Statistical and computational trade-offs in Bayesian inference

Tamara Broderick
ITT Career Development
Assistant Professor, MIT

With: Nick Boyd, Trevor Campbell, Ryan Giordano, Joseph Gonzalez, Jonathan H. Huggins, Stefanie Jegelka, Brian Kulis, Michael I. Jordan, Rachael Meager, Xinghao Pan, Andre Wibisono, Ashia C. Wilson

Statistical and computation trade-offs

- Bayesian inference

Statistical and computation trade-offs

- Bayesian inference
  - Complex, modular models

Statistical and computation trade-offs

- Bayesian inference
  - Complex, modular models; posterior distribution

Statistical and computation trade-offs

- Bayesian inference $p(\theta)$
  - Complex, modular models; posterior distribution

Statistical and computation trade-offs

- Bayesian inference $p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

Statistical and computation trade-offs

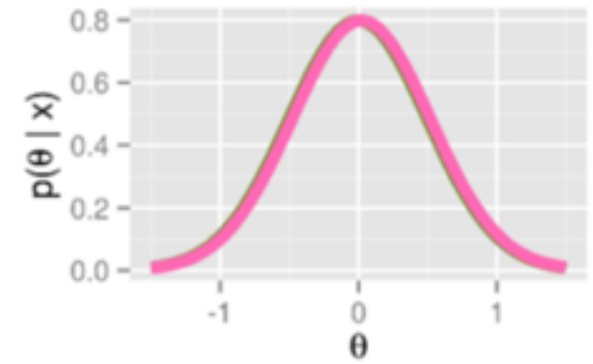
- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

- Challenge: Approximating the posterior can be computationally expensive

Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

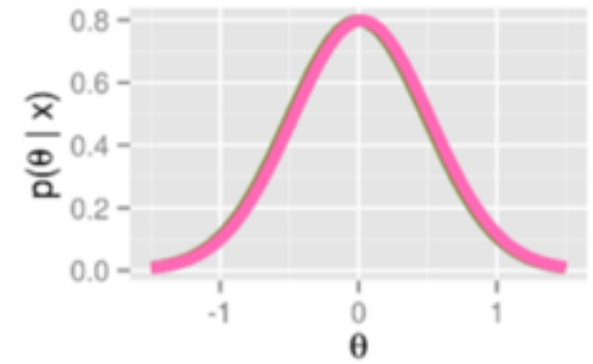
- Challenge: Approximating the posterior can be computationally expensive



Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

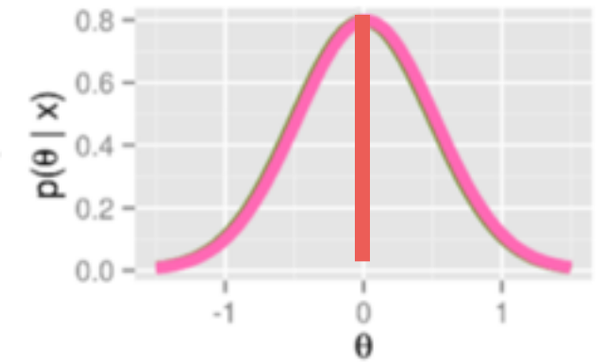
- Challenge: Approximating the posterior can be computationally expensive
  - Do we need the whole posterior?



Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

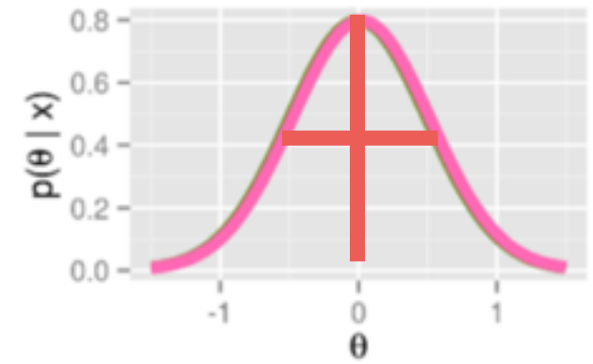
- Challenge: Approximating the posterior can be computationally expensive
  - Do we need the whole posterior?



Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

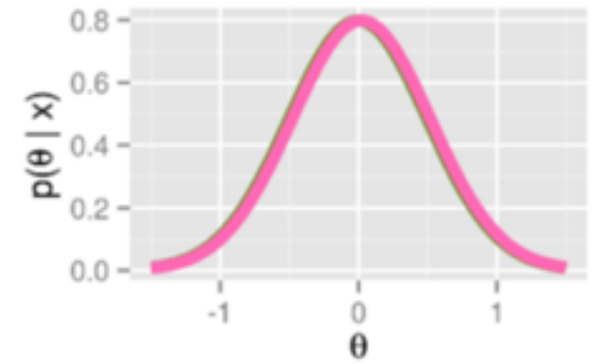
- Challenge: Approximating the posterior can be computationally expensive
  - Do we need the whole posterior?



Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

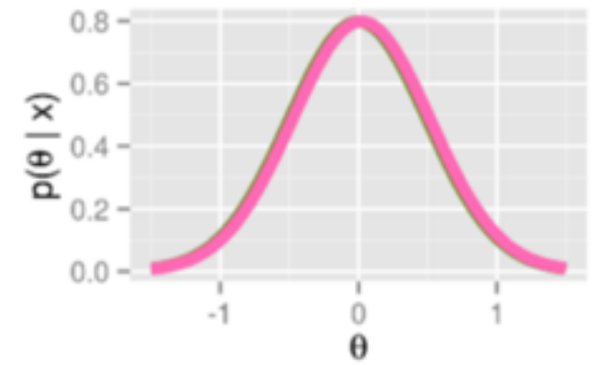
- Challenge: Approximating the posterior can be computationally expensive
  - Do we need the whole posterior?



Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

- Challenge: Approximating the posterior can be computationally expensive
  - Do we need the whole posterior?

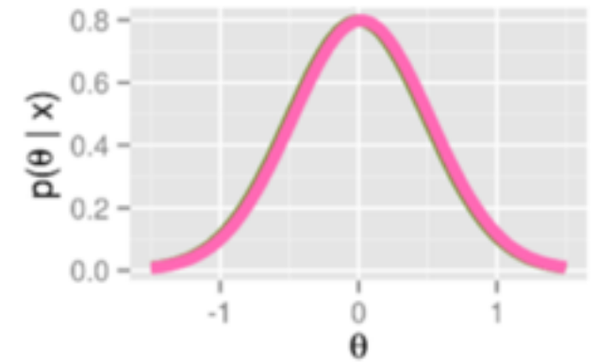


- Challenge: Express prior beliefs in a distribution

Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

- Challenge: Approximating the posterior can be computationally expensive
  - Do we need the whole posterior?

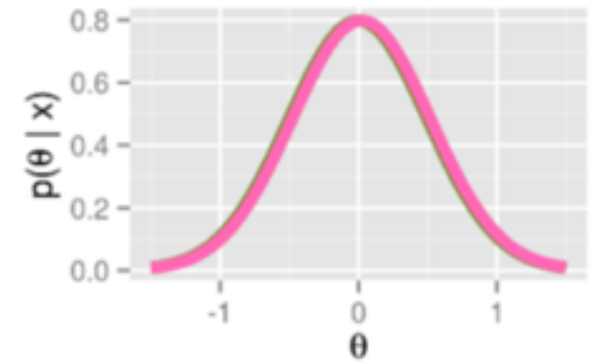


- Challenge: Express prior beliefs in a distribution (time-consuming)

Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

- Challenge: Approximating the posterior can be computationally expensive
  - Do we need the whole posterior?



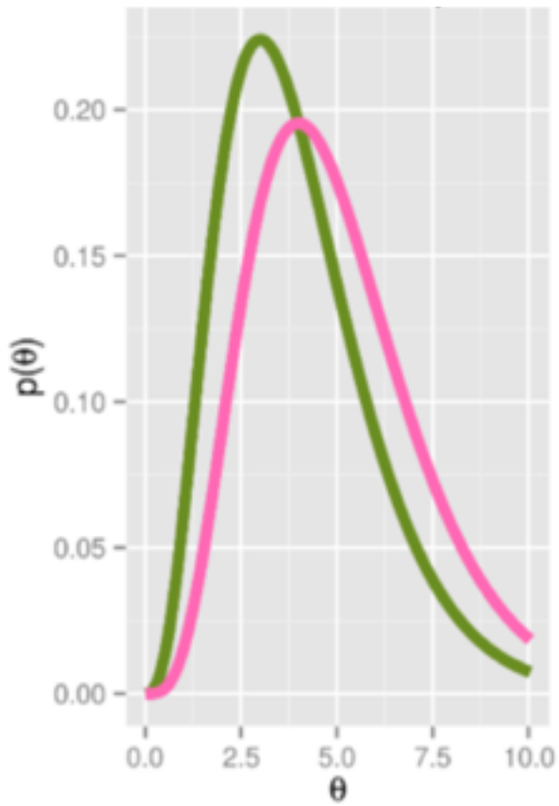
- Challenge: Express prior beliefs in a distribution (time-consuming; subjective)

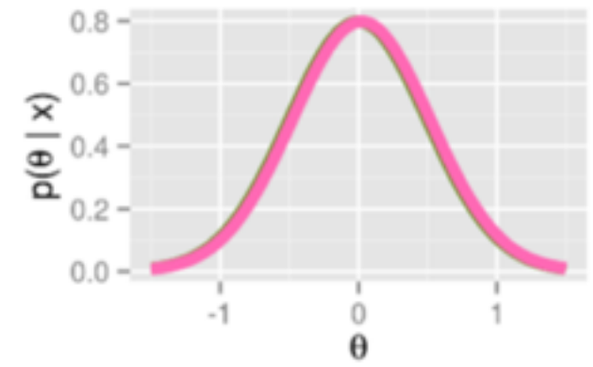
Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

- Challenge: Approximating the posterior can be computationally expensive
  - Do we need the whole posterior?

Some reasonable priors





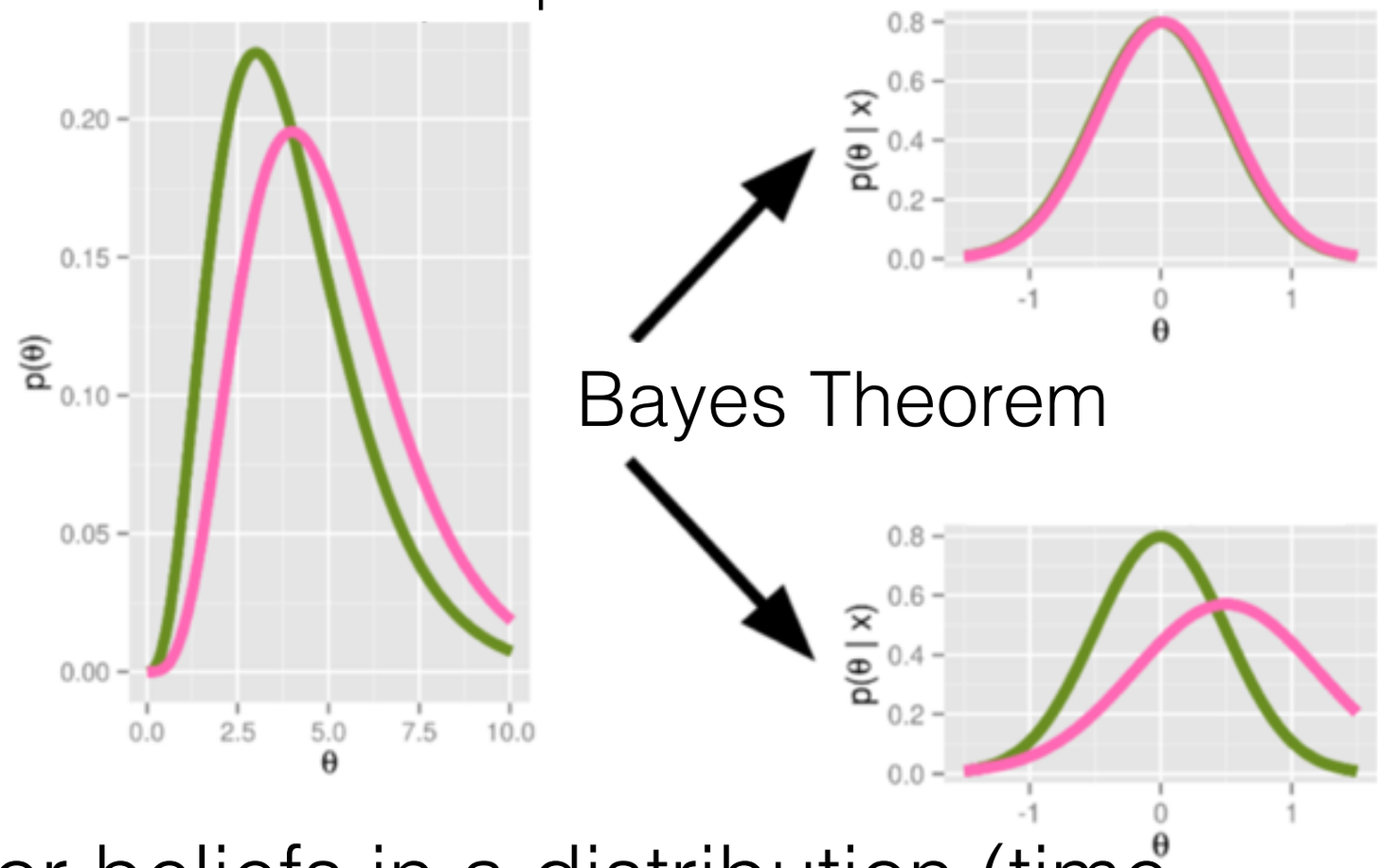
- Challenge: Express prior beliefs in a distribution (time-consuming; subjective)

Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

- Challenge: Approximating the posterior can be computationally expensive
  - Do we need the whole posterior?

Some reasonable priors

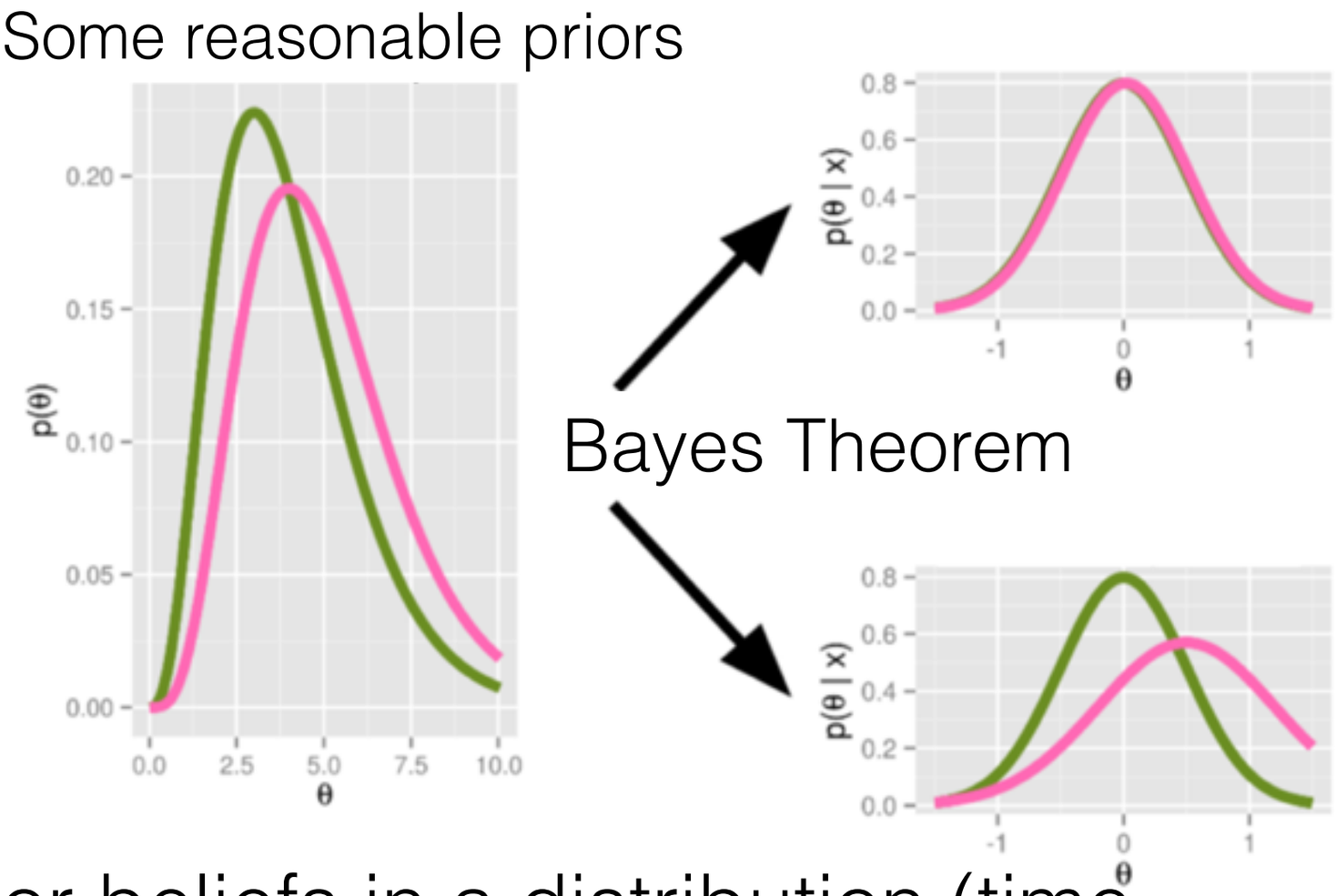


- Challenge: Express prior beliefs in a distribution (time-consuming; subjective)

Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

- Challenge: Approximating the posterior can be computationally expensive
  - Do we need the whole posterior?



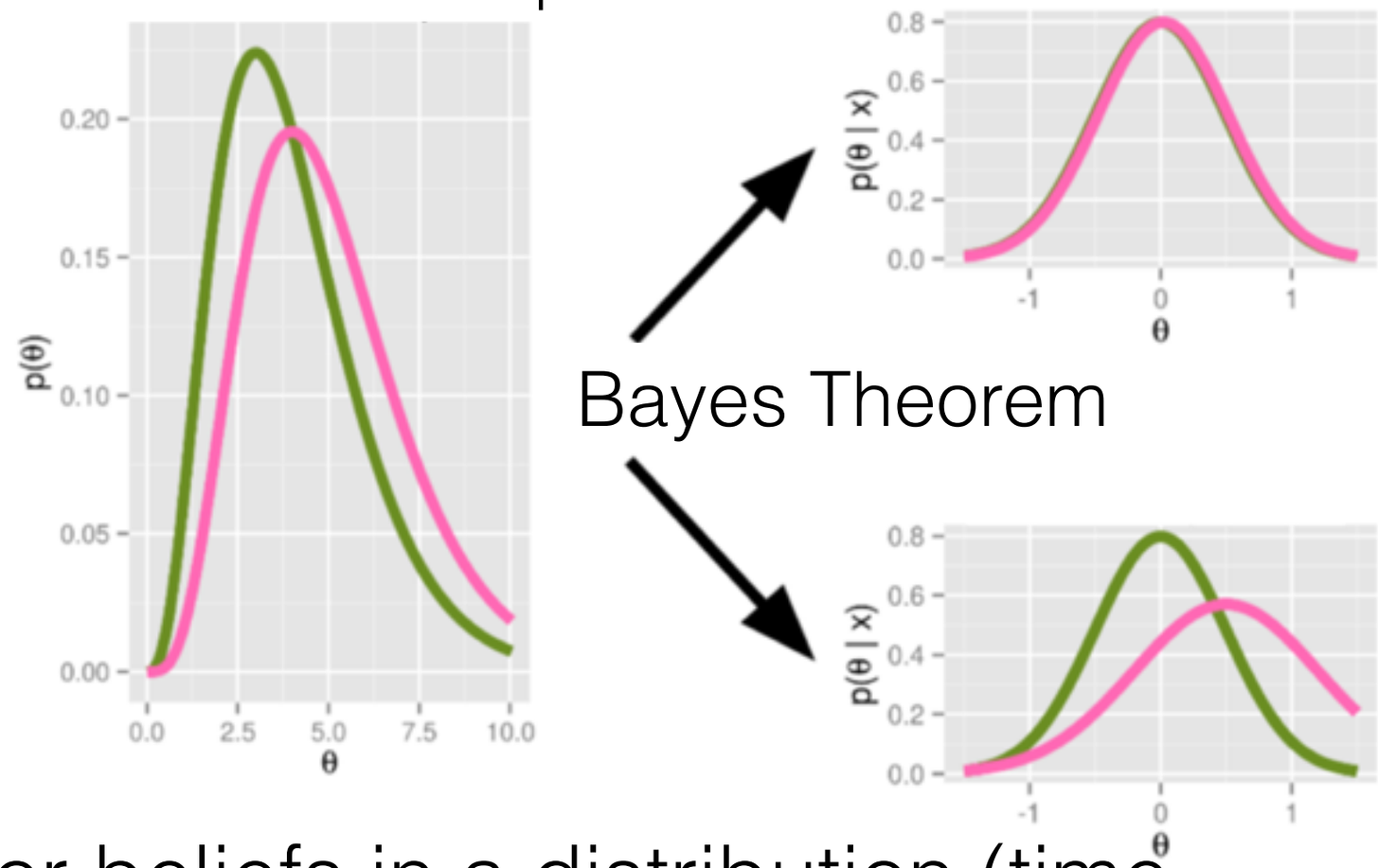
- Challenge: Express prior beliefs in a distribution (time-consuming; subjective; complex models)

Statistical and computation trade-offs

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution

- Challenge: Approximating the posterior can be computationally expensive
  - Do we need the whole posterior?

Some reasonable priors



- Challenge: Express prior beliefs in a distribution (time-consuming; subjective; complex models)
  - Global vs. local robustness

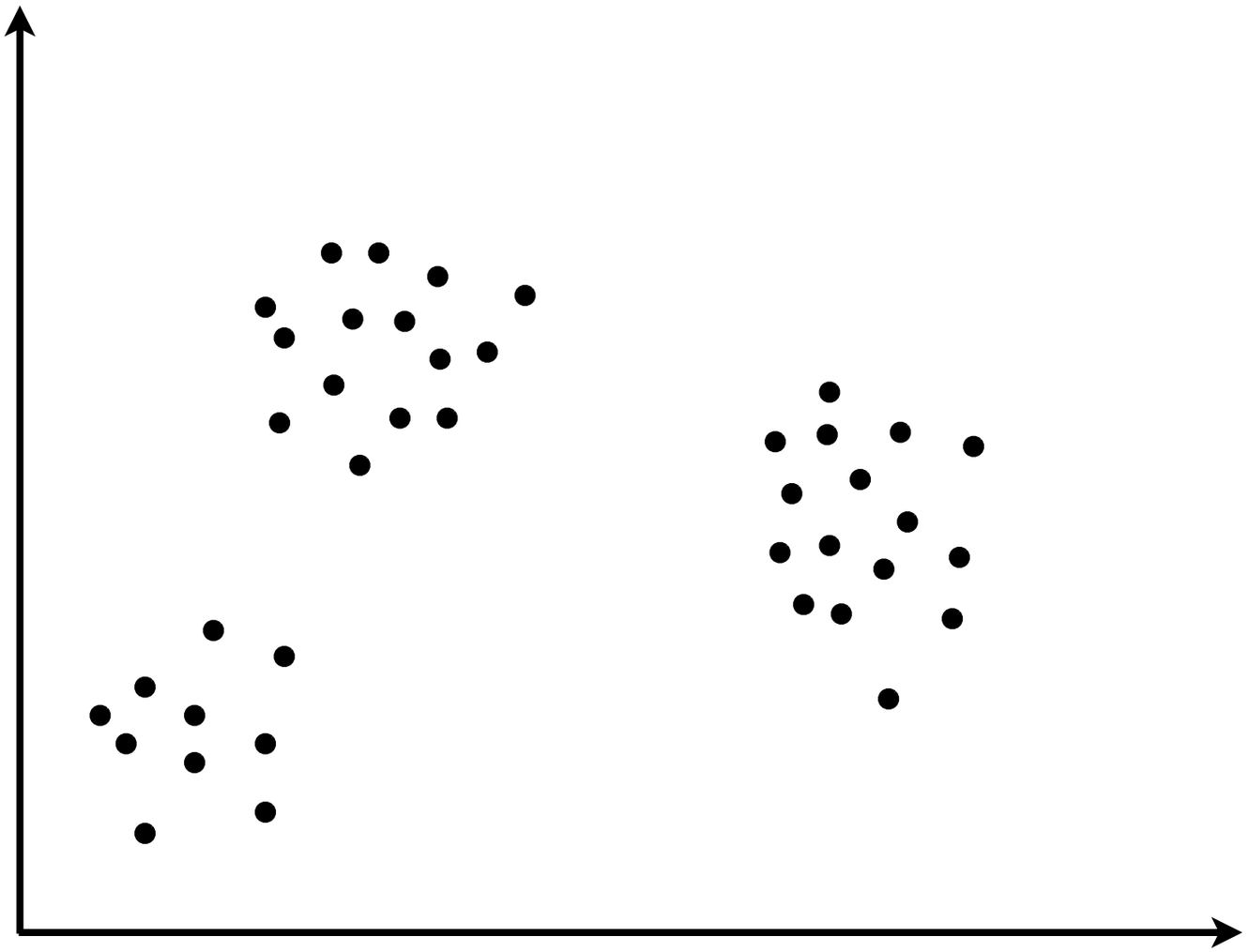
Roadmap

- Posterior approximation trade-offs
  - Point estimates
  - Uncertainties
- Robustness trade-offs
  - Local robustness quantification
- Theoretical guarantees on quality

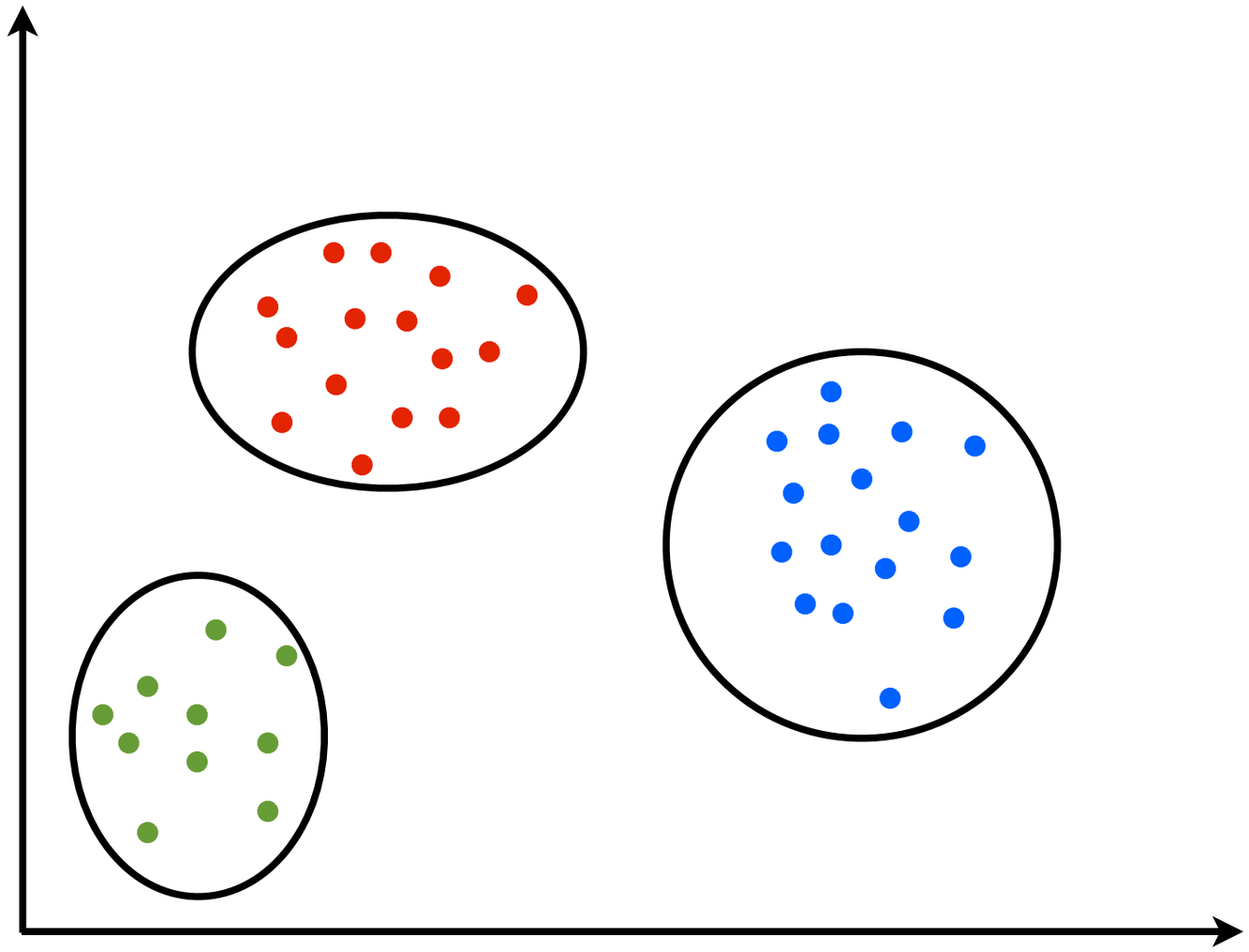
Roadmap

- Posterior approximation trade-offs
  - Point estimates
  - Uncertainties
- Robustness trade-offs
  - Local robustness quantification
- Theoretical guarantees on quality

Clustering

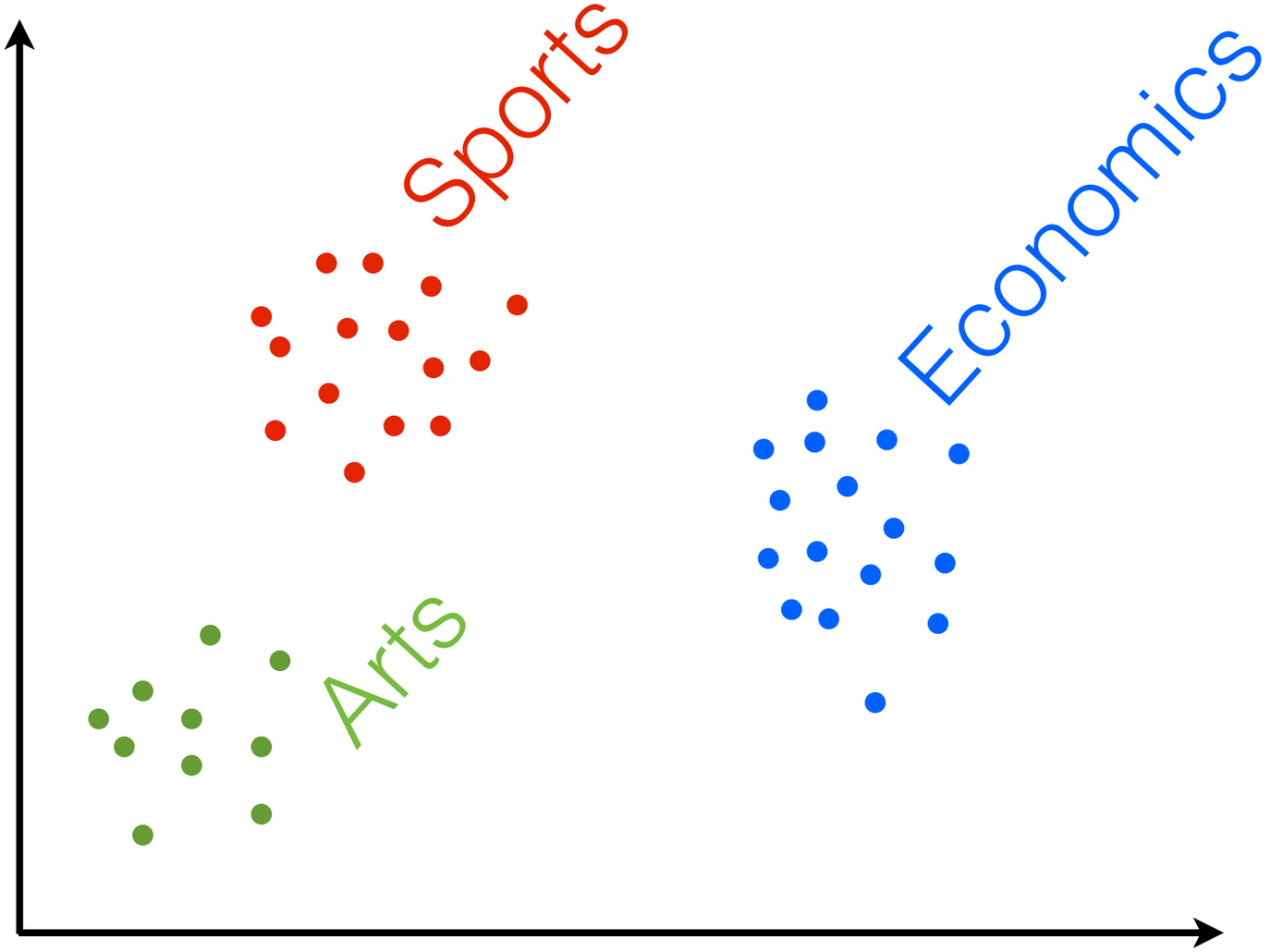


Clustering



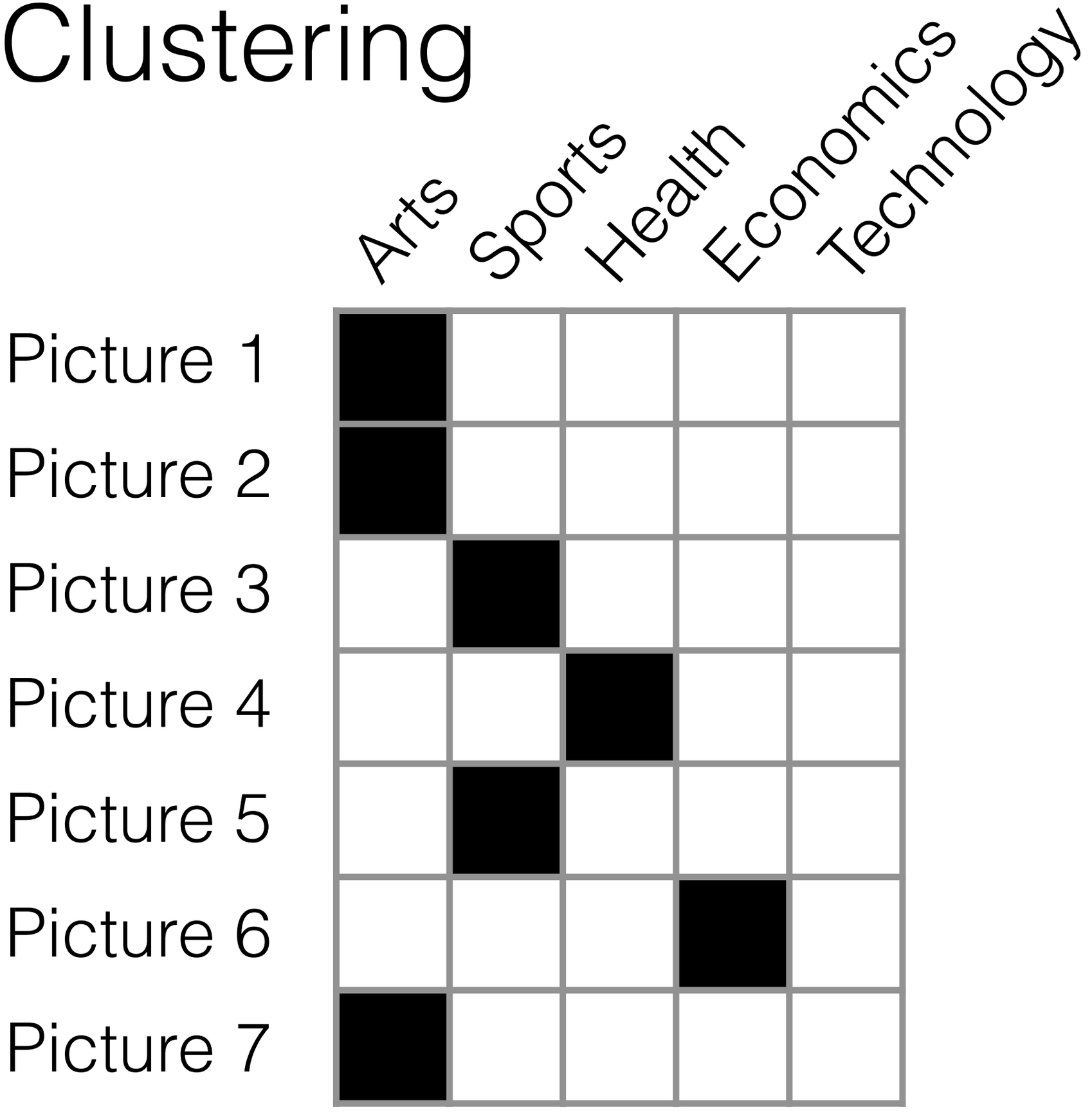
"Clusters"

Clustering

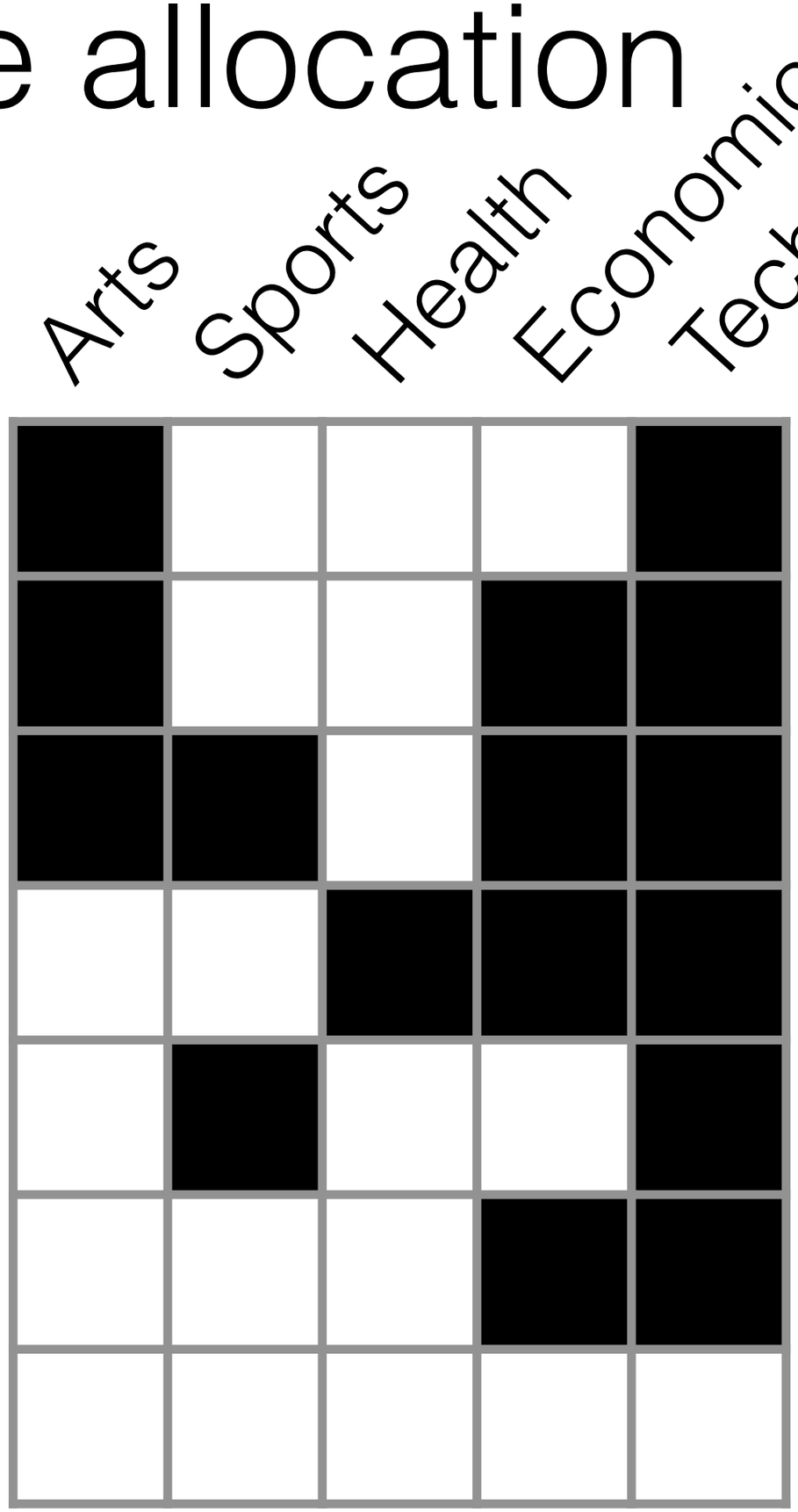


"Clusters"

Clustering

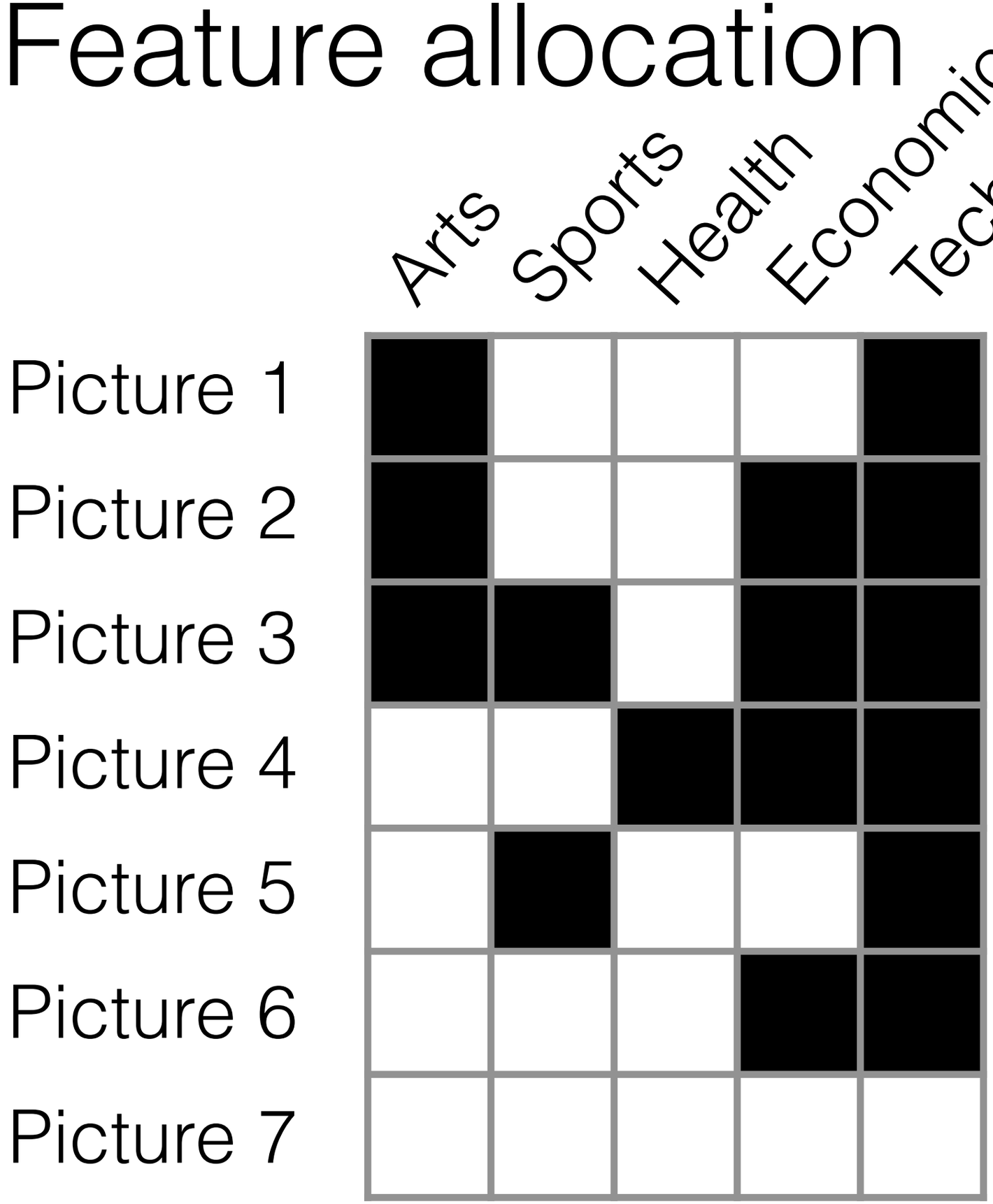

Arts
Sports
Health
Economics
Technology
Picture 1
Picture 2
Picture 3
Picture 4
Picture 5
Picture 6
Picture 7

Feature allocation



"Features"

Feature allocation



Many other possible latent structures in data

How do we learn latent structure?

How do we learn latent structure?

K-means

How do we learn latent structure?

K-means

- Fast

How do we learn latent structure?

K-means

- Fast
- Can parallelize

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes

- Modular (general latent structure)

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes

- Modular (general latent structure)
- Flexible (K can grow as data grows)

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes

- Modular (general latent structure)
- Flexible (K can grow as data grows)
- Coherent treatment of uncertainty

How do we learn latent structure?

K-means
- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes
- Modular (general latent structure)
- Flexible (K can grow as data grows)
- Coherent treatment of uncertainty

But...
- Nowadays: can have petabytes of data
- Practitioners turn to what runs

MAD-Bayes perspectives

- Bayesian nonparametrics assists the optimization-based inference community

MAD-Bayes perspectives

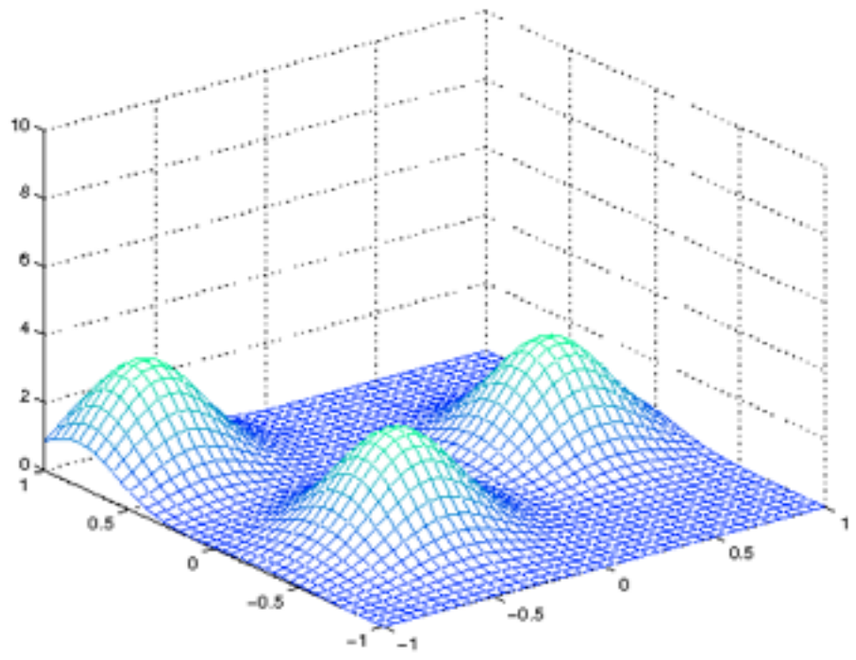
- Bayesian nonparametrics assists the optimization-based inference community
  - New, modular, flexible, nonparametric objectives & regularizers

MAD-Bayes perspectives

- Bayesian nonparametrics assists the optimization-based inference community
  - New, modular, flexible, nonparametric objectives & regularizers
  - Alternative perspective: fast initialization

MAD-Bayes perspectives

- Bayesian nonparametrics assists the optimization-based inference community
  - New, modular, flexible, nonparametric objectives & regularizers
  - Alternative perspective: fast initialization



- Inspiration
  - Consider a finite Gaussian mixture model
  - The steps of the EM algorithm limit to the steps of the K-means algorithm as the Gaussian variance is taken to 0

MAD-Bayes

The MAD-Bayes idea

- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective

MAD-Bayes

The MAD-Bayes idea

- Start with **nonparametric Bayes** model
- Take a similar **limit** to get a **K-means-like objective**

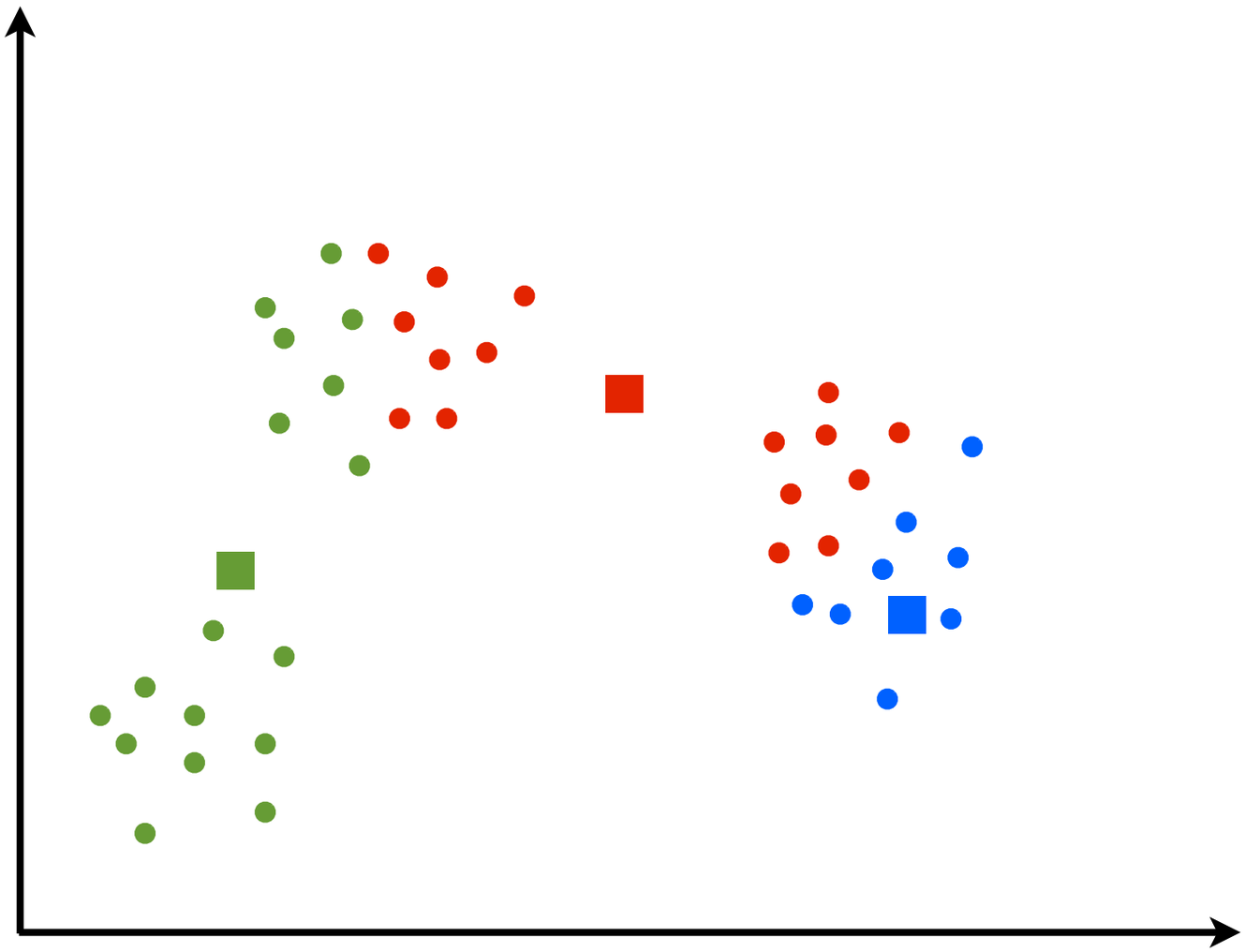
MAD-Bayes

The MAD-Bayes idea

- Start with nonparametric Bayes model
- Take a similar limit to get a **K-means-like objective**

