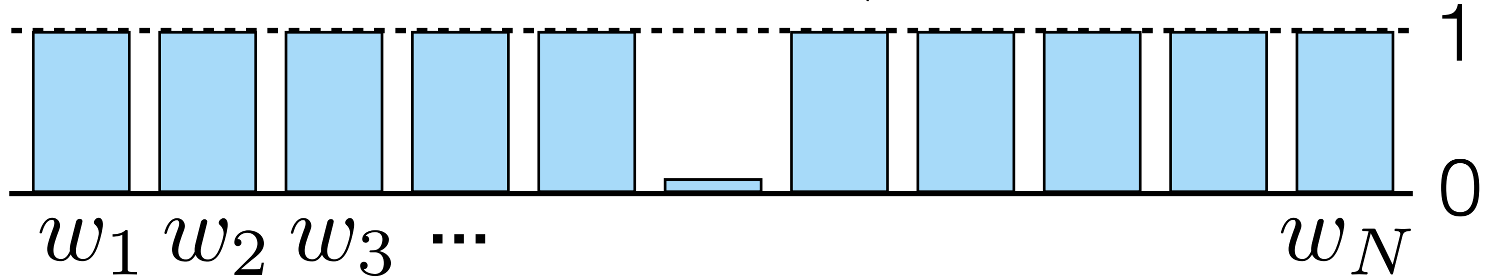
- A data analysis: $\hat{\theta}(w) := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} w_n f(\theta, d_n) + R(\theta)$

estimator datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest $\phi(w)$: E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



- Each dropped data subset corresponds to a different w
- $\phi(w) \approx \phi^{\mathrm{lin}}(w)$:= first-order Taylor expansion in w

Setup & the Approximation

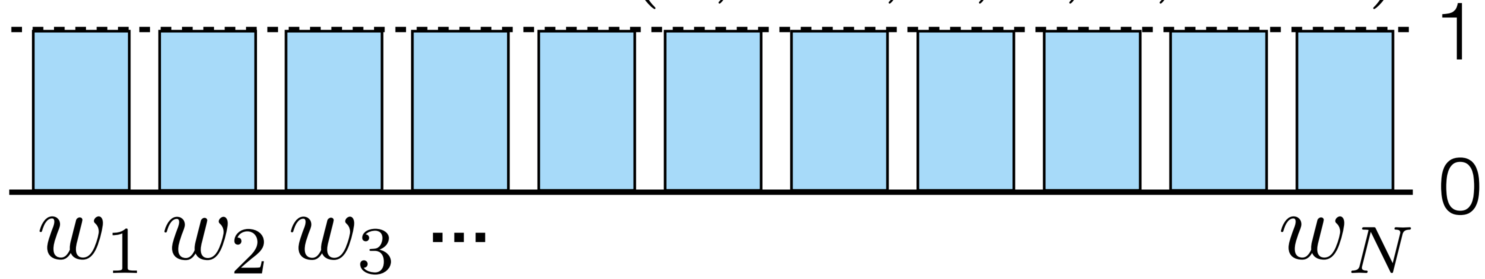
loss parameters penalty

- A data analysis: $\hat{\theta}(w) := \operatorname*{argmin}_{\theta} \sum_{n=1}^{N} w_n f(\theta, d_n) + R(\theta)$