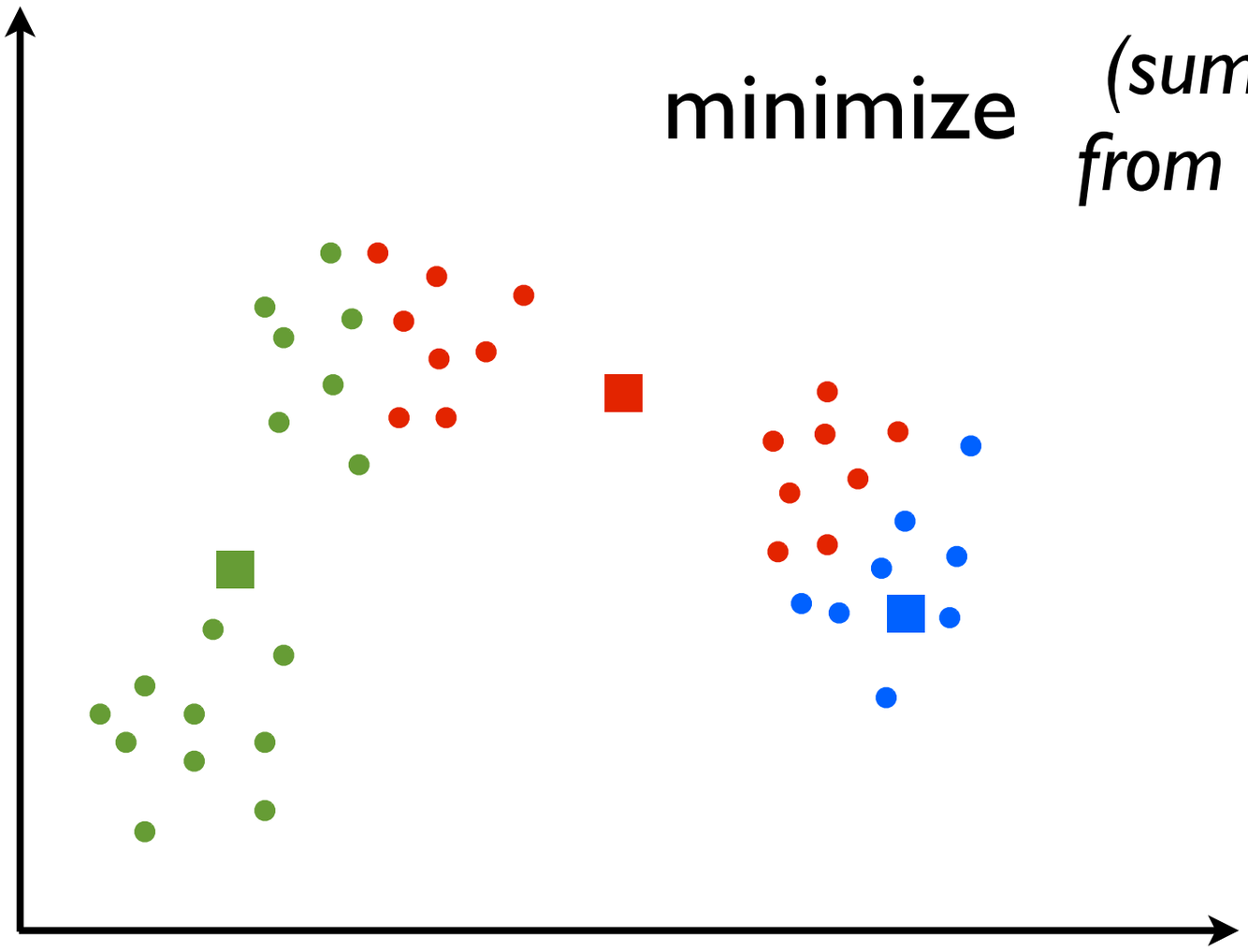
K-means

K-means clustering problem



K-means

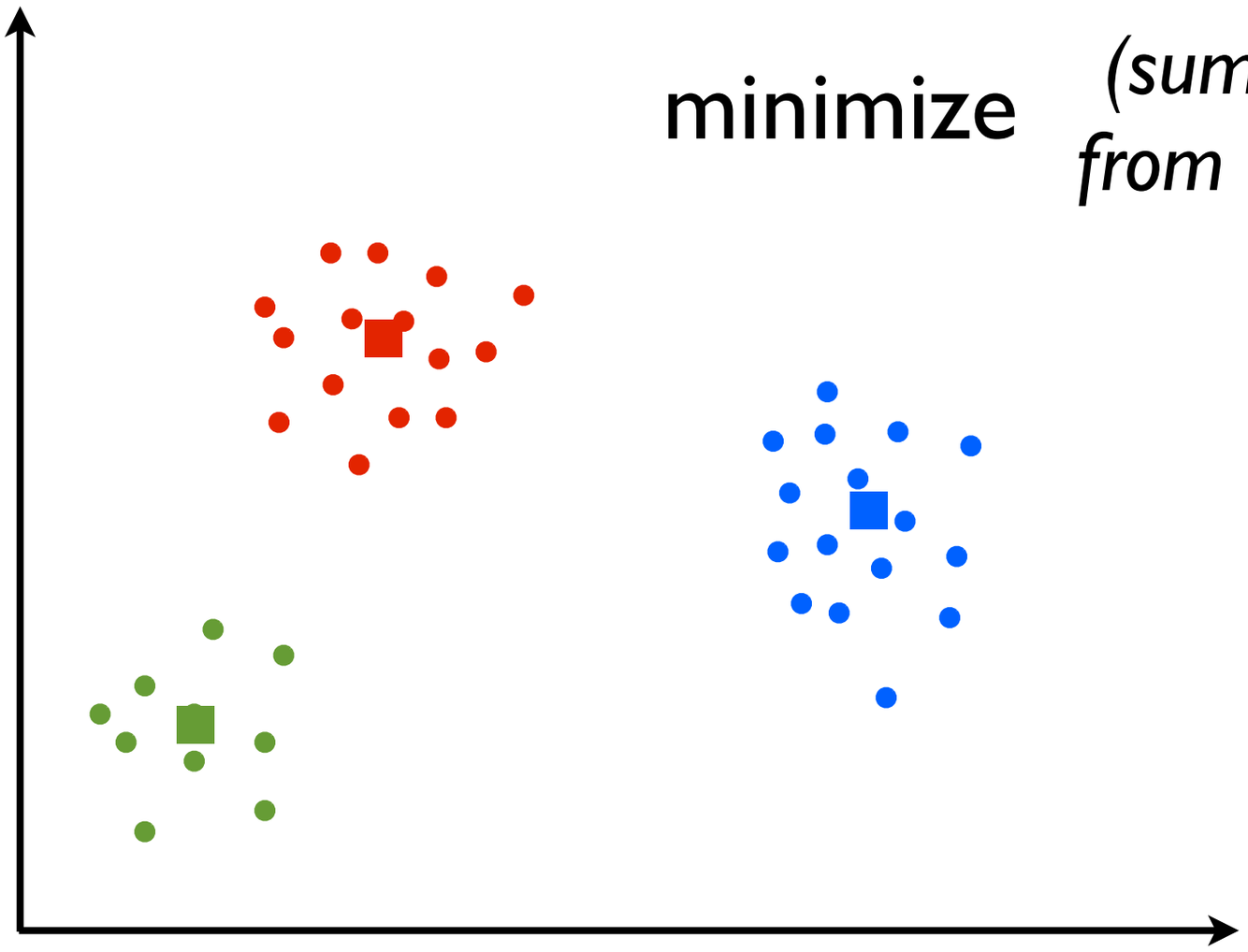
K-means clustering problem



minimize *(sum of square distances from data points to cluster centers)*

K-means

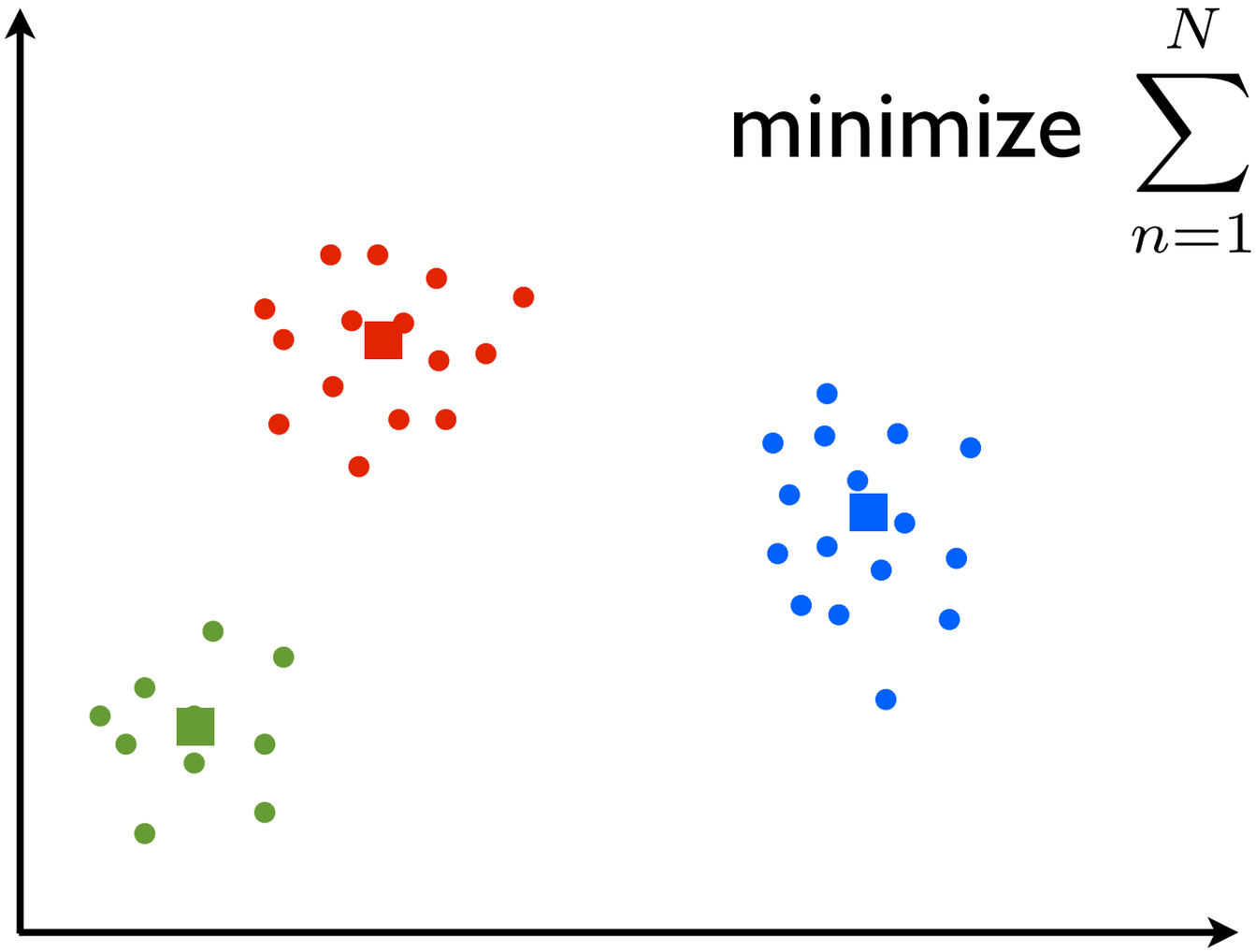
K-means clustering problem



minimize *(sum of square distances from data points to cluster centers)*

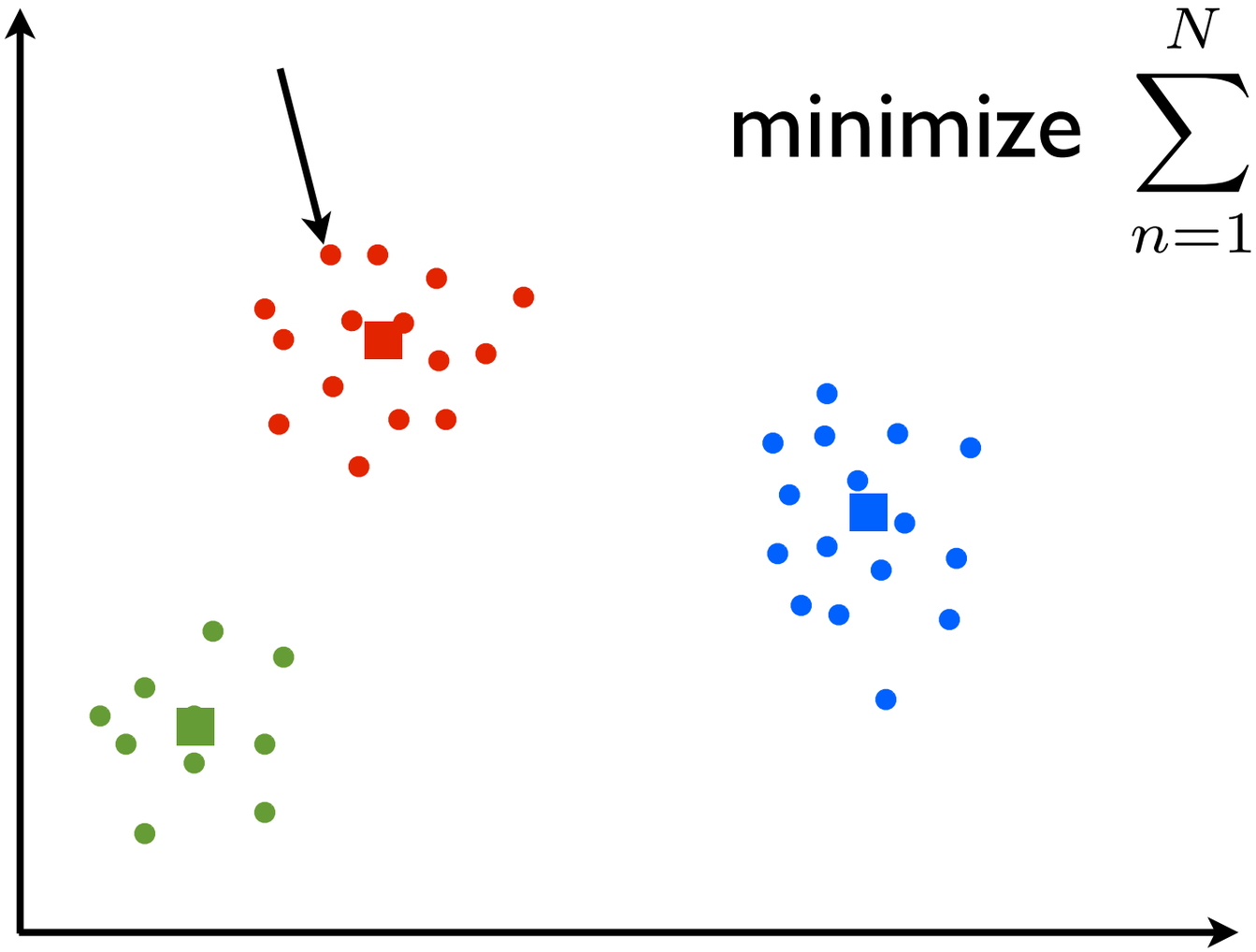
K-means

K-means clustering problem



$$\text{minimize} \sum_{n=1}^{N} \|x_n - \text{center}_n\|^2$$

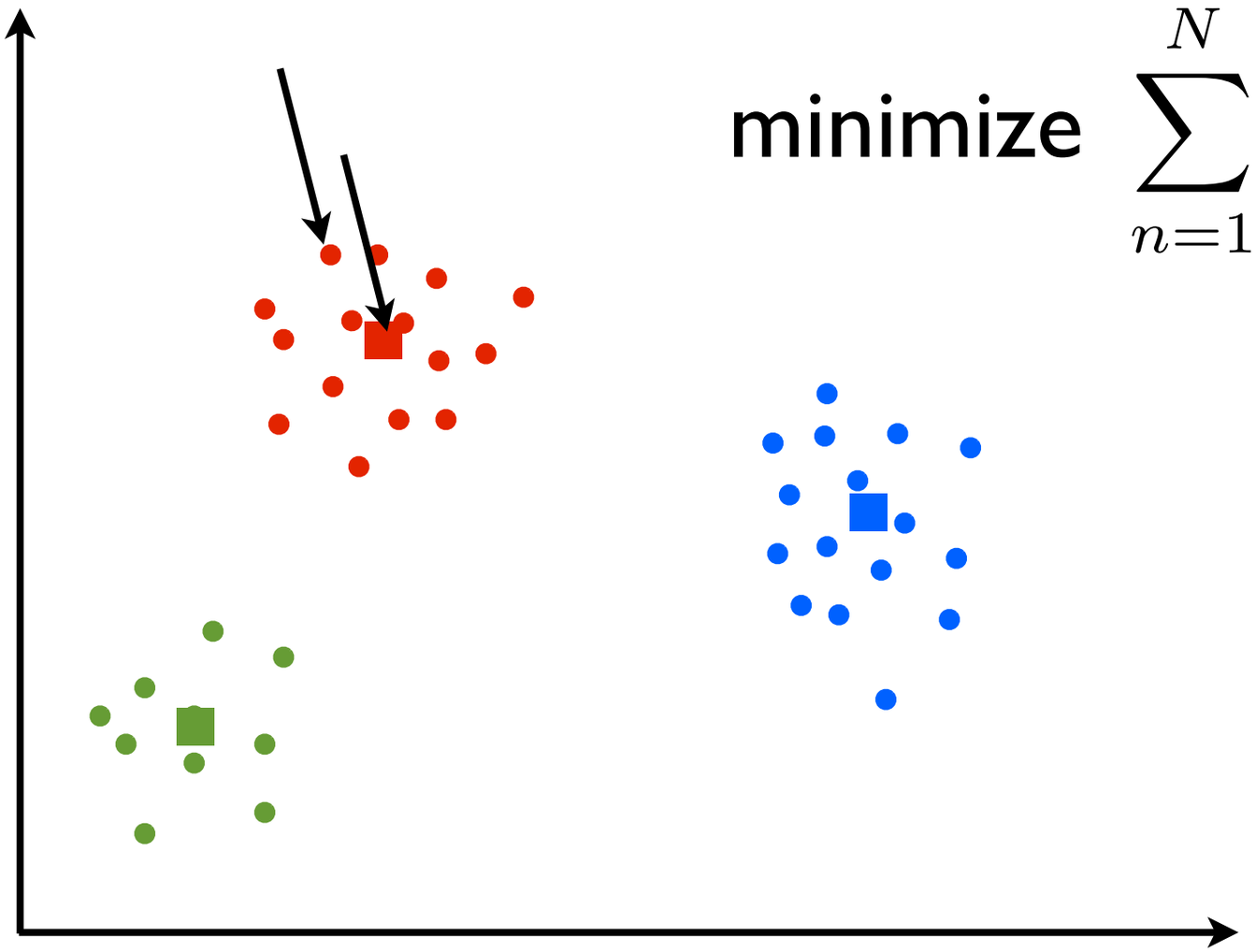
K-means



K-means clustering problem

$$\text{minimize} \sum_{n=1}^{N} \|x_n - \text{center}_n\|^2$$

K-means



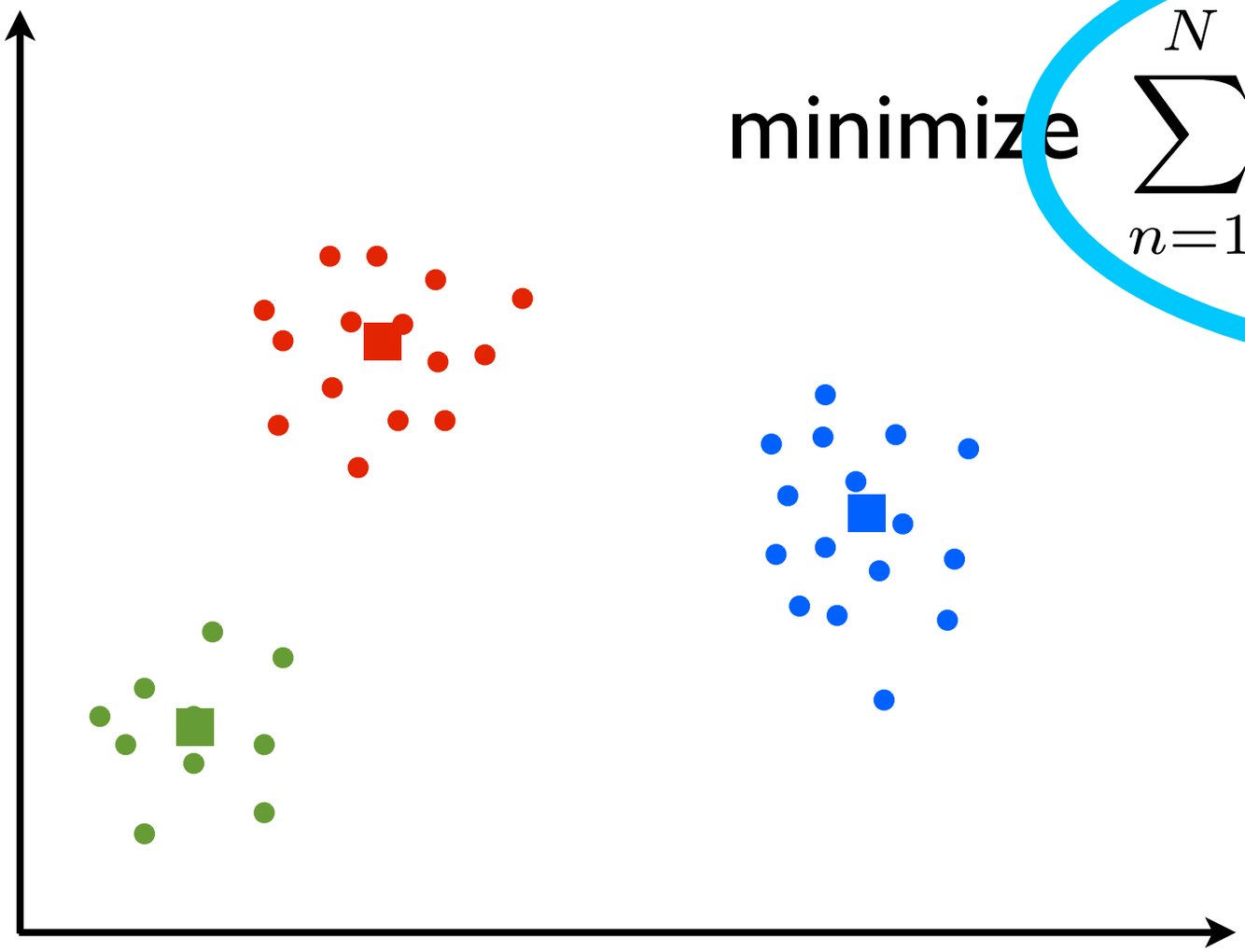
K-means clustering problem

$$\text{minimize} \sum_{n=1}^{N} \|x_n - \text{center}_n\|^2$$

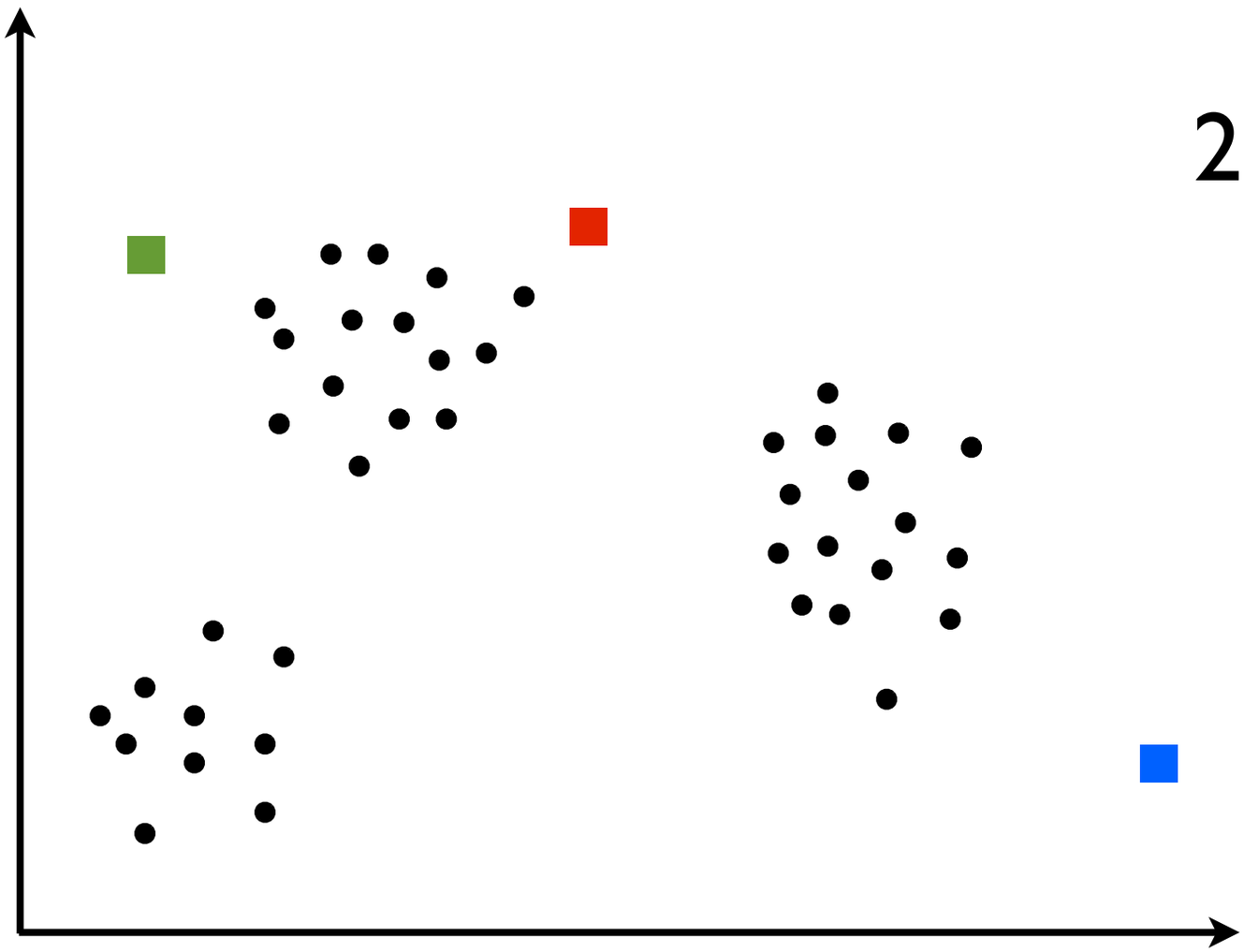
K-means

K-means objective

minimize $\sum_{n=1}^{N} \|x_n - \text{center}_n\|^2$



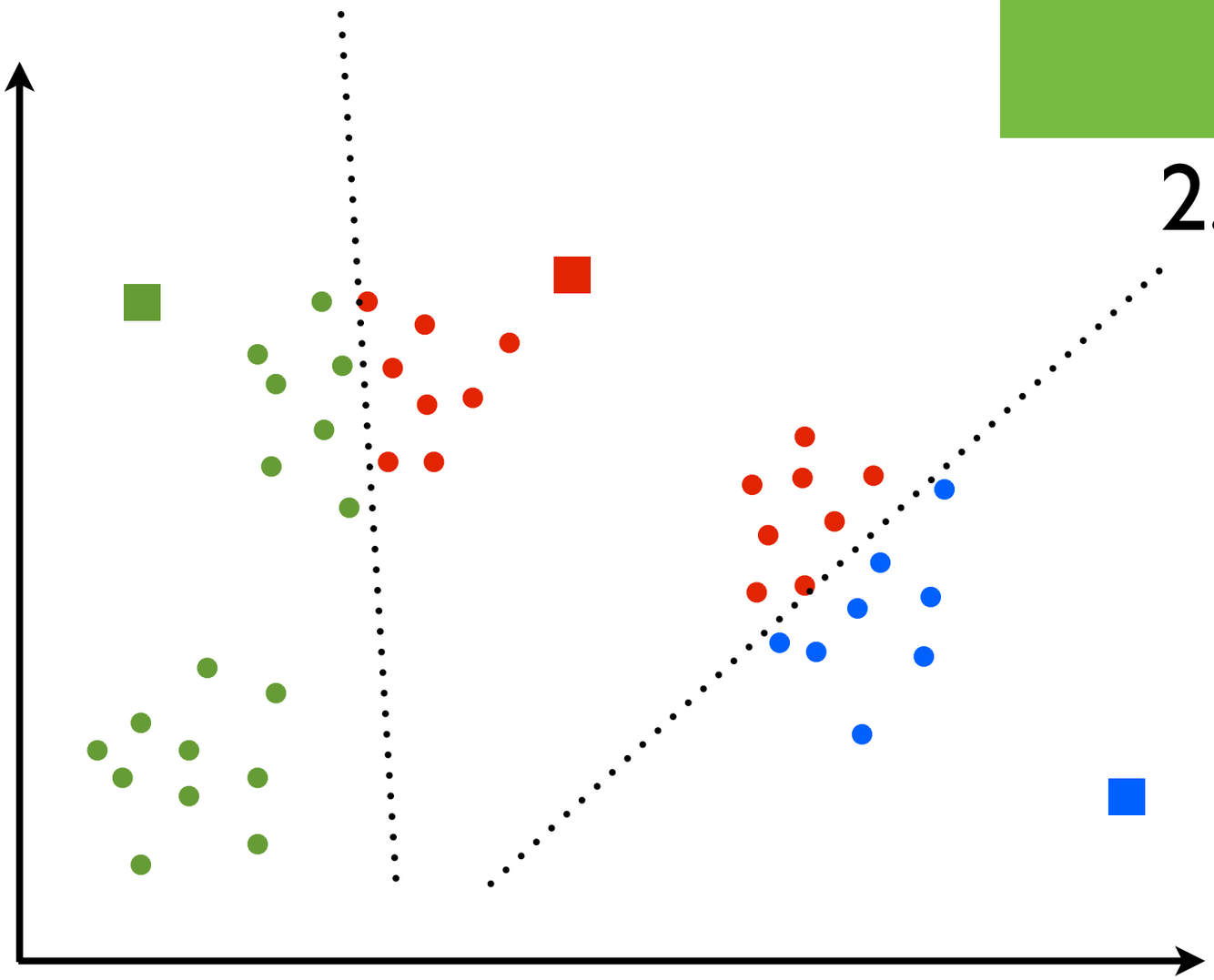
Lloyd's algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

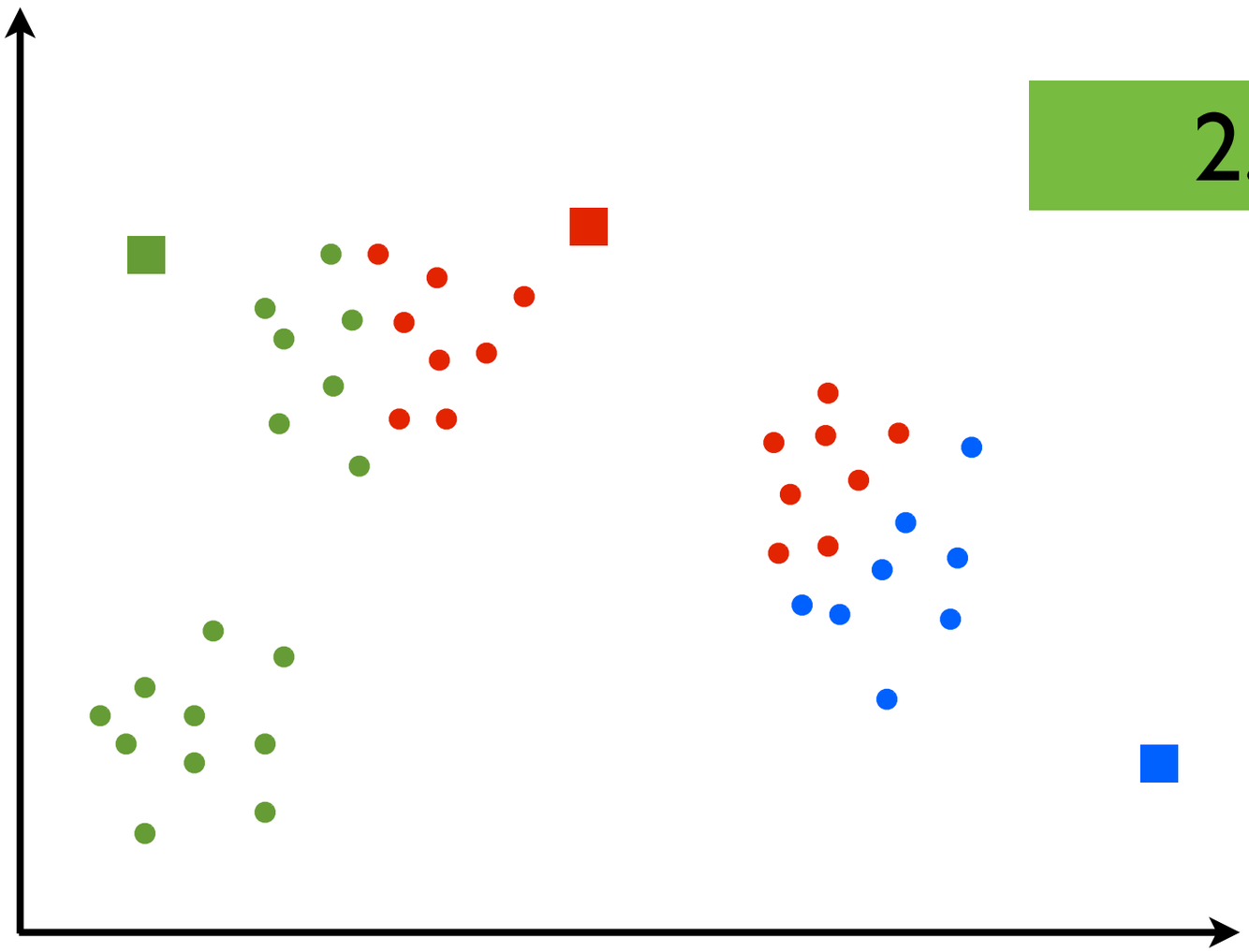
Lloyd's algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

Lloyd's algorithm

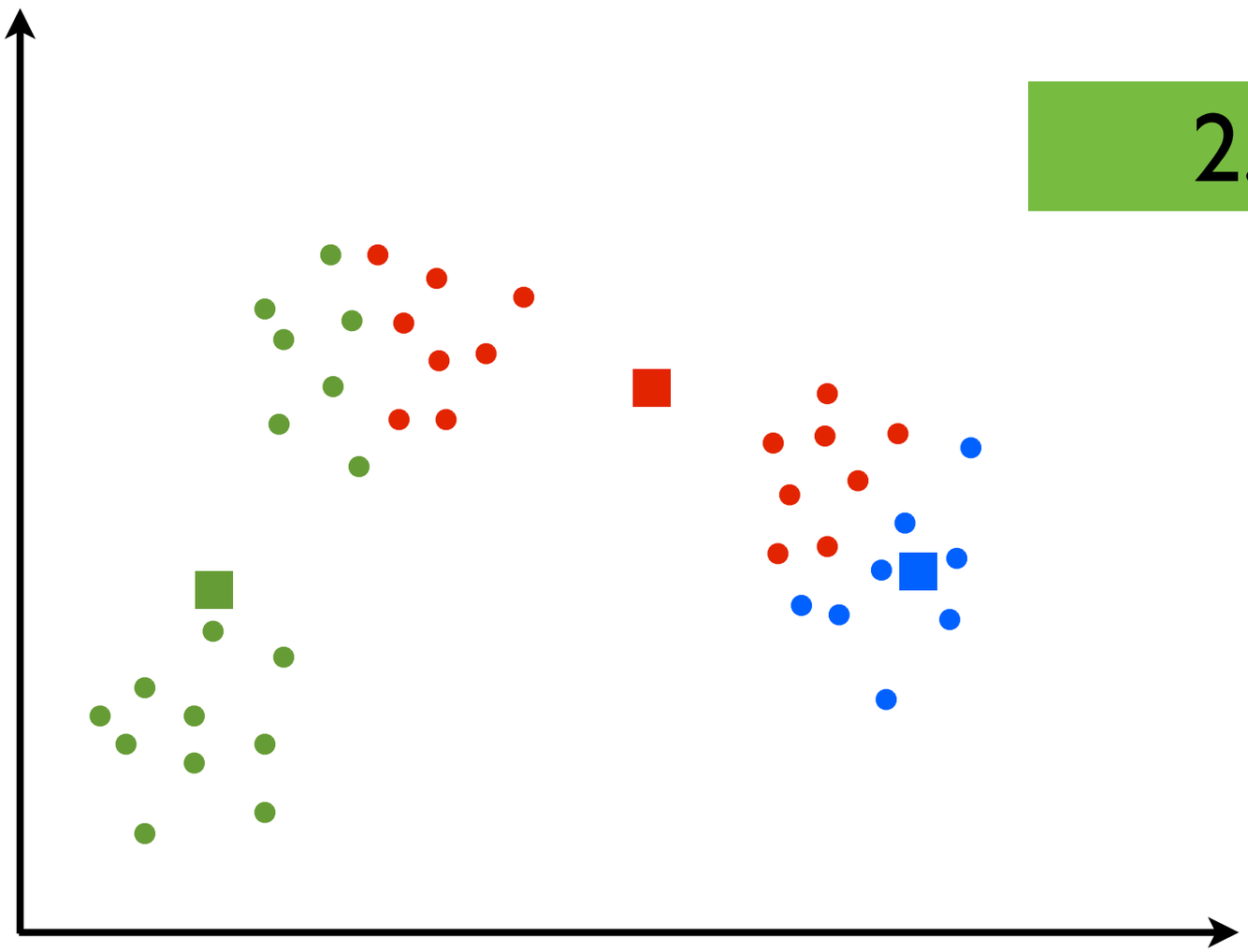


Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster

2. Update cluster means

Lloyd's algorithm

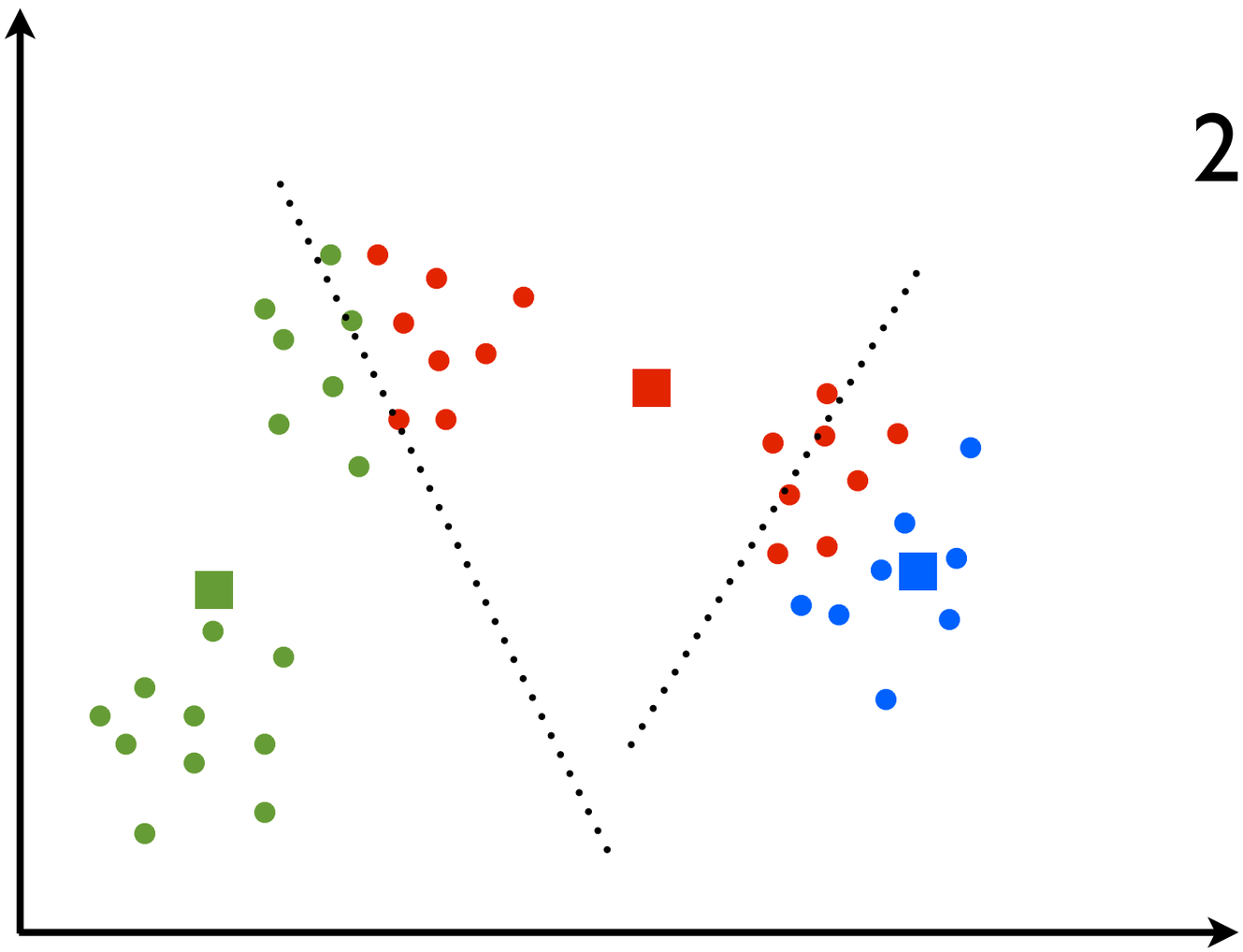


Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster

2. Update cluster means

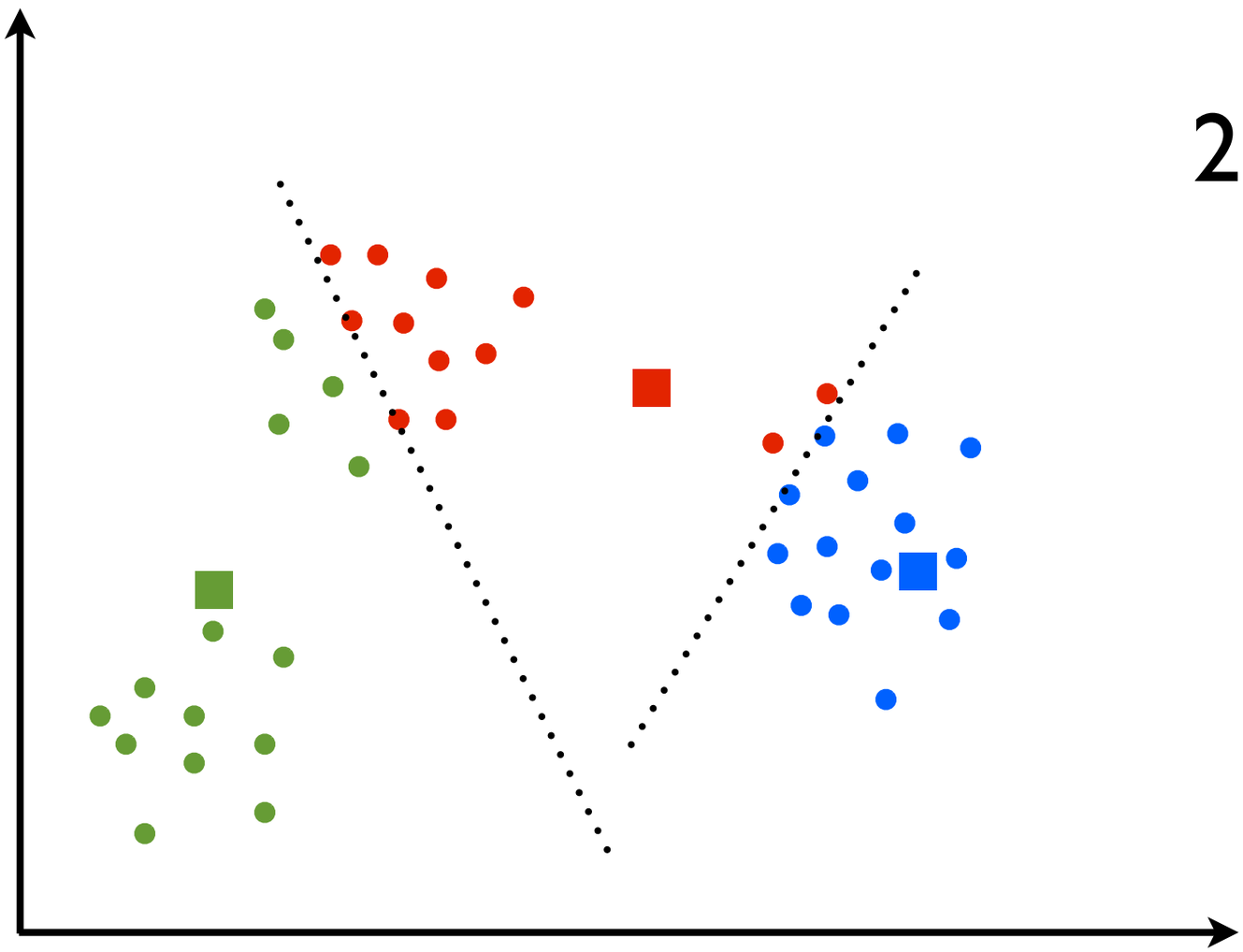
Lloyd's algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

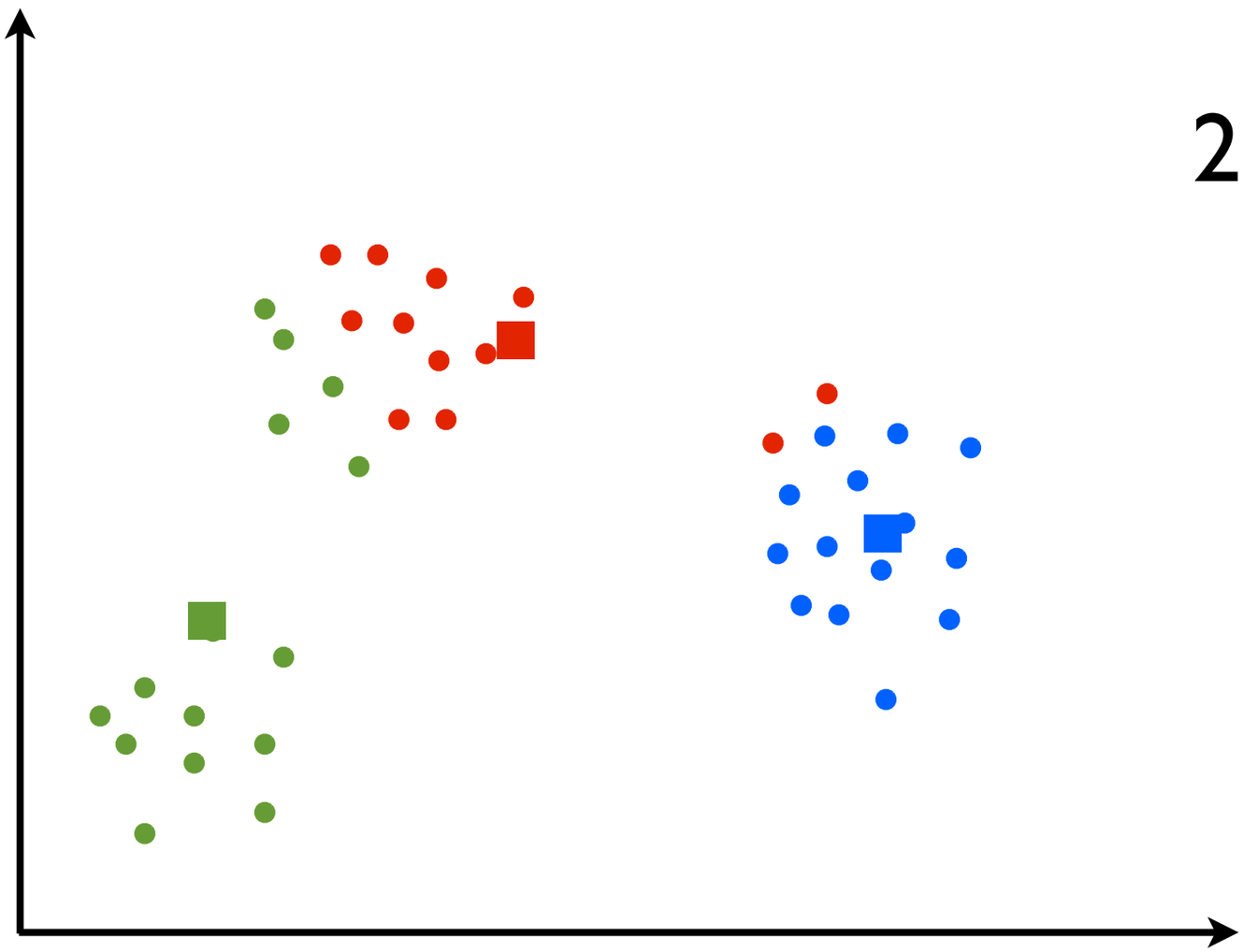
Lloyd's algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

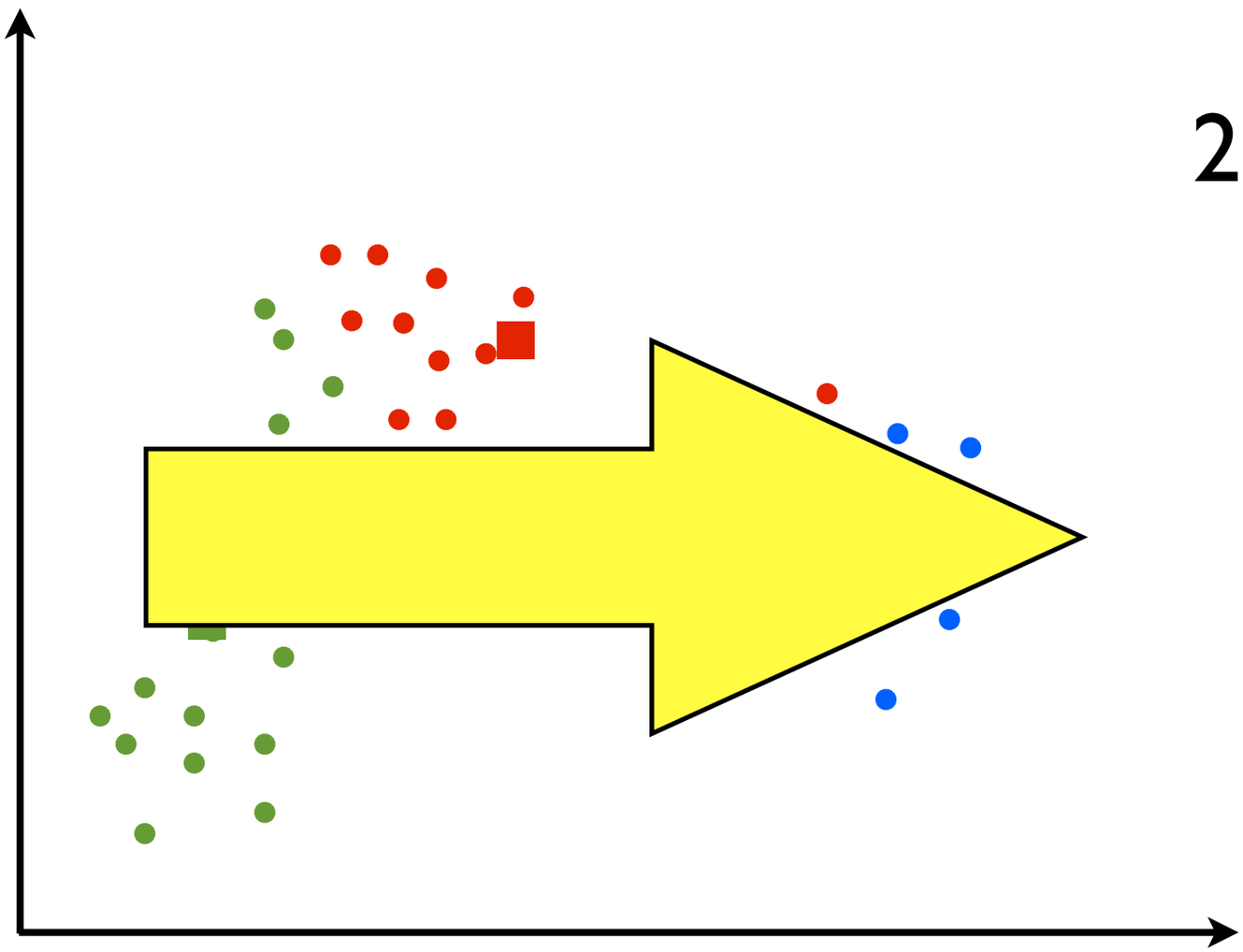
Lloyd's algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

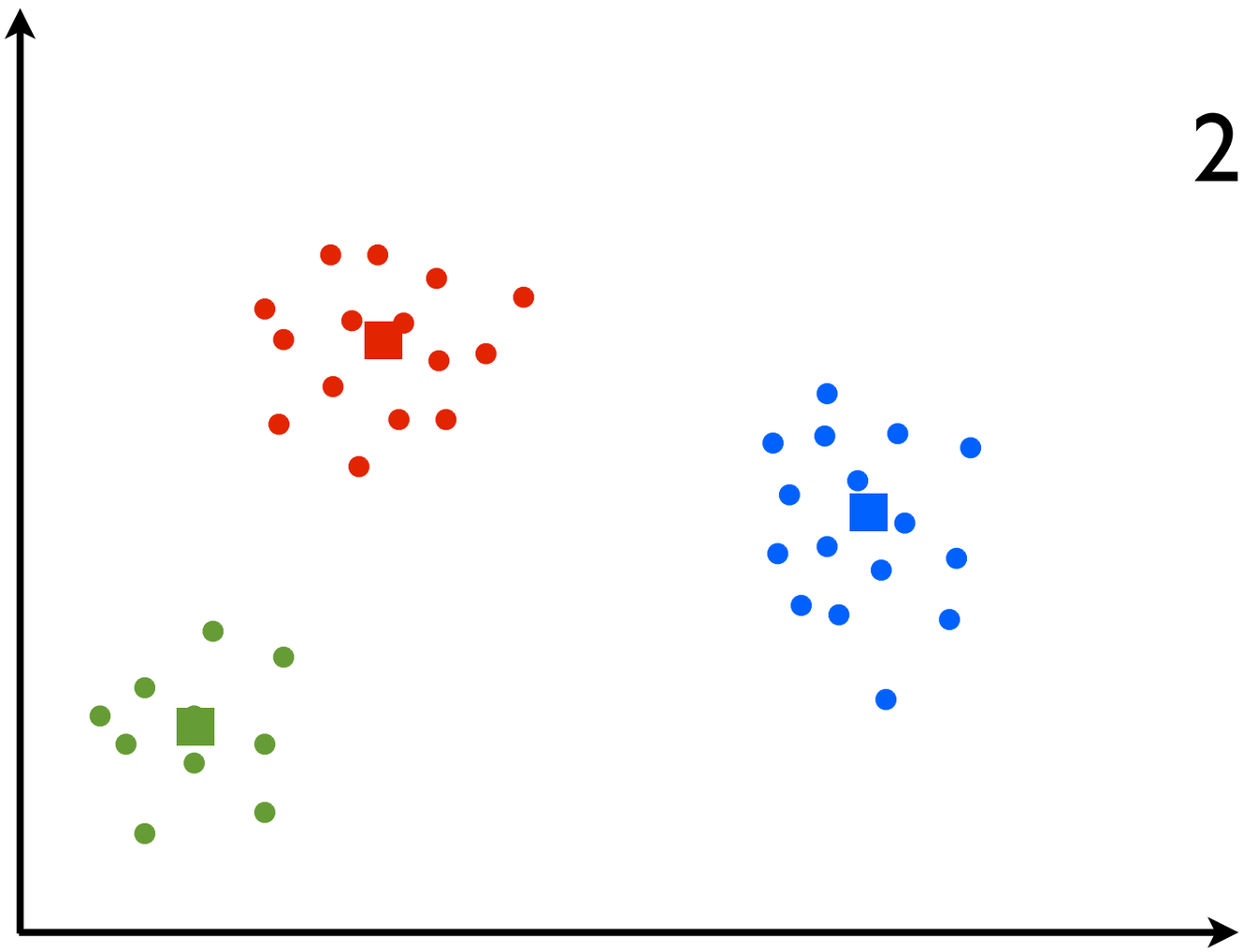
Lloyd's algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