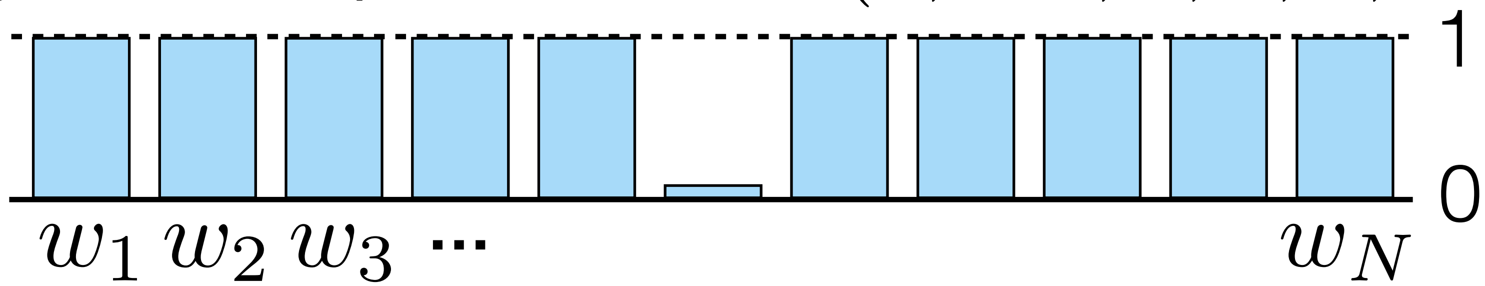
estimator

datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest $\phi(w)$: E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



- Each dropped data subset corresponds to a different w
- $\phi(w) \approx \phi^{\text{lin}}(w) :=$ first-order Taylor expansion in w
  - Finding worst data to drop: linear in N + cost of "sort"

Setup & the Approximation

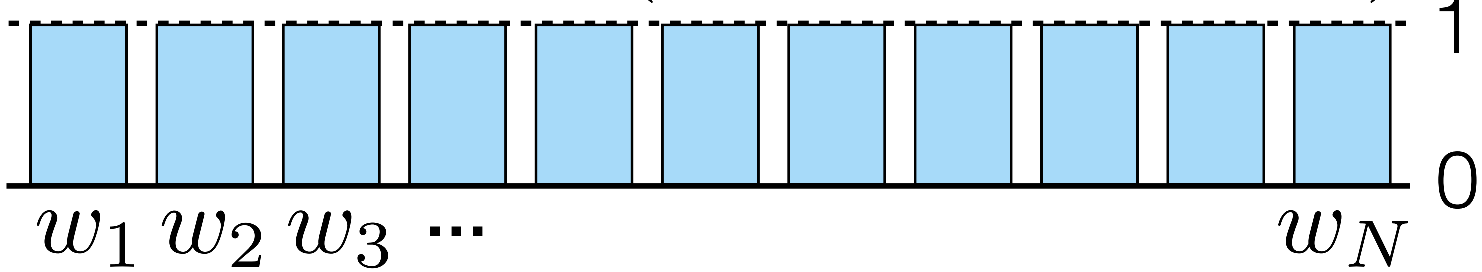
loss parameters penalty

- A data analysis: $\hat{\theta}(w) := \underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} w_n f(\theta, d_n) + R(\theta)$

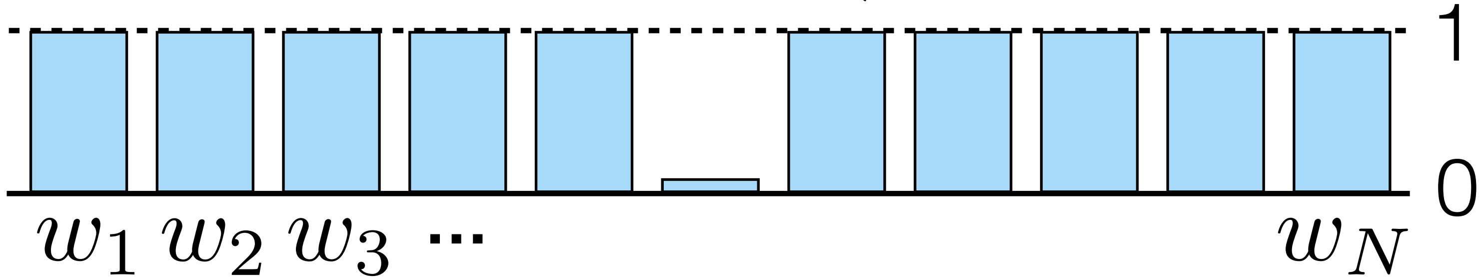
estimator

datum; e.g. (x_n, y_n)