Lloyd's algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

MAD-Bayes

The MAD-Bayes idea

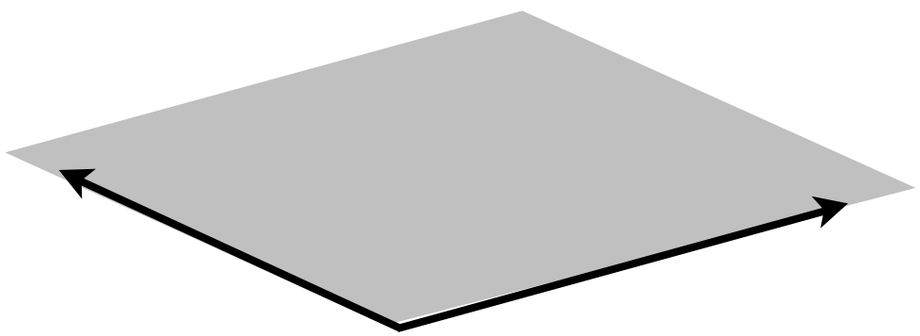
- Start with nonparametric Bayes model
- Take a similar limit to get a **K-means-like objective**

MAD-Bayes

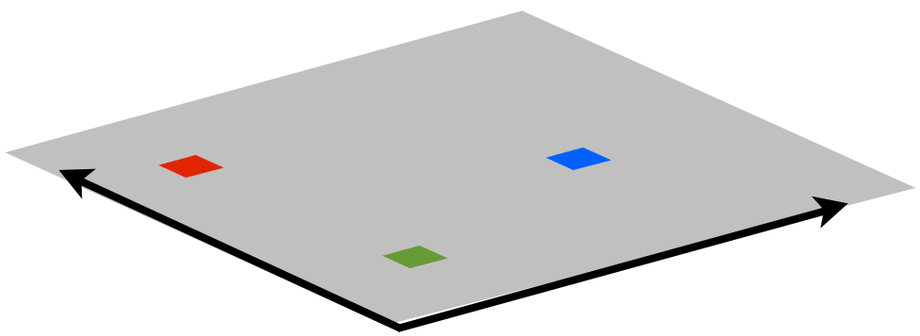
The MAD-Bayes idea

- Start with **nonparametric Bayes** model
- Take a similar limit to get a K-means-like objective

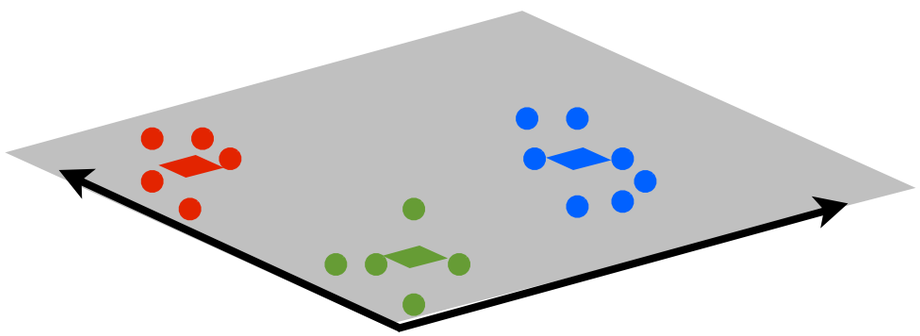
Bayesian model



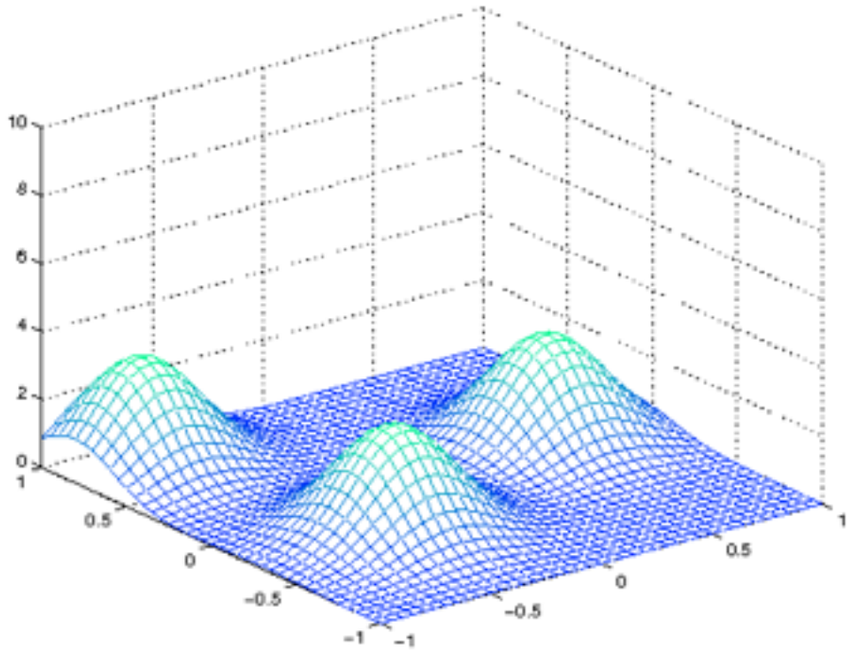
Bayesian model



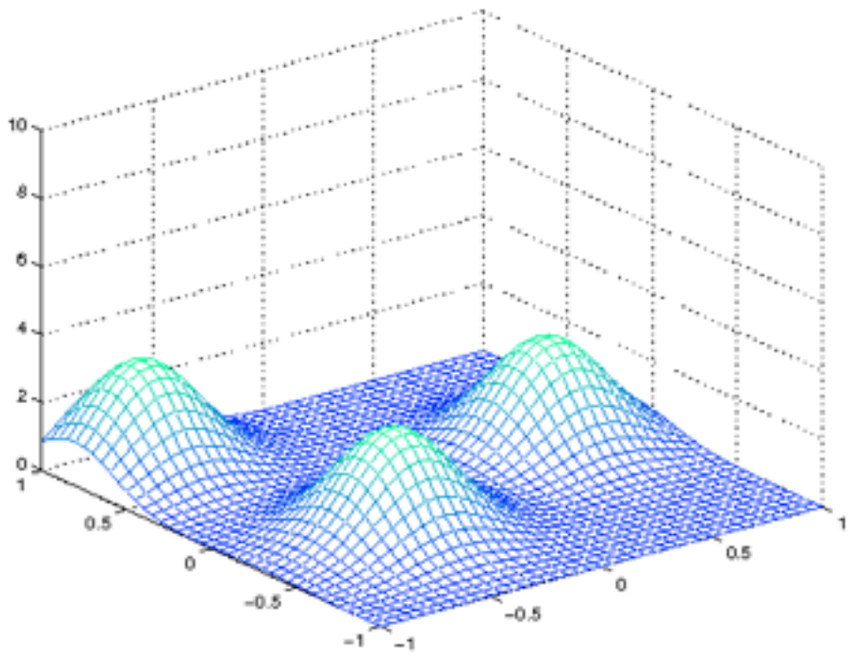
Bayesian model



Bayesian model



Bayesian model



Nonparametric

- number of parameters can grow with the number of data points

MAD-Bayes

The MAD-Bayes idea

- Start with **nonparametric Bayes** model
- Take a similar limit to get a K-means-like objective

MAD-Bayes

The MAD-Bayes idea

- Start with nonparametric Bayes model
- Take a similar **limit** to get a K-means-like objective

MAD-Bayes

MAD-Bayes

- *Maximum a Posteriori* (MAP) is an optimization problem

$$\text{argmax}_{\text{parameters}} \mathbb{P}(\text{parameters}|\text{data})$$

MAD-Bayes

- *Maximum a Posteriori* (MAP) is an optimization problem

$$\text{argmax}_{\text{parameters}} \mathbb{P}(\text{parameters}|\text{data})$$

- We take a limit of the objective (posterior) and get one like K-means

MAD-Bayes

- *Maximum a Posteriori* (MAP) is an optimization problem

$$\text{argmax}_{\text{parameters}} \mathbb{P}(\text{parameters}|\text{data})$$

- We take a limit of the objective (posterior) and get one like K-means
  - ◇ "Small-variance asymptotics"

MAD-Bayes

Bayesian posterior

K-means-like objectives

MAD-Bayes

Bayesian posterior

K-means-like objectives

Mixture of K Gaussians  K-means

MAD-Bayes

<u>Bayesian posterior</u>

<u>K-means-like objectives</u>

Mixture of K Gaussians → K-means

Dirichlet process mixture → Unbounded number of clusters

MAD-Bayes

Bayesian posterior → K-means-like objectives

Mixture of K Gaussians → K-means

Dirichlet process mixture → Unbounded number of clusters

Hierarchical Dirichlet process → Multiple data sets share cluster centers

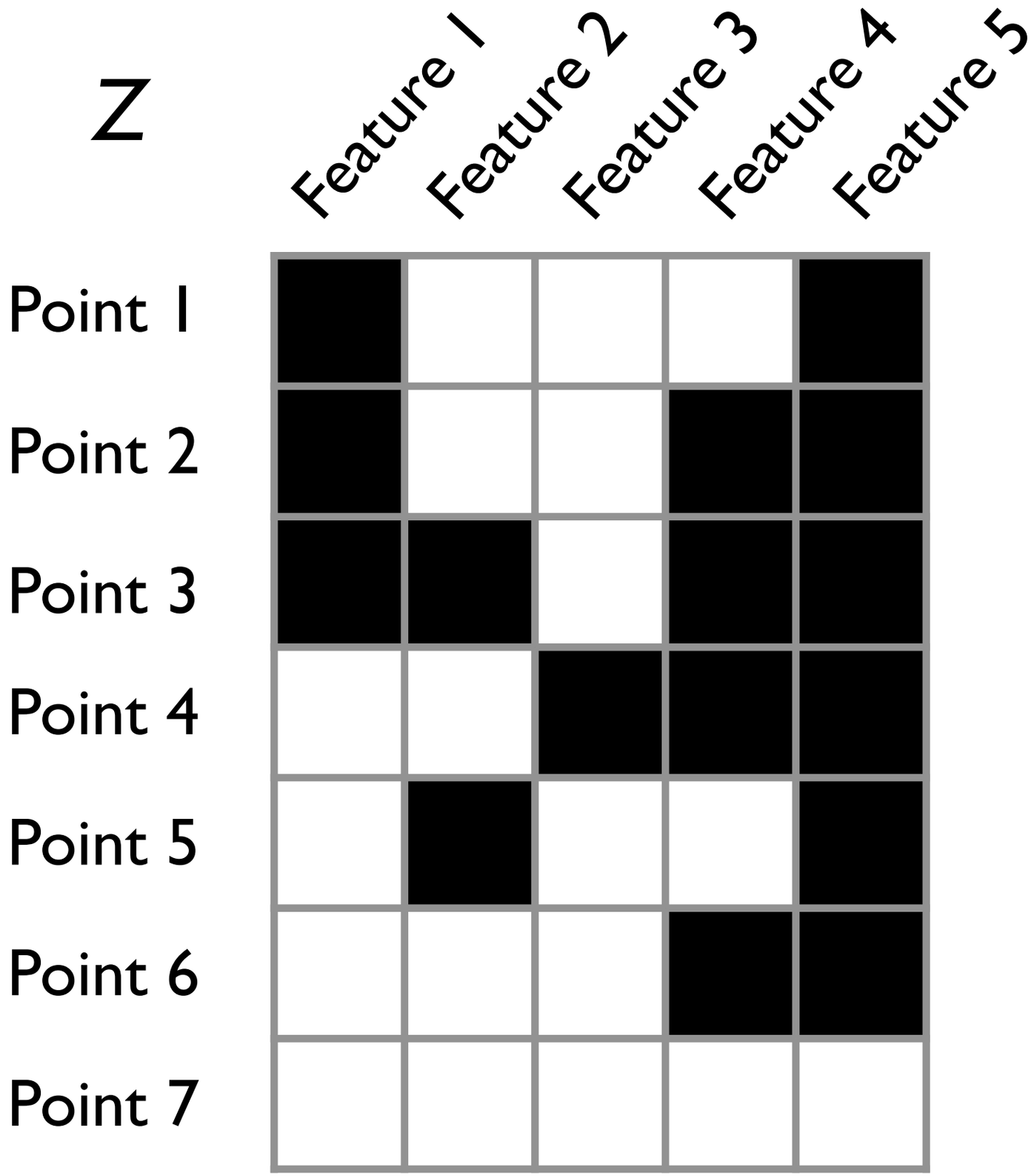
MAD-Bayes

Bayesian posterior	K-means-like objectives
Mixture of K Gaussians	K-means
Dirichlet process mixture	Unbounded number of clusters
Hierarchical Dirichlet process	Multiple data sets share cluster centers
⋮	⋮

MAD-Bayes

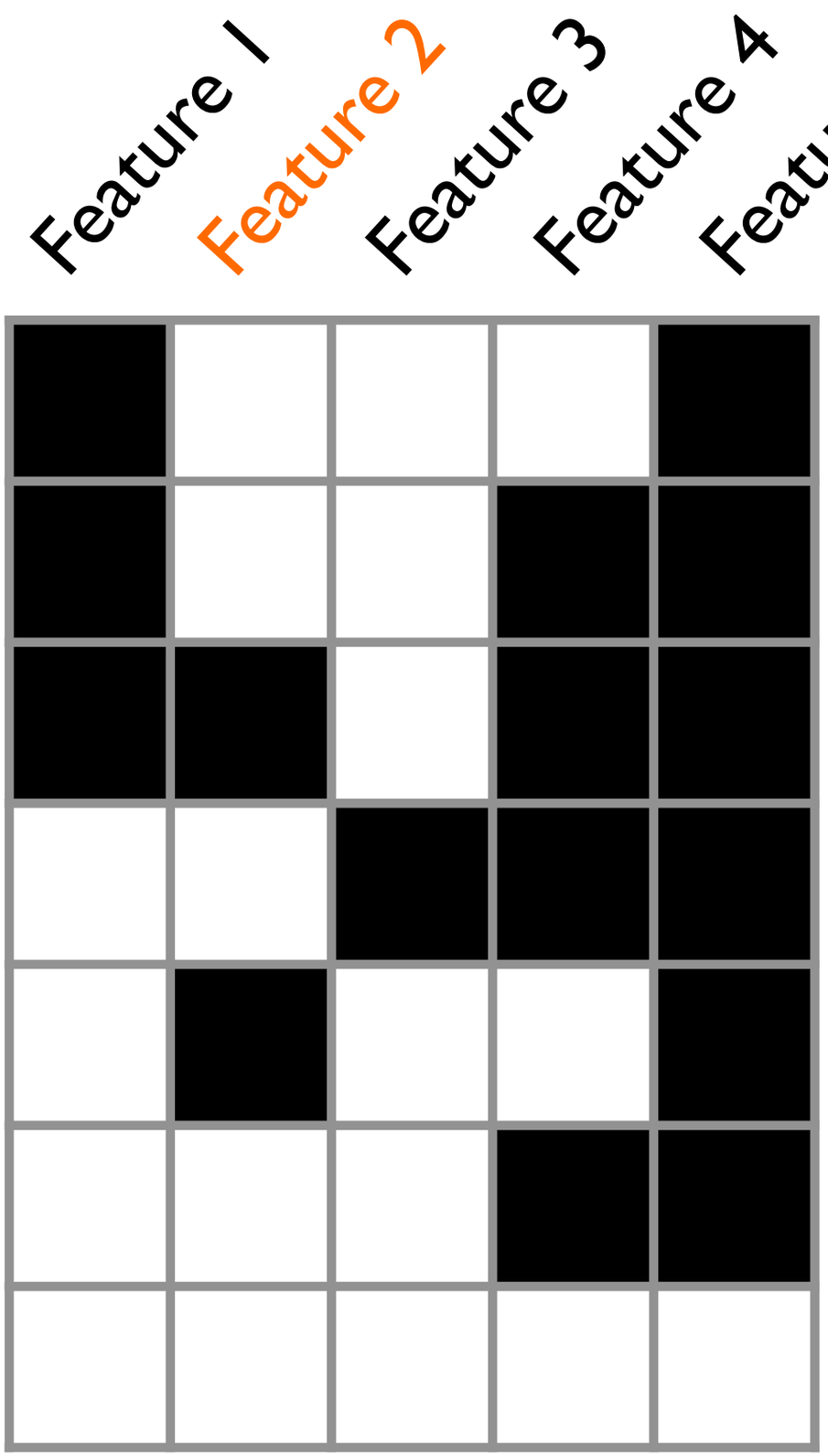
Bayesian posterior	K-means-like objectives
Mixture of K Gaussians	K-means
Dirichlet process mixture	Unbounded number of clusters
Hierarchical Dirichlet process	Multiple data sets share cluster centers
⋮	⋮
Beta process	Features

Features

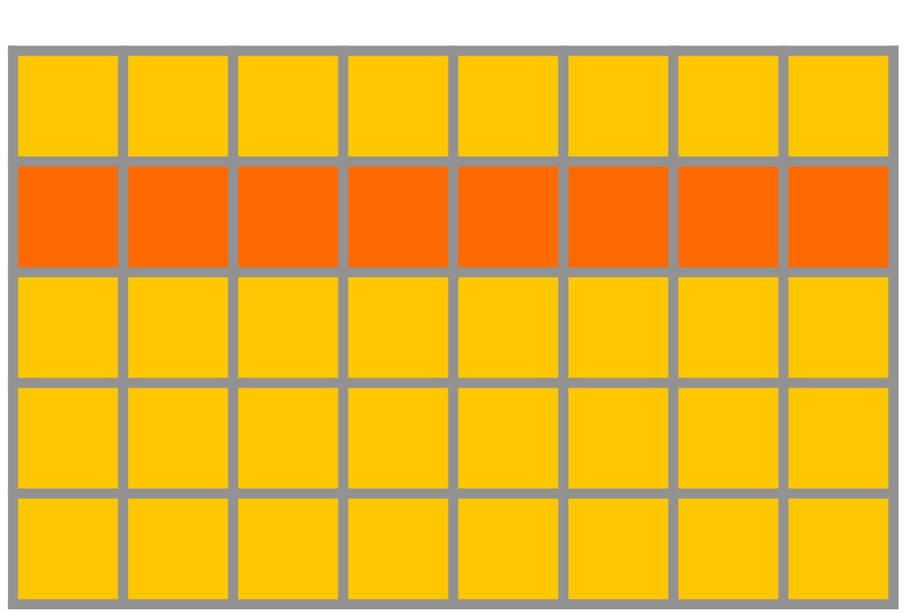

Z
Feature 1
Feature 2
Feature 3
Feature 4
Feature 5
Point 1
Point 2
Point 3
Point 4
Point 5
Point 6
Point 7

Features

Z

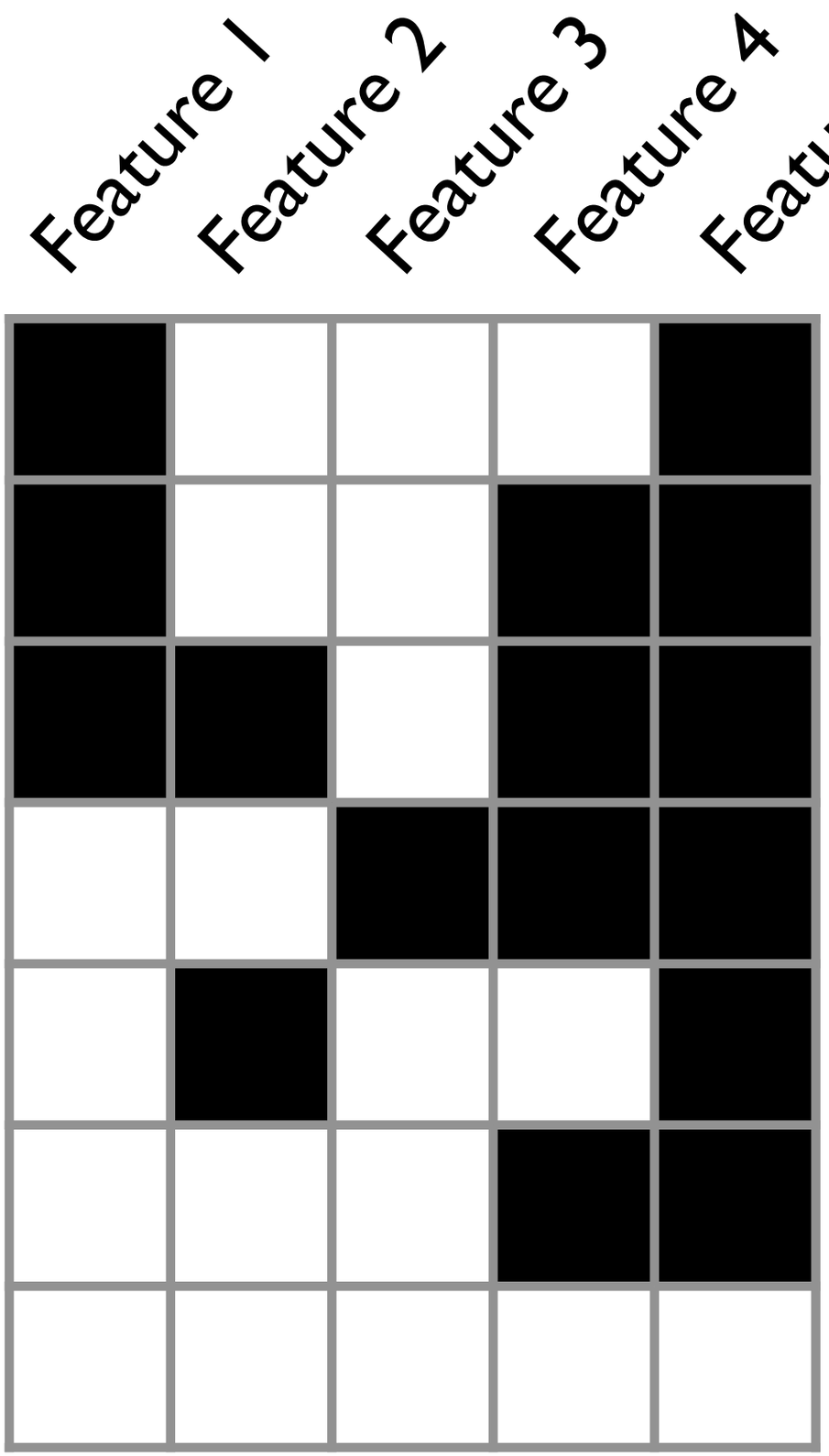

Feature 1
Feature 2
Feature 3
Feature 4
Feature 5
Point 1
Point 2
Point 3
Point 4
Point 5
Point 6
Point 7

A

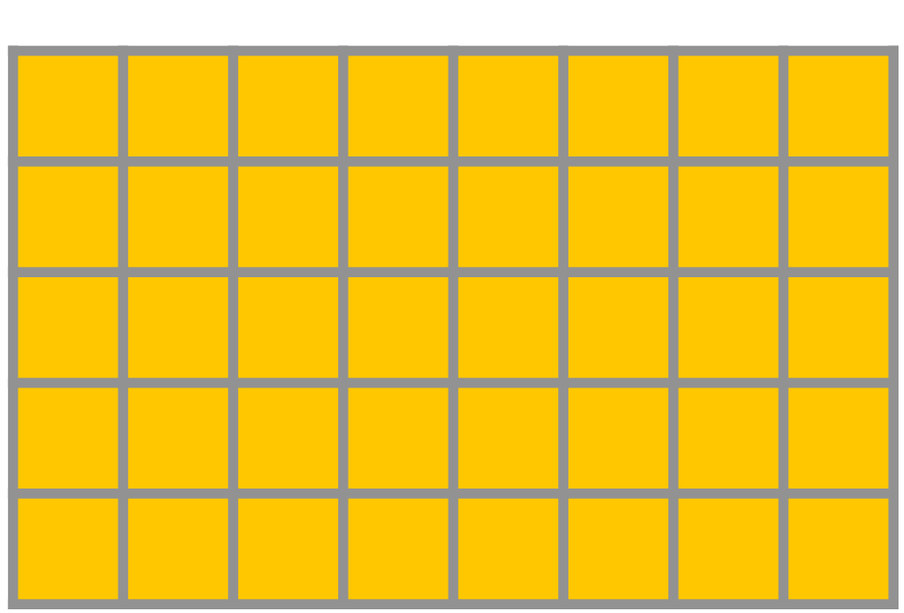


Features

Z


Feature 1
Feature 2
Feature 3
Feature 4
Feature 5
Point 1
Point 2
Point 3
Point 4
Point 5
Point 6
Point 7

A



MAD-Bayes

Bayesian posterior

$$\mathbb{P}(Z, A|X)$$

$$\propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X-ZA)'(X-ZA))\right\}$$

$$\cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k}-1)!(N-S_{N,k})!}{N!}$$

$$\cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.$$

MAD-Bayes

Bayesian posterior

$$\mathbb{P}(Z, A|X)$$

$$\propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X-ZA)'(X-ZA))\right\}$$

$$\cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k}-1)!(N-S_{N,k})!}{N!}$$

$$\cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.$$

MAD-Bayes

Bayesian posterior

$$\mathbb{P}(Z, A|X)$$

$$\propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X-ZA)'(X-ZA))\right\}$$

$$\cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k}-1)!(N-S_{N,k})!}{N!}$$

$$\cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.$$

MAD-Bayes

Bayesian posterior

$$\mathbb{P}(Z, A|X)$$

$$\propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X - ZA)'(X - ZA))\right\}$$

$$\cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k}-1)!(N-S_{N,k})!}{N!}$$

$$\cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.$$

MAD-Bayes

Bayesian posterior

$$\mathbb{P}(Z, A|X)$$

$$\propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X - ZA)'(X - ZA))\right\}$$

$$\cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k}-1)!(N-S_{N,k})!}{N!}$$

$$\cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.$$

MAD-Bayes

Bayesian posterior

$$\mathbb{P}(Z, A|X)$$

$$\propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X - ZA)'(X - ZA))\right\}$$

$$\cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k}-1)!(N-S_{N,k})!}{N!}$$

$$\cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.$$

MAD-Bayes

Bayesian posterior

$$\mathbb{P}(Z, A|X)$$

$$\propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X - ZA)'(X - ZA))\right\}$$

$$\cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k}-1)!(N-S_{N,k})!}{N!}$$

$$\cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.$$

MAD-Bayes

BP-means objective

$$\text{argmin}_{K^+,Z,A} \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$

MAD-Bayes

BP-means objective

$$\text{argmin}_{K^+,Z,A} \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$

MAD-Bayes

BP-means objective

$$\text{argmin}_{K^+,Z,A} \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$

MAD-Bayes

BP-means objective

$$\text{argmin}_{K^+,Z,A} \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$

BP-means algorithm

Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to features
   - Create a new feature if it lowers the objective
2. Update feature means $A \leftarrow (Z'Z)^{-1}Z'X$

MAD-Bayes

BP-means objective

$$\text{argmin}_{K^+,Z,A} \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$

BP-means algorithm

Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to features
   - Create a new feature if it lowers the objective
2. Update feature means $A \leftarrow (Z'Z)^{-1}Z'X$

MAD-Bayes

BP-means objective

$$\text{argmin}_{K^+,Z,A} \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$

BP-means algorithm

Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to features
   - Create a new feature if it lowers the objective
2. Update feature means $A \leftarrow (Z'Z)^{-1}Z'X$

MAD-Bayes

BP-means objective

$$\text{argmin}_{K^+,Z,A} \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$

BP-means algorithm

Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to features
   - Create a new feature if it lowers the objective
2. Update feature means $A \leftarrow (Z'Z)^{-1}Z'X$

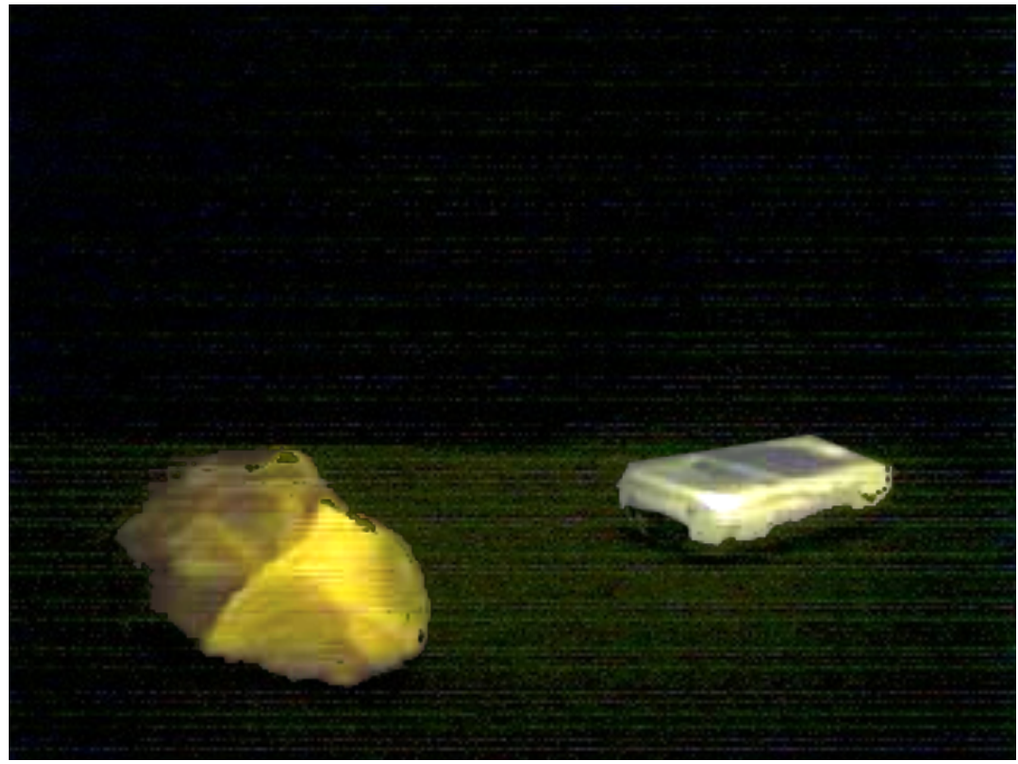
MAD-Bayes

Griffiths & Ghahramani (2006) computer vision problem "tabletop data"

[Griffiths, Ghahramani 2006]

MAD-Bayes

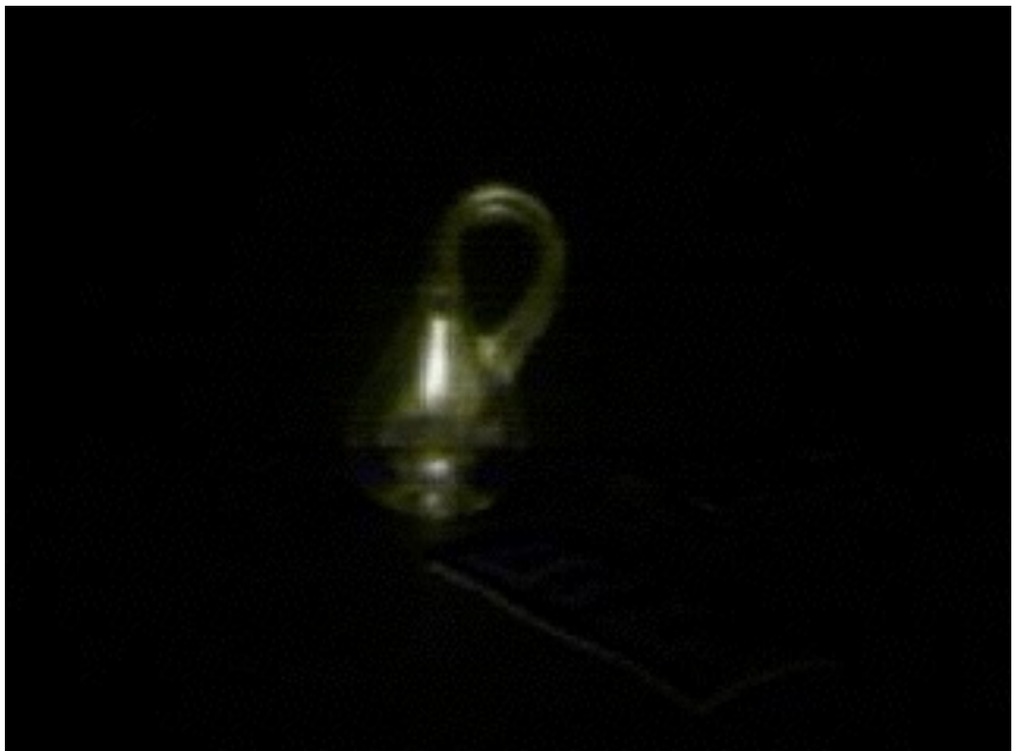
Griffiths & Ghahramani (2006) computer vision problem "tabletop data"



MAD-Bayes

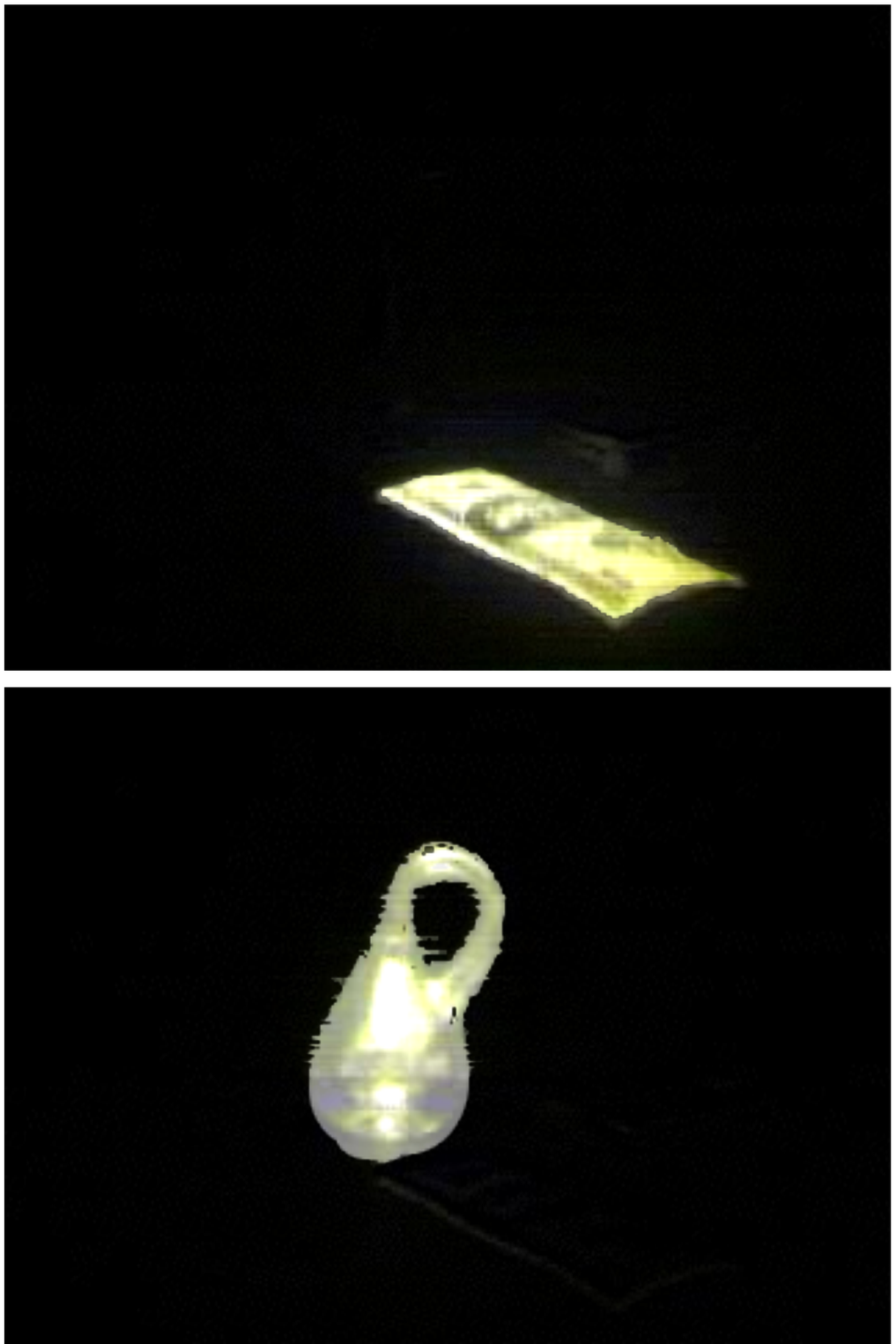
BP-means features: table and four objects






MAD-Bayes

BP-means features: table and four objects



MAD-Bayes

Griffiths & Ghahramani (2006) computer vision problem "tabletop data"

Bayesian posterior
Gibbs sampler

BP-means algorithm

$8.5 * 10^3$ sec

0.36 sec

Still faster by order of magnitude
if restart 1000 times

MAD-Bayes

Parallelism and optimistic concurrency control

	DP-means alg.	BP-means alg.
# data points	**134M**	**8M**
time per iteration	**5.5 min**	**4.3 min**

MAD-Bayes

Bayesian posterior	K-means-like objectives
Mixture of K Gaussians	K-means
Dirichlet process mixture	Unbounded number of clusters
Hierarchical Dirichlet process	Multiple data sets share cluster centers
⋮	⋮
Beta process	Features

MAD-Bayes Conclusions

- We provide new optimization objectives and regularizers
  - In fact, general means of obtaining more
  - Straightforward, fast algorithms

MAD-Bayes Conclusions

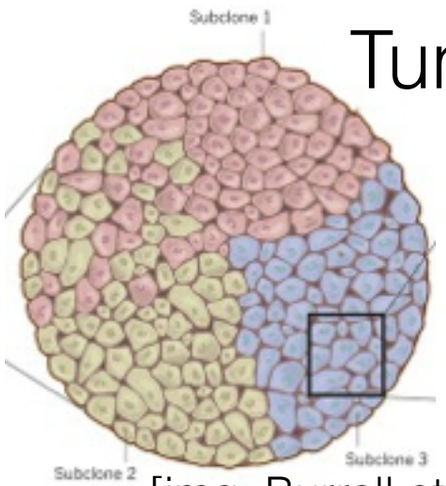
- We provide new optimization objectives and regularizers
  - In fact, general means of obtaining more
  - Straightforward, fast algorithms

- Applications:

MAD-Bayes Conclusions

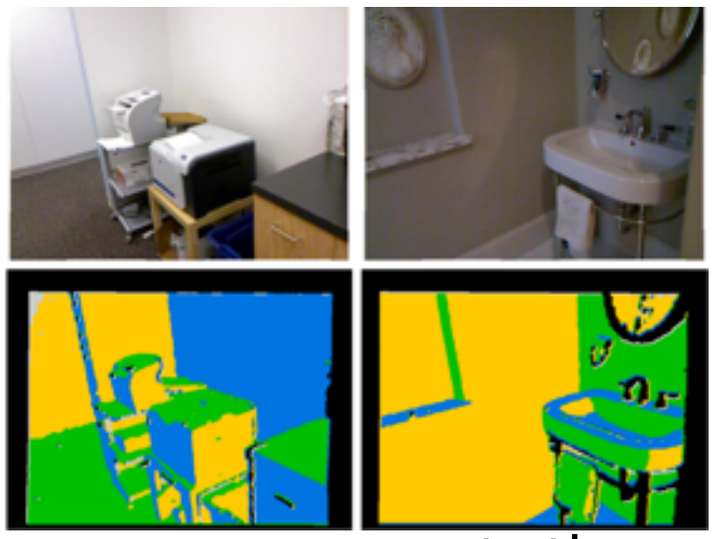
- We provide new optimization objectives and regularizers
  - In fact, general means of obtaining more
  - Straightforward, fast algorithms

- Applications:



Tumor heterogeneity
[Xu et al 2015]

[img: Burrell et al 2013]



Real-time scene segmentation
[Straub et al 2016]

Protein matching and alignment
[Green 2015]

Change points in disease progression
[Huggins et al 2015]

Roadmap

- Posterior approximation trade-offs
  - Point estimates
  - Uncertainties
- Robustness trade-offs
  - Local robustness quantification
- Theoretical guarantees on quality

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes)
  - Uncertainties
- Robustness trade-offs
  - Local robustness quantification
- Theoretical guarantees on quality

Roadmap

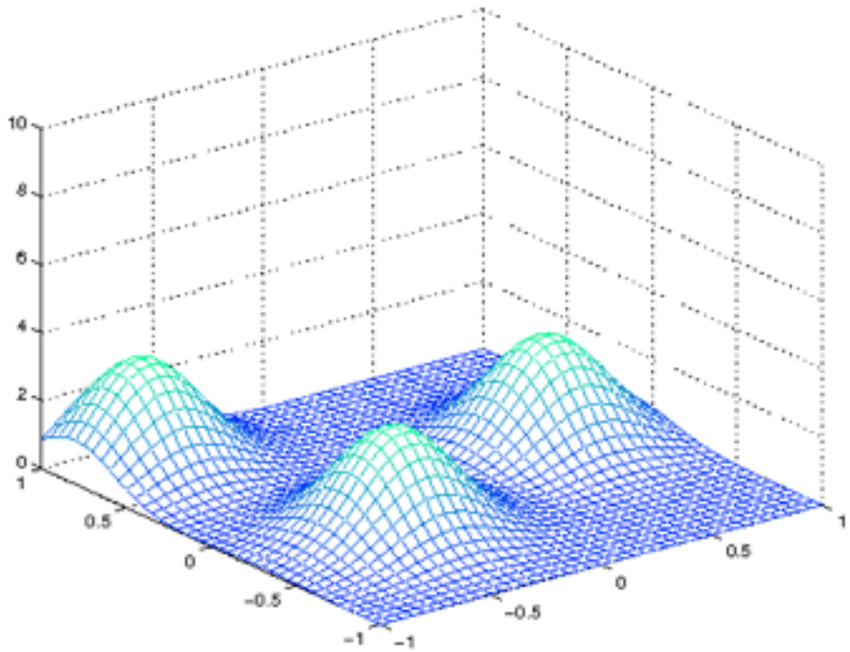
- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes)
  - Uncertainties
- Robustness trade-offs
  - Local robustness quantification
- Theoretical guarantees on quality

Variational Bayes

Variational Bayes

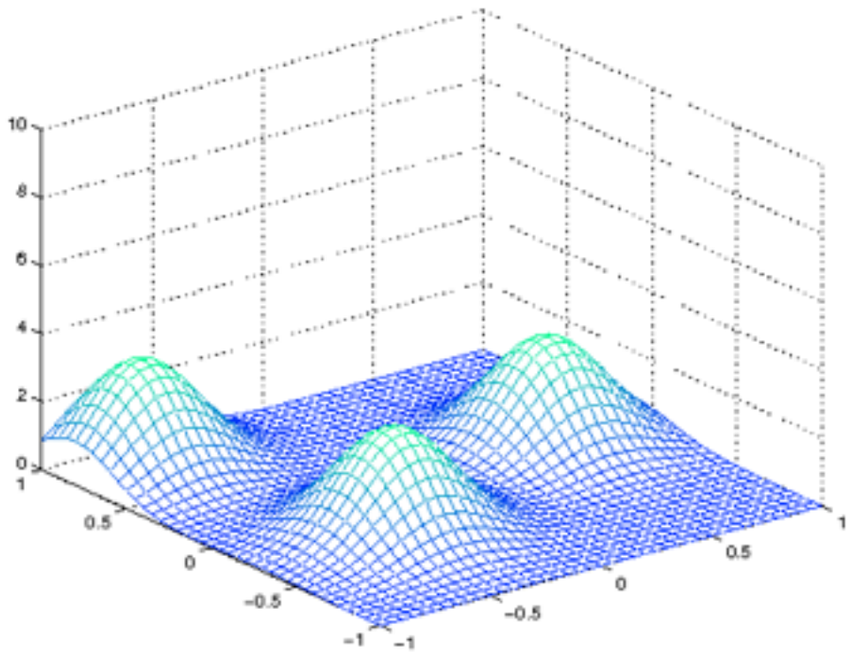
- Variational Bayes (VB)

Variational Bayes



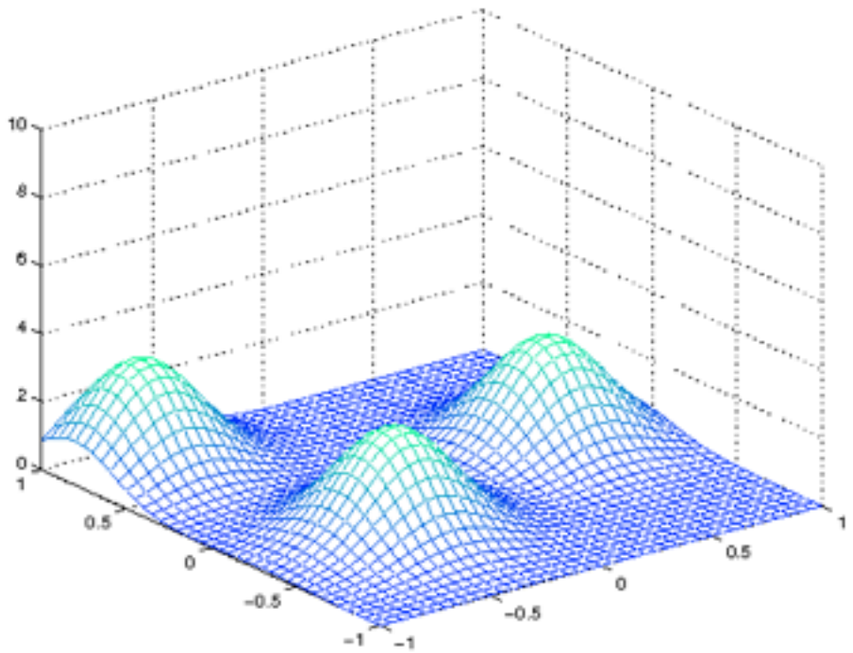
- Variational Bayes (VB)

Variational Bayes

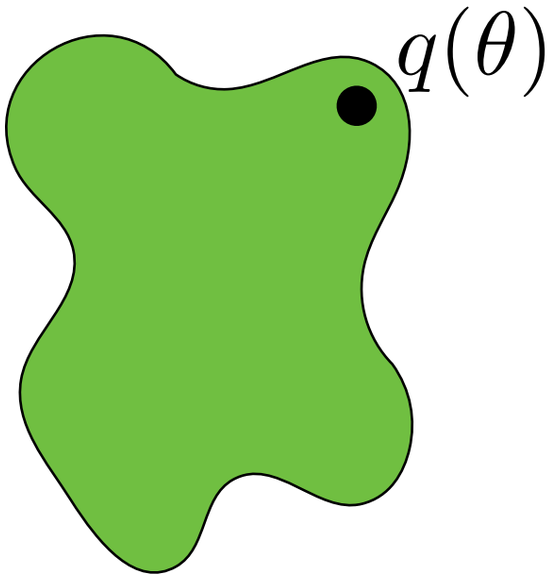


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$

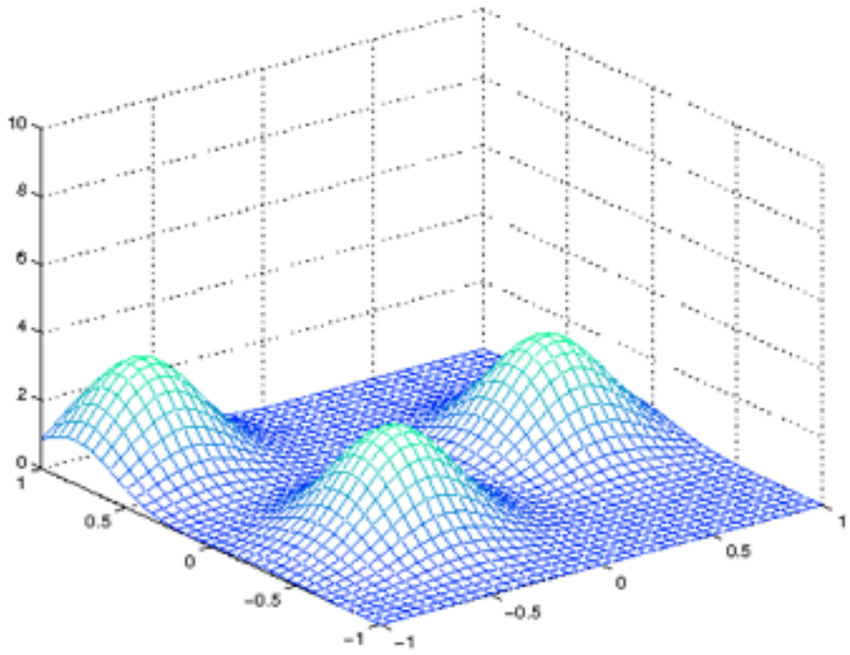
Variational Bayes



- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$

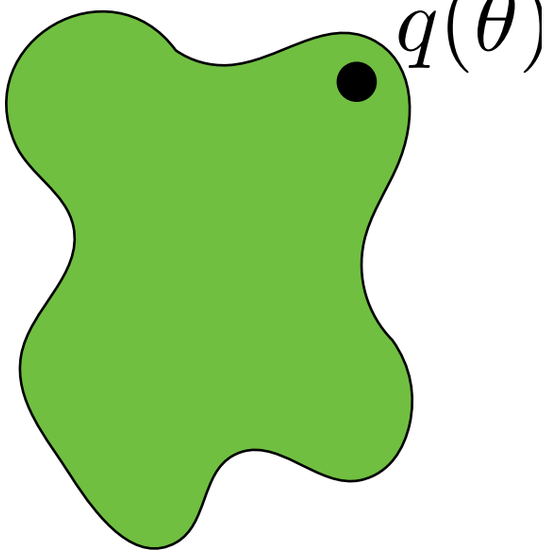


Variational Bayes

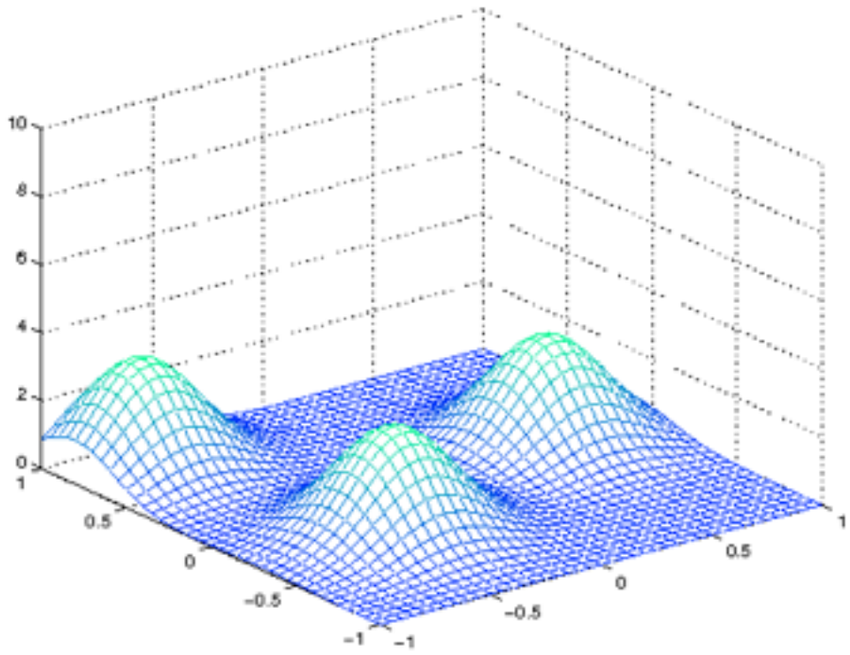


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$

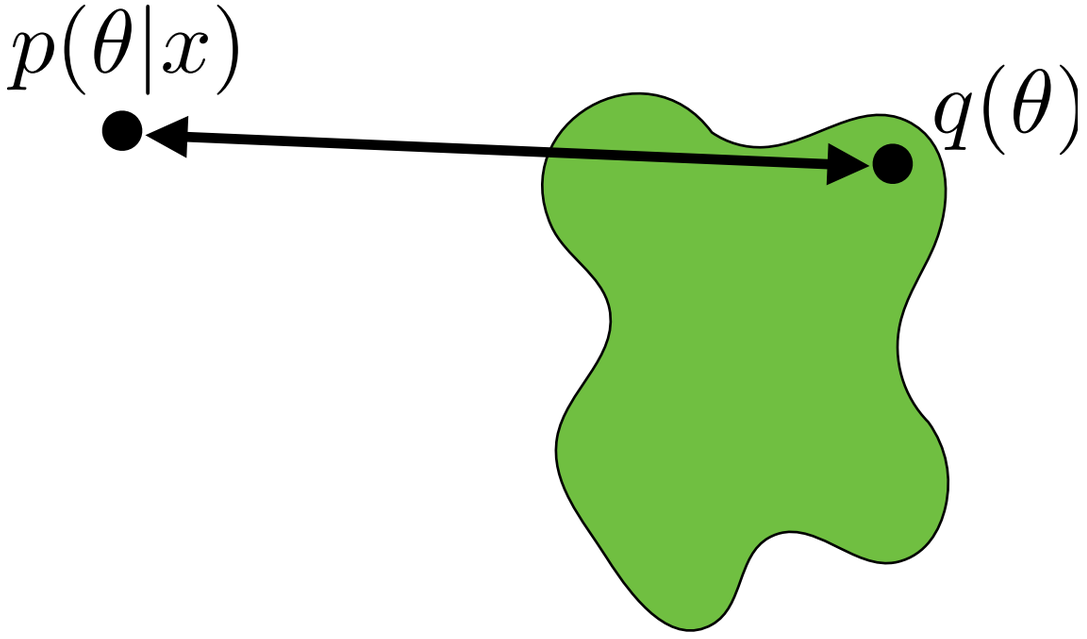
$p(\theta|x)$



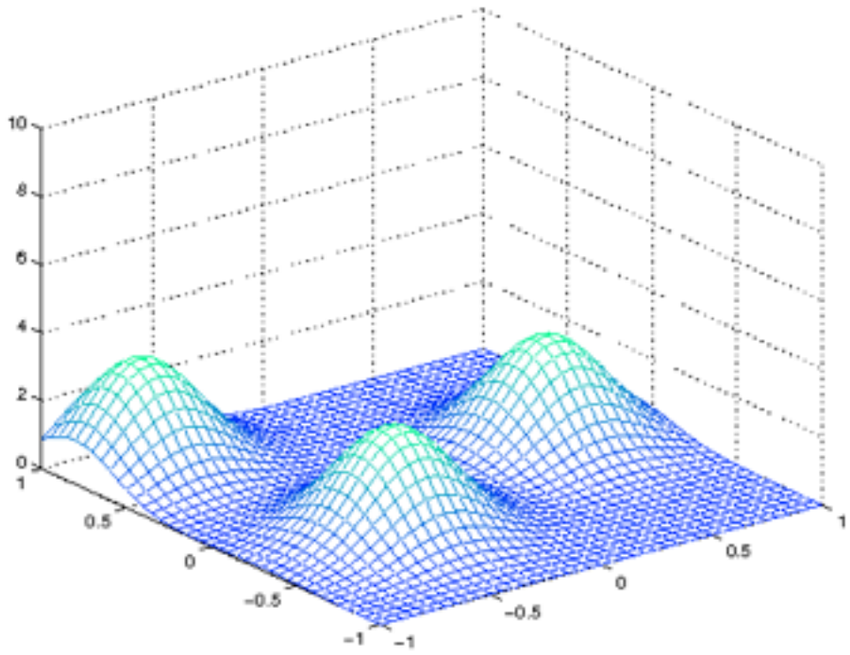
Variational Bayes



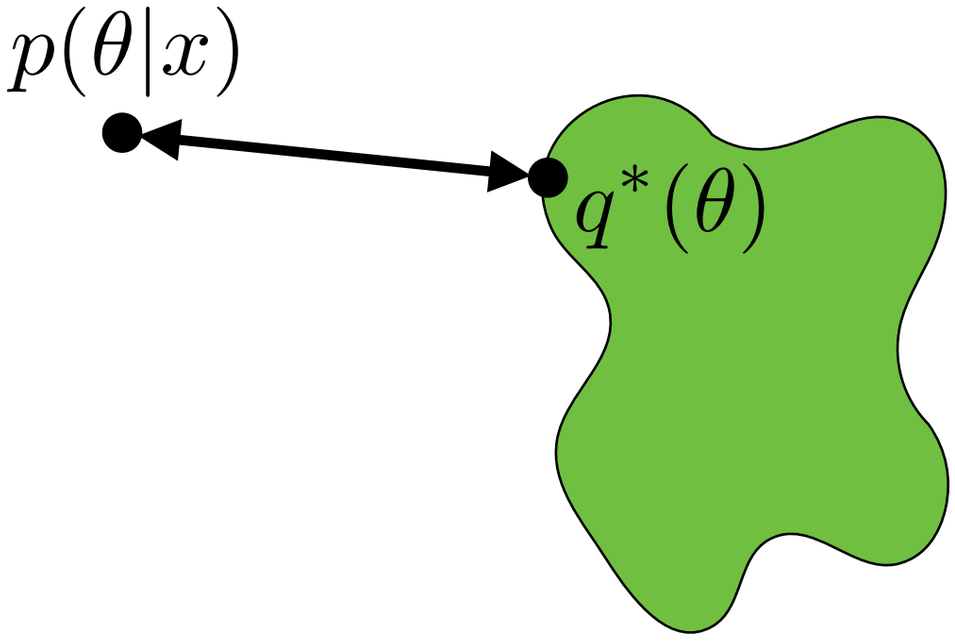
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$



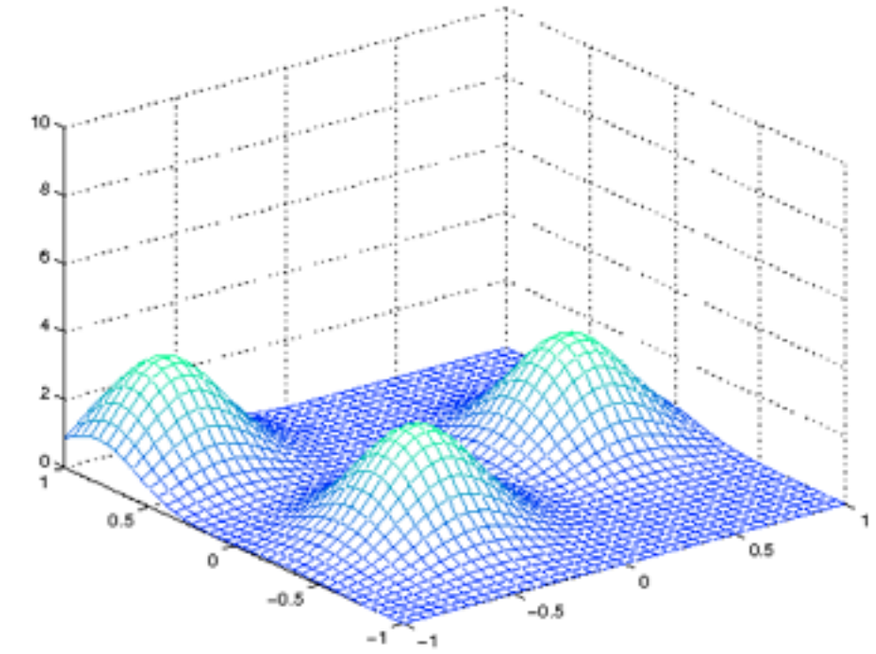
Variational Bayes

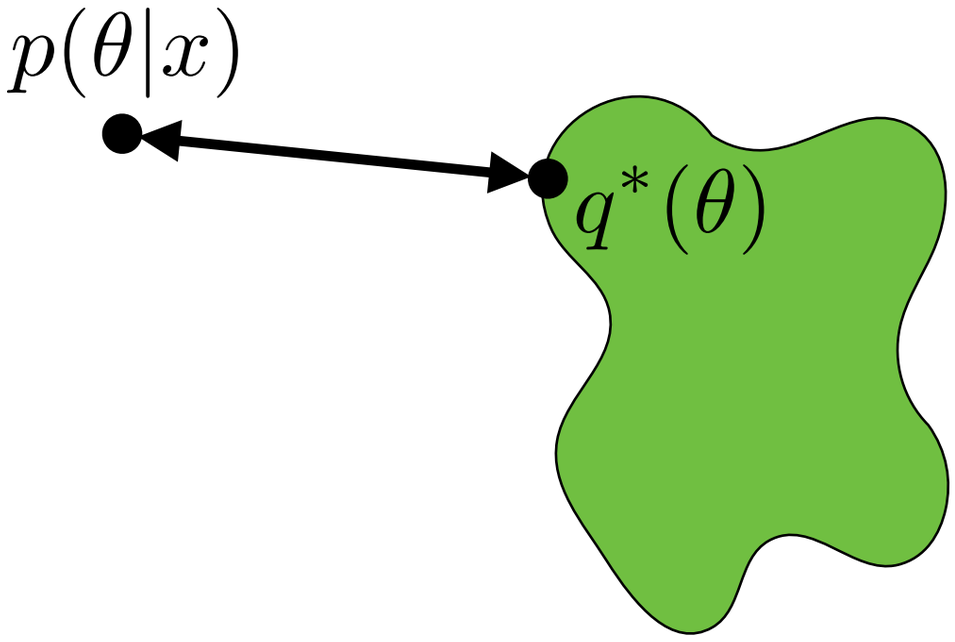


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$



Variational Bayes

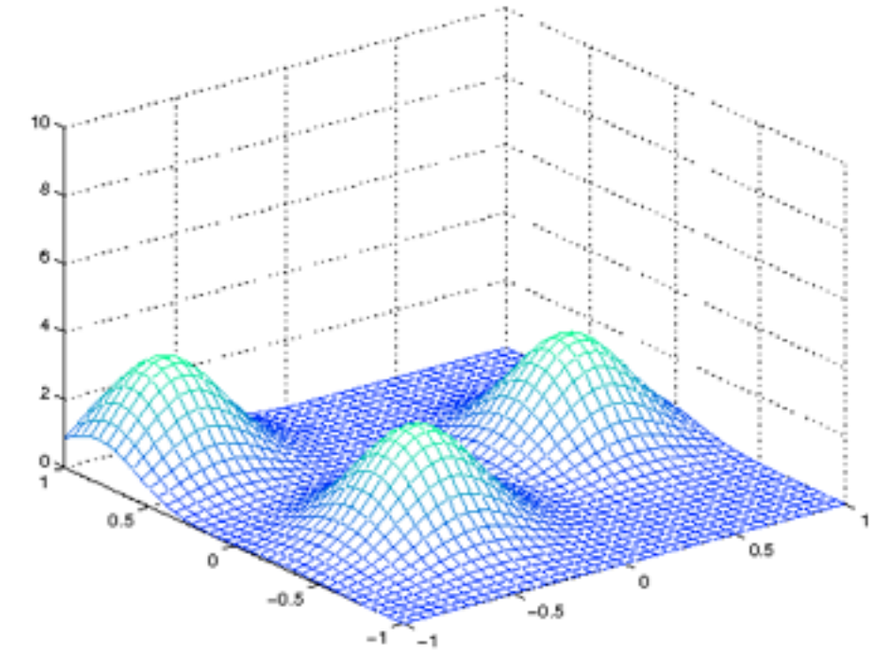


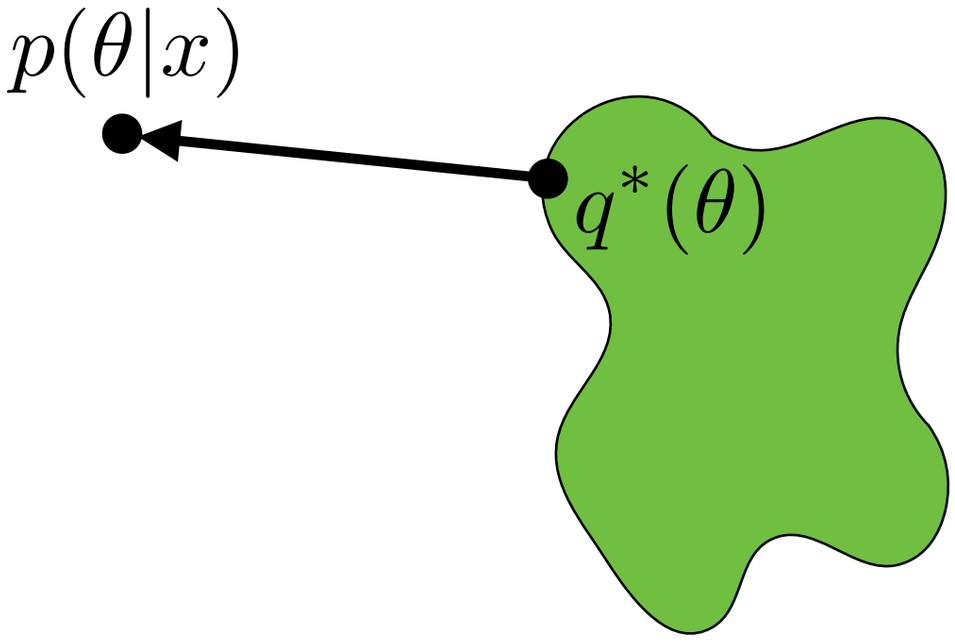


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Leibler (KL) divergence:

$$KL(q\|p(\cdot|x))$$

Variational Bayes

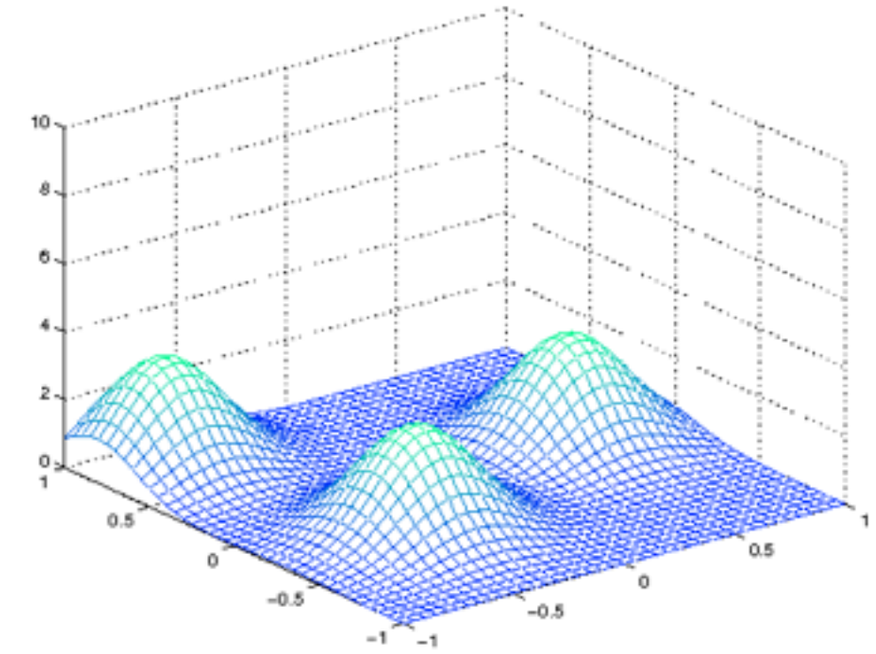


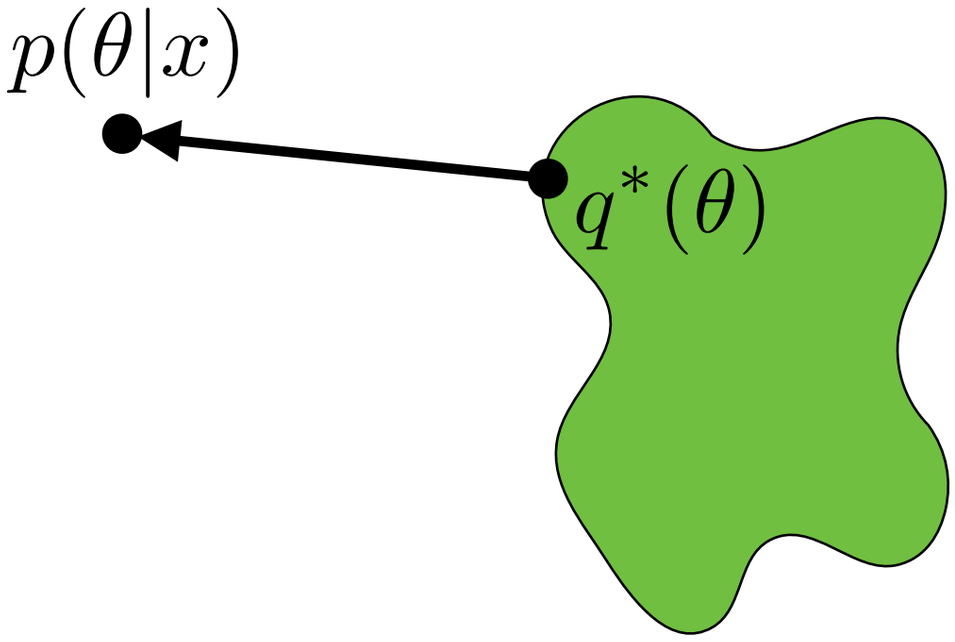


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Leibler (KL) divergence:

$$KL(q\|p(\cdot|x))$$

Variational Bayes



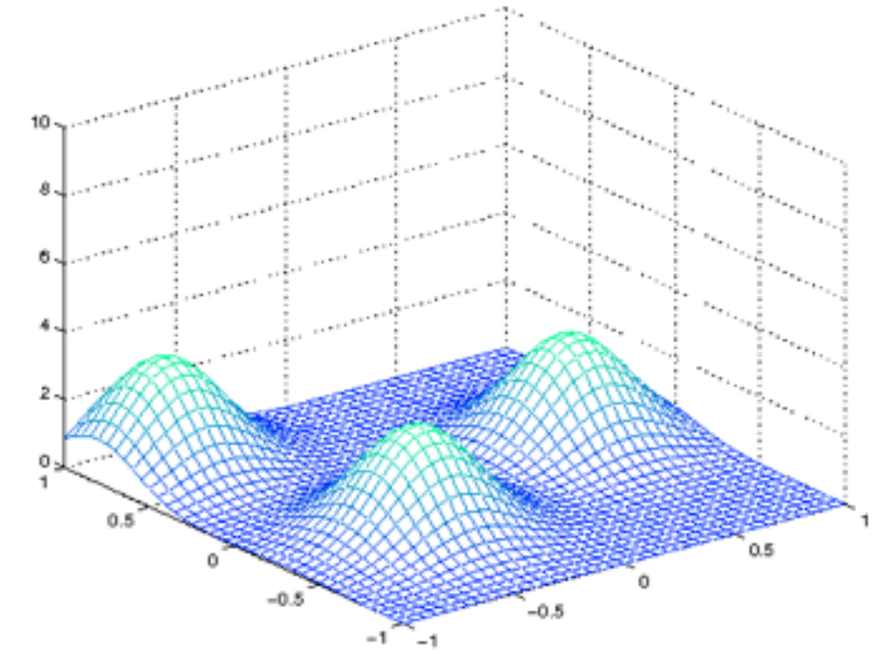


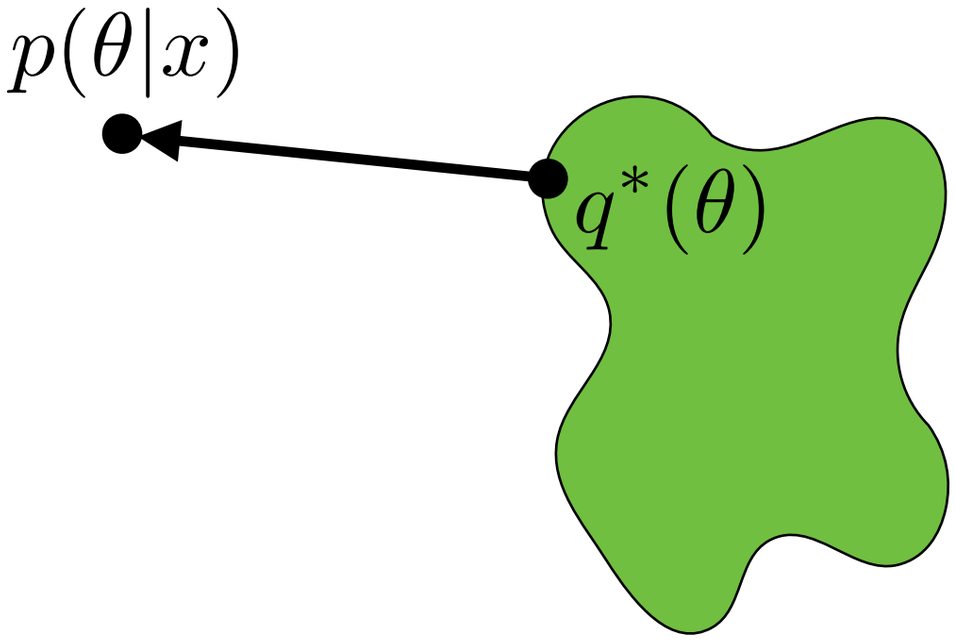
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Leibler (KL) divergence:

$$KL(q\|p(\cdot|x))$$

- VB practical success

Variational Bayes



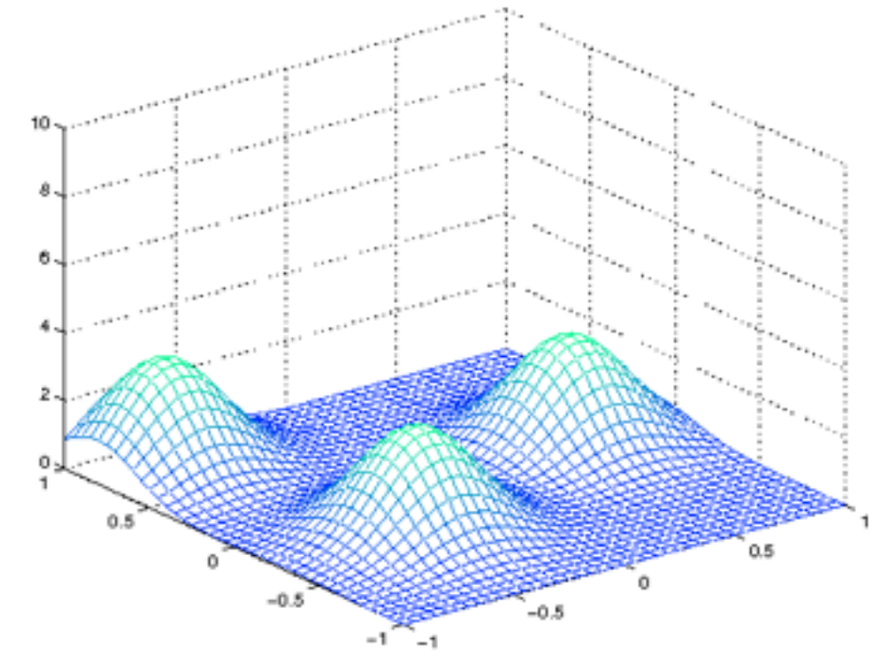


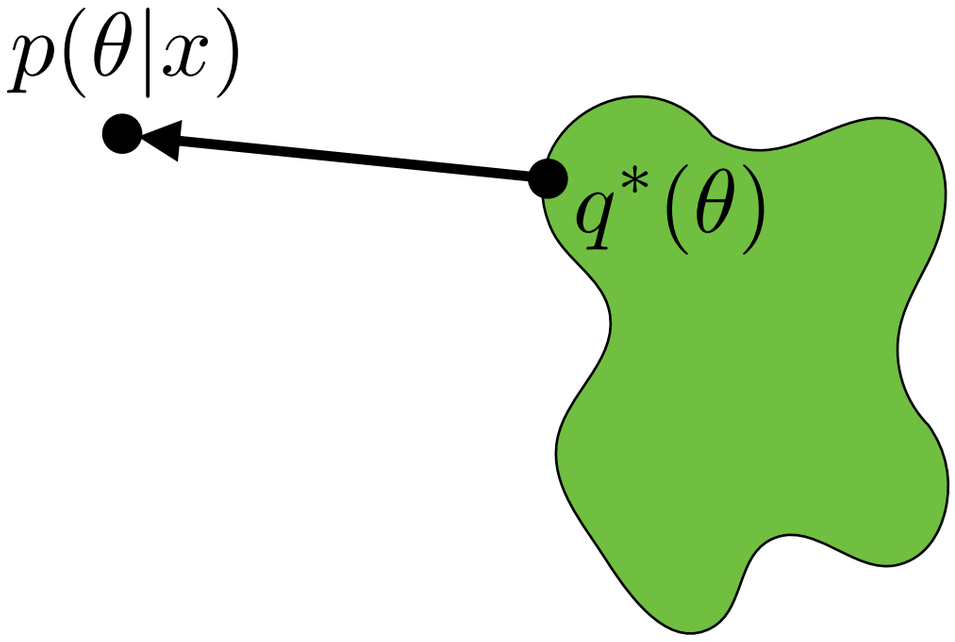
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Leibler (KL) divergence:

$$KL(q\|p(\cdot|x))$$

- VB practical success
  - point estimates and prediction

Variational Bayes





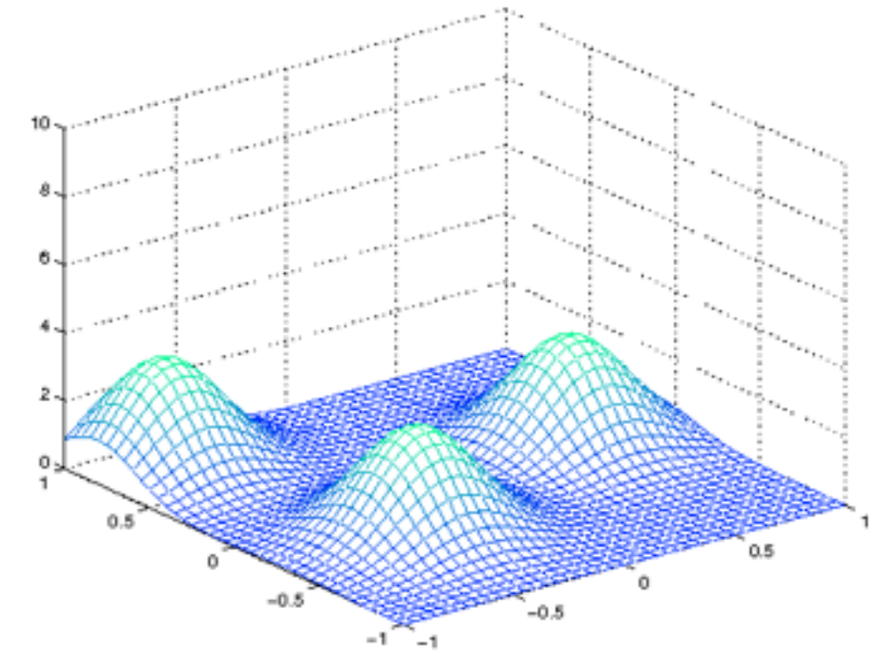
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Leibler (KL) divergence:

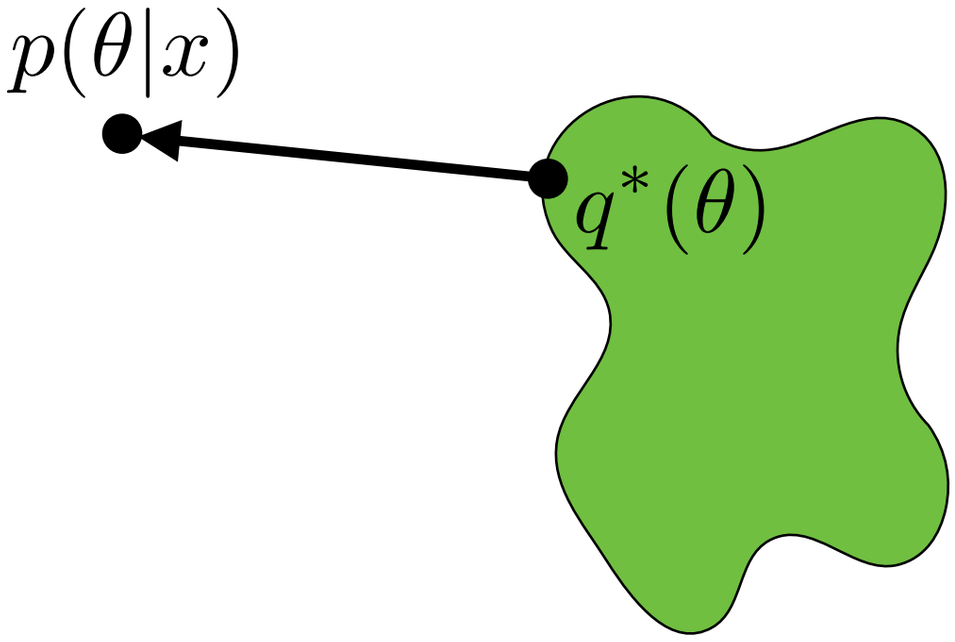
$$KL(q\|p(\cdot|x))$$

- VB practical success
  - point estimates and prediction
  - fast

[Broderick, Boyd, Wibisono, Wilson, Jordan 2013]

Variational Bayes





- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Leibler (KL) divergence:

$$KL(q\|p(\cdot|x))$$

- VB practical success
  - point estimates and prediction
  - fast, streaming, distributed

[Broderick, Boyd, Wibisono, Wilson, Jordan 2013]

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes)
  - Uncertainties
- Robustness trade-offs
  - Local robustness quantification
- Theoretical guarantees on quality

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes, variational Bayes)
  - Uncertainties
- Robustness trade-offs
  - Local robustness quantification
- Theoretical guarantees on quality

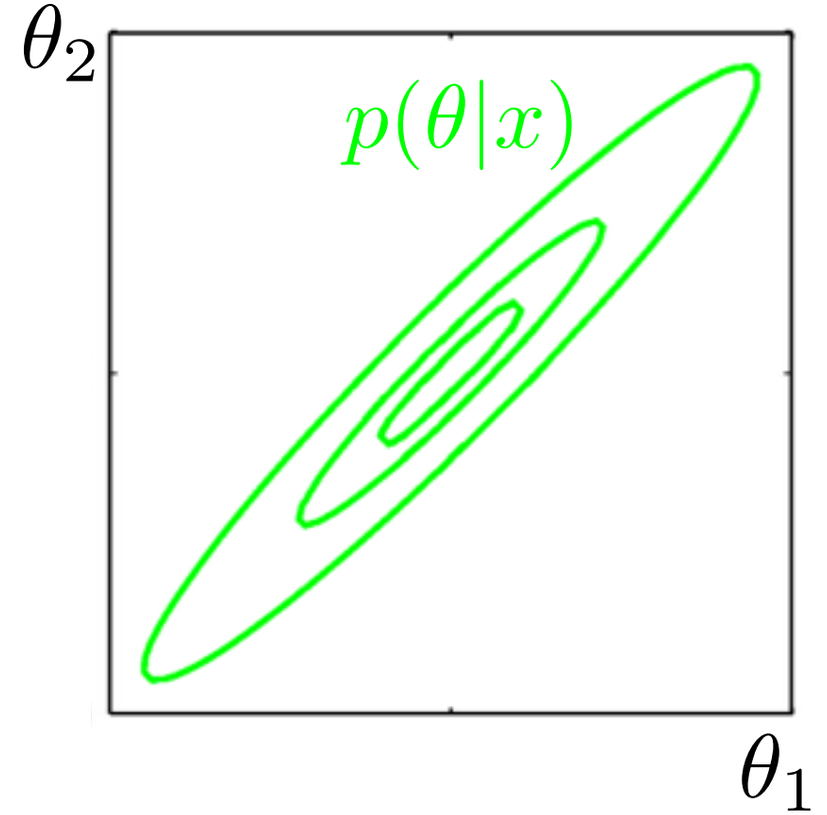
What about uncertainty?

What about uncertainty?

- Variational Bayes

What about uncertainty?

- Variational Bayes

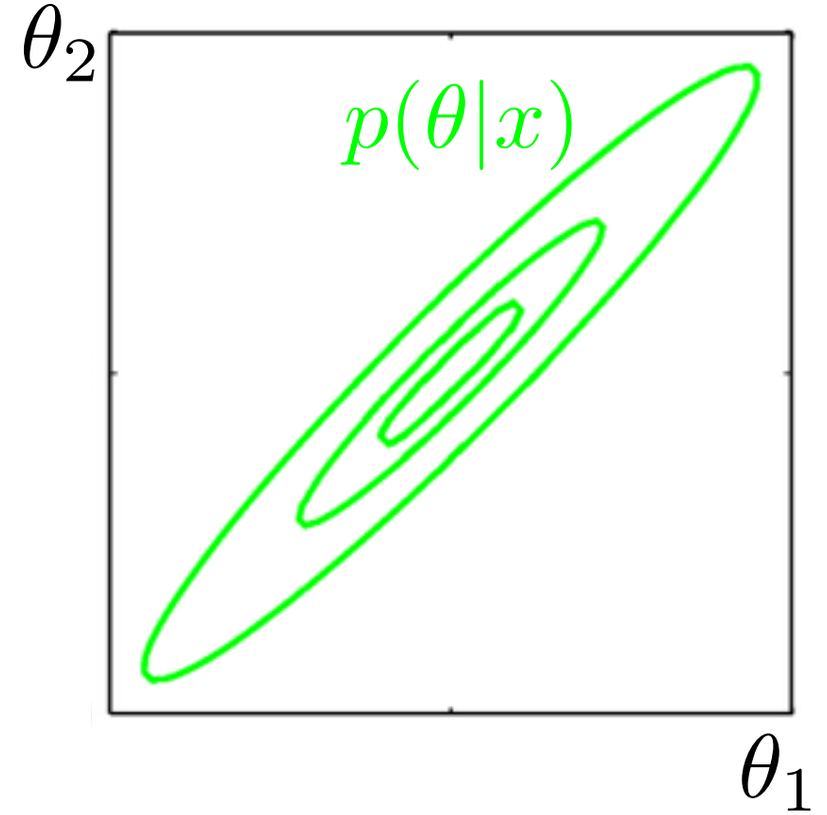


[Bishop 2006]

What about uncertainty?

- Variational Bayes

$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$



[Bishop 2006]

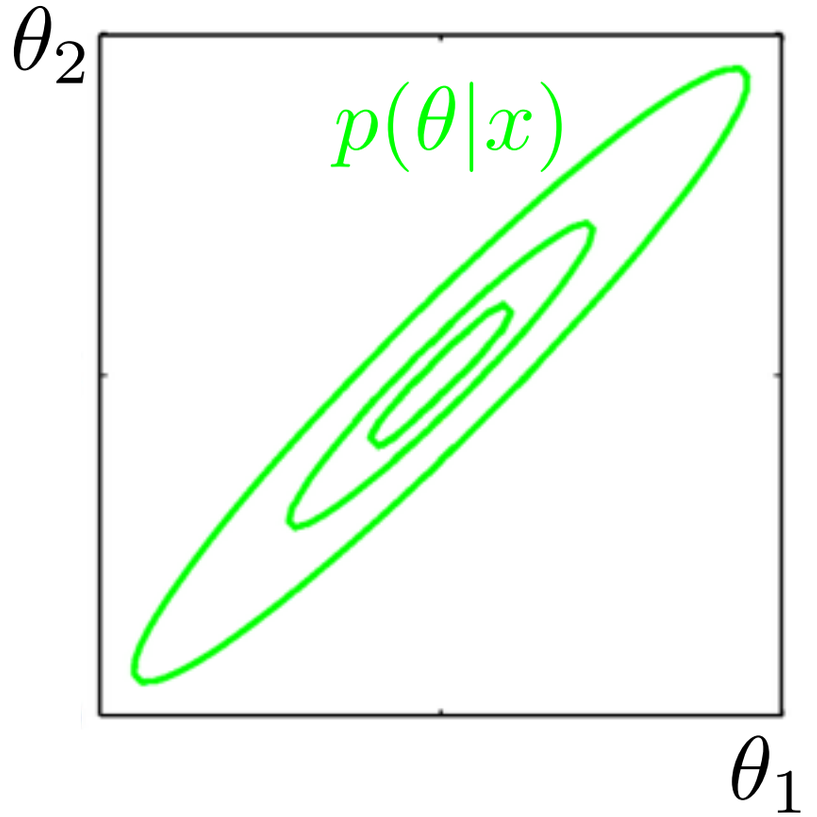
What about uncertainty?

- Variational Bayes

$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- Mean-field variational Bayes (MFVB)

$$q(\theta) = \prod_{j=1}^{J} q(\theta_j)$$



[Bishop 2006]

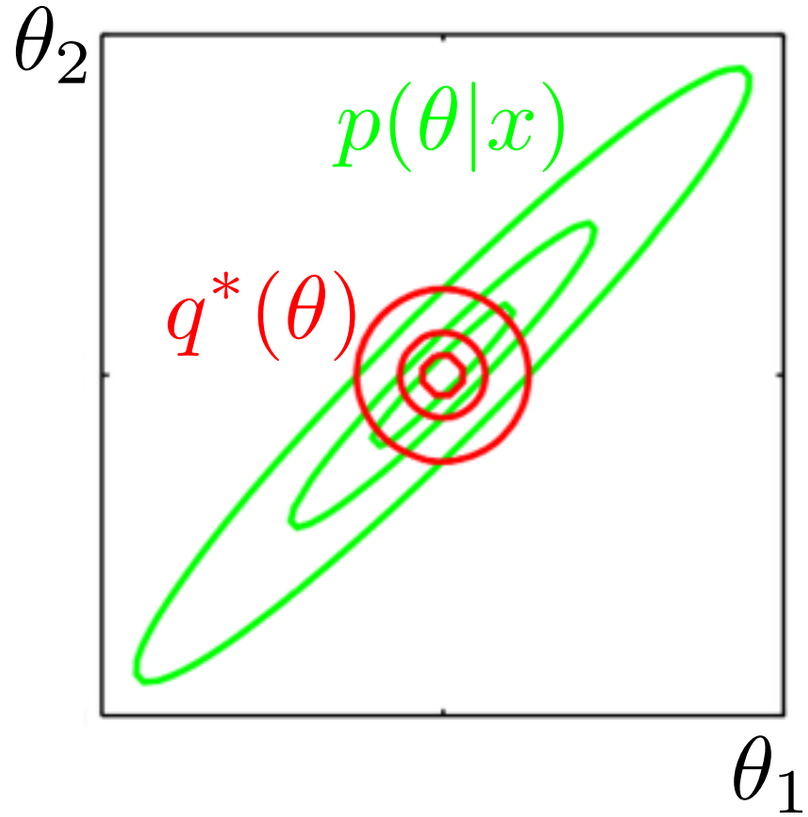
What about uncertainty?

- Variational Bayes

$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- Mean-field variational Bayes (MFVB)

$$q(\theta) = \prod_{j=1}^{J} q(\theta_j)$$



[Bishop 2006]

What about uncertainty?

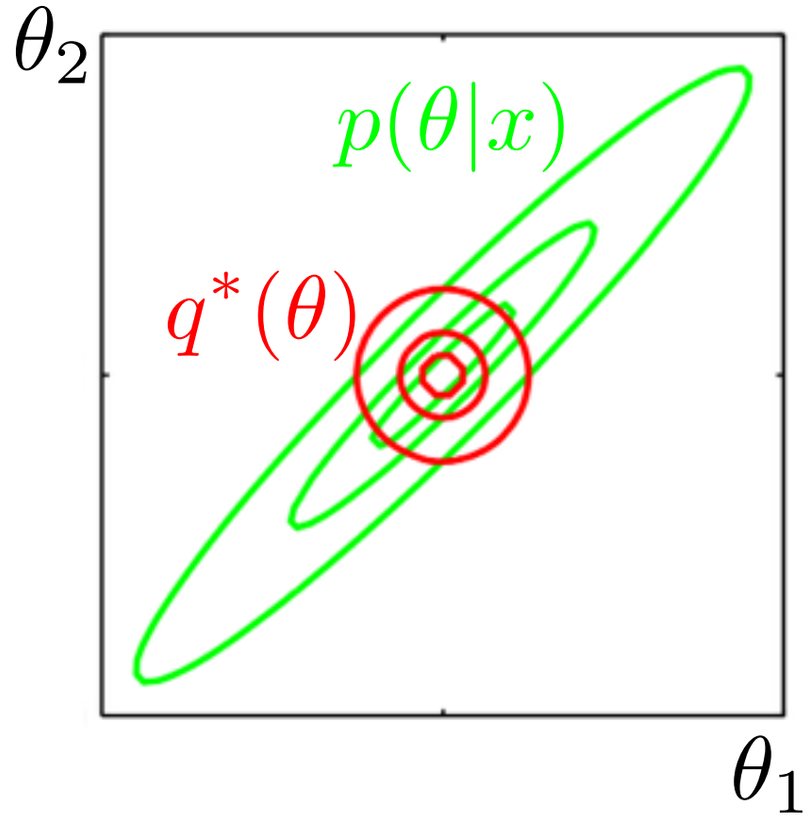
- Variational Bayes

$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- Mean-field variational Bayes (MFVB)

$$q(\theta) = \prod_{j=1}^{J} q(\theta_j)$$

- Underestimates variance (sometimes severely)



[Bishop 2006]

What about uncertainty?

- Variational Bayes

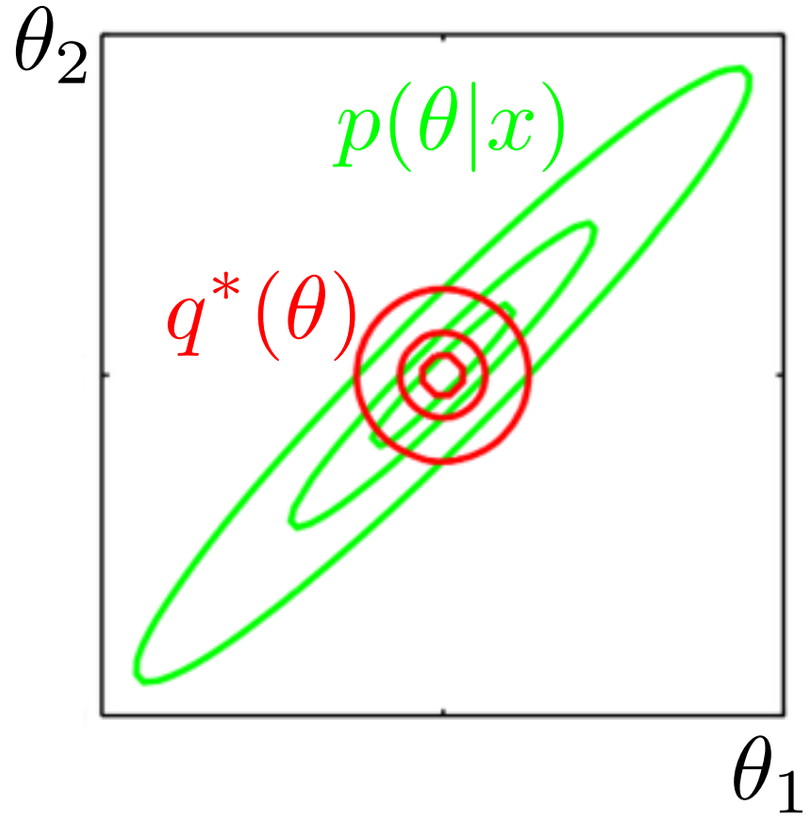
$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- Mean-field variational Bayes (MFVB)

$$q(\theta) = \prod_{j=1}^{J} q(\theta_j)$$

- Underestimates variance (sometimes severely)

- No covariance estimates



[Bishop 2006]

What about uncertainty?

- Variational Bayes

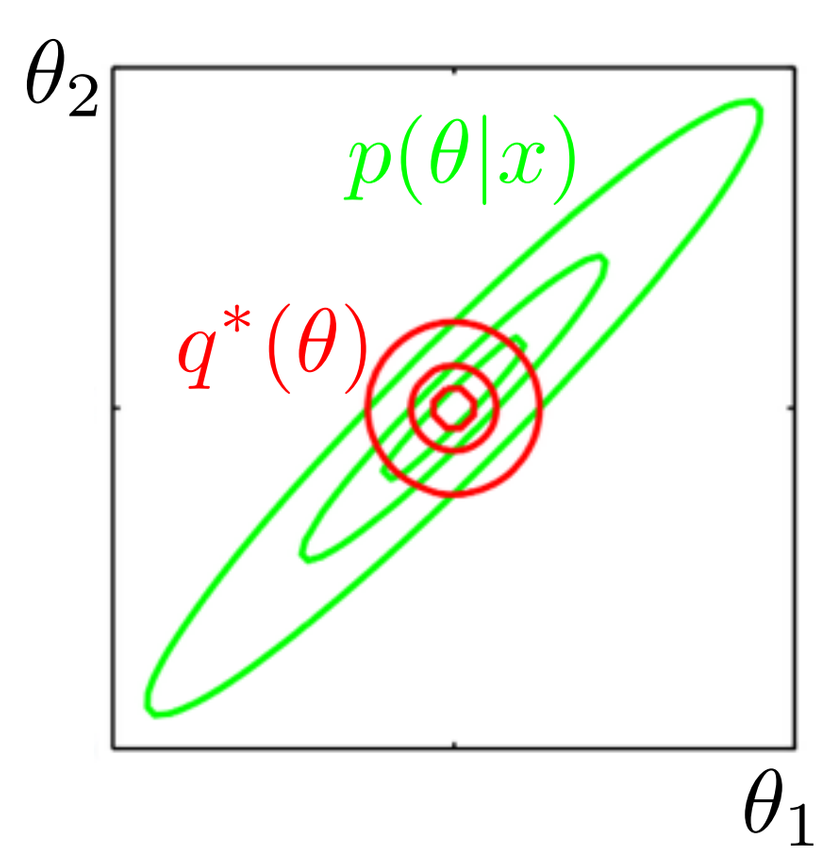
$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- Mean-field variational Bayes (MFVB)

$$q(\theta) = \prod_{j=1}^{J} q(\theta_j)$$

- Underestimates variance (sometimes severely)

- No covariance estimates



[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011]

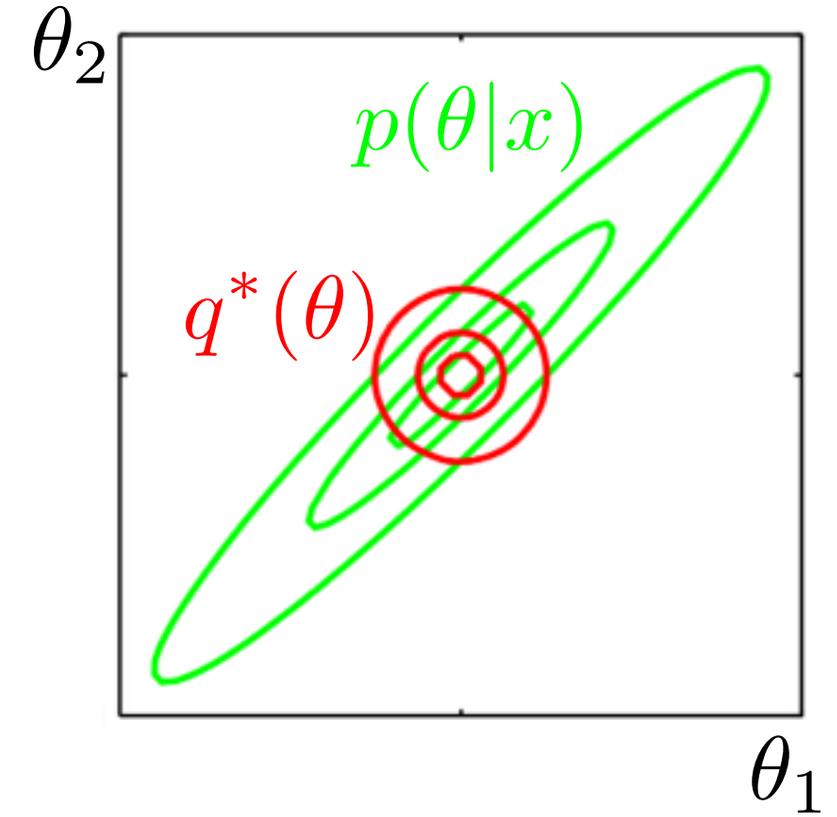
What about uncertainty?