  - Actually any Z-estimator works (e.g. MAP, VB, multistage)
- A quantity of interest $\phi(w)$: E.g. posterior mean, posterior credible interval endpoint, confidence interval endpoint
- Original problem: $w = (1, \ldots, 1, 1, 1, \ldots 1)$



- Dropping a data point: $w = (1, \ldots, 1, 0, 1, \ldots, 1)$



- Each dropped data subset corresponds to a different w
- $\phi(w) \approx \phi^{\mathrm{lin}}(w) :=$ first-order Taylor expansion in w
  - Finding worst data to drop: linear in N + cost of “sort”
  - Can automate with automatic differentiation tools

What makes an analysis non-robust?

What makes an analysis non-robust?

- **It's not just non-significance**

What makes an analysis non-robust?

- **It’s not just non-significance**
  - Oregon Medicaid, Finkelstein et al 2012, >21,000 data

What makes an analysis non-robust?

- **It's not just non-significance**
  - Oregon Medicaid, Finkelstein et al 2012, >21,000 data
  - $p < 0.01$ for a positive effect of lottery on health

What makes an analysis non-robust?

- **It's not just non-significance**
  - Oregon Medicaid, Finkelstein et al 2012, >21,000 data
  - $p < 0.01$ for a positive effect of lottery on health
    - We find: drop 11 points (0.05%) to change signific.

What makes an analysis non-robust?

- **It's not just non-significance**
  - Oregon Medicaid, Finkelstein et al 2012, >21,000 data
  - $p < 0.01$ for a positive effect of lottery on health
    - We find: drop 11 points (0.05%) to change signific.
- **It's not just that everything is non-robust**

What makes an analysis non-robust?

- **It's not just non-significance**
  - Oregon Medicaid, Finkelstein et al 2012, >21,000 data
  - $p < 0.01$ for a positive effect of lottery on health
    - We find: drop 11 points (0.05%) to change signific.
- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 data

What makes an analysis non-robust?

- **It's not just non-significance**
  - Oregon Medicaid, Finkelstein et al 2012, >21,000 data
  - $p < 0.01$ for a positive effect of lottery on health
    - We find: drop 11 points (0.05%) to change signific.
- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 data
  - We find: drop >4% data to change sign and/or signific.

What makes an analysis non-robust?

- **It's not just non-significance**
  - Oregon Medicaid, Finkelstein et al 2012, >21,000 data
  - $p < 0.01$ for a positive effect of lottery on health
    - We find: drop 11 points (0.05%) to change signific.
- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 data
  - We find: drop >4% data to change sign and/or signific.
- **Removing outliers isn't a panacea**

What makes an analysis non-robust?

- **It’s not just non-significance**
  - Oregon Medicaid, Finkelstein et al 2012, >21,000 data
  - $p < 0.01$ for a positive effect of lottery on health
    - We find: drop 11 points (0.05%) to change signific.
- **It’s not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 data
  - We find: drop >4% data to change sign and/or signific.
- **Removing outliers isn’t a panacea**
  - Angelucci & De Giorgi 2009 look at spillover effect on non-poor households & remove largest responses

What makes an analysis non-robust?

- **It's not just non-significance**
  - Oregon Medicaid, Finkelstein et al 2012, >21,000 data
  - $p < 0.01$ for a positive effect of lottery on health
    - We find: drop 11 points (0.05%) to change signific.
- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 data
  - We find: drop >4% data to change sign and/or signific.
- **Removing outliers isn't a panacea**
  - Angelucci & De Giorgi 2009 look at spillover effect on non-poor households & remove largest responses
  - We find: can drop 3 points of >4,000 & change signific.

What makes an analysis non-robust?