- Variational Bayes

$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- Mean-field variational Bayes (MFVB)

$$q(\theta) = \prod_{j=1}^{J} q(\theta_j)$$



- Underestimates variance (sometimes severely)

- No covariance estimates

[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011]
[Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2015]

Linear response

Linear response

- Cumulant-generating function

Linear response

- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta}$$

Linear response

- Cumulant-generating function

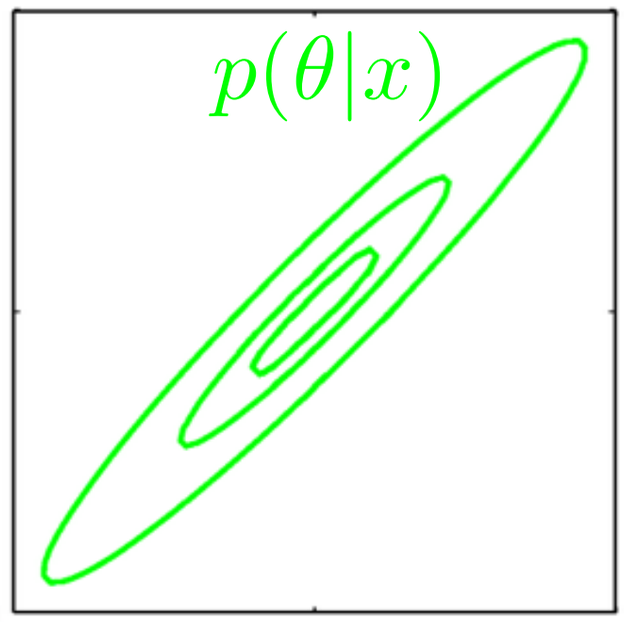
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

Linear response

- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance



[Bishop 2006]

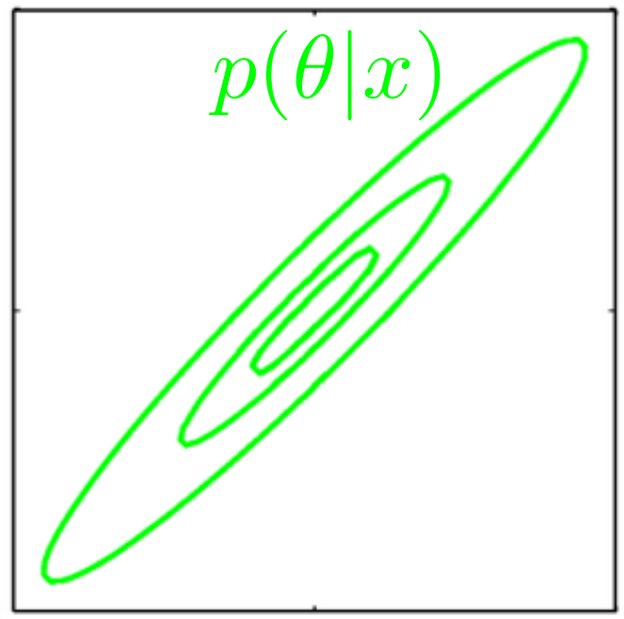
Linear response

- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance

$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0}$$



[Bishop 2006]

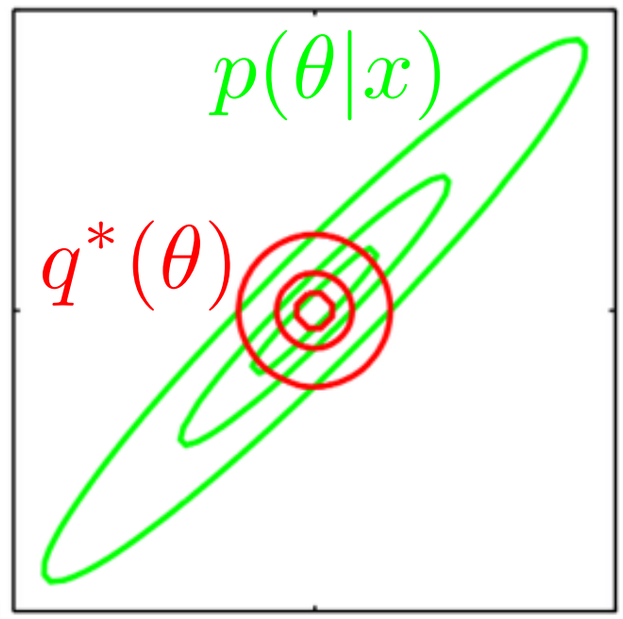
Linear response

- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0}$$



[Bishop 2006]

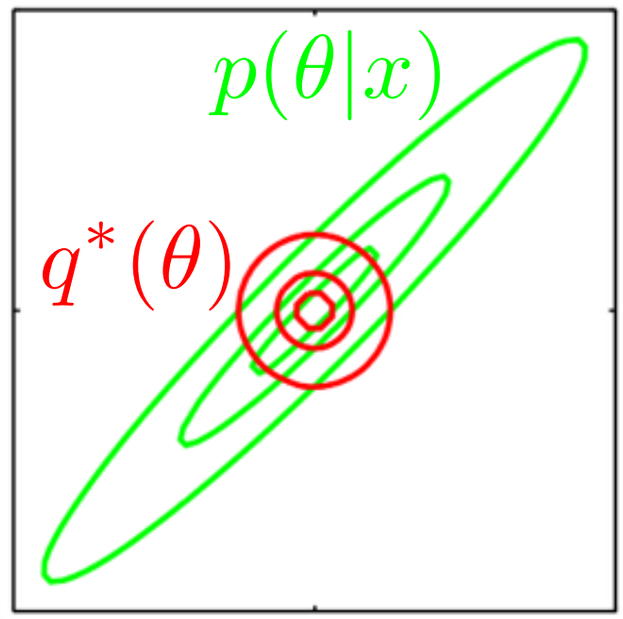
Linear response

- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$



[Bishop 2006]

Linear response

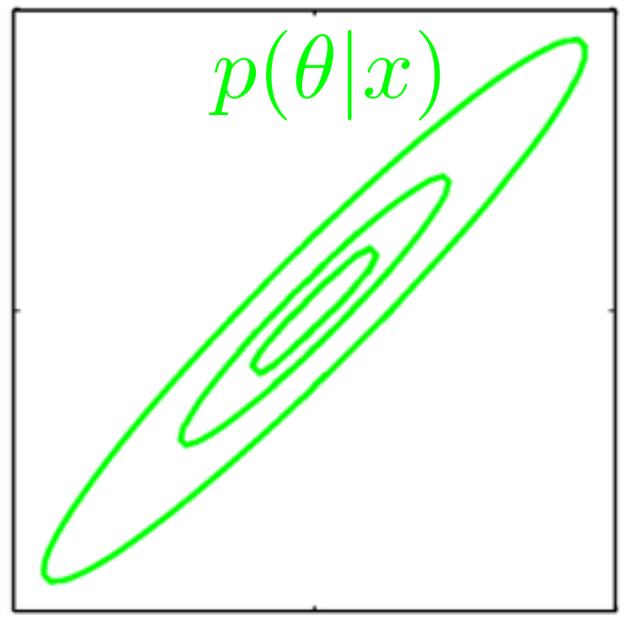
- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"



[Bishop 2006]

Linear response

- Cumulant-generating function

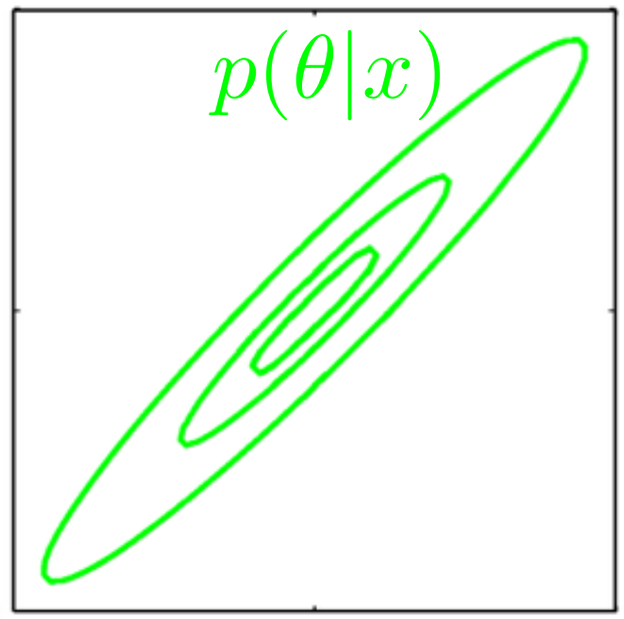
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$$\log p(\theta|x)$$



[Bishop 2006]

Linear response

- Cumulant-generating function

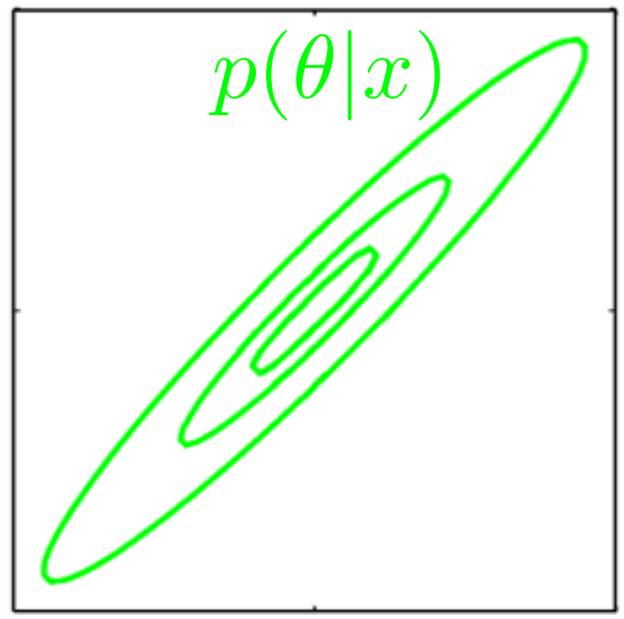
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$$\log p(\theta|x) + t^T \theta$$



[Bishop 2006]

Linear response

- Cumulant-generating function

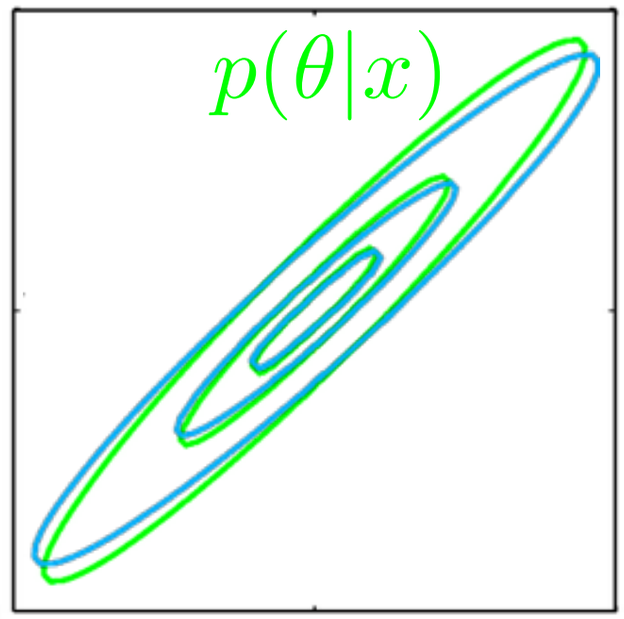
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$$\log p(\theta|x) + t^T \theta$$



[Bishop 2006]

Linear response

- Cumulant-generating function

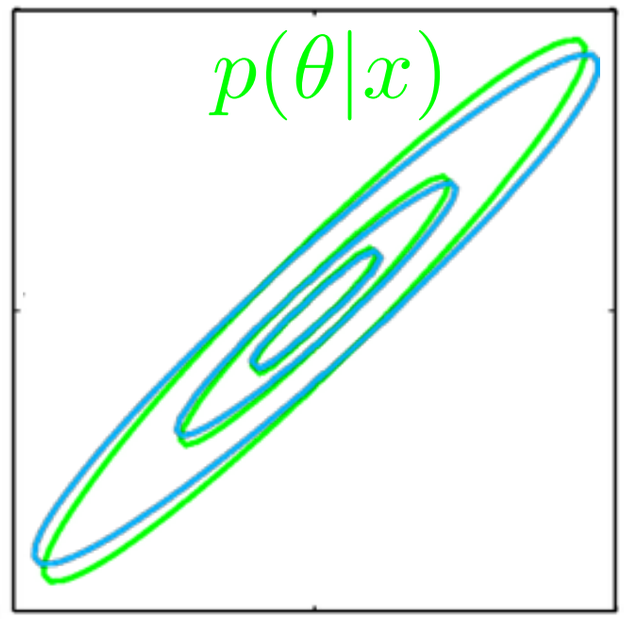
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$$\log p_t(\theta) := \log p(\theta|x) + t^T \theta$$



[Bishop 2006]

Linear response

- Cumulant-generating function

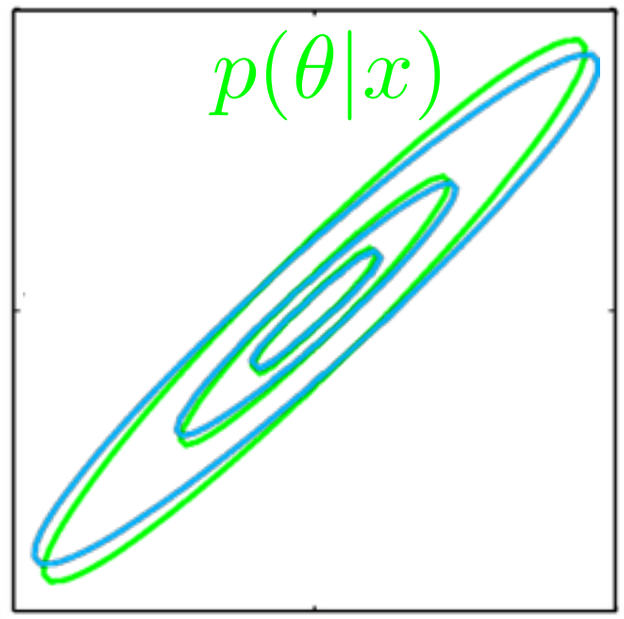
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$$\log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t)$$



[Bishop 2006]

Linear response

- Cumulant-generating function

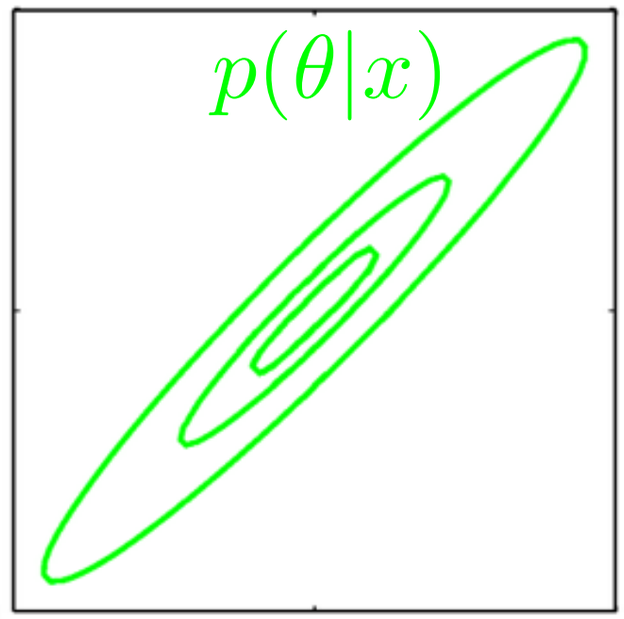
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$$\log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t)$$



[Bishop 2006]

Linear response

- Cumulant-generating function

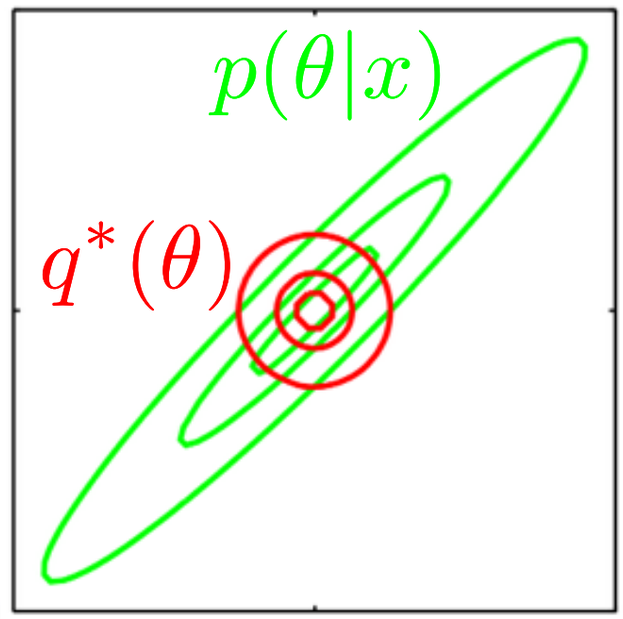
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t)$, MFVB q_t^*



[Bishop 2006]

Linear response

- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

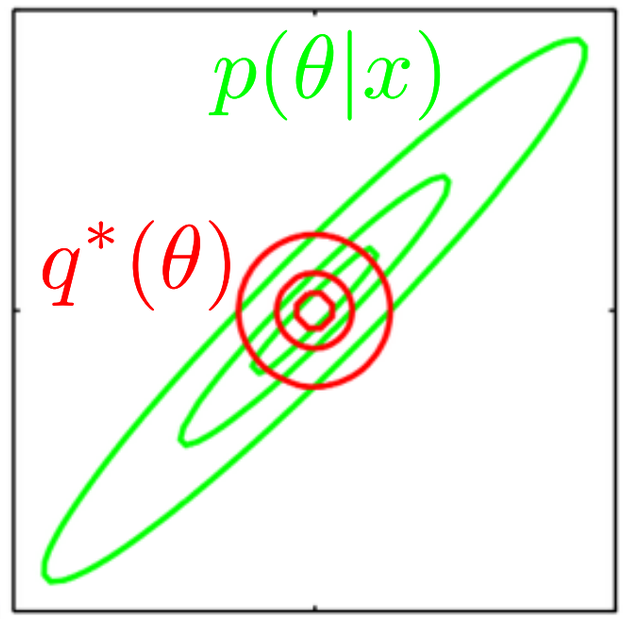
- True posterior covariance vs MFVB covariance

$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t)$, MFVB q_t^*

- The LRVB approximation



[Bishop 2006]

Linear response

- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

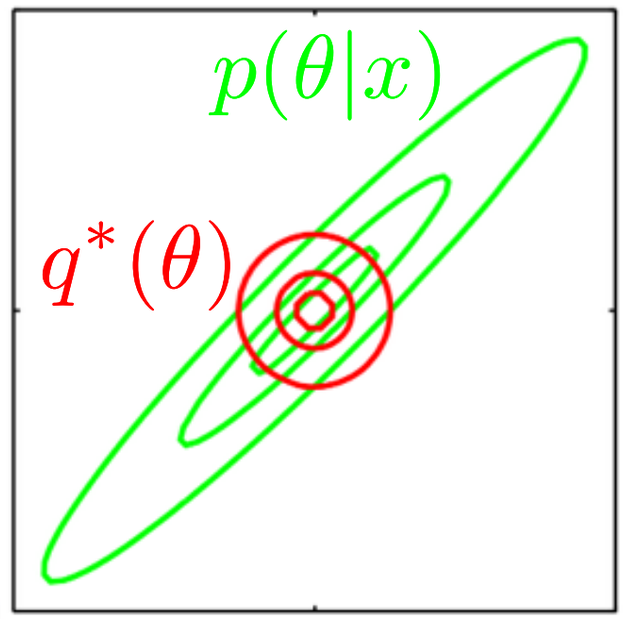
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t)$, MFVB q_t^*

- The LRVB approximation

$$\Sigma = \frac{d}{dt^T} \left[\frac{d}{dt} C_{p(\cdot|x)}(t) \right] \Bigg|_{t=0}$$



[Bishop 2006]

Linear response

- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

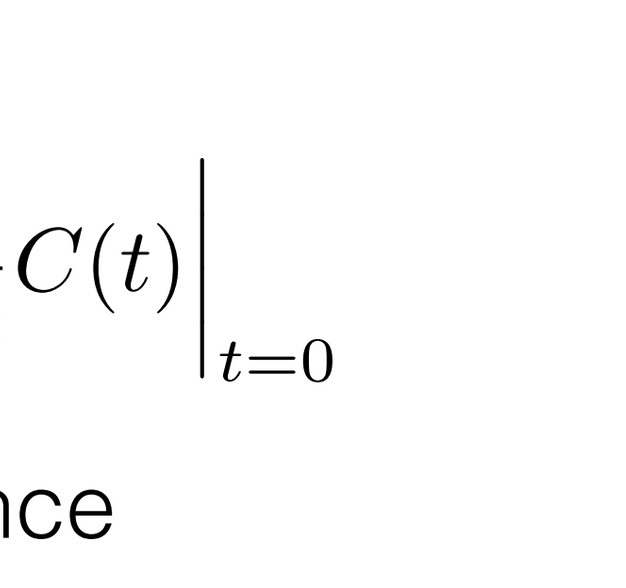
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t)$, MFVB q_t^*

- The LRVB approximation

$$\Sigma = \left. \frac{d}{dt^T} \mathbb{E}_{p_t} \theta \right|_{t=0}$$



[Bishop 2006]

Linear response

- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

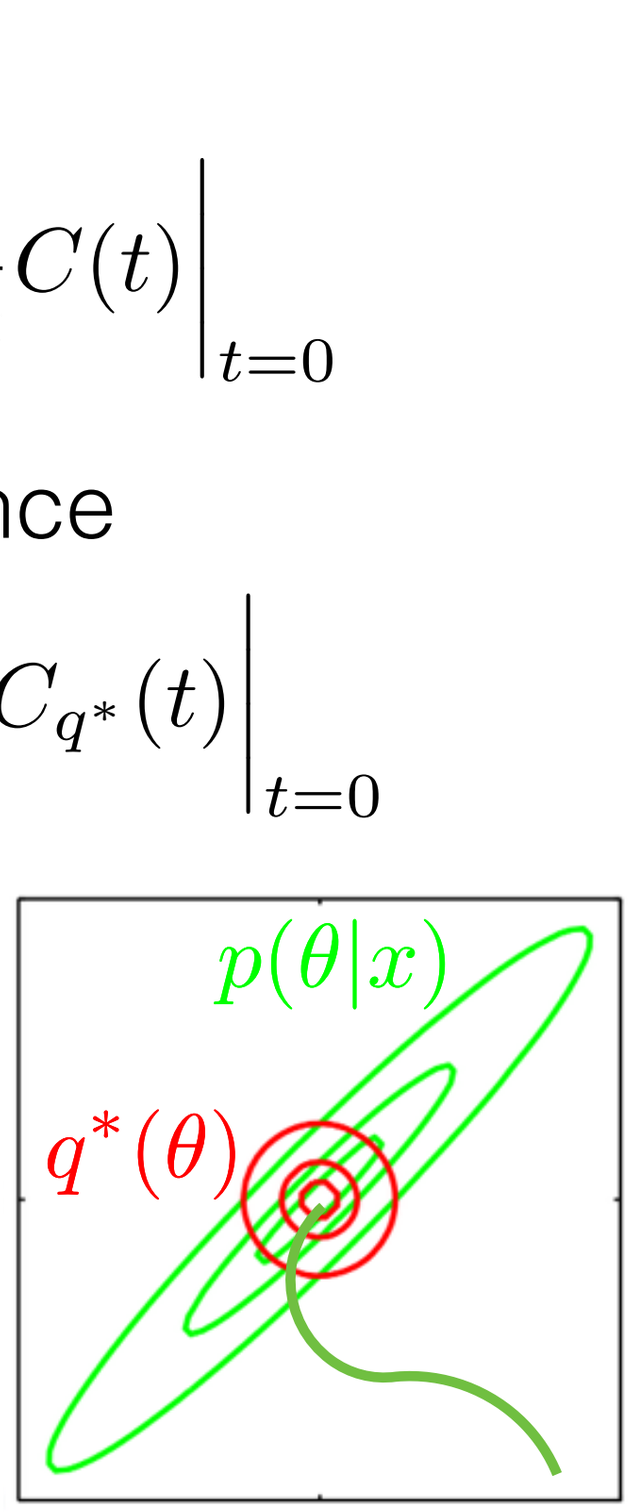
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t)$, MFVB q_t^*

- The LRVB approximation

$$\Sigma = \left. \frac{d}{dt^T} \mathbb{E}_{p_t} \theta \right|_{t=0}$$



[Bishop 2006]

Linear response

- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

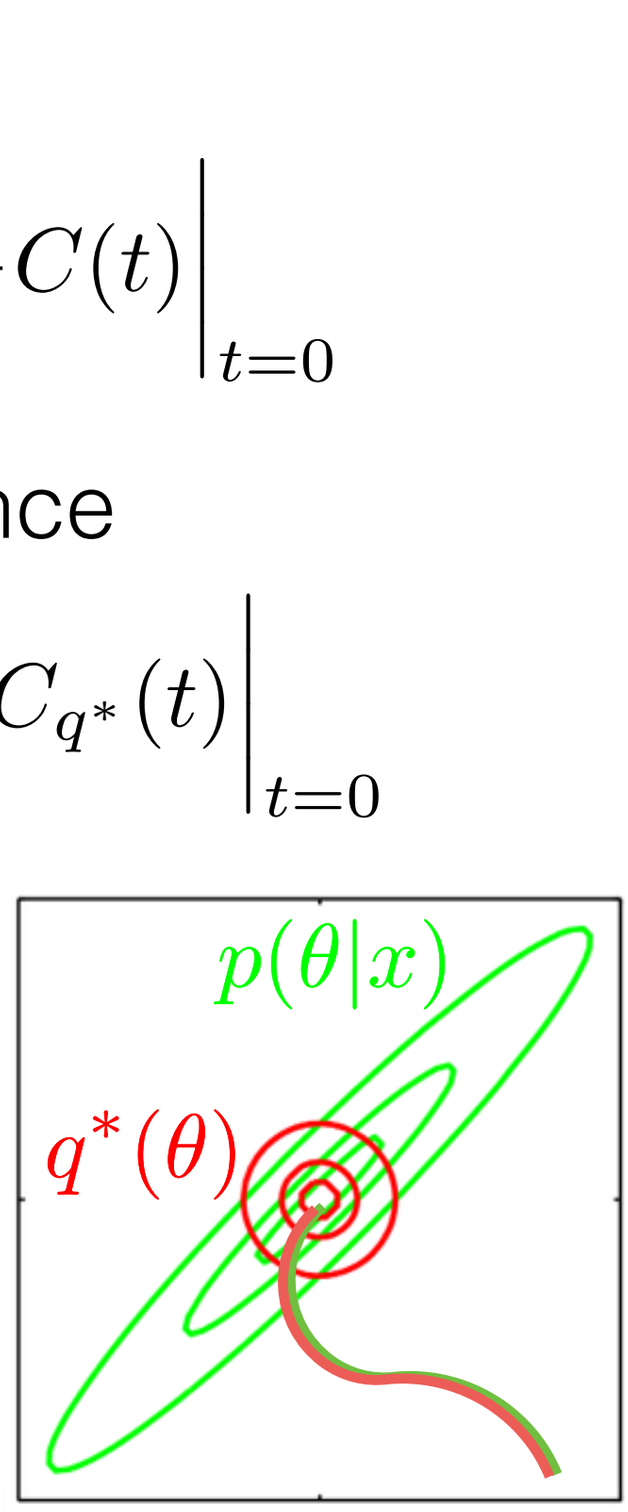
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t)$, MFVB q_t^*

- The LRVB approximation

$$\Sigma = \left. \frac{d}{dt^T} \mathbb{E}_{p_t} \theta \right|_{t=0}$$



[Bishop 2006]

Linear response

- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

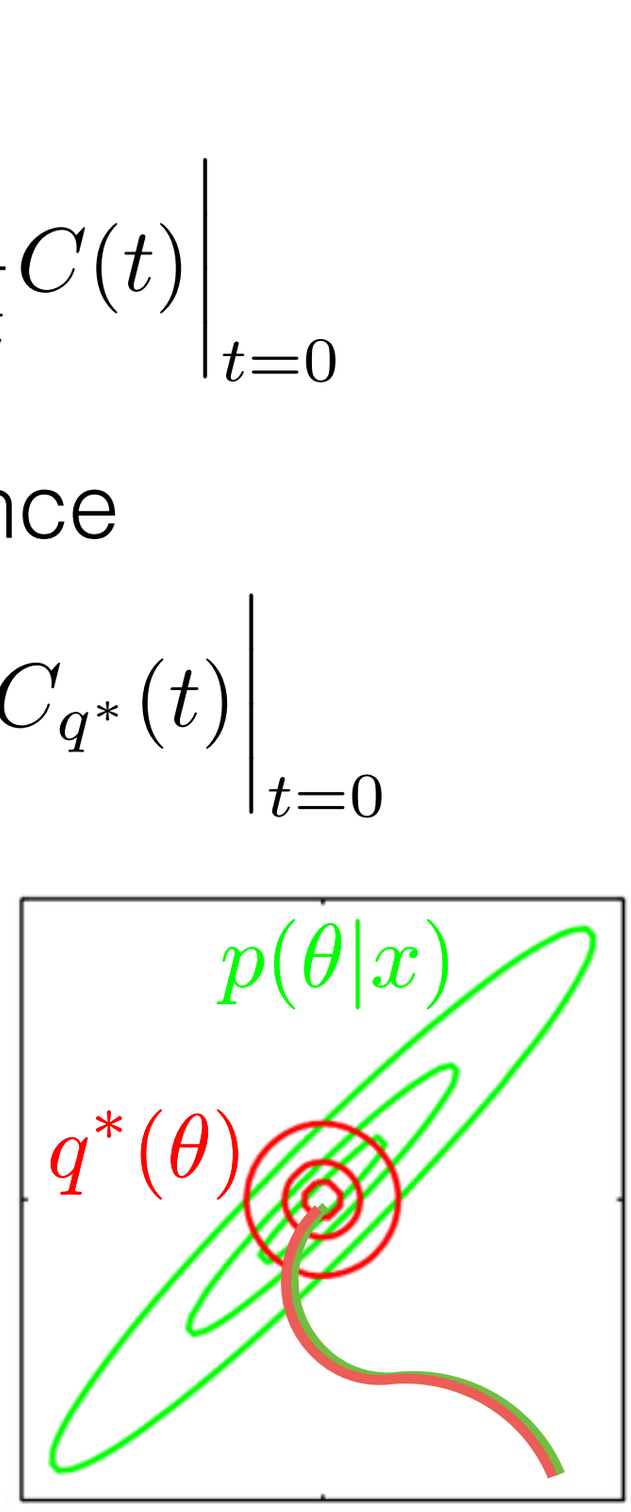
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t)$, MFVB q_t^*

- The LRVB approximation

$$\Sigma = \left. \frac{d}{dt^T} \mathbb{E}_{p_t} \theta \right|_{t=0} \approx \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0}$$



[Bishop 2006]

Linear response

- Cumulant-generating function

$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

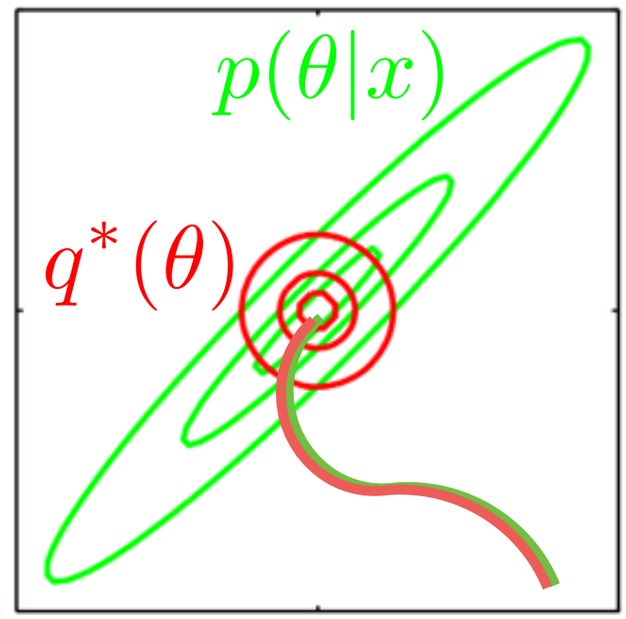
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t)$, MFVB q_t^*

- The LRVB approximation

$$\Sigma = \left. \frac{d}{dt^T} \mathbb{E}_{p_t} \theta \right|_{t=0} \approx \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0} =: \hat{\Sigma}$$



[Bishop 2006]

LRVB estimator

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0}$

LRVB estimator

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0}$
- Suppose q_t exponential family

LRVB estimator

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t

LRVB estimator

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t

$$\hat{\Sigma} =$$

LRVB estimator

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t

$$\hat{\Sigma} = \left(\left. \frac{\partial^2 KL}{\partial m \partial m^T} \right|_{m=m^*} \right)^{-1}$$

LRVB estimator

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t

$$\hat{\Sigma} = \left(\left. \frac{\partial^2 KL}{\partial m \partial m^T} \right|_{m=m^*} \right)^{-1} = (I - VH)^{-1} V$$

LRVB estimator

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t

$$\hat{\Sigma} = \left(\left. \frac{\partial^2 KL}{\partial m \partial m^T} \right|_{m=m^*} \right)^{-1} = (I - VH)^{-1} V$$

- Symmetric and positive definite at local min of KL

LRVB estimator

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t

$$\hat{\Sigma} = \left(\left. \frac{\partial^2 KL}{\partial m \partial m^T} \right|_{m=m^*} \right)^{-1} = (I - VH)^{-1} V$$

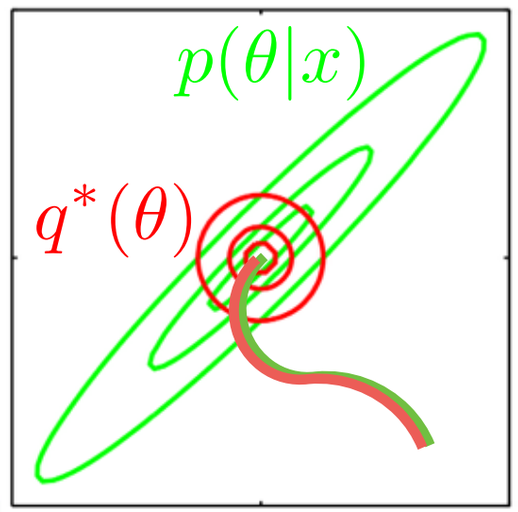
- Symmetric and positive definite at local min of KL
- The LRVB assumption: $\mathbb{E}_{p_t} \theta \approx \mathbb{E}_{q_t^*} \theta$

LRVB estimator

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t

$$\hat{\Sigma} = \left(\left. \frac{\partial^2 KL}{\partial m \partial m^T} \right|_{m=m^*} \right)^{-1} = (I - VH)^{-1} V$$

- Symmetric and positive definite at local min of KL
- The LRVB assumption: $\mathbb{E}_{p_t} \theta \approx \mathbb{E}_{q_t^*} \theta$



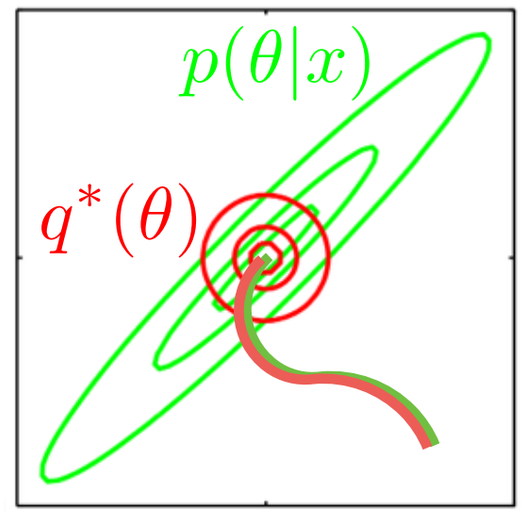
[Bishop 2006]

LRVB estimator

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t

$$\hat{\Sigma} = \left(\left. \frac{\partial^2 KL}{\partial m \partial m^T} \right|_{m=m^*} \right)^{-1} = (I - VH)^{-1} V$$

- Symmetric and positive definite at local min of KL
- The LRVB assumption: $\mathbb{E}_{p_t} \theta \approx \mathbb{E}_{q_t^*} \theta$
- LRVB estimate is exact when MFVB gives exact mean (e.g. multivariate normal)



[Bishop 2006]

Microcredit Experiment

Microcredit Experiment

- Simplified from Meager (2015)

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

profit → y_{kn}

Microcredit Experiment

- Simplified from Meager (2015)
- K microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\quad , \quad)$$

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k \qquad\qquad , \quad)$$

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \quad)$$

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

1 if microcredit

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \quad)$$

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

1 if microcredit

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \quad)$$

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

1 if microcredit

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

1 if microcredit

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

1 if microcredit

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

- Priors and hyperpriors:

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

1 if microcredit

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

- Priors and hyperpriors:

$$\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, C\right)$$

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

profit

1 if microcredit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

- Priors and hyperpriors:

$$\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, C\right)$$

$$\sigma_k^{-2} \overset{iid}{\sim} \Gamma(a, b)$$

Microcredit Experiment

- Simplified from Meager (2015)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

profit

1 if microcredit

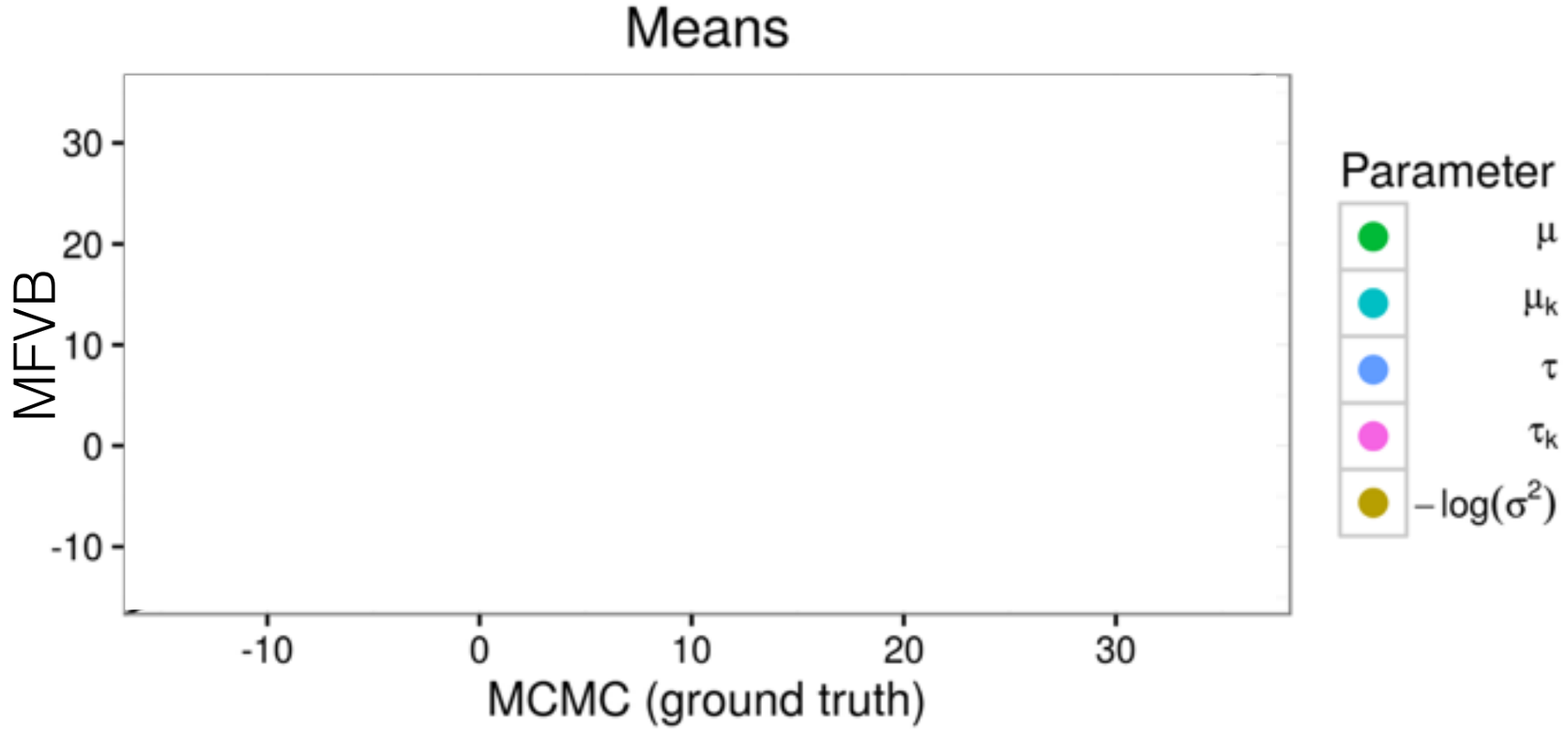
$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

- Priors and hyperpriors:

$$\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, C\right) \qquad \begin{pmatrix} \mu \\ \tau \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu_0 \\ \tau_0 \end{pmatrix}, \Lambda^{-1}\right)$$

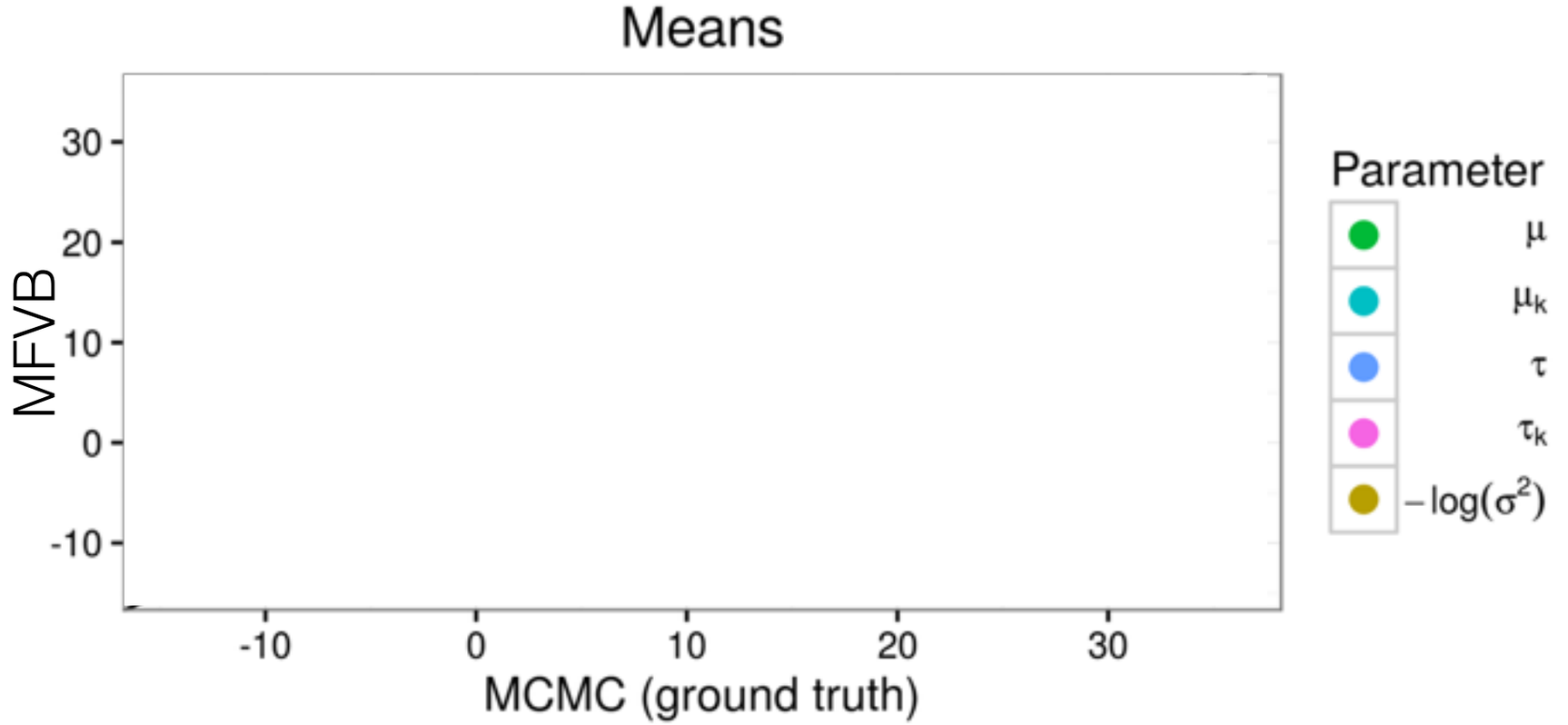
$$\sigma_k^{-2} \overset{iid}{\sim} \Gamma(a, b) \qquad C \sim \text{Sep\&LKJ}(\eta, c, d)$$

Microcredit Experiment



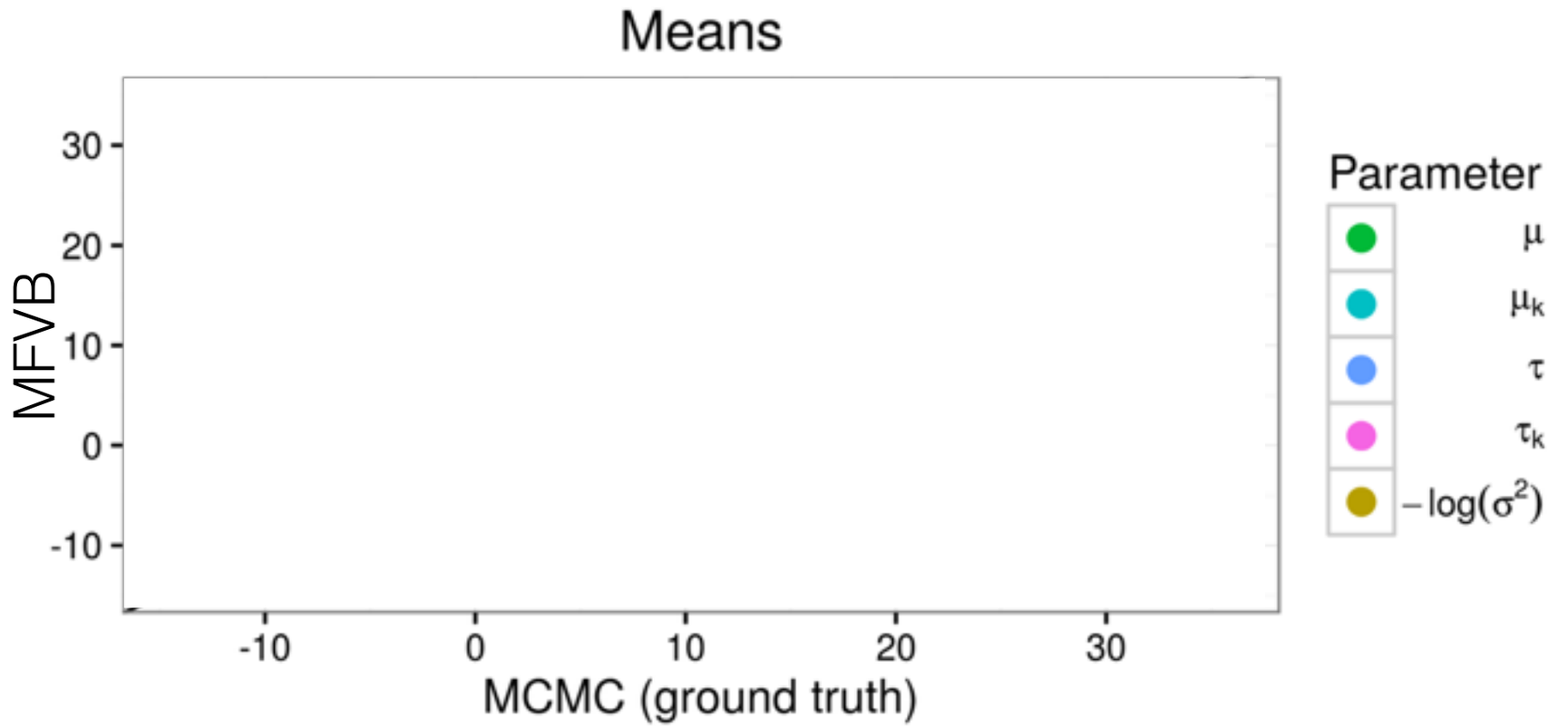
Microcredit Experiment

- *One set* of 2500 MCMC draws:
  45 minutes



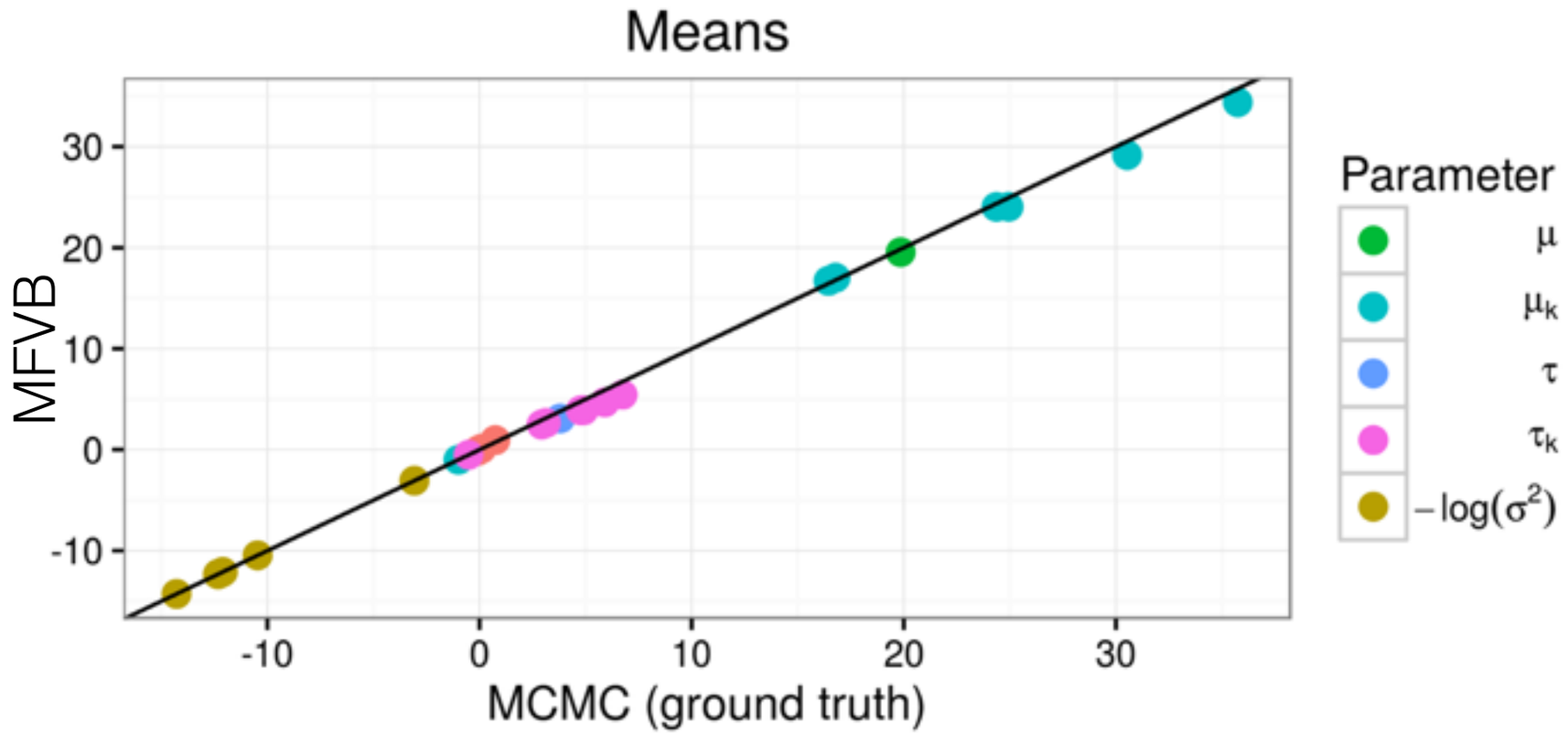
Microcredit Experiment

- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**



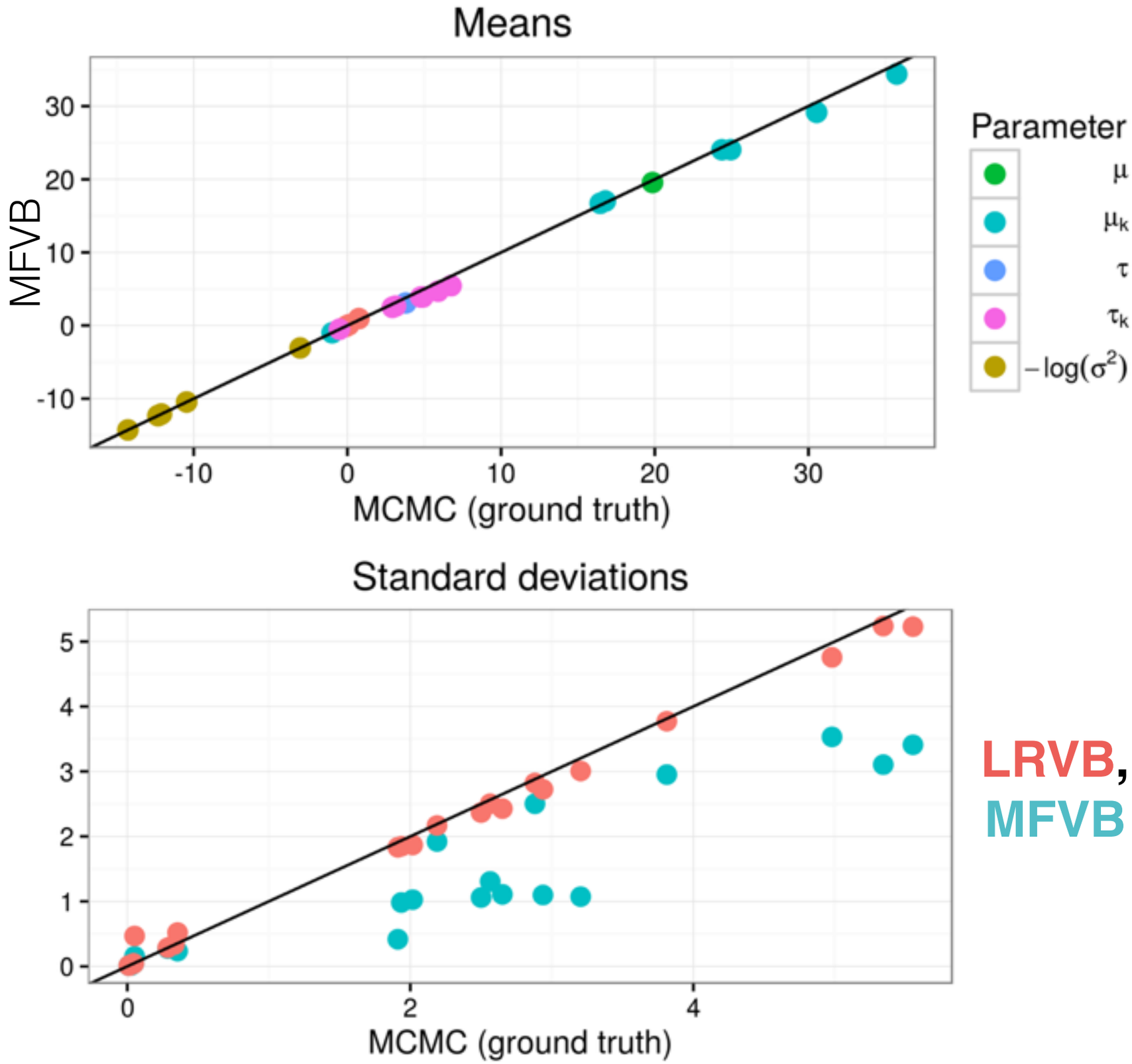
Microcredit Experiment

- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**



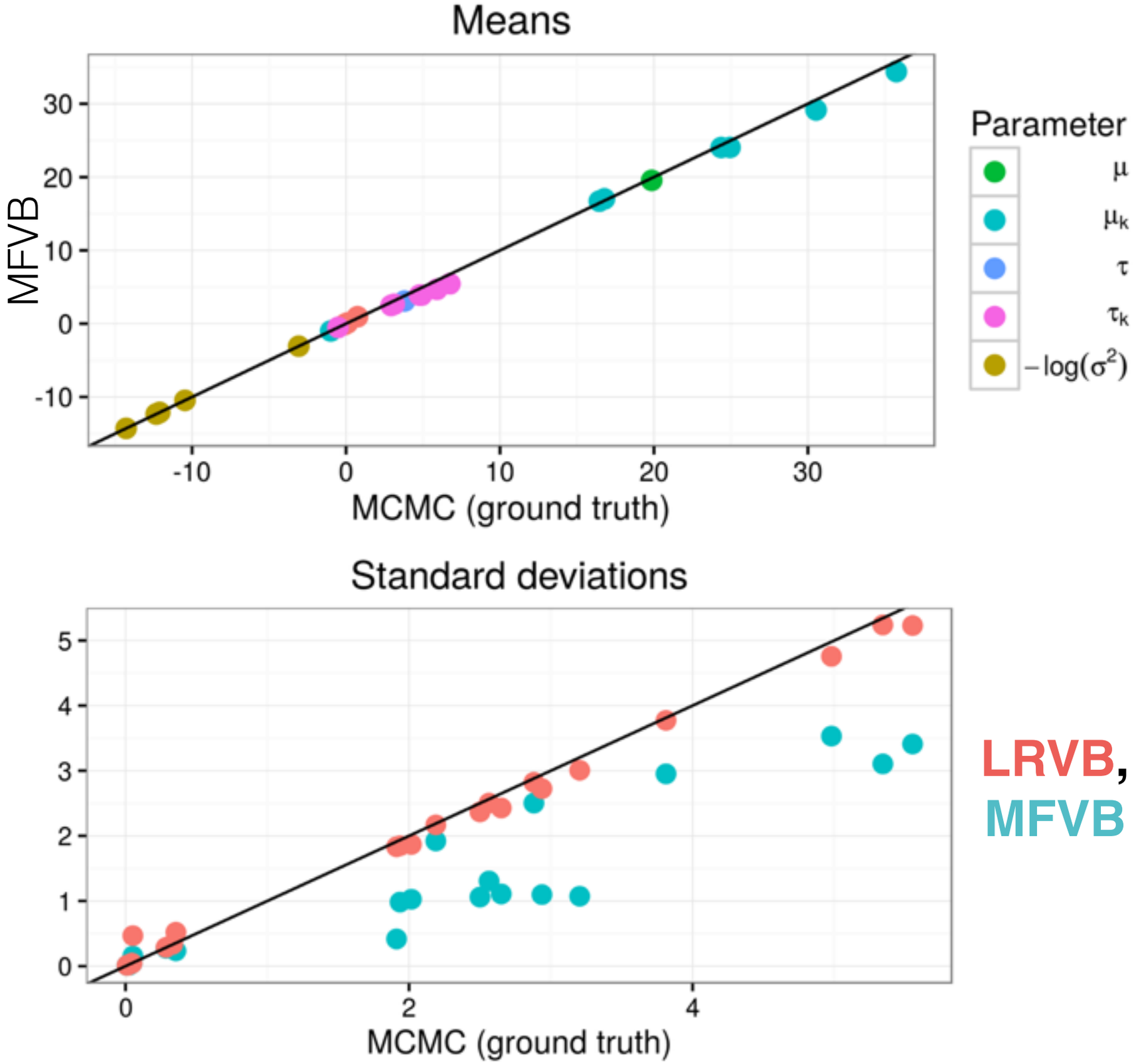
Microcredit Experiment

- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**



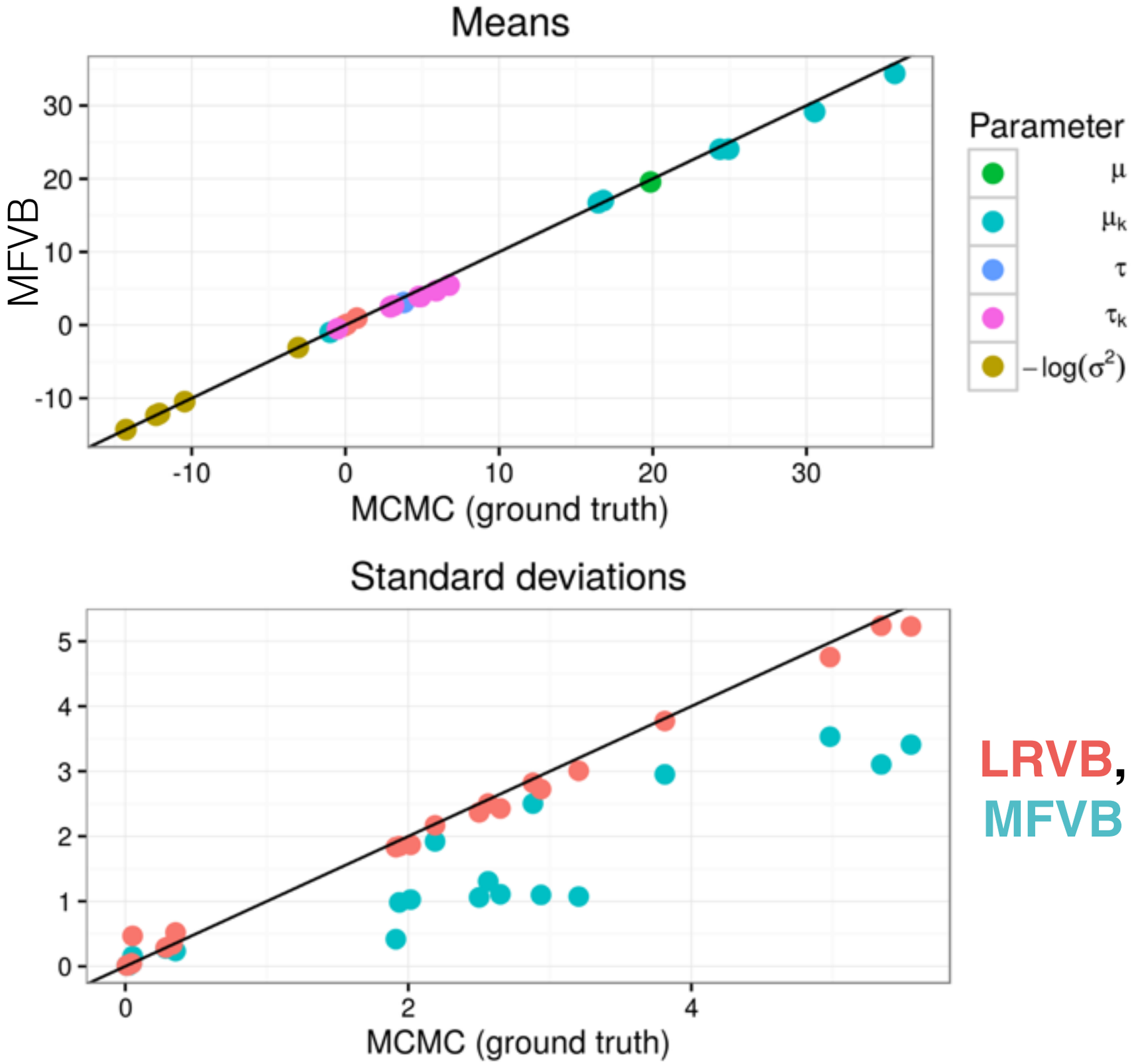
Microcredit Experiment

- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**
- τ mean (MFVB): 3.08 USD PPP



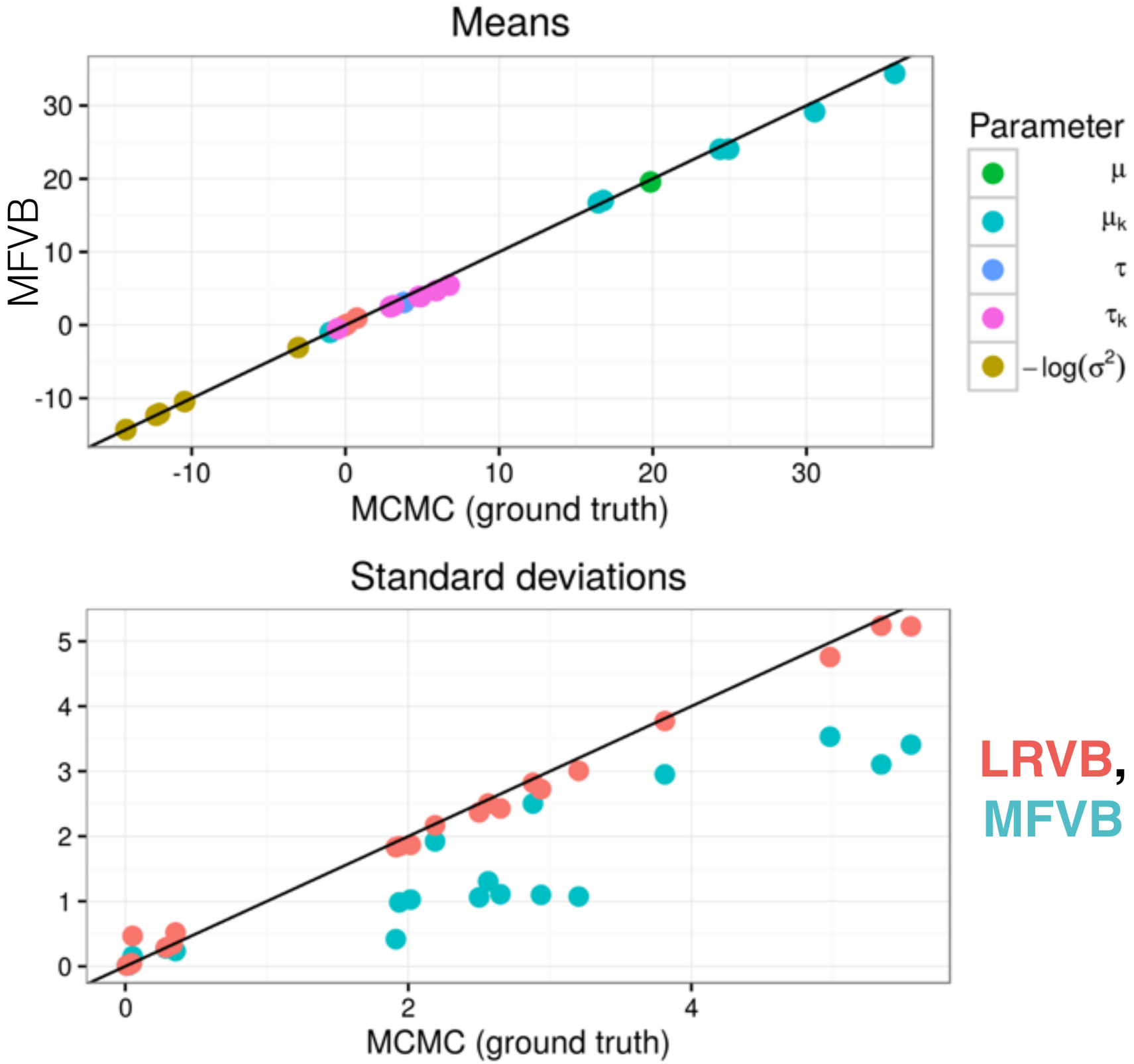
Microcredit Experiment

- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**
- τ mean (MFVB): 3.08 USD PPP
- τ std dev (LRVB): 1.83 USD PPP



Microcredit Experiment

- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**
- τ mean (MFVB): 3.08 USD PPP
- τ std dev (LRVB): 1.83 USD PPP
- Mean is 1.68 std dev from 0



Scaling the matrix inverse

Scaling the matrix inverse

- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1}V$

Scaling the matrix inverse

- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1}V$

- Decomposition of parameter vector

Scaling the matrix inverse

- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1}V$

- Decomposition of parameter vector

$$\theta = (\alpha^T, z^T)^T$$

Scaling the matrix inverse

- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1}V$

- Decomposition of parameter vector

$$\theta = (\alpha^T, z^T)^T \qquad H = \begin{array}{|c|c|} \hline H_\alpha & H_{\alpha z} \\ \hline H_{z\alpha} & H_z \\ \hline \end{array}$$

Scaling the matrix inverse

- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1}V$
- Decomposition of parameter vector

$$\theta = (\alpha^T, z^T)^T \qquad H = \begin{array}{|c|c|}\hline H_\alpha & H_{\alpha z} \\ \hline H_{z\alpha} & H_z \\ \hline \end{array}$$

- Schur complement

Scaling the matrix inverse

- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1}V$

- Decomposition of parameter vector

$$\theta = (\alpha^T, z^T)^T \qquad H = \begin{array}{|c|c|}\hline H_\alpha & H_{\alpha z} \\ \hline H_{z\alpha} & H_z \\ \hline \end{array}$$

- Schur complement

$$\hat{\Sigma}_\alpha = \left(I_\alpha - V_\alpha H_\alpha - V_\alpha H_{\alpha z}\left(I_z - V_z H_z\right)^{-1} V_z H_{z\alpha}\right)^{-1} V_\alpha$$

Scaling the matrix inverse

- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1}V$

- Decomposition of parameter vector

$$\theta = (\alpha^T, z^T)^T \qquad H = \begin{array}{|c|c|}\hline H_\alpha & H_{\alpha z} \\ \hline H_{z\alpha} & H_z \\ \hline \end{array}$$

- Schur complement

$$\hat{\Sigma}_\alpha = \left(I_\alpha - V_\alpha H_\alpha - V_\alpha H_{\alpha z}\left(I_z - V_z H_z\right)^{-1} V_z H_{z\alpha}\right)^{-1} V_\alpha$$

Scaling the matrix inverse

- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1}V$

- Decomposition of parameter vector

$$\theta = (\alpha^T, z^T)^T \qquad H = \begin{array}{|c|c|}\hline H_\alpha & H_{\alpha z} \\ \hline H_{z\alpha} & H_z \\ \hline \end{array}$$

- Schur complement

$$\hat{\Sigma}_\alpha = \left(I_\alpha - V_\alpha H_\alpha - V_\alpha H_{\alpha z}\left(I_z - V_z H_z\right)^{-1} V_z H_{z\alpha}\right)^{-1} V_\alpha$$

Scaling the matrix inverse

- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1}V$

- Decomposition of parameter vector

$$\theta = (\alpha^T, z^T)^T \qquad H = \begin{array}{|c|c|}\hline H_\alpha & H_{\alpha z} \\ \hline H_{z\alpha} & H_z \\ \hline \end{array}$$

- Schur complement

$$\hat{\Sigma}_\alpha = \left(I_\alpha - V_\alpha H_\alpha - V_\alpha H_{\alpha z}\left(I_z - V_z H_z\right)^{-1} V_z H_{z\alpha}\right)^{-1} V_\alpha$$

- Sparsity patterns

Scaling the matrix inverse

- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1}V$

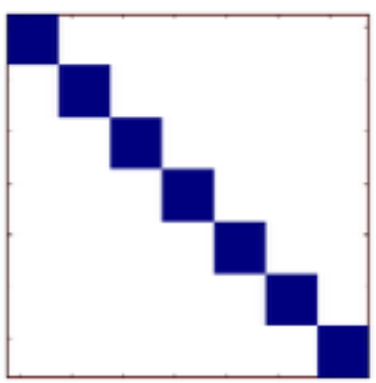
- Decomposition of parameter vector

$$\theta = (\alpha^T, z^T)^T \qquad H = \begin{array}{|c|c|}\hline H_\alpha & H_{\alpha z} \\ \hline H_{z\alpha} & H_z \\ \hline \end{array}$$

- Schur complement

$$\hat{\Sigma}_\alpha = \left(I_\alpha - V_\alpha H_\alpha - V_\alpha H_{\alpha z}\left(I_z - V_z H_z\right)^{-1} V_z H_{z\alpha}\right)^{-1} V_\alpha$$

- Sparsity patterns

V 

Scaling the matrix inverse

- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1} V$

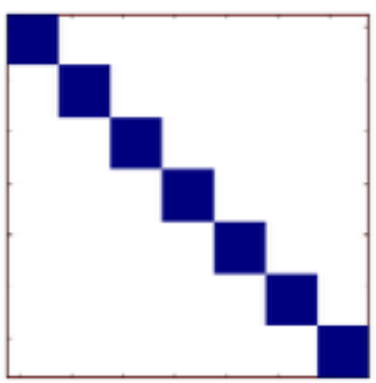
- Decomposition of parameter vector

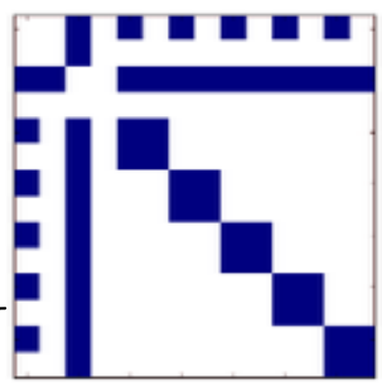
$$\theta = (\alpha^T, z^T)^T \qquad H = \begin{array}{|c|c|} \hline H_\alpha & H_{\alpha z} \\ \hline H_{z\alpha} & H_z \\ \hline \end{array}$$

- Schur complement

$$\hat{\Sigma}_\alpha = \left(I_\alpha - V_\alpha H_\alpha - V_\alpha H_{\alpha z} \left(I_z - V_z H_z\right)^{-1} V_z H_{z\alpha}\right)^{-1} V_\alpha$$

- Sparsity patterns

V 

H 

Scaling the matrix inverse

- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1} V$

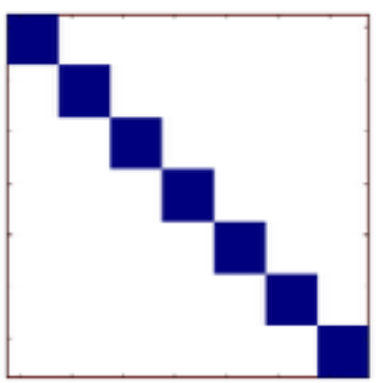
- Decomposition of parameter vector

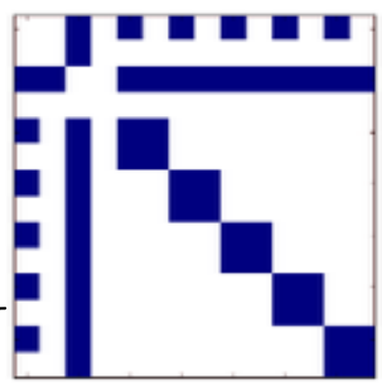
$$\theta = (\alpha^T, z^T)^T \qquad H = \begin{array}{|c|c|} \hline H_\alpha & H_{\alpha z} \\ \hline H_{z\alpha} & H_z \\ \hline \end{array}$$

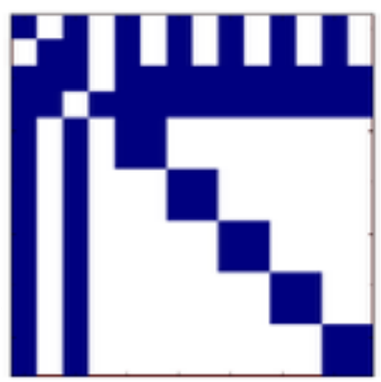
- Schur complement

$$\hat{\Sigma}_\alpha = \left(I_\alpha - V_\alpha H_\alpha - V_\alpha H_{\alpha z} \left(I_z - V_z H_z\right)^{-1} V_z H_{z\alpha}\right)^{-1} V_\alpha$$

- Sparsity patterns

V 

H 

$I - VH$ 

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes, variational Bayes)
  - Uncertainties
- Robustness trade-offs
  - Local robustness quantification
- Theoretical guarantees on quality

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes, variational Bayes)
  - Uncertainties (Linear response)
- Robustness trade-offs
  - Local robustness quantification
- Theoretical guarantees on quality

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes, variational Bayes)
  - Uncertainties (Linear response)
- Robustness trade-offs
  - Local robustness quantification
- Theoretical guarantees on quality

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes, variational Bayes)
  - Uncertainties (Linear response)
- Robustness trade-offs
  - Local robustness quantification
- Theoretical guarantees on quality

Robustness quantification

- Bayes Theorem

$$p(\theta|x)$$

$$\propto_\theta p(x|\theta)p(\theta)$$

Robustness quantification

- Bayes Theorem

$$p(\theta|x,\alpha)$$

$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

Robustness quantification

- Bayes Theorem

$$p_\alpha(\theta) := p(\theta|x,\alpha)$$

$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

Robustness quantification

- Bayes Theorem

$$p_\alpha(\theta) := p(\theta|x,\alpha)$$
$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

- Sensitivity

Robustness quantification

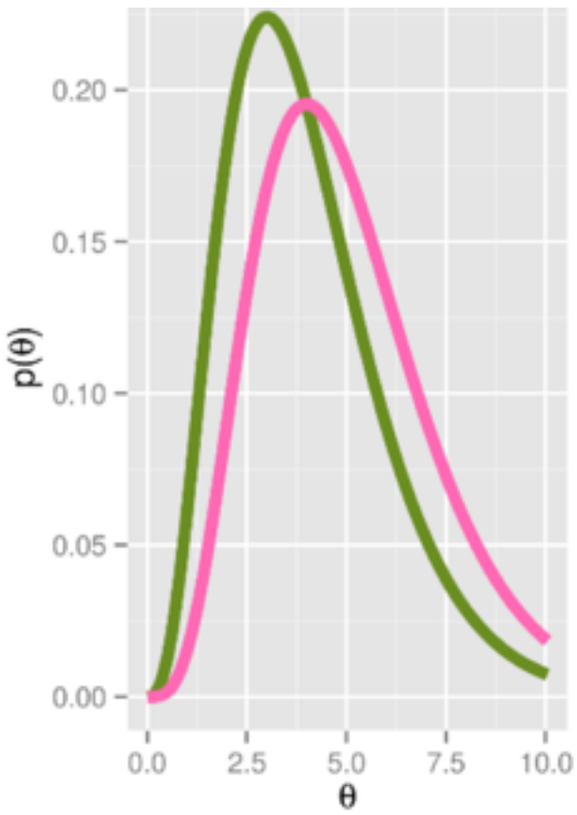
- Bayes Theorem

$$p_\alpha(\theta) := p(\theta|x,\alpha)$$

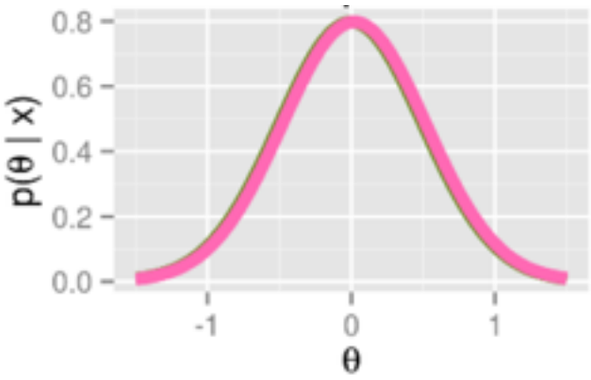
$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

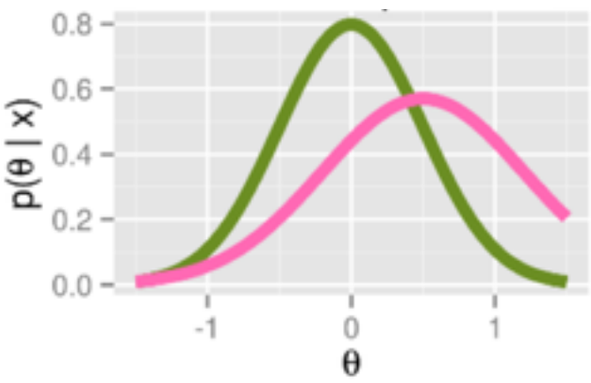
- Sensitivity

Some reasonable priors



Bayes Theorem





Robustness quantification

- Bayes Theorem

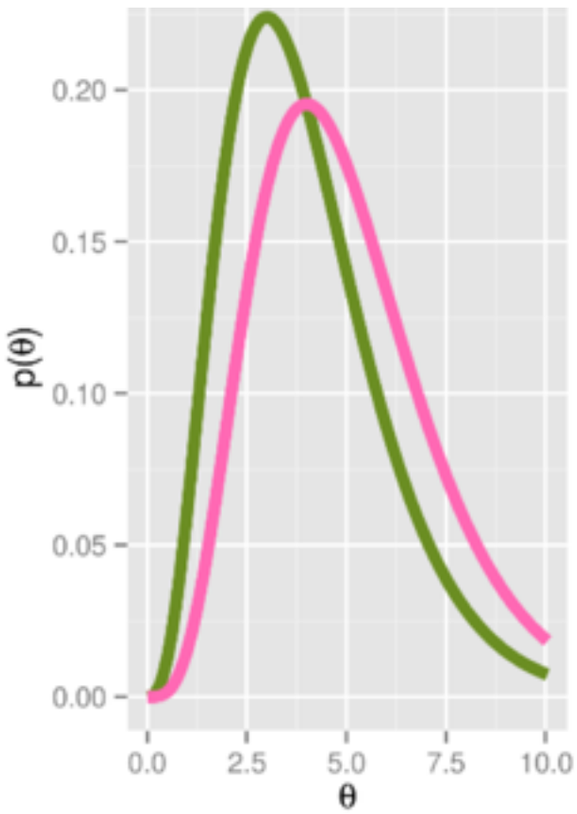
$$p_\alpha(\theta) := p(\theta|x,\alpha)$$

$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

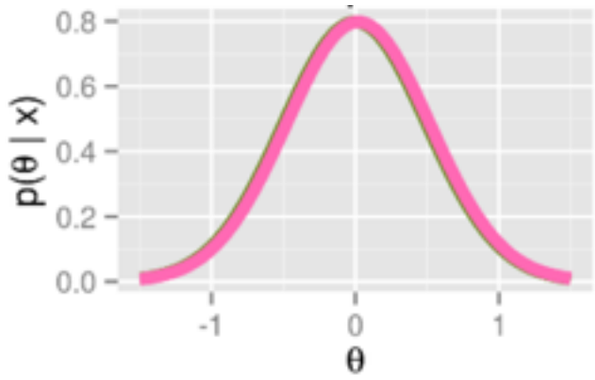
- Sensitivity

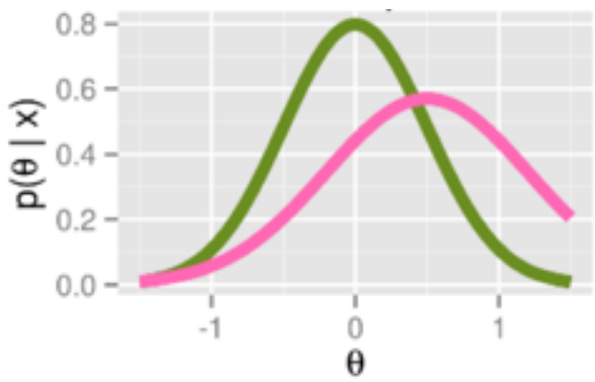
$$\mathbb{E}_{p_\alpha}[g(\theta)]$$

Some reasonable priors



Bayes Theorem





Robustness quantification

- Bayes Theorem

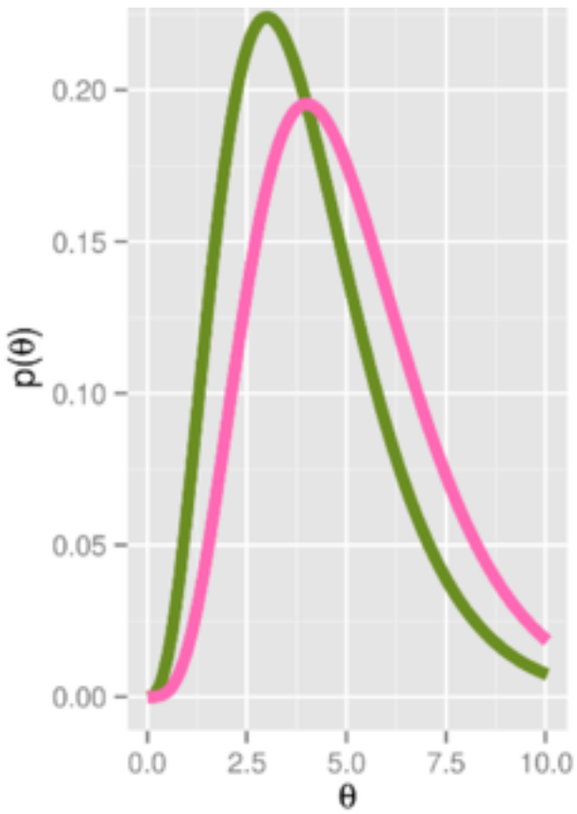
$$p_\alpha(\theta) := p(\theta|x,\alpha)$$

$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

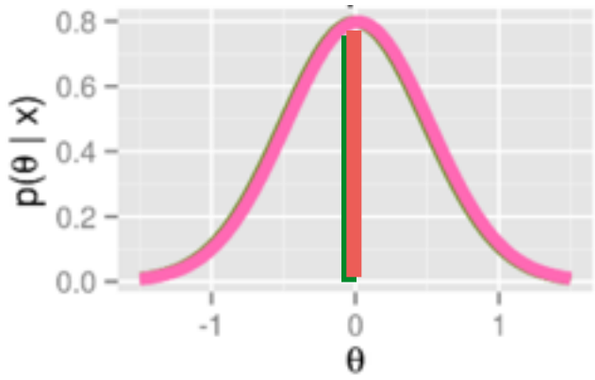
- Sensitivity

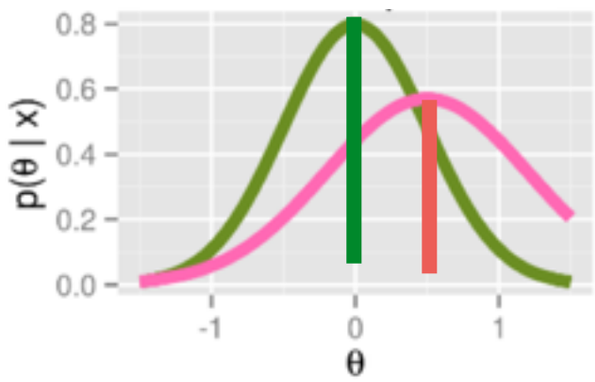
$$\mathbb{E}_{p_\alpha}[g(\theta)]$$

Some reasonable priors



Bayes Theorem





Robustness quantification

- Bayes Theorem

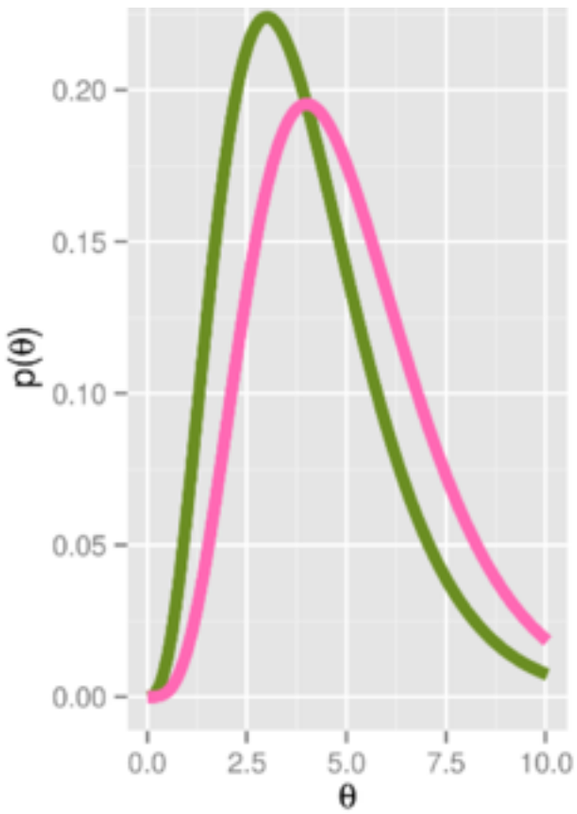
$$p_\alpha(\theta) := p(\theta|x,\alpha)$$

$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

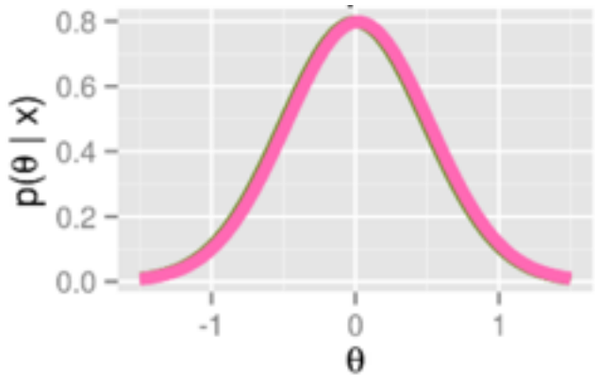
- Sensitivity

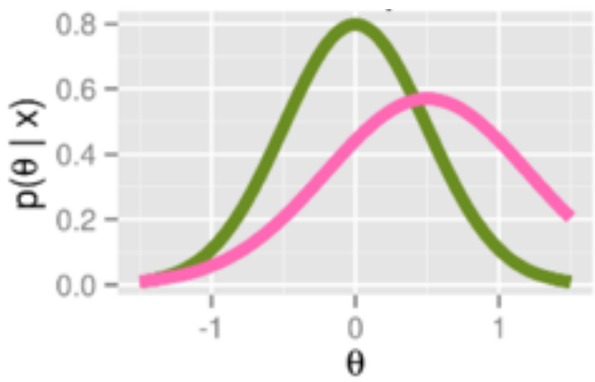
$$\mathbb{E}_{p_\alpha}[g(\theta)]$$

Some reasonable priors



Bayes Theorem





Robustness quantification

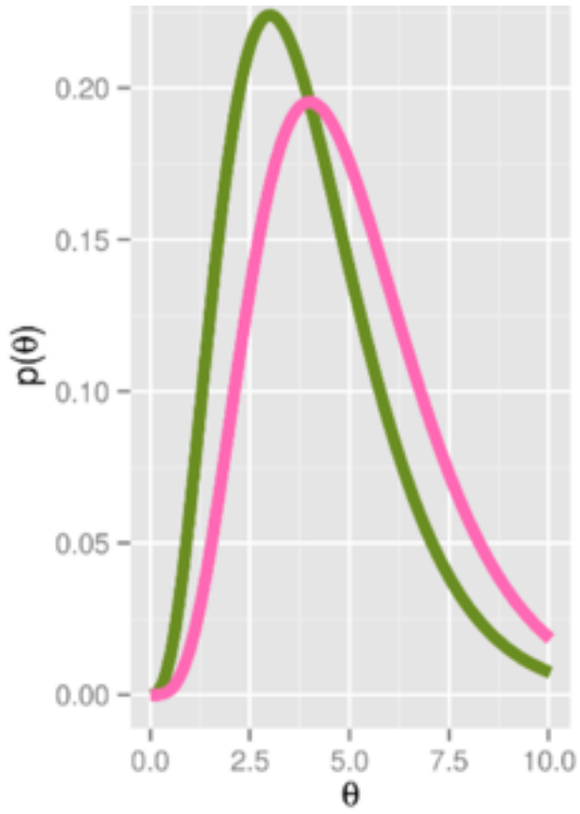
- Bayes Theorem

$$p_\alpha(\theta) := p(\theta|x,\alpha)$$
$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

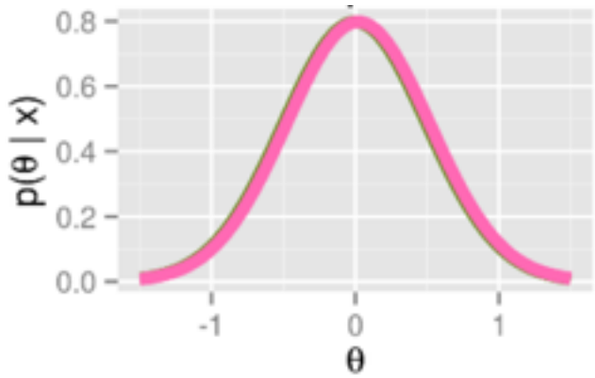
- Sensitivity

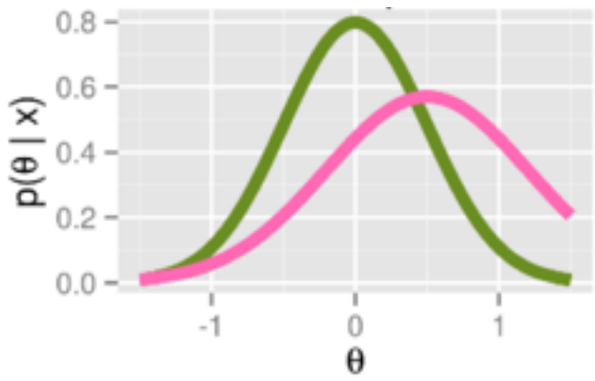
$$S := \left.\frac{d\mathbb{E}_{p_\alpha}[g(\theta)]}{d\alpha}\right|_\alpha \Delta\alpha$$

Some reasonable priors



Bayes Theorem





Robustness quantification

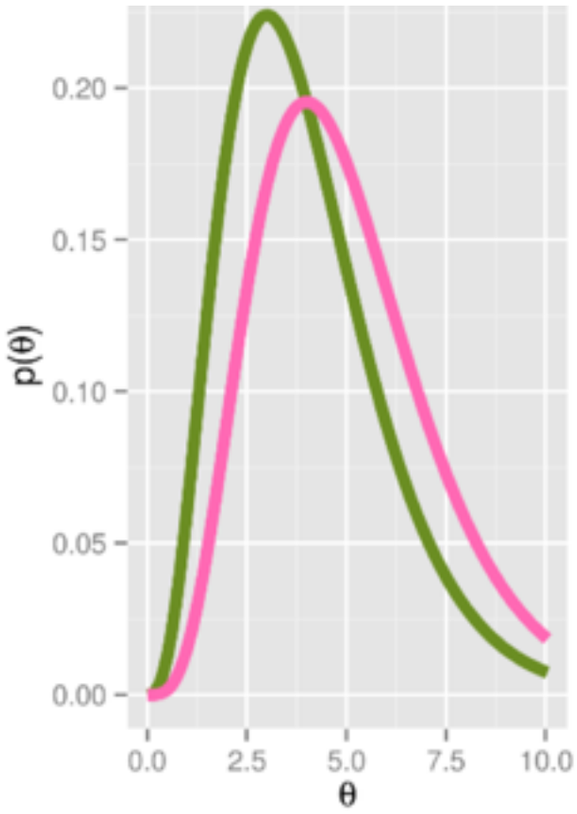
- Bayes Theorem

$$p_\alpha(\theta) := p(\theta|x,\alpha)$$
$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

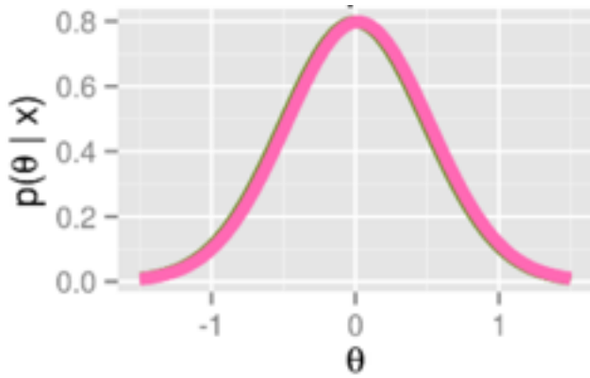
- Sensitivity

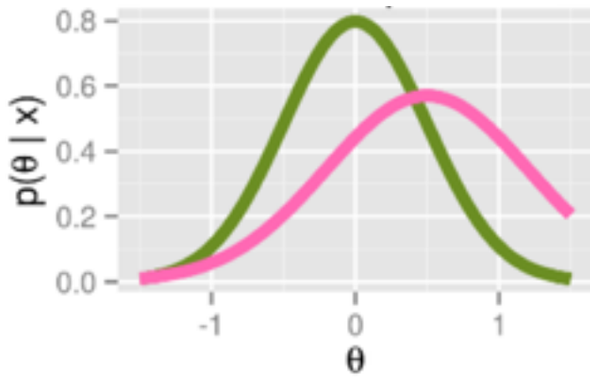
$$S := \left.\frac{d\mathbb{E}_{p_\alpha}[g(\theta)]}{d\alpha}\right|_\alpha \Delta\alpha$$

Some reasonable priors



Bayes Theorem





Robustness quantification

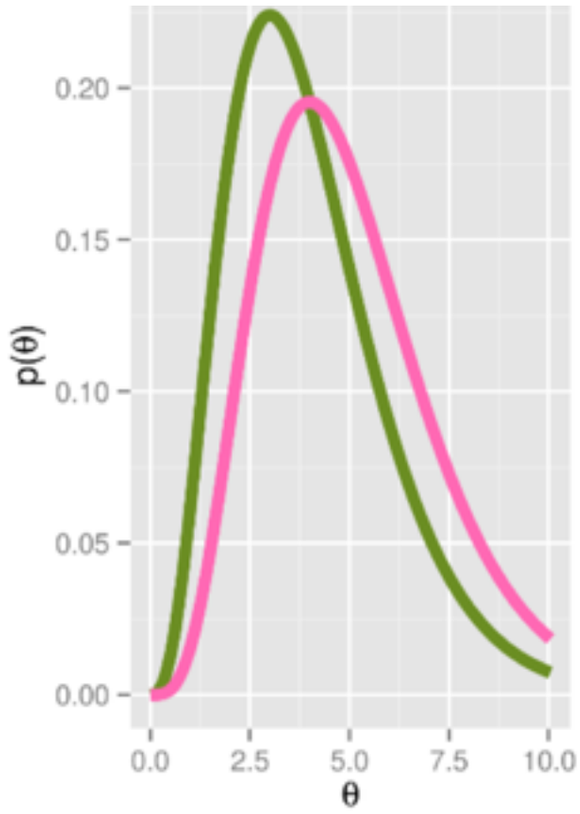
- Bayes Theorem

$$p_\alpha(\theta) := p(\theta|x,\alpha)$$
$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

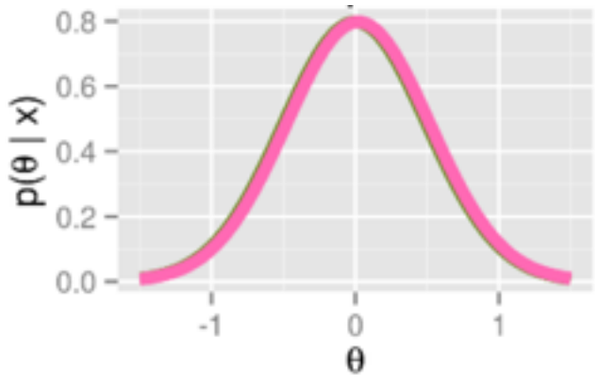
- Sensitivity

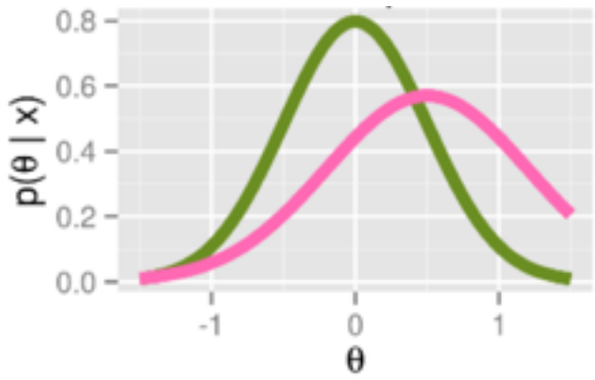
$$S := \left.\frac{d\mathbb{E}_{p_\alpha}[g(\theta)]}{d\alpha}\right|_\alpha \Delta\alpha$$
$$\approx \left.\frac{d\mathbb{E}_{q^*_\alpha}[g(\theta)]}{d\alpha}\right|_\alpha \Delta\alpha =: \hat{S}$$

Some reasonable priors



Bayes Theorem





Robustness quantification

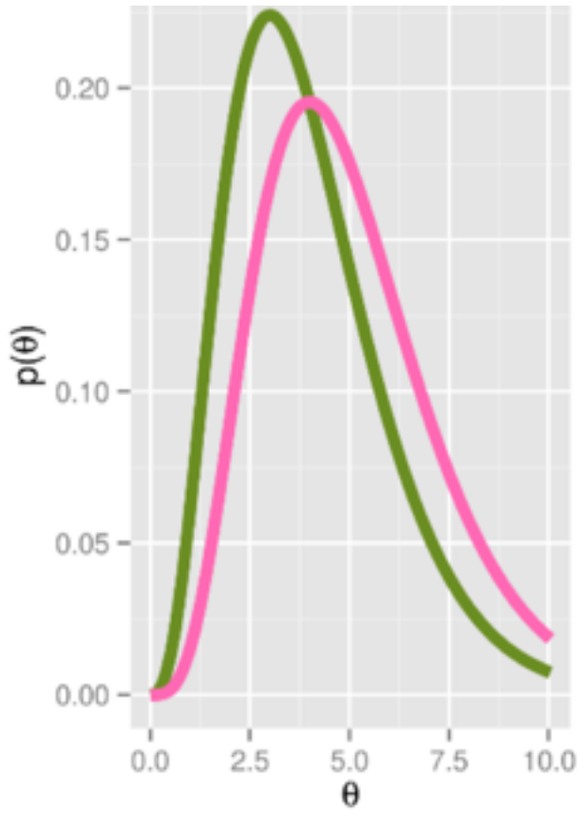
- Bayes Theorem

$$p_\alpha(\theta) := p(\theta|x,\alpha)$$
$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

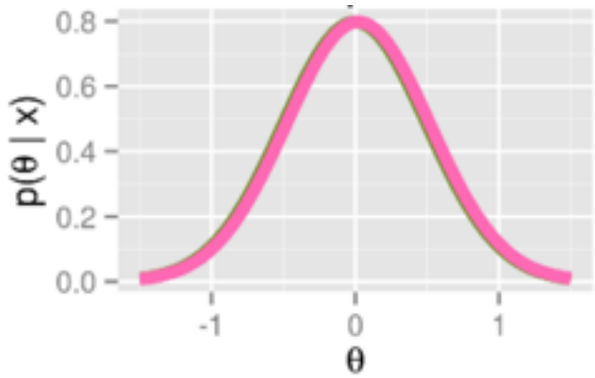
- Sensitivity

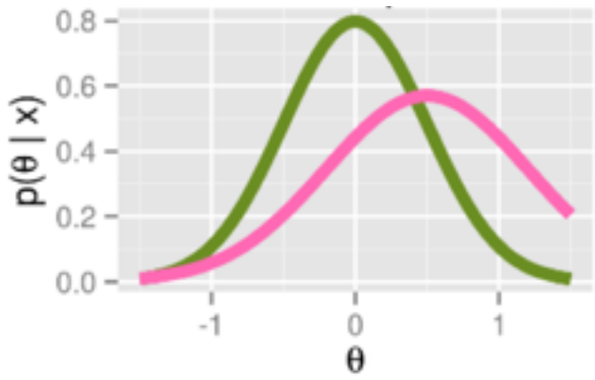
$$S := \left.\frac{d\mathbb{E}_{p_\alpha}[g(\theta)]}{d\alpha}\right|_\alpha \Delta\alpha$$
$$\approx \left.\frac{d\mathbb{E}_{q^*_\alpha}[g(\theta)]}{d\alpha}\right|_\alpha \Delta\alpha =: \hat{S}$$

Some reasonable priors



Bayes Theorem





LRVB estimator

Robustness quantification

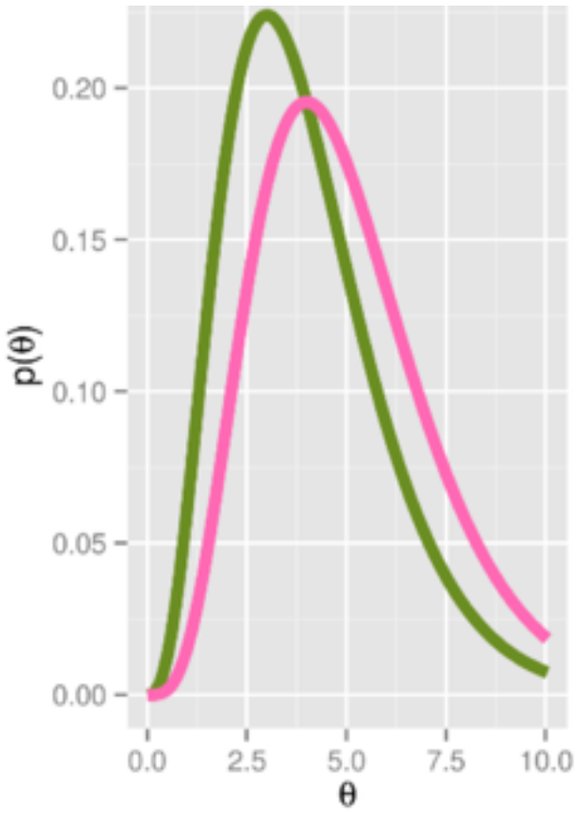
- Bayes Theorem

$$p_\alpha(\theta) := p(\theta|x,\alpha)$$
$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

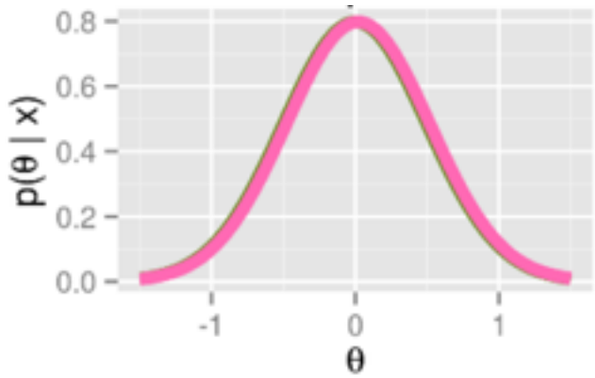
- Sensitivity

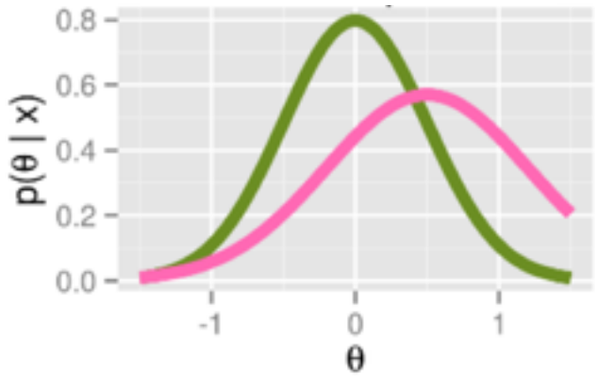
$$S := \left.\frac{d\mathbb{E}_{p_\alpha}[g(\theta)]}{d\alpha}\right|_\alpha \Delta\alpha$$
$$\approx \left.\frac{d\mathbb{E}_{q^*_\alpha}[g(\theta)]}{d\alpha}\right|_\alpha \Delta\alpha =: \hat{S}$$

Some reasonable priors



Bayes Theorem





← LRVB estimator

- When q^*_α in exponential family

Robustness quantification

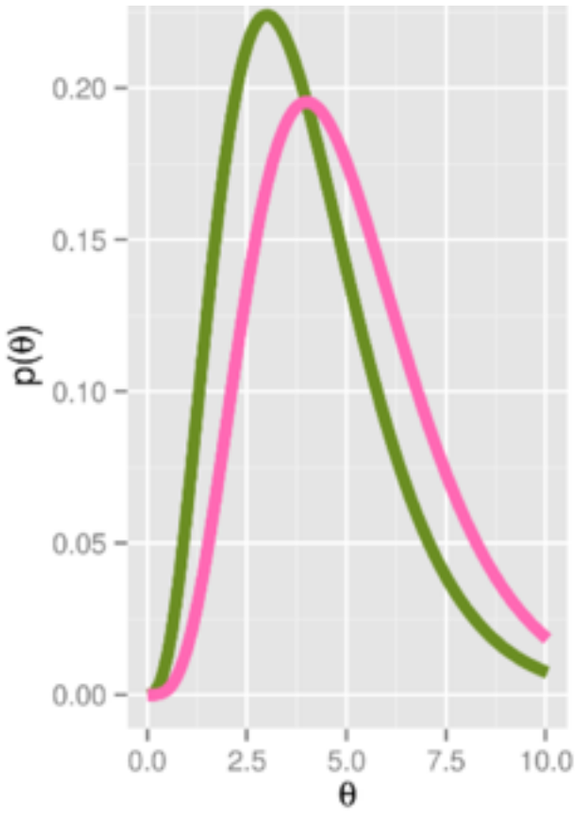
- Bayes Theorem

$$p_\alpha(\theta) := p(\theta|x,\alpha)$$
$$\propto_\theta p(x|\theta)p(\theta|\alpha)$$

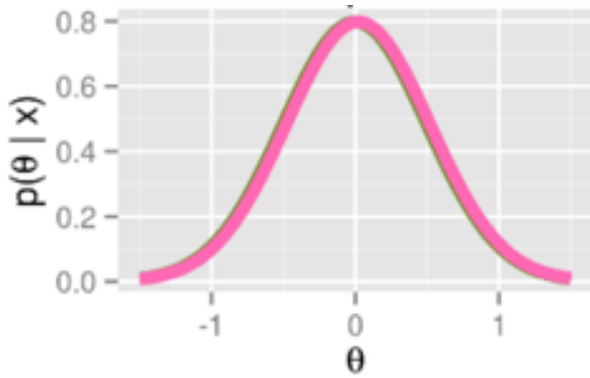
- Sensitivity

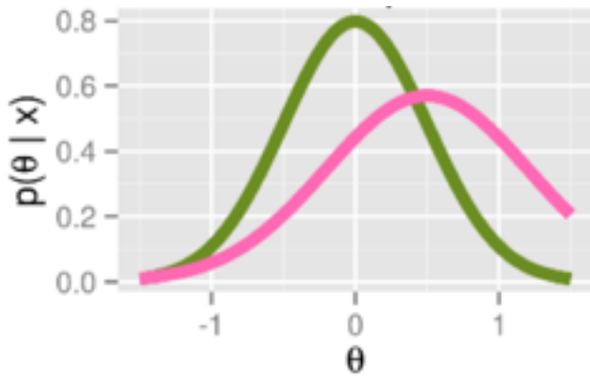
$$S := \left.\frac{d\mathbb{E}_{p_\alpha}[g(\theta)]}{d\alpha}\right|_\alpha \Delta\alpha$$
$$\approx \left.\frac{d\mathbb{E}_{q^*_\alpha}[g(\theta)]}{d\alpha}\right|_\alpha \Delta\alpha =: \hat{S}$$

Some reasonable priors



Bayes Theorem





← LRVB estimator

- When q^*_α in exponential family

$$\hat{S} = A\left(\left.\frac{\partial^2 KL}{\partial m \partial m^T}\right|_{m=m^*}\right)^{-1} B$$

Microcredit Experiment

- Simplified from Meager (2015)
- K microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