- **It's not just non-significance**
  - Oregon Medicaid, Finkelstein et al 2012, >21,000 data
  - $p < 0.01$ for a positive effect of lottery on health
    - We find: drop 11 points (0.05%) to change signific.
- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 data
  - We find: drop >4% data to change sign and/or signific.
- **Removing outliers isn't a panacea**
  - Angelucci & De Giorgi 2009 look at spillover effect on non-poor households & remove largest responses
  - We find: can drop 3 points of >4,000 & change signific.
- **It's not just heavy tails or reporting means**

What makes an analysis non-robust?

- **It's not just non-significance**
  - Oregon Medicaid, Finkelstein et al 2012, >21,000 data
  - $p < 0.01$ for a positive effect of lottery on health
    - We find: drop 11 points (0.05%) to change signific.
- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 data
  - We find: drop >4% data to change sign and/or signific.
- **Removing outliers isn't a panacea**
  - Angelucci & De Giorgi 2009 look at spillover effect on non-poor households & remove largest responses
  - We find: can drop 3 points of >4,000 & change signific.
- **It's not just heavy tails or reporting means**
  - We run Gaussian linear model simulations & find both robust/non-robust cases; **issue is signal-to-noise**

Try it out!

- We present a way to check if there is a very small fraction of data you can drop to change decisions
- **Paper**: Broderick, Giordano, Meager "An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?" (alphabetical)

`arxiv.org/abs/2011.14999`

Try it out!

- We present a way to check if there is a very small fraction of data you can drop to change decisions
- **Paper**: Broderick, Giordano, Meager “An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?” (alphabetical)

  `arxiv.org/abs/2011.14999`

  - Paper above focused on optimization. MCMC soon!

Try it out!

- We present a way to check if there is a very small fraction of data you can drop to change decisions
- **Paper**: Broderick, Giordano, Meager "An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?" (alphabetical)

  `arxiv.org/abs/2011.14999`

  - Paper above focused on optimization. MCMC soon!
- **Code, etc**: `github.com/rgiordan/zaminfluence`

Try it out!

- We present a way to check if there is a very small fraction of data you can drop to change decisions
- **Paper**: Broderick, Giordano, Meager “An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?” (alphabetical)
  `arxiv.org/abs/2011.14999`
  - Paper above focused on optimization. MCMC soon!
- **Code, etc**: `github.com/rgiordan/zaminfluence`
- **Our check can flag p-hacking**:
  - michaelwiebe.com/blog/2021/01/amip
  - rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html

Try it out!

- We present a way to check if there is a very small fraction of data you can drop to change decisions
- **Paper**: Broderick, Giordano, Meager "An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?" (alphabetical)

  `arxiv.org/abs/2011.14999`

  - Paper above focused on optimization. MCMC soon!
- **Code, etc**: `github.com/rgiordan/zaminfluence`
- **Our check can flag p-hacking**:
  - michaelwiebe.com/blog/2021/01/amip
  - rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html
- **Our check is just one part of a trustworthy workflow**:
  - Broderick, Gelman, Meager, Smith, Zheng "Toward a Taxonomy of Trust for Probabilistic Machine Learning" *ArXiv:2112.03270*

Try it out!

- We present a way to check if there is a very small fraction of data you can drop to change decisions
- **Paper**: Broderick, Giordano, Meager "An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?" (alphabetical)

  `arxiv.org/abs/2011.14999`
  - Paper above focused on optimization. MCMC soon!
- **Code, etc**: `github.com/rgiordan/zaminfluence`
- **Our check can flag p-hacking**:
  - michaelwiebe.com/blog/2021/01/amip
  - rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html
- **Our check is just one part of a trustworthy workflow**:
  - Broderick, Gelman, Meager, Smith, Zheng "Toward a Taxonomy of Trust for Probabilistic Machine Learning" *ArXiv:2112.03270*
- **Variational Bayes covariance correction**:
  - Giordano, Broderick, Jordan. Covariances, Robustness, and Variational Bayes, *JMLR* 2018. (Also Giordano, Broderick, Jordan, *NeurIPS* 2015.)