1 if microcredit

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

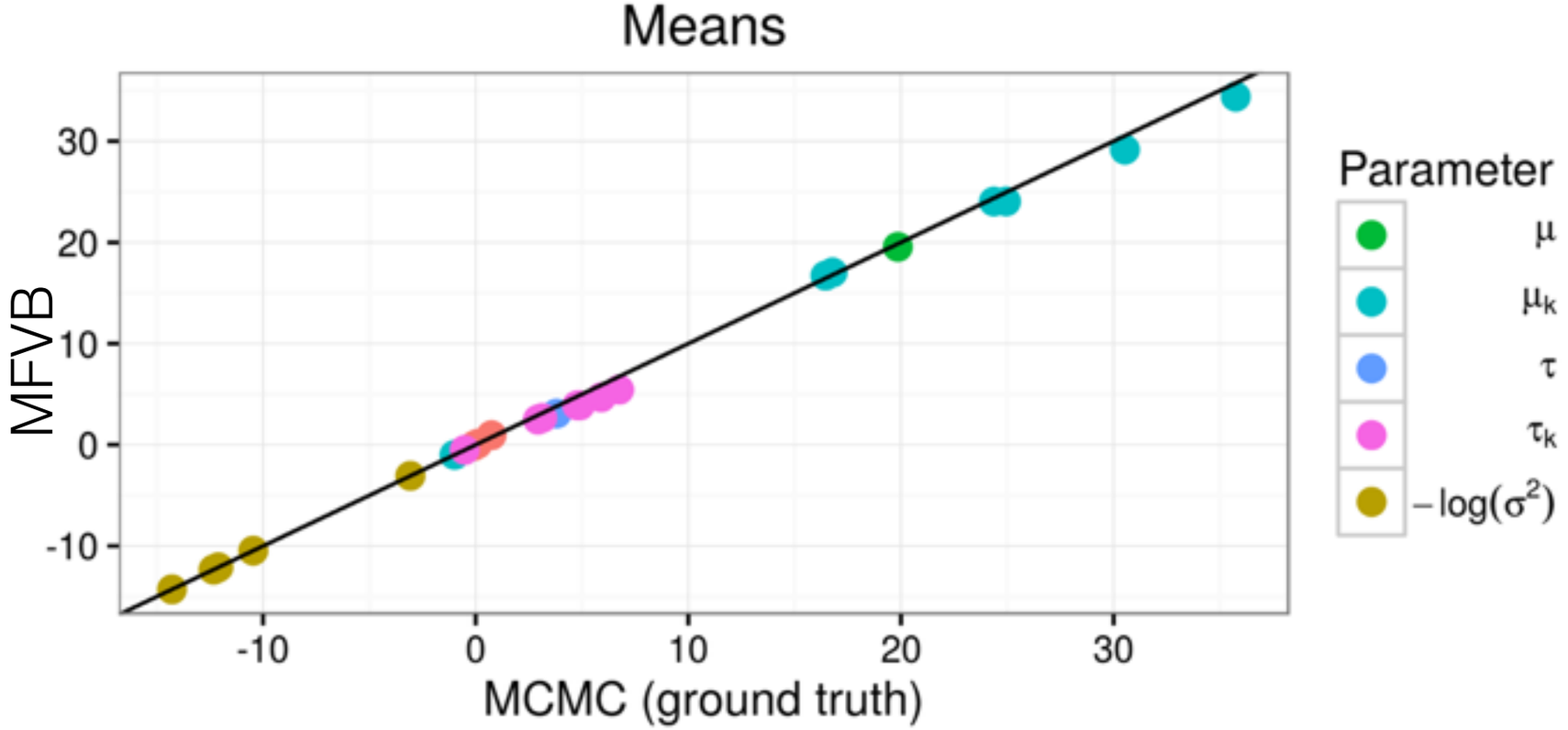
- Priors and hyperpriors:

$$\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, C\right) \qquad \begin{pmatrix} \mu \\ \tau \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu_0 \\ \tau_0 \end{pmatrix}, \Lambda^{-1}\right)$$

$$\sigma_k^{-2} \overset{iid}{\sim} \Gamma(a, b) \qquad C \sim \text{Sep\&LKJ}(\eta, c, d)$$

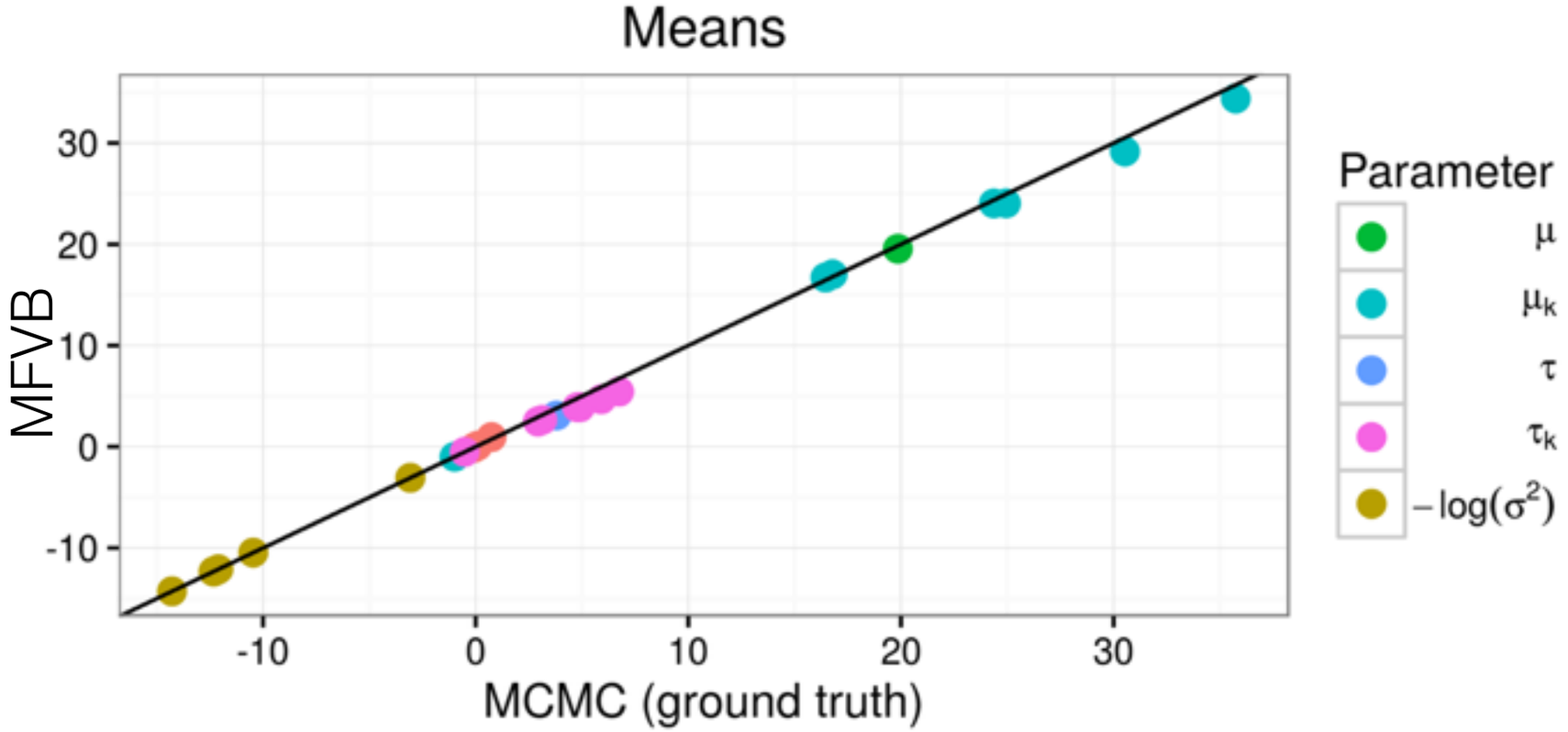
Microcredit Experiment

Microcredit Experiment



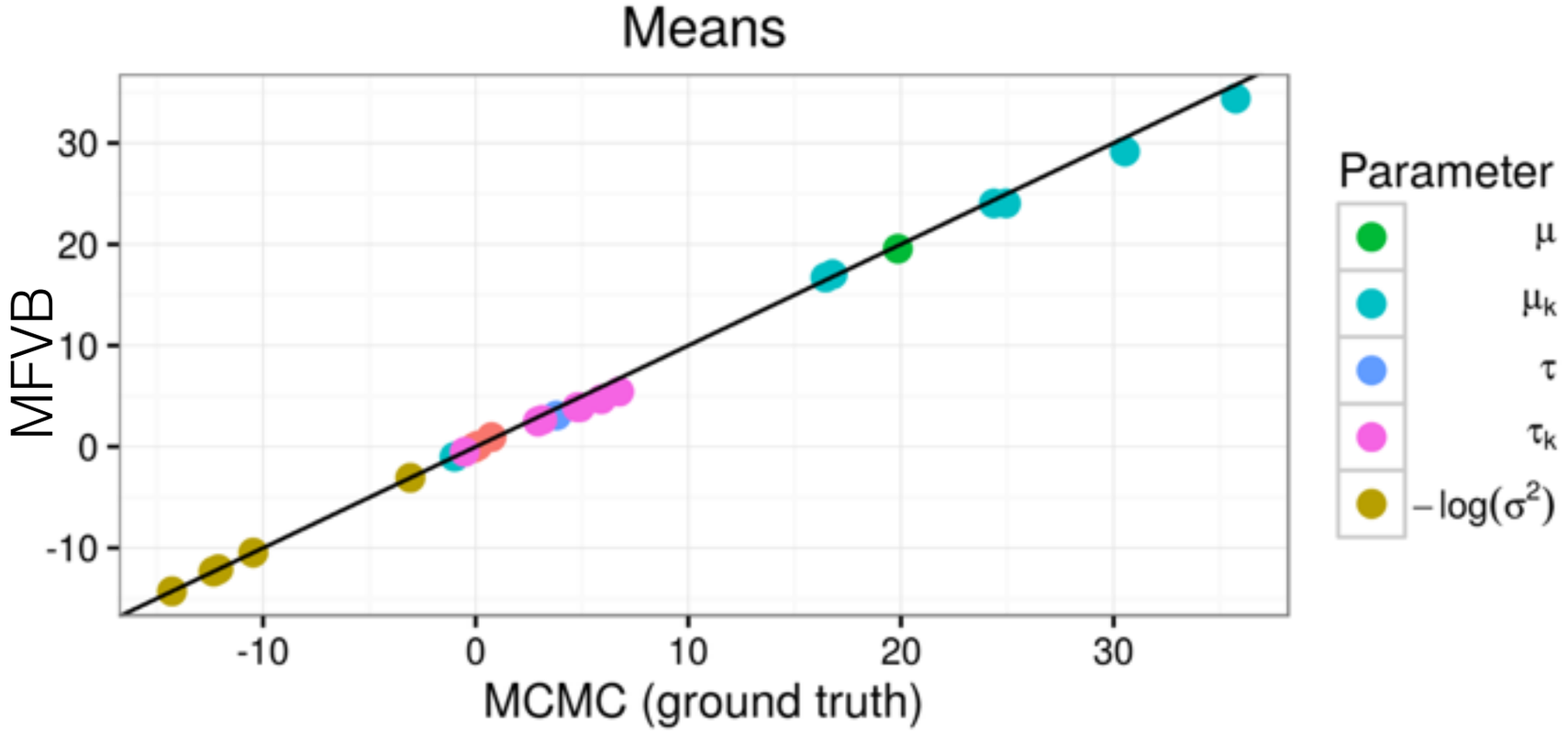
Microcredit Experiment

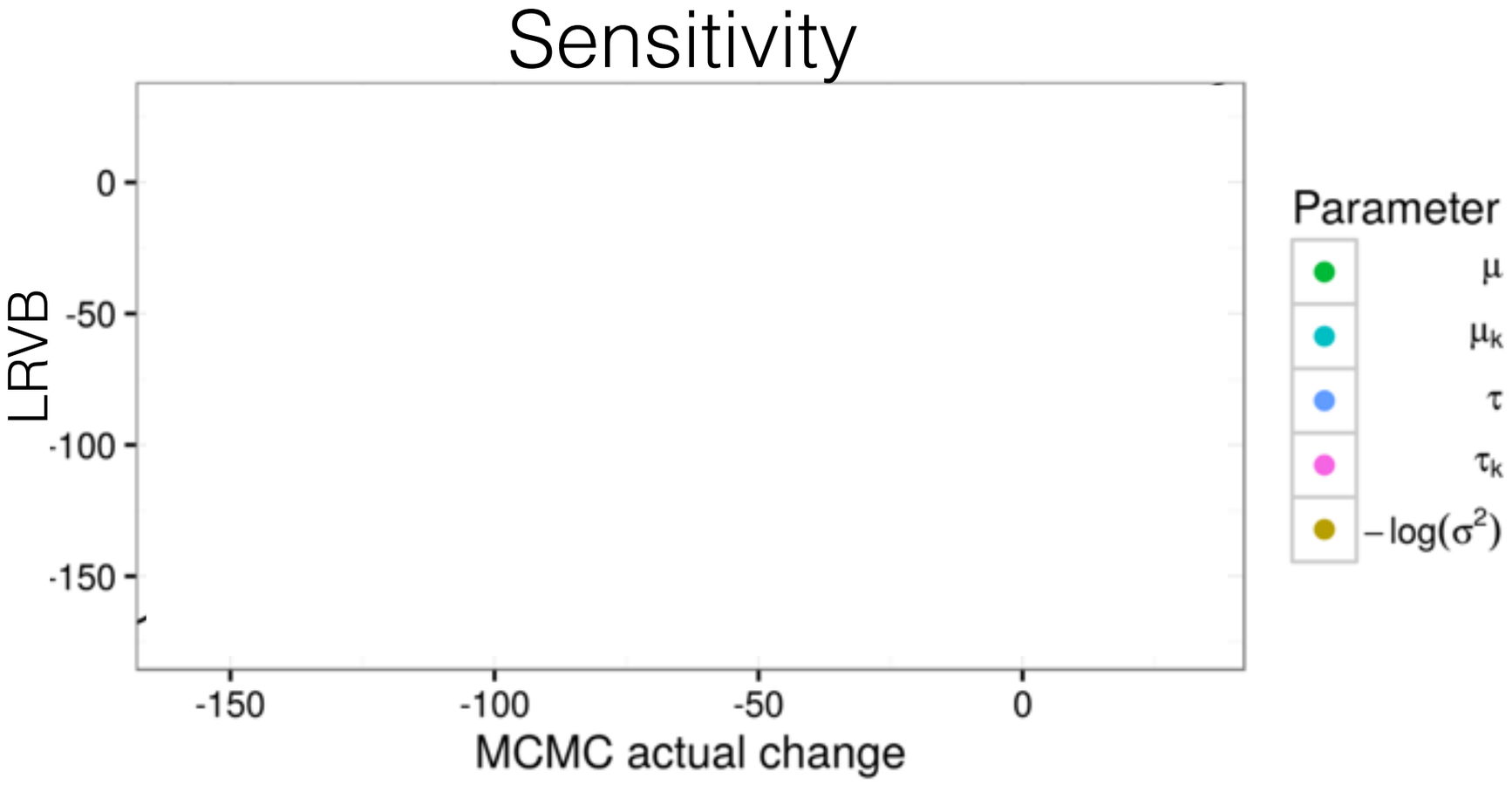
- Perturb Λ_{11}: 0.03 ➔ 0.04



Microcredit Experiment

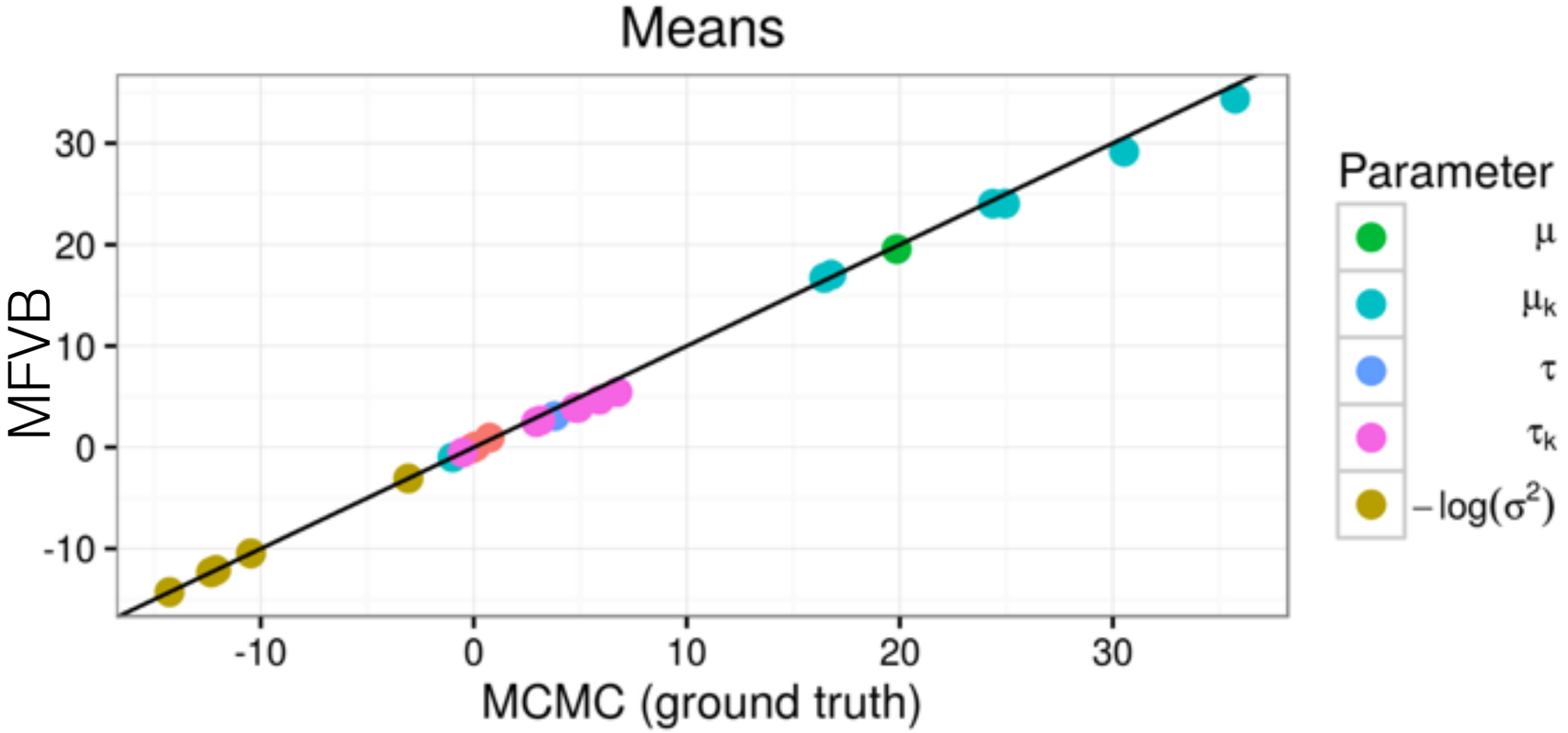
- Perturb Λ_{11}: 0.03 ➔ 0.04

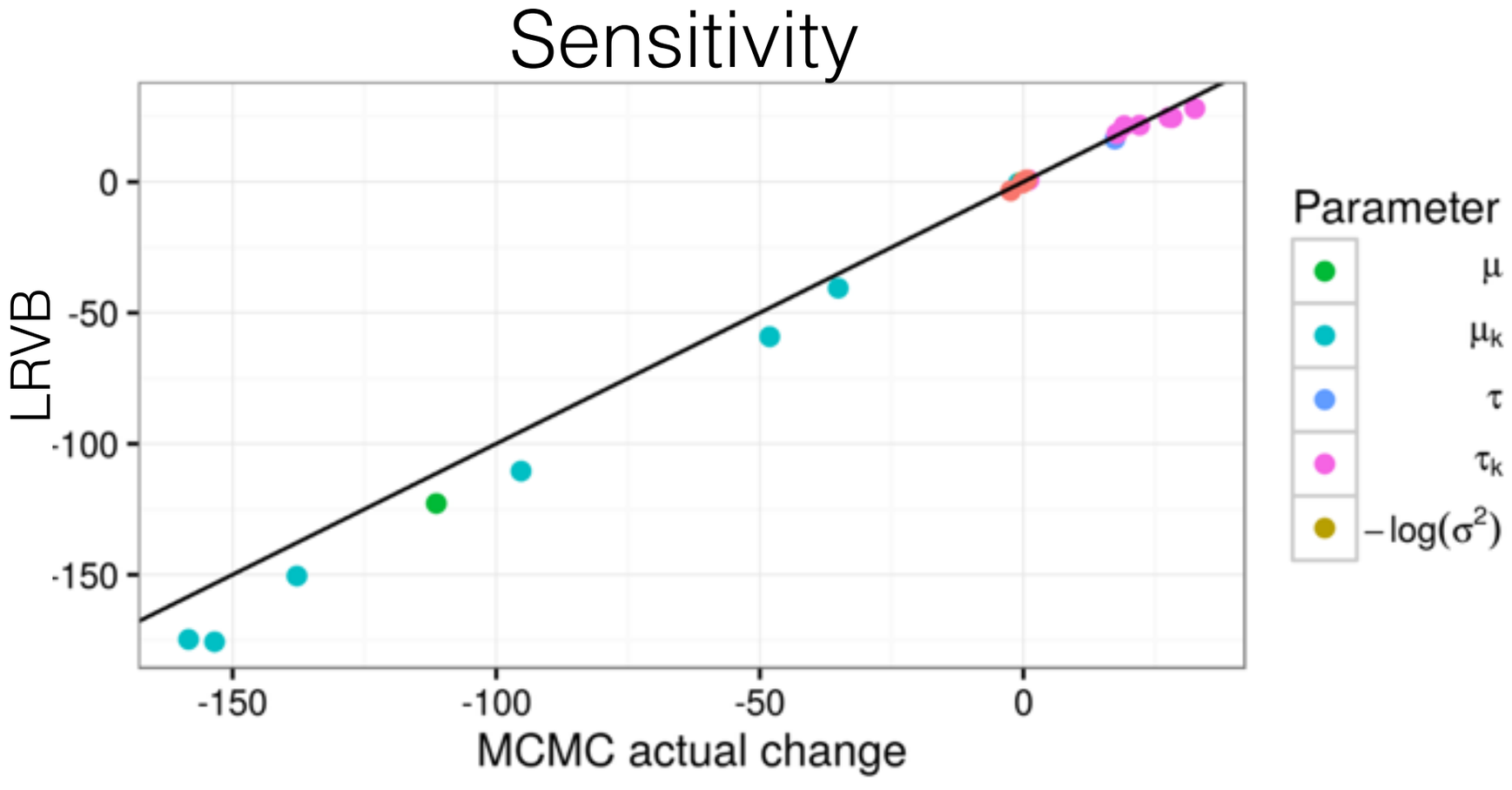




Microcredit Experiment

- Perturb Λ_{11}: 0.03 ➔ 0.04





Microcredit Experiment

Microcredit Experiment

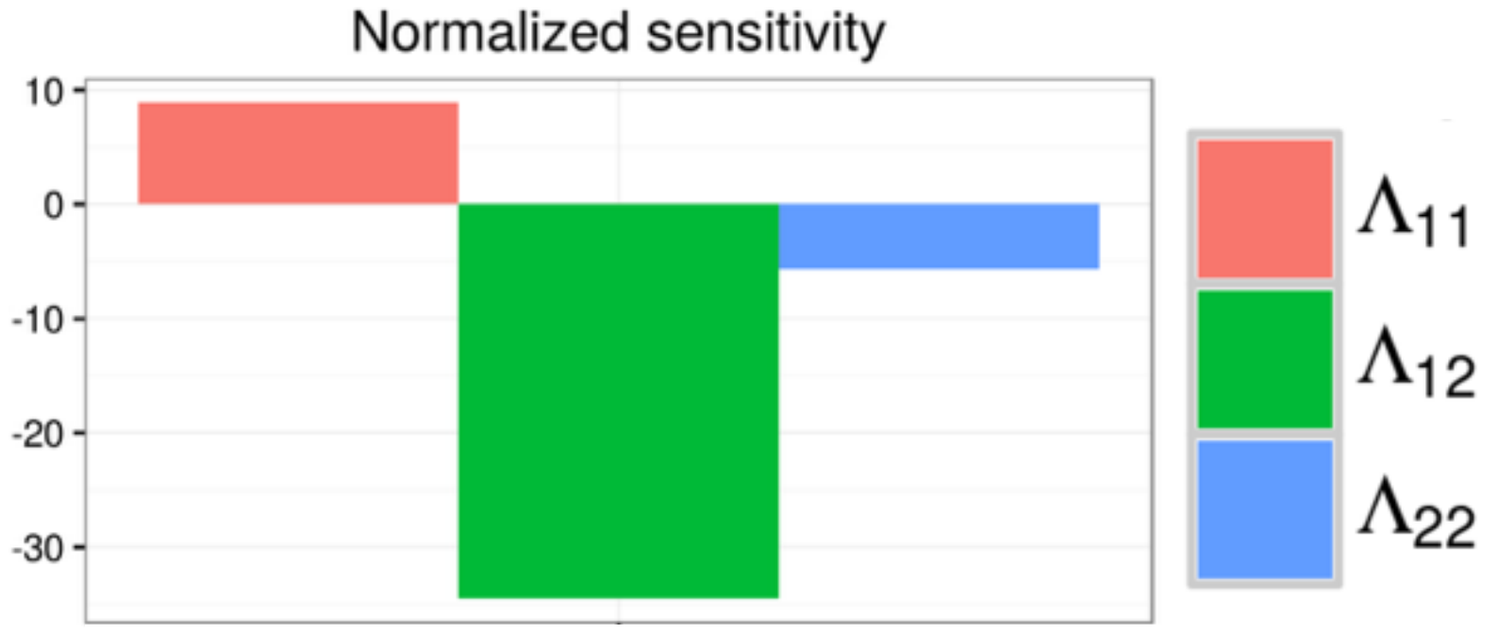
- Sensitivity of the expected microcredit effect (τ)

Microcredit Experiment

- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of τ std devs

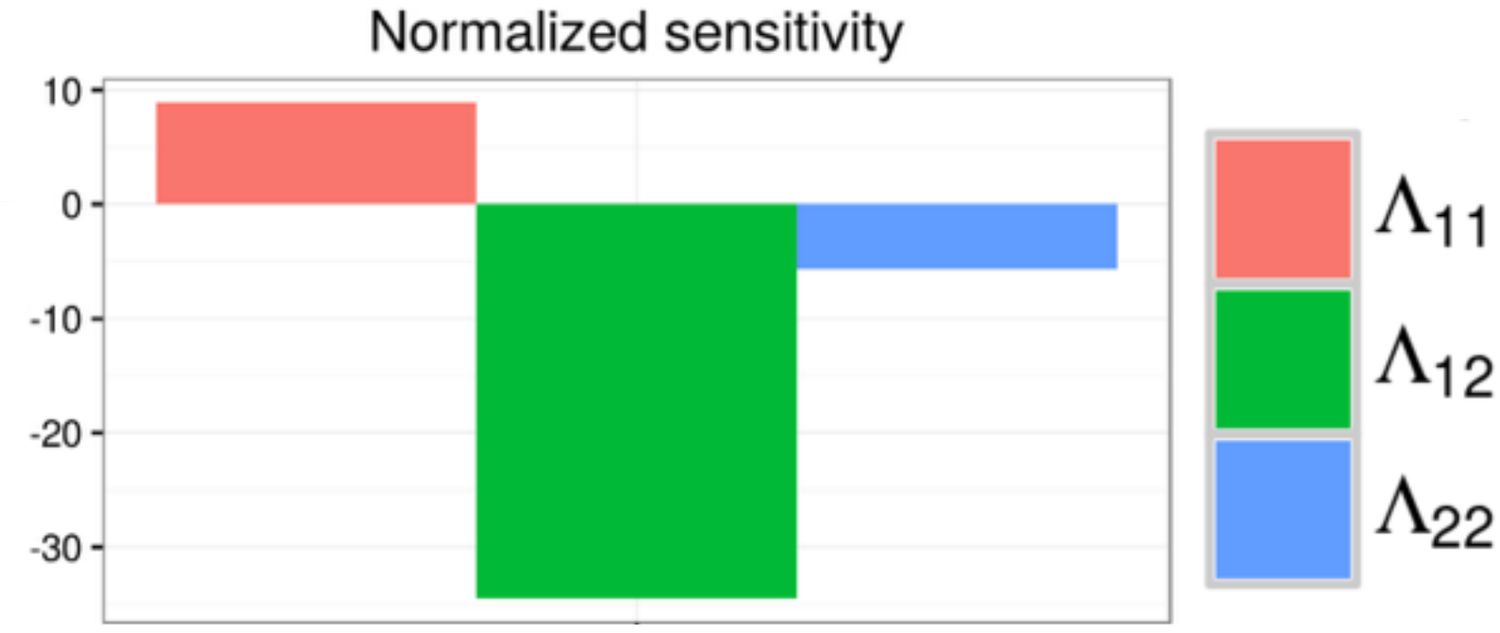
Microcredit Experiment

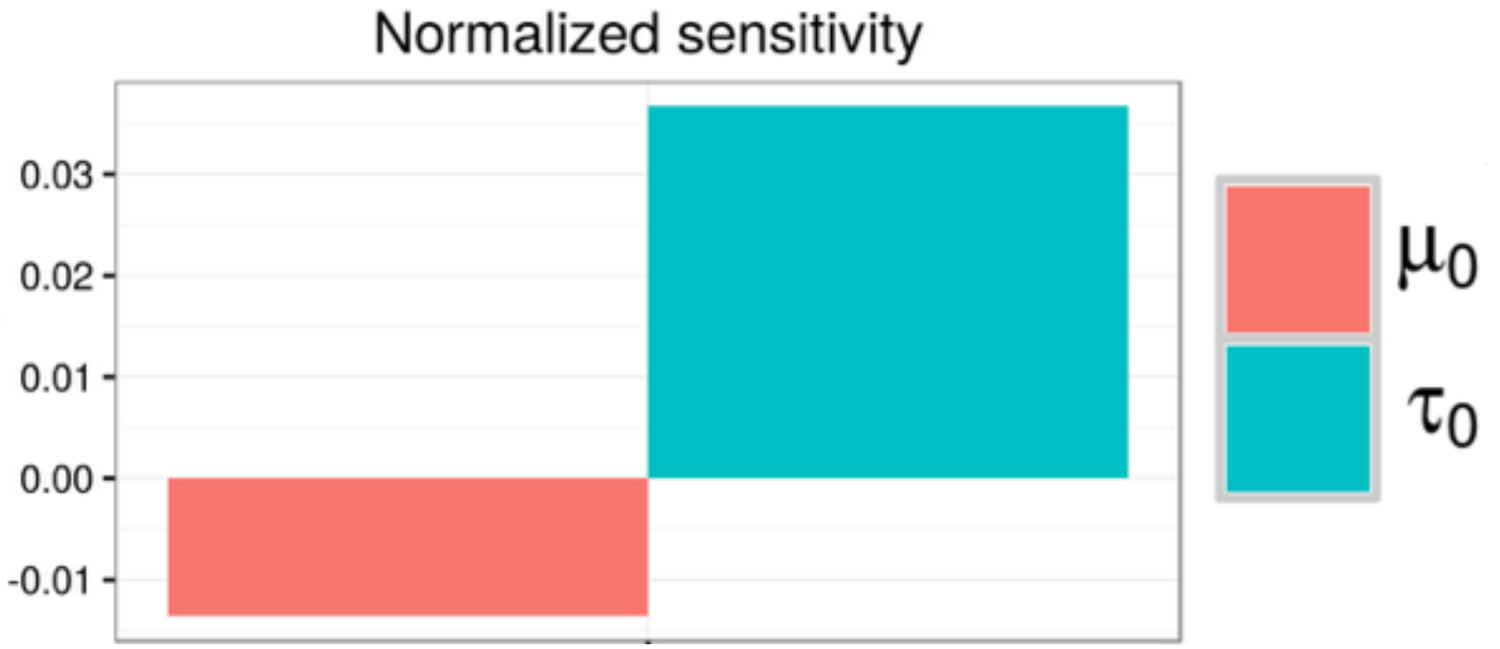
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of τ std devs



Microcredit Experiment

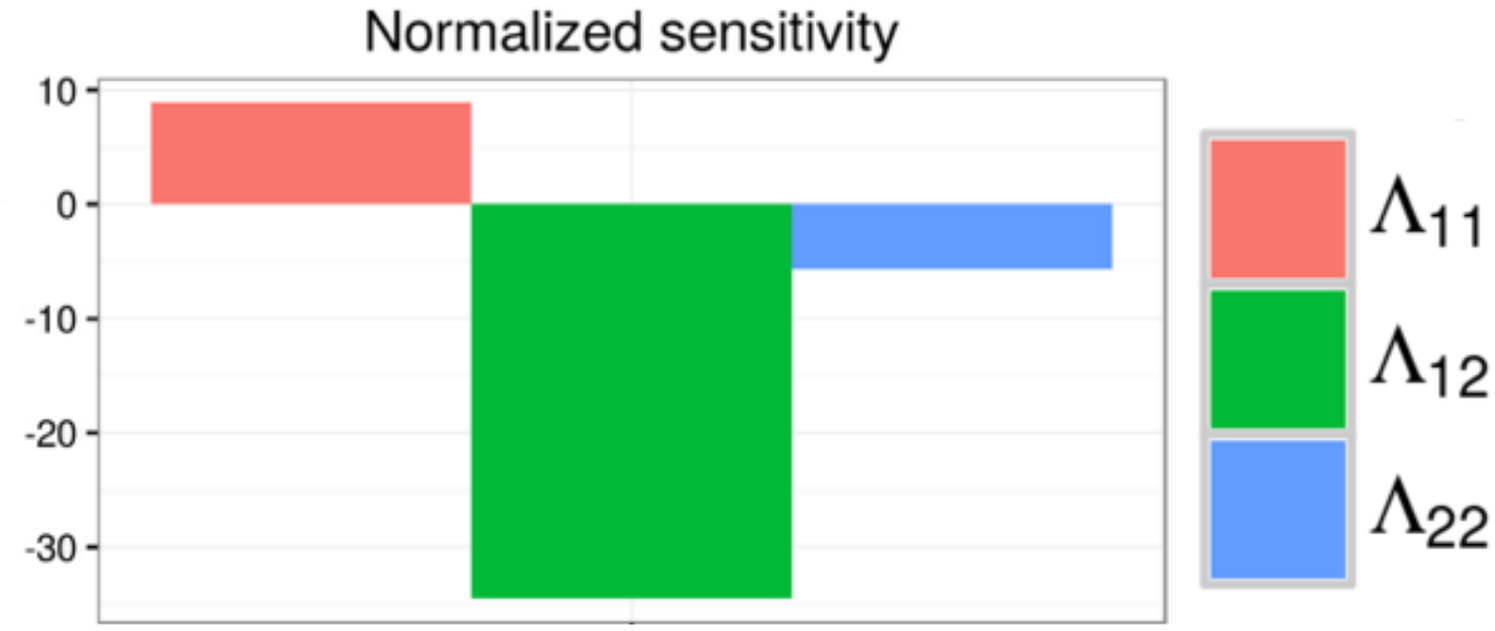
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of τ std devs

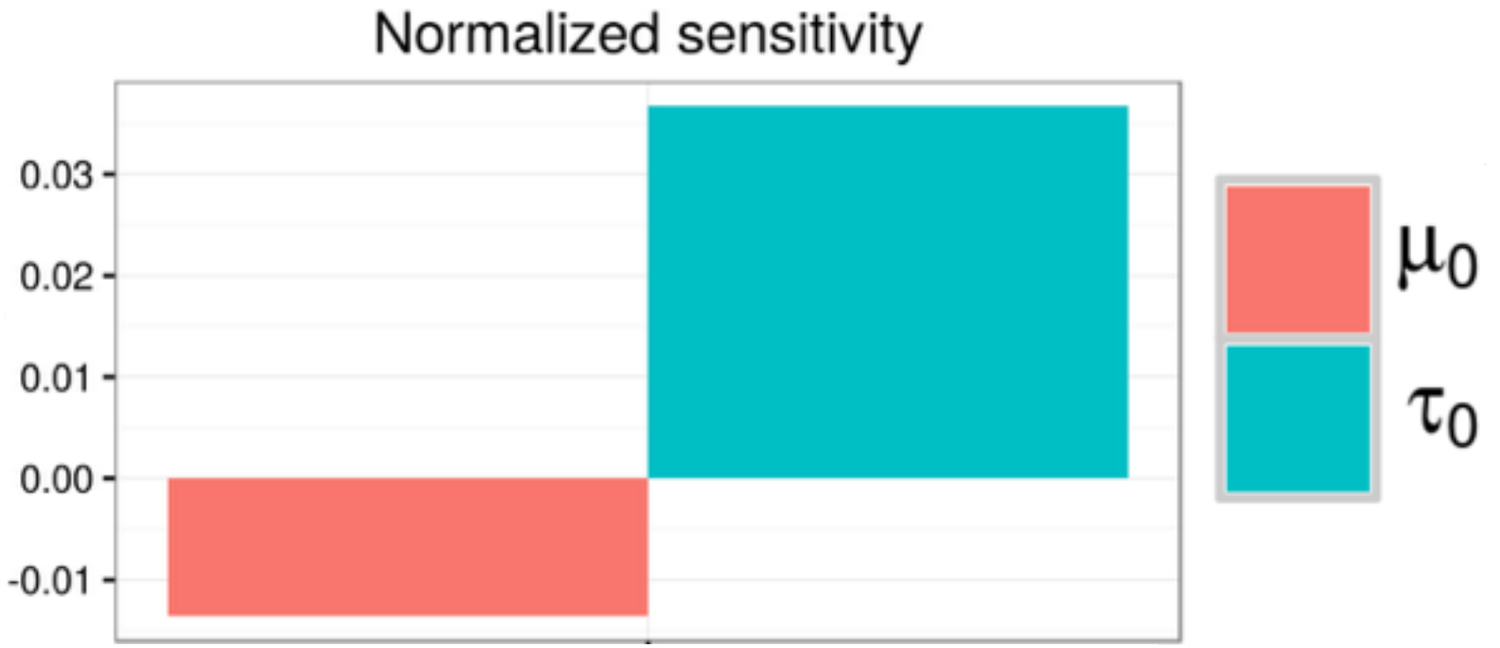




Microcredit Experiment

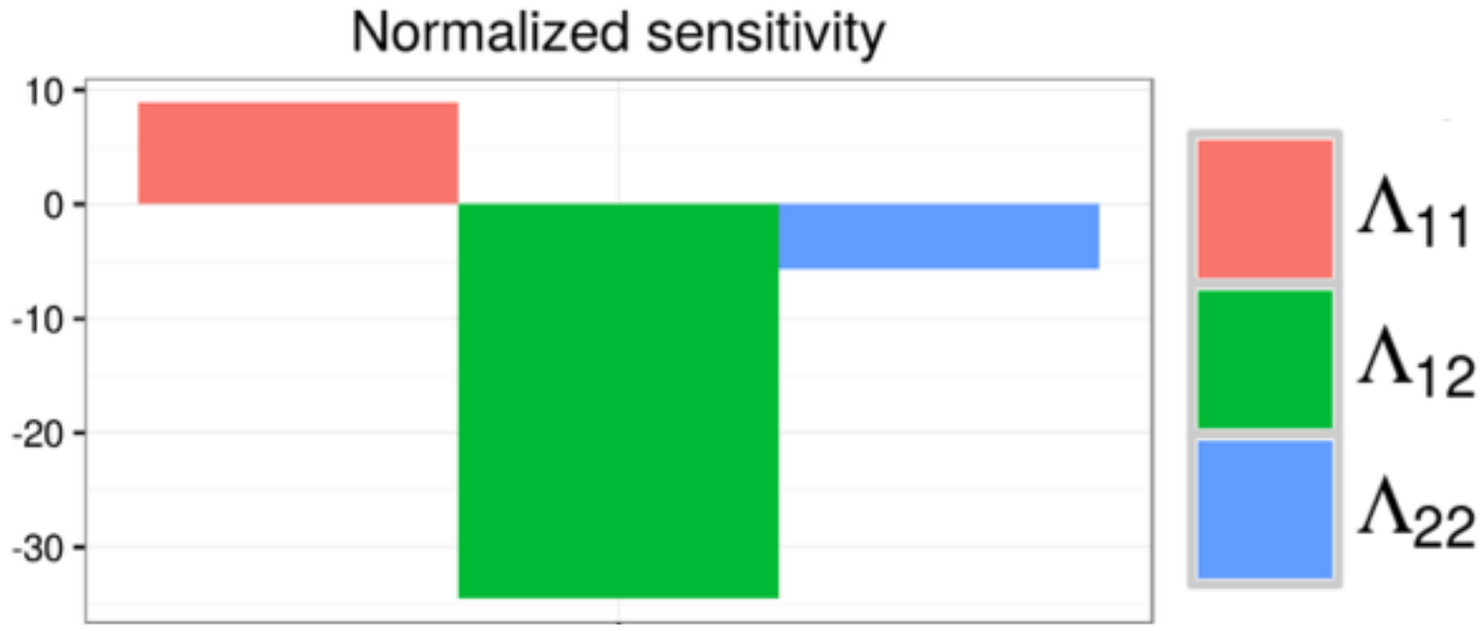
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of τ std devs
- τ mean (MFVB): 3.08 USD PPP
- τ std dev (LRVB): 1.83 USD PPP
- Mean is 1.68 std dev from 0

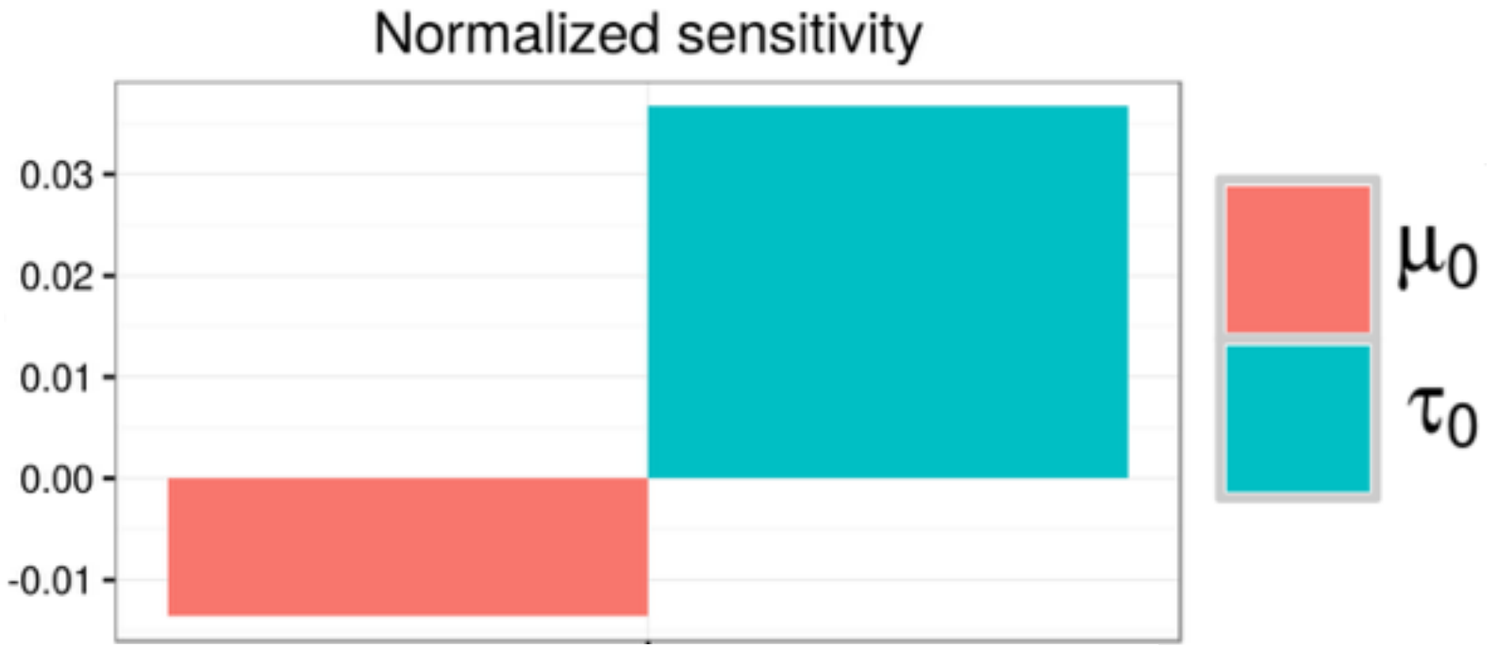




Microcredit Experiment

- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of τ std devs
- τ mean (MFVB): 3.08 USD PPP
- τ std dev (LRVB): 1.83 USD PPP
- Mean is 1.68 std dev from 0
- Λ_{11} += 0.04

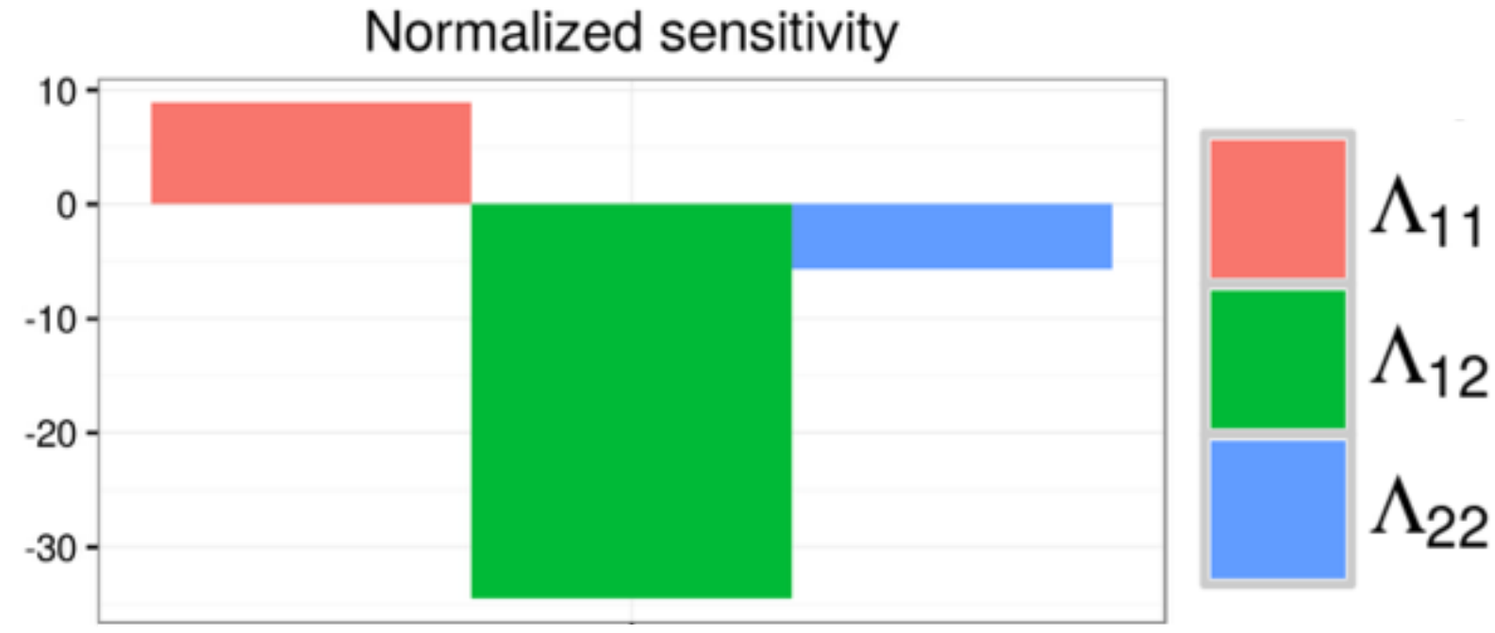


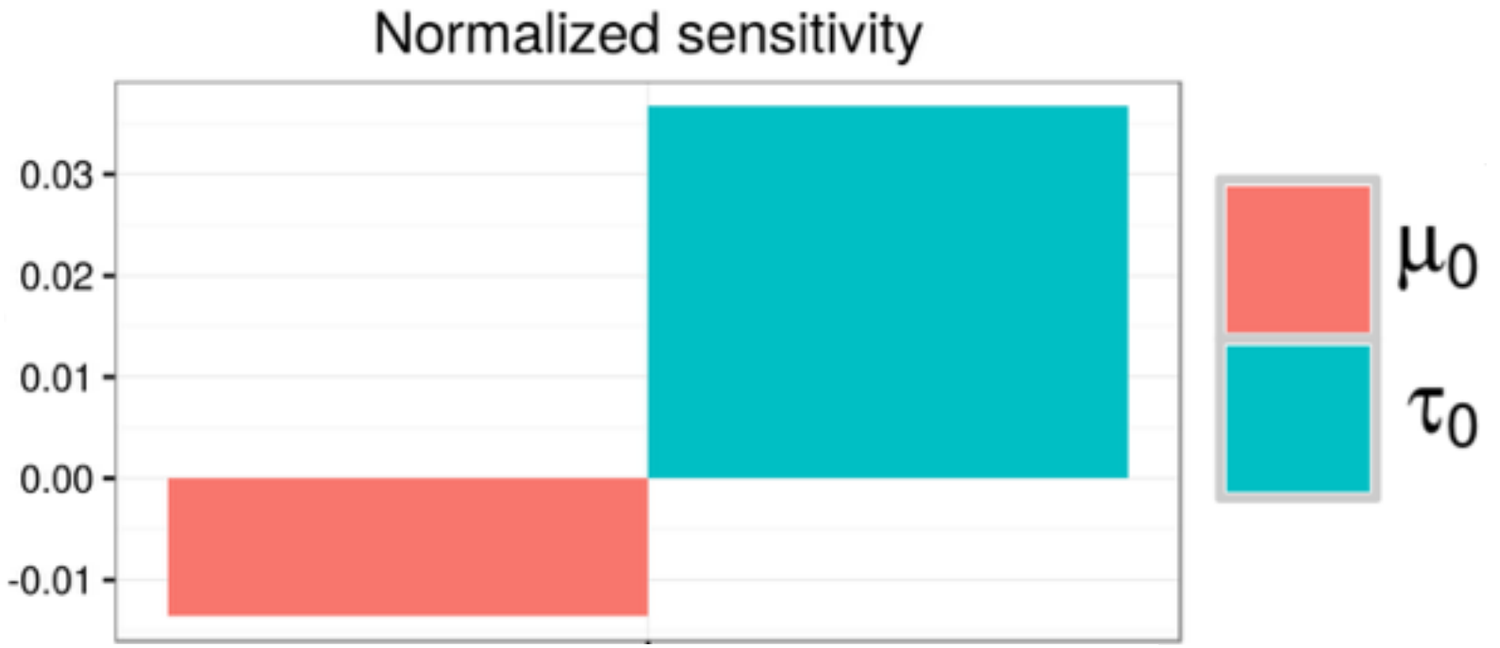


Microcredit Experiment

- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of τ std devs
- τ mean (MFVB): 3.08 USD PPP
- τ std dev (LRVB): 1.83 USD PPP
- Mean is 1.68 std dev from 0
- Λ_{11} += 0.04

⇒ Mean > 2 std dev





Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes, variational Bayes)
  - Uncertainties (Linear response)
- Robustness trade-offs
  - Local robustness quantification
- Theoretical guarantees on quality

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes, variational Bayes)
  - Uncertainties (Linear response)
- Robustness trade-offs
  - Local robustness quantification (Linear response)
- Theoretical guarantees on quality

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes, variational Bayes)
  - Uncertainties (Linear response)
- Robustness trade-offs
  - Local robustness quantification (Linear response)
- Theoretical guarantees on quality

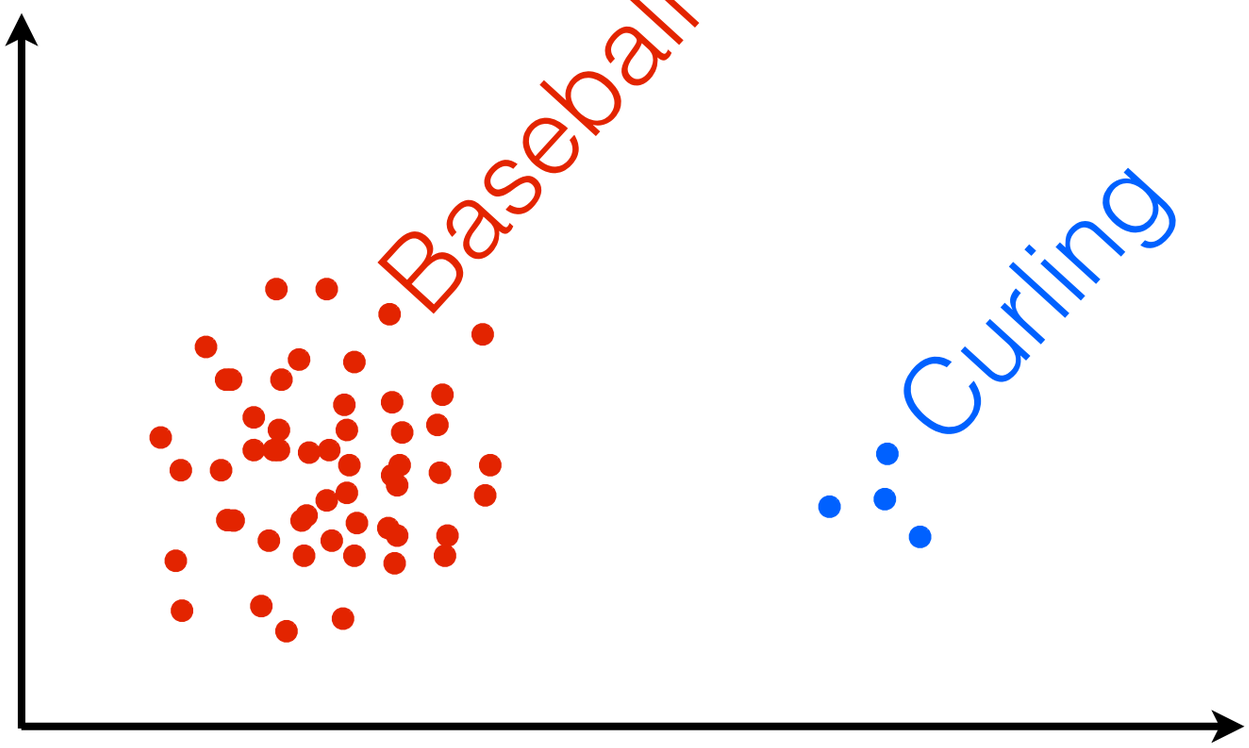
Coresets

- Pre-process data to get a smaller, weighted data set

[Agarwal et al 2005, Feldman & Langberg 2011]

Coresets

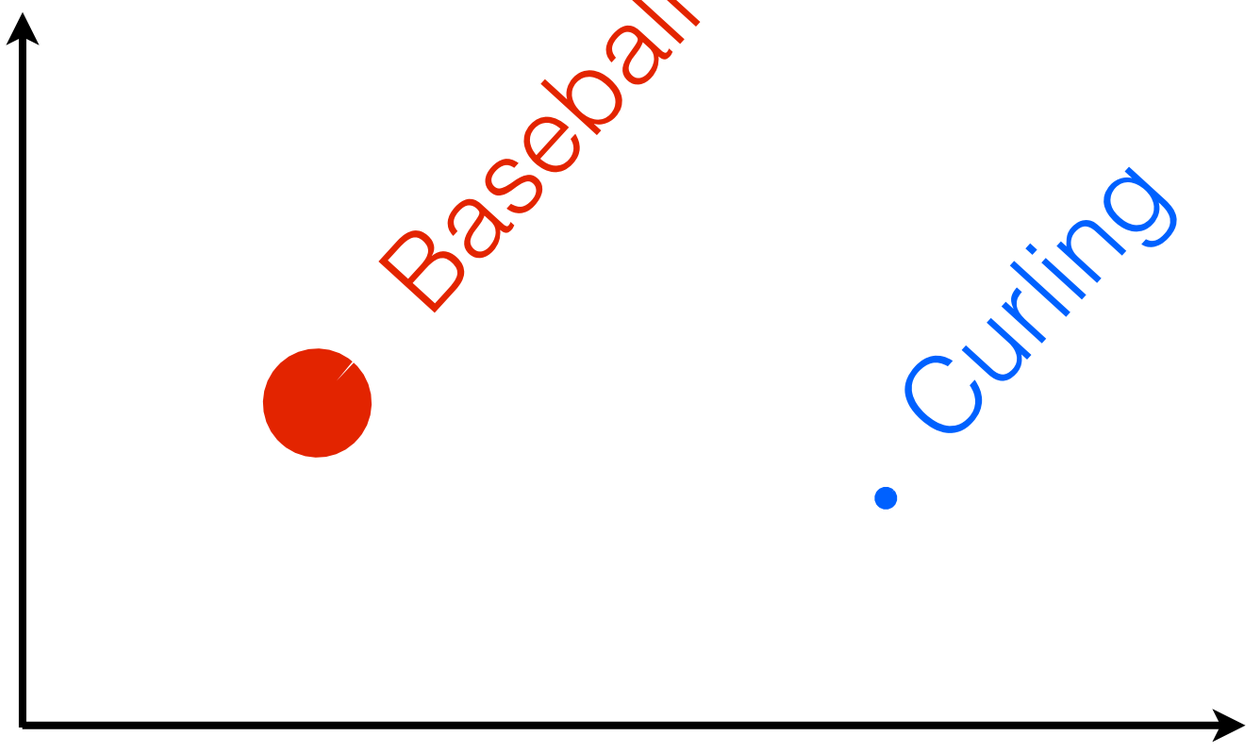
- Pre-process data to get a smaller, weighted data set



[Agarwal et al 2005, Feldman & Langberg 2011]

Coresets

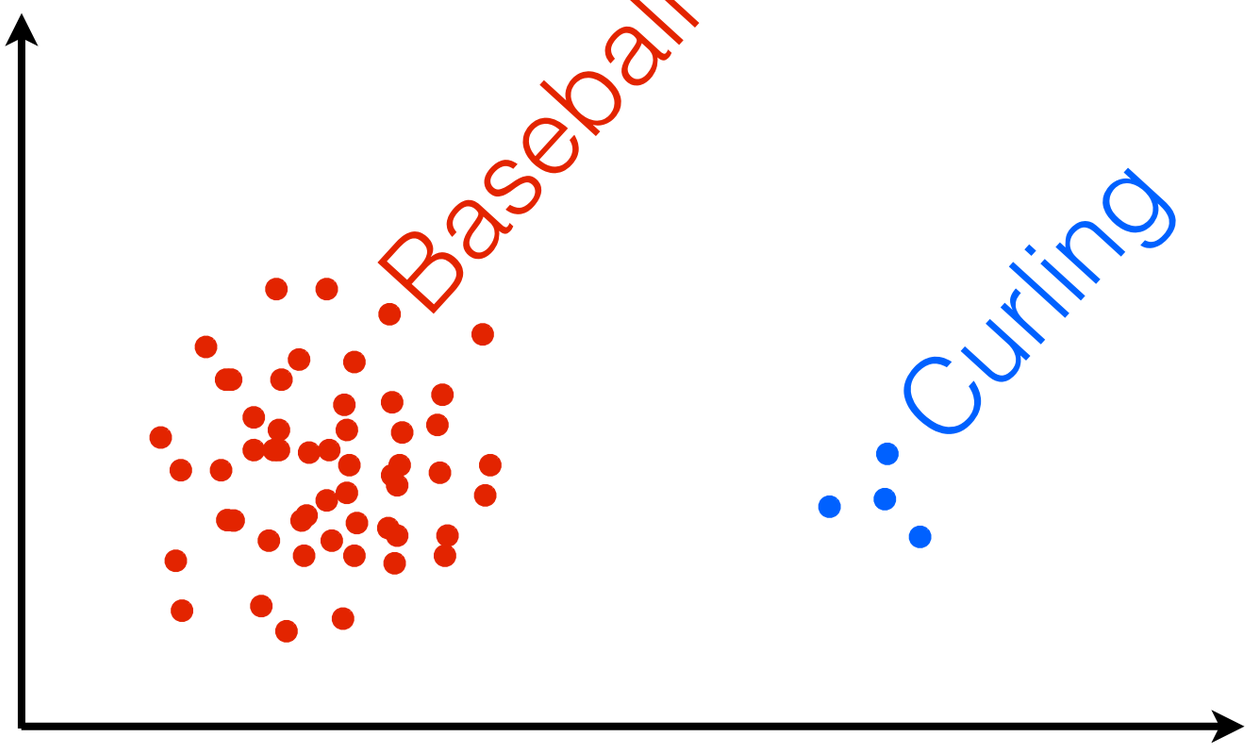
- Pre-process data to get a smaller, weighted data set



[Agarwal et al 2005, Feldman & Langberg 2011]

Coresets

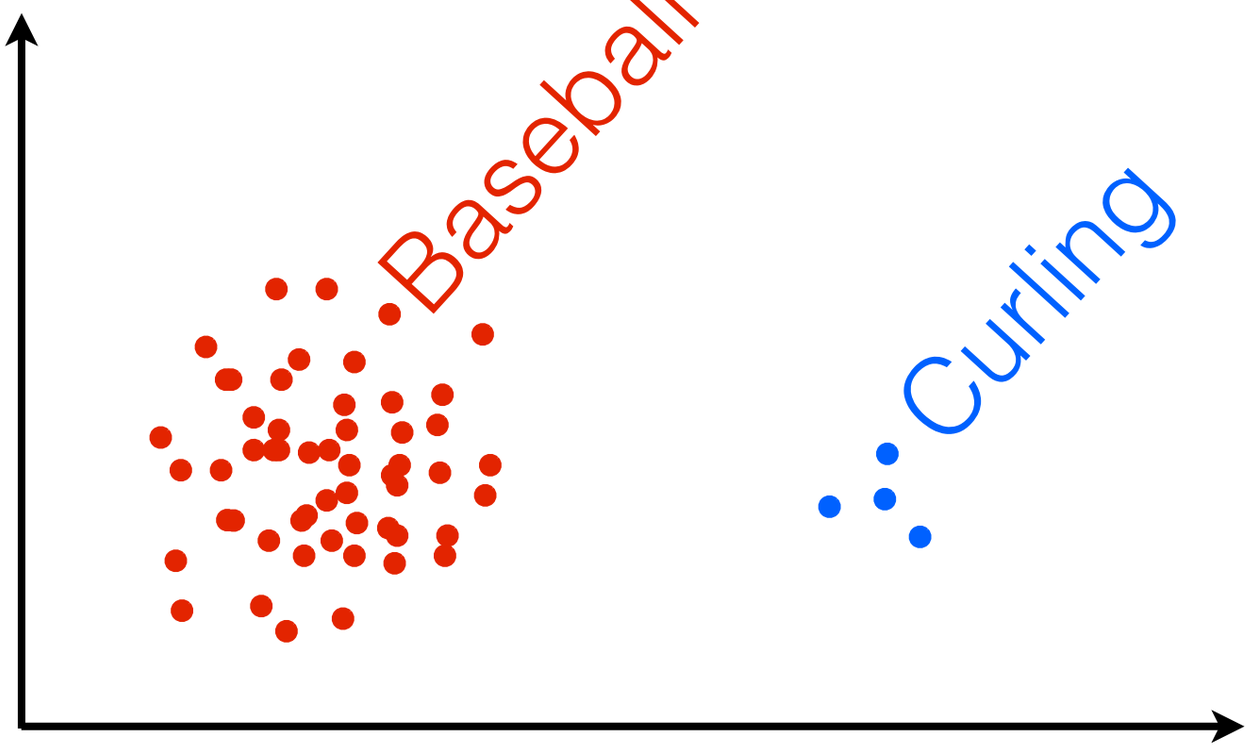
- Pre-process data to get a smaller, weighted data set



[Agarwal et al 2005, Feldman & Langberg 2011]

Coresets

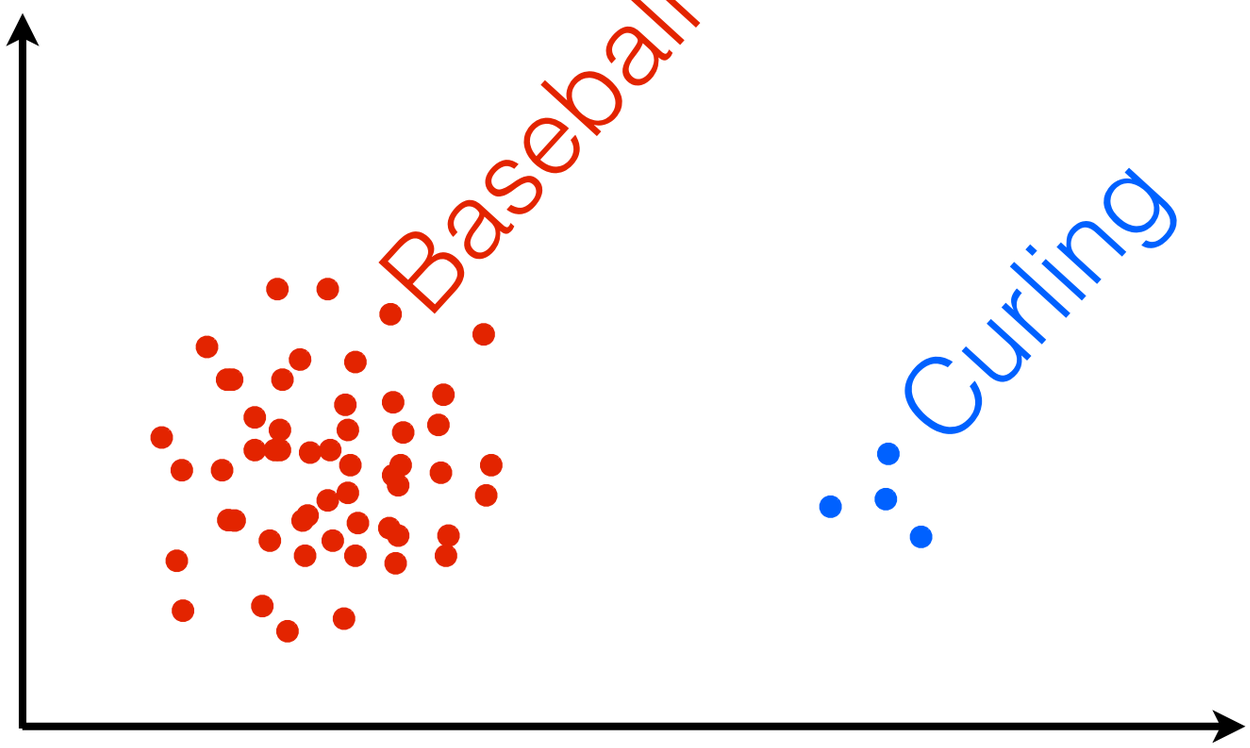
- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality



[Agarwal et al 2005, Feldman & Langberg 2011]

Coresets

- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality

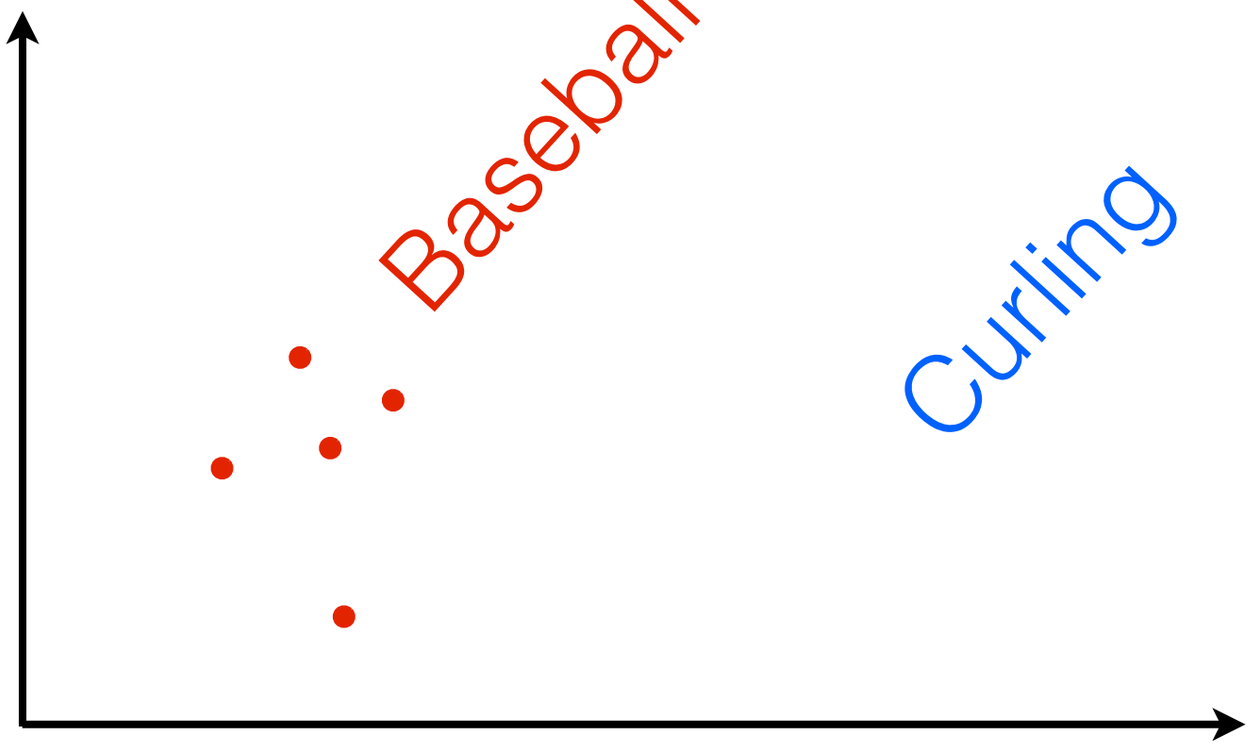


- Cf. data squashing, subsampling

[Agarwal et al 2005, Feldman & Langberg 2011, Madigan et al 2000]

Coresets

- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality

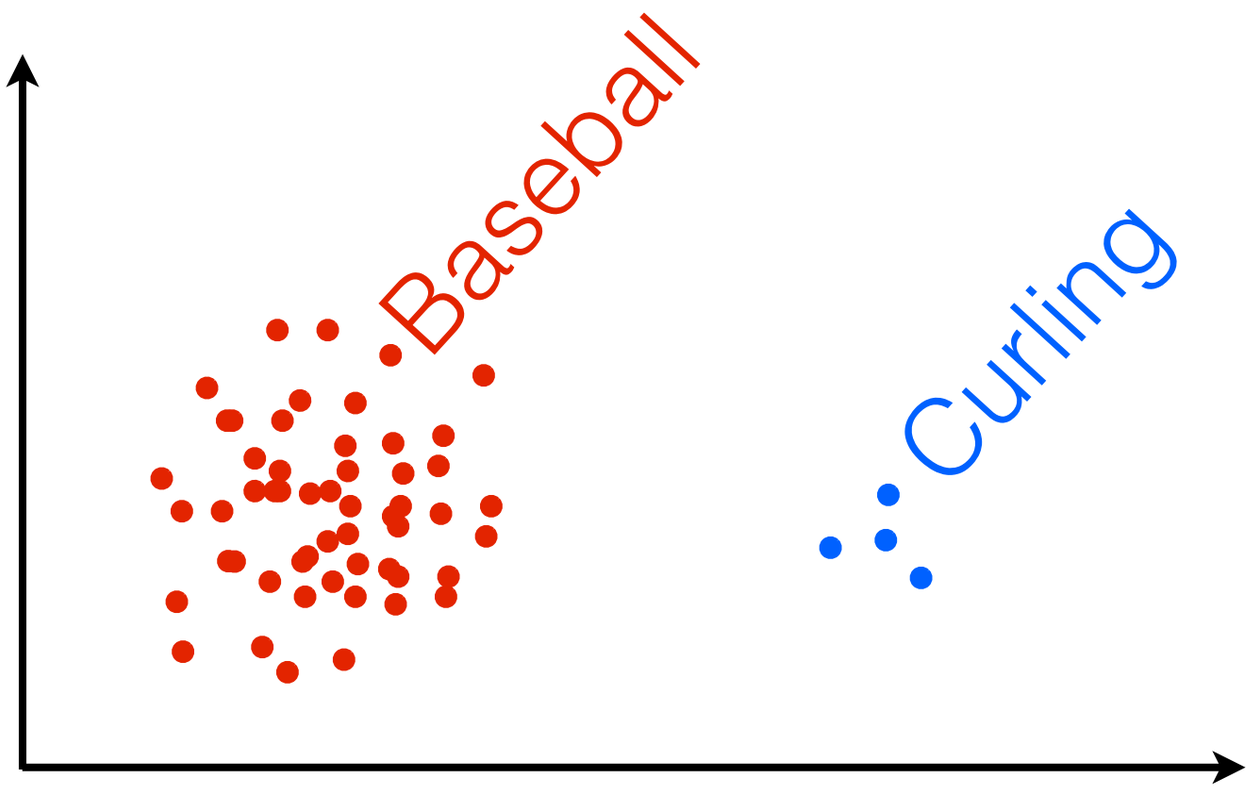


- Cf. data squashing, subsampling

[Agarwal et al 2005, Feldman & Langberg 2011, Madigan et al 2000]

Coresets

- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality

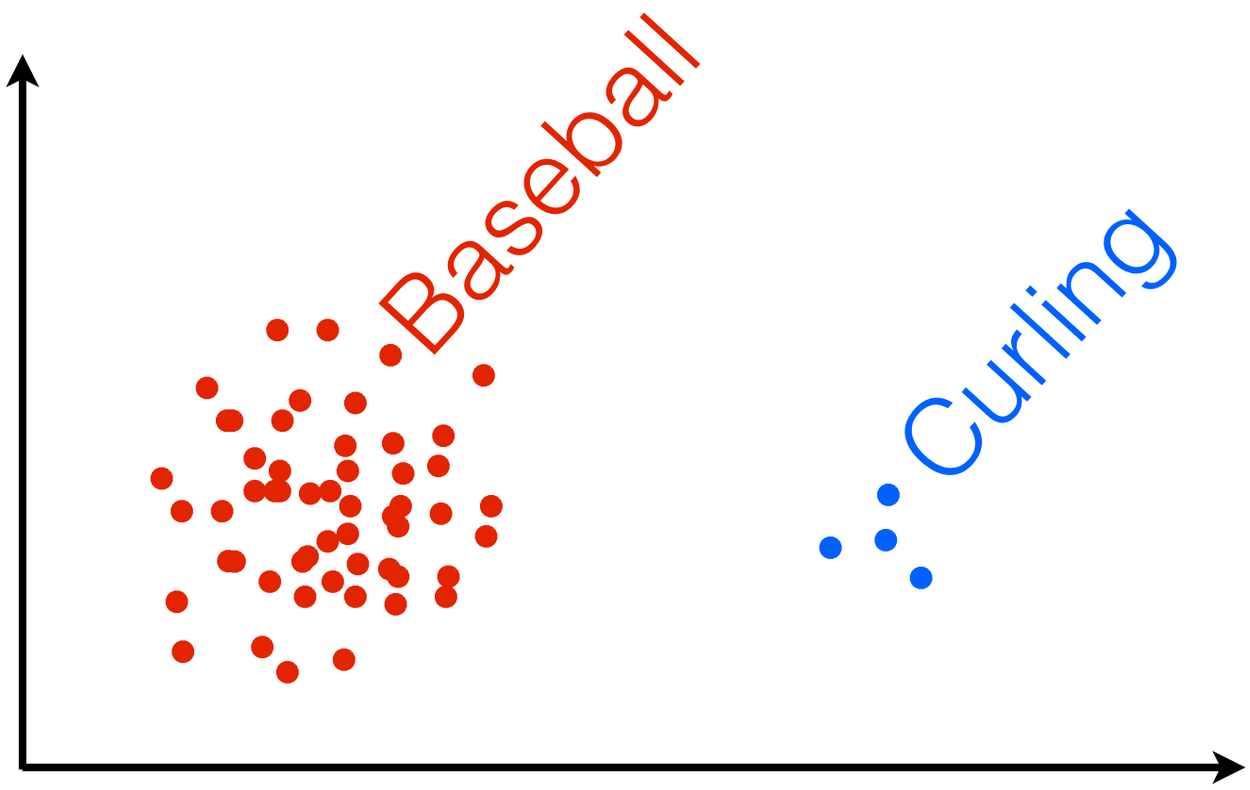


- Cf. data squashing, subsampling
- We develop: coresets for Bayes

[Agarwal et al 2005, Feldman & Langberg 2011, Madigan et al 2000, Huggins, Campbell, Broderick 2016]

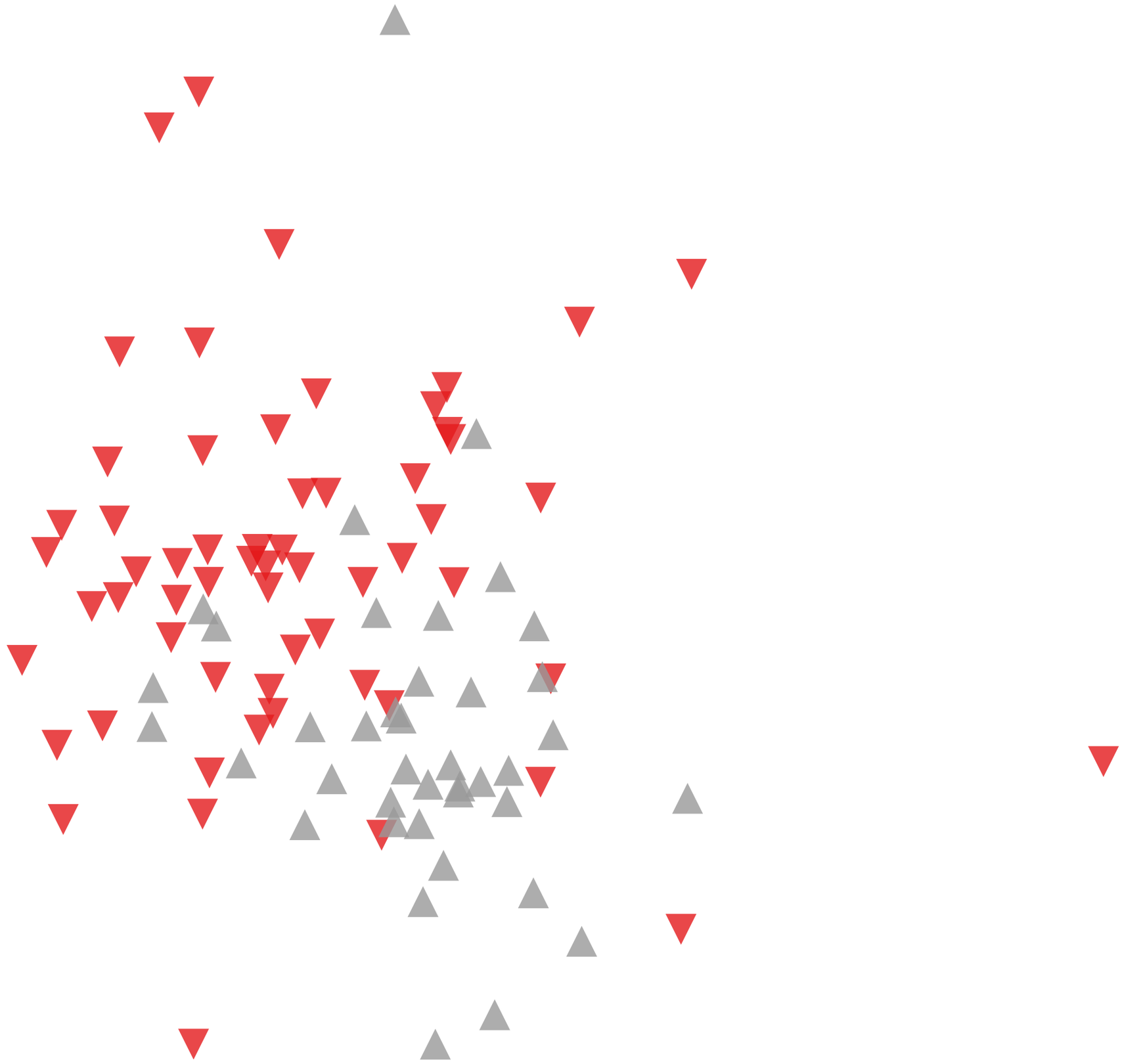
Coresets

- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality

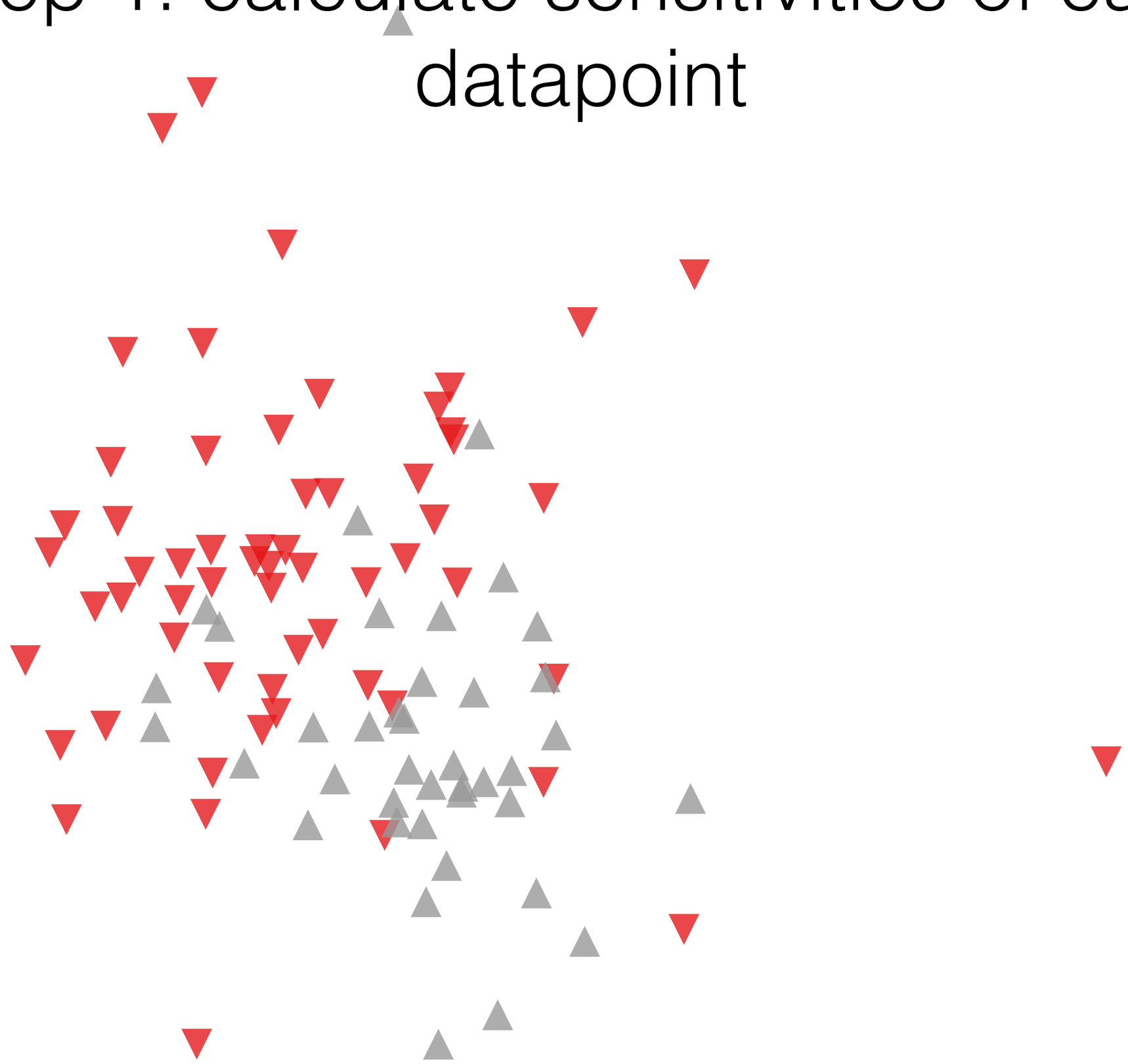


- Cf. data squashing, subsampling
- We develop: coresets for Bayes
  - Logistic regression

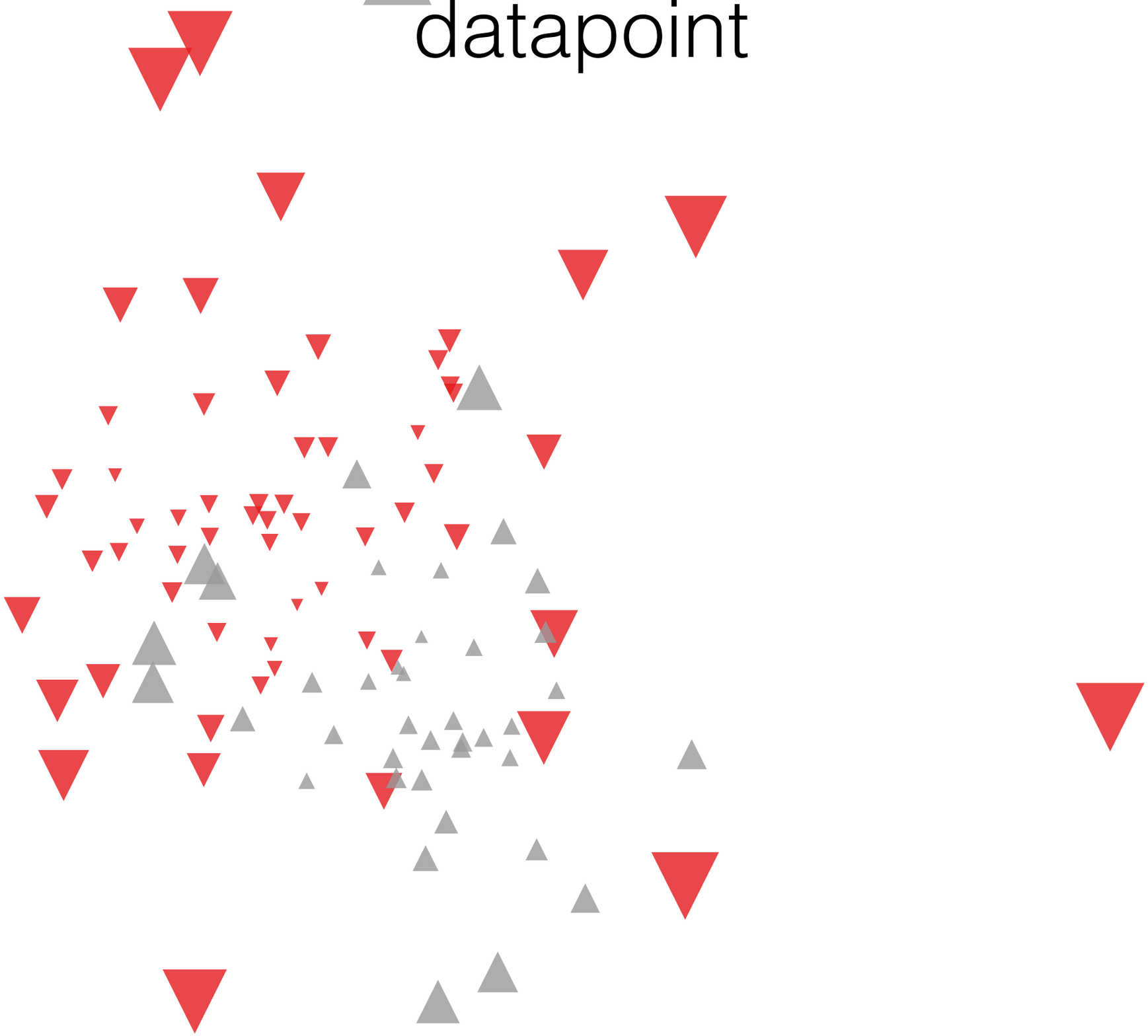
[Agarwal et al 2005, Feldman & Langberg 2011, Madigan et al 2000; Huggins, Campbell, Broderick 2016]



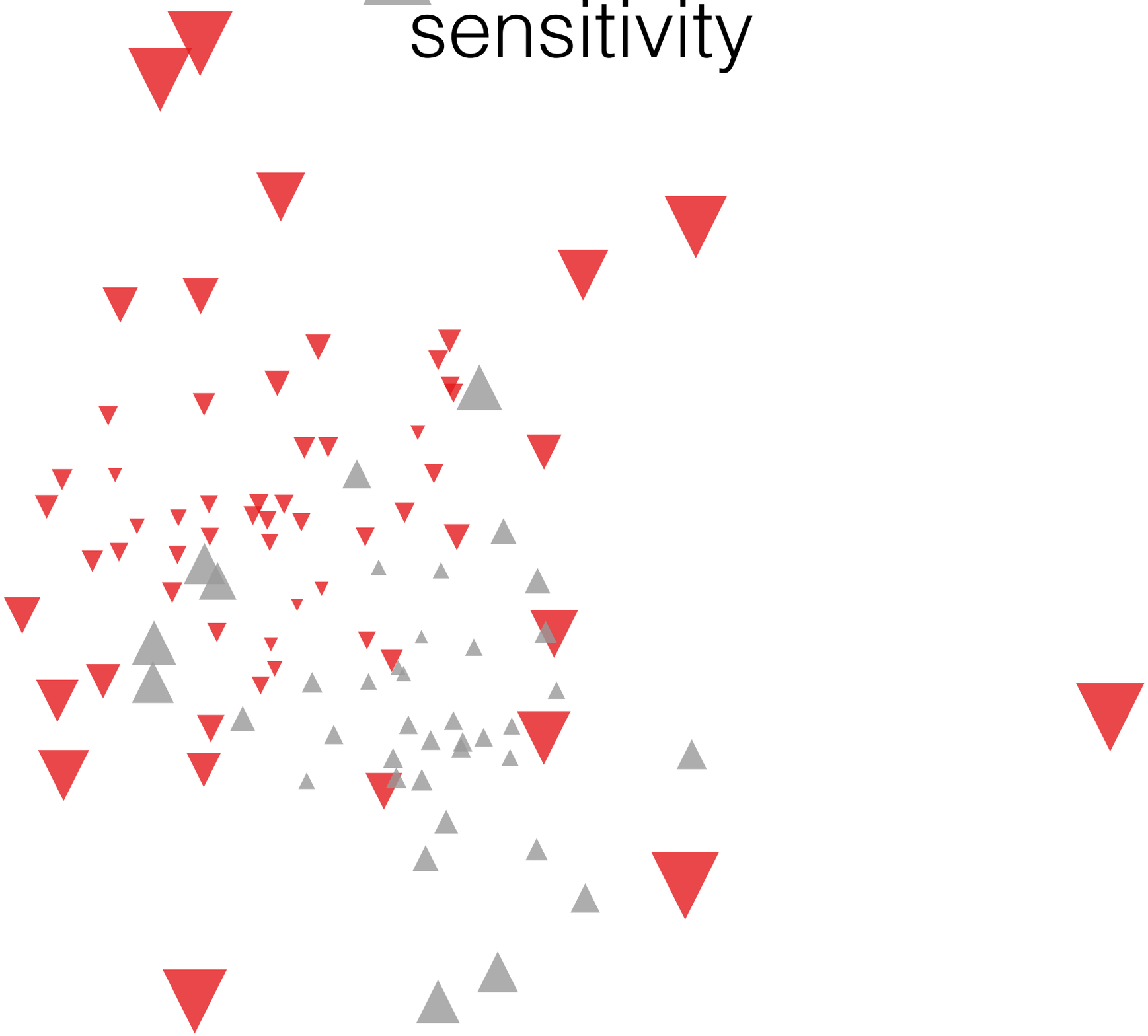
Step 1: calculate sensitivities of each datapoint



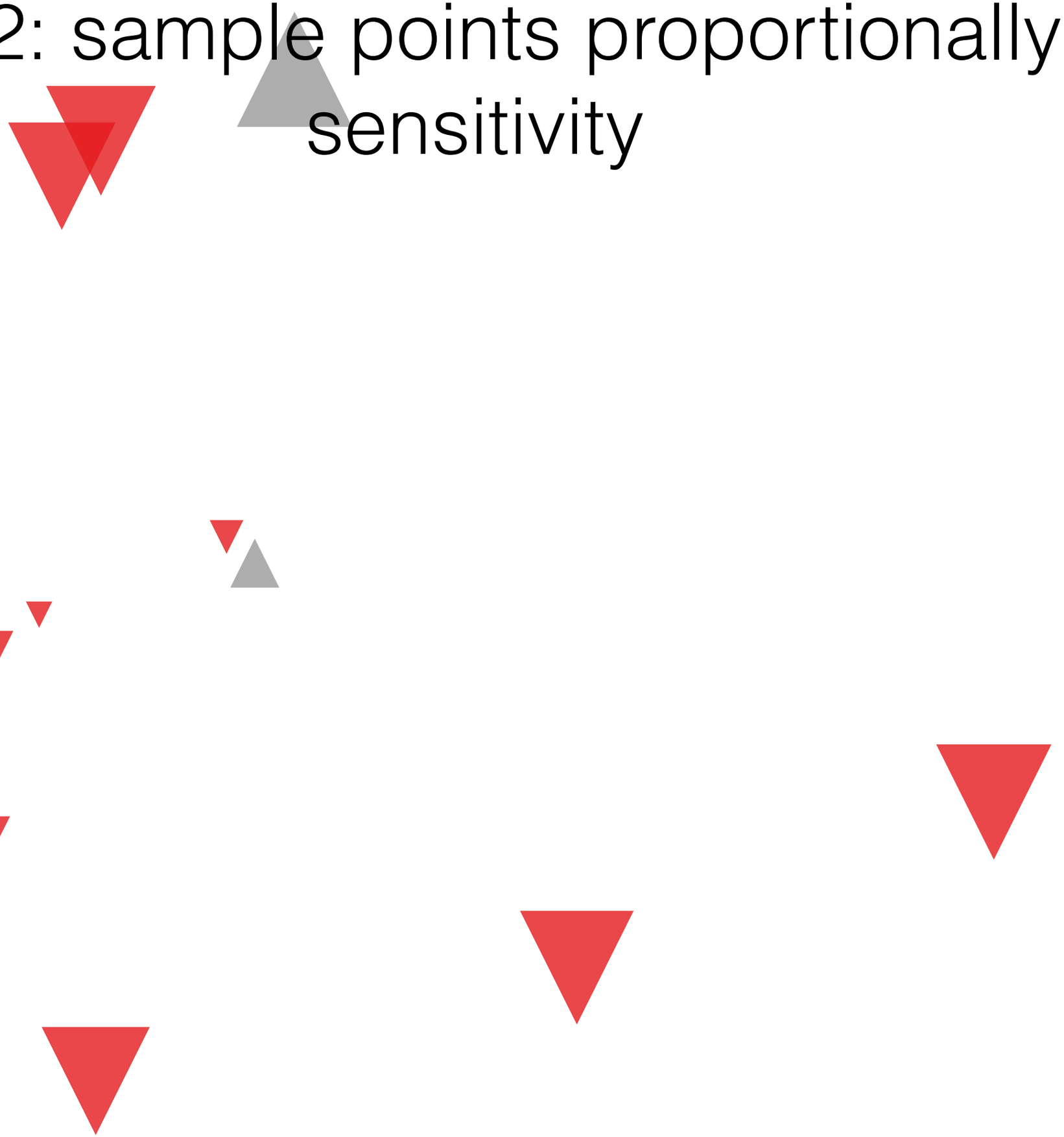
Step 1: calculate sensitivities of each datapoint



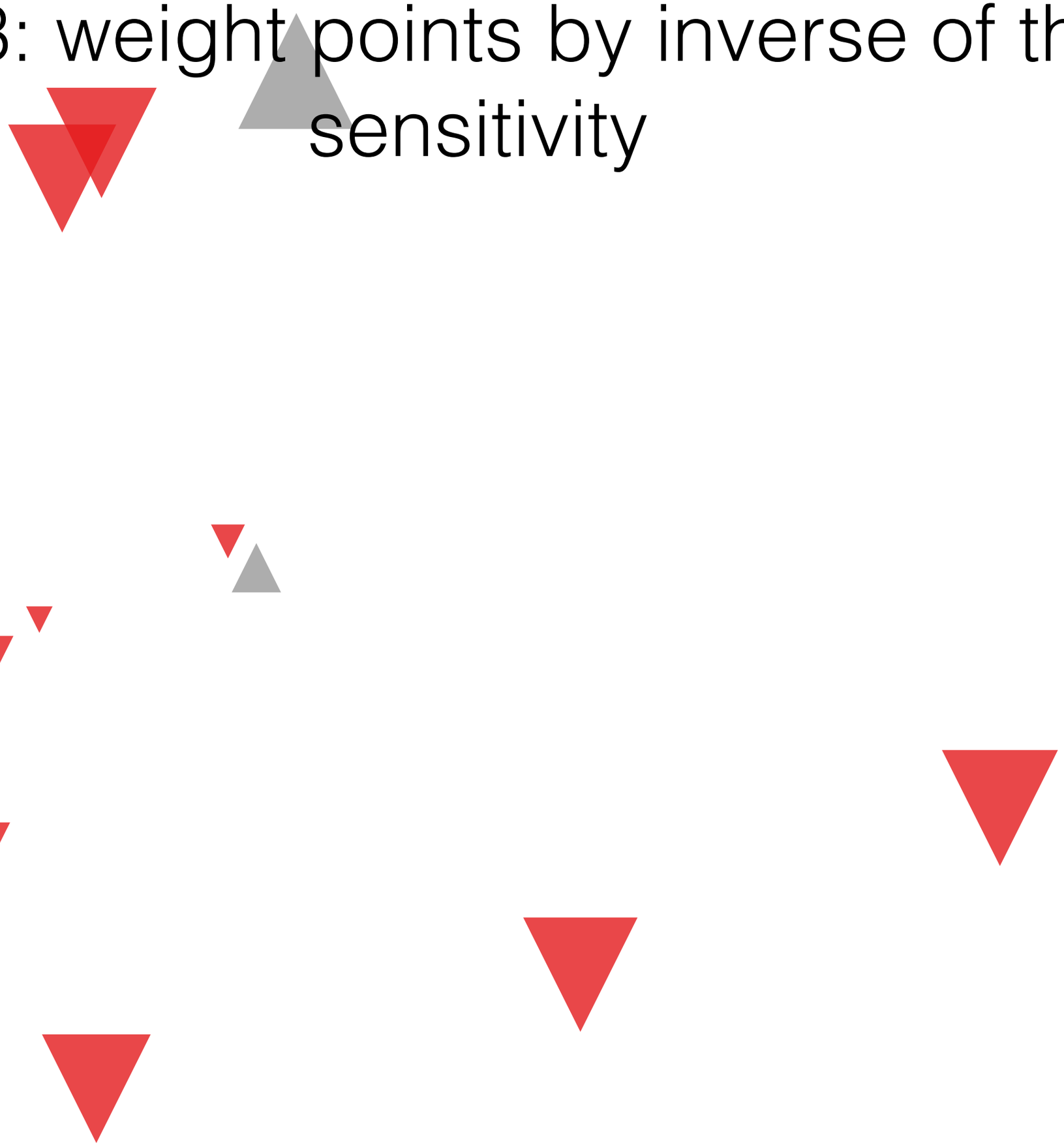
Step 2: sample points proportionally to sensitivity



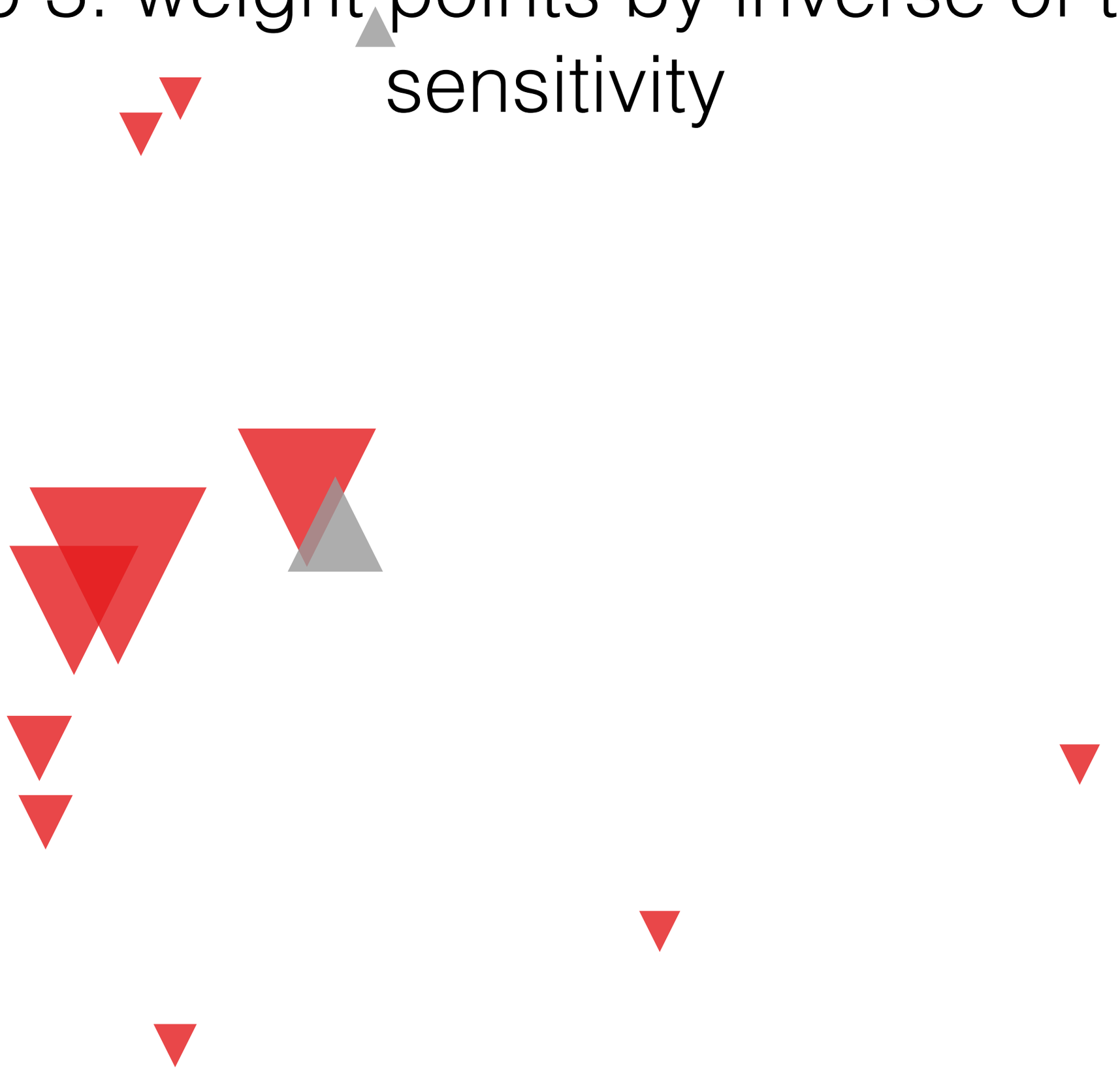
Step 2: sample points proportionally to sensitivity



Step 3: weight points by inverse of their sensitivity



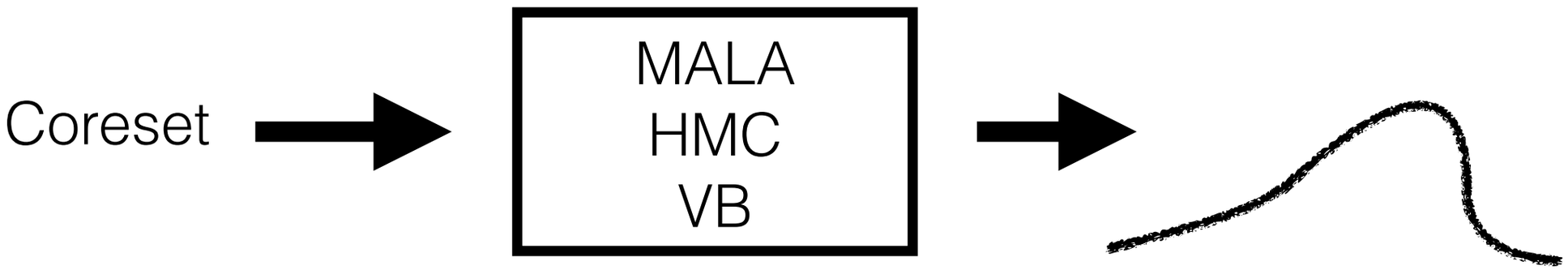
Step 3: weight points by inverse of their sensitivity



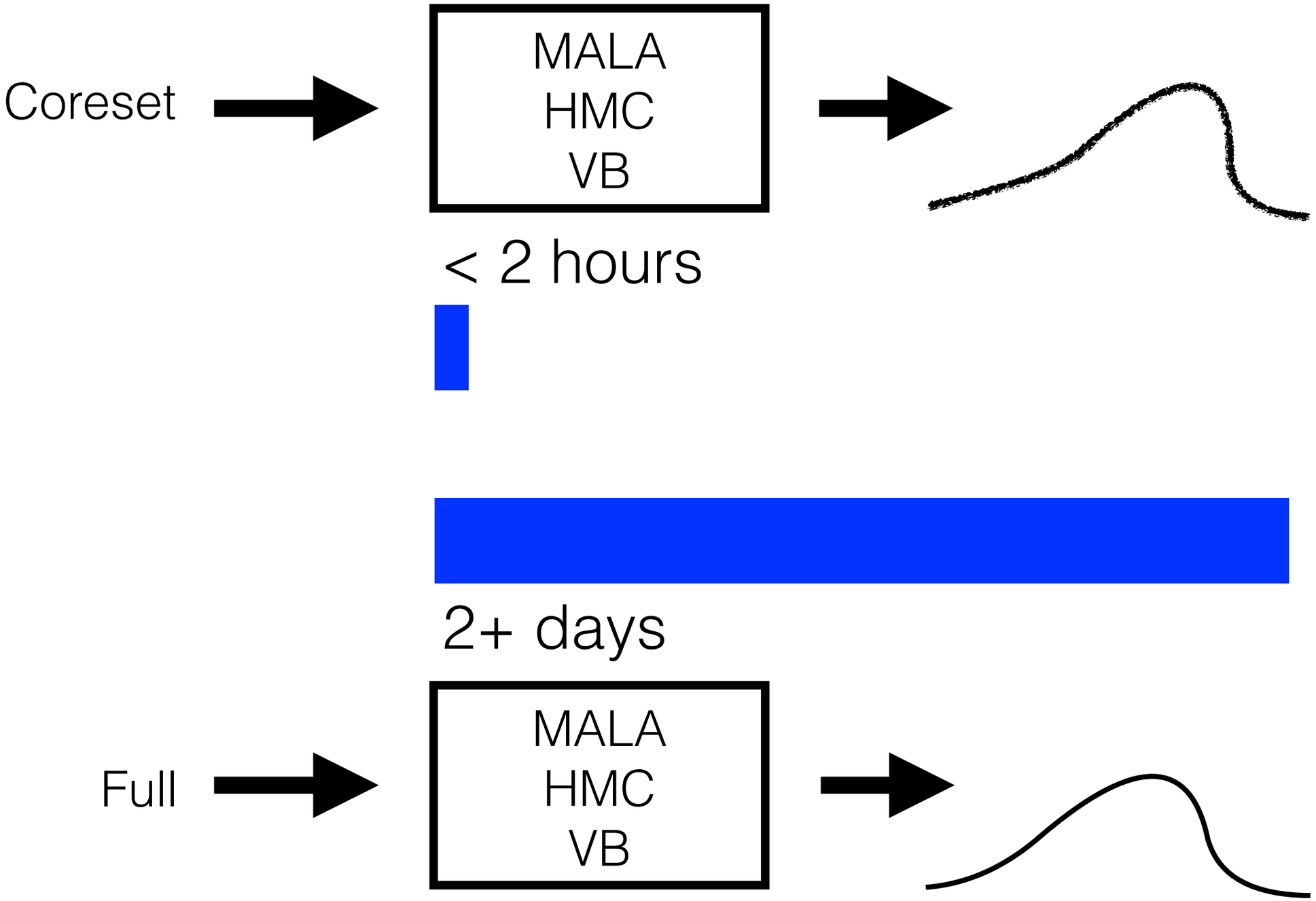
Step 4: input weighted points to existing approximate posterior algorithm

Step 4: input weighted points to existing approximate posterior algorithm

Step 4: input weighted points to existing approximate posterior algorithm



Step 4: input weighted points to existing approximate posterior algorithm



Results

Results

Thm sketch (HCB). Choose ε > 0, δ ∈ (0,1). Our algorithm runs in *O*(*N*) time and creates $\text{coreset-size} \sim \text{const} \cdot \epsilon^{-2} + \log(1/\delta)$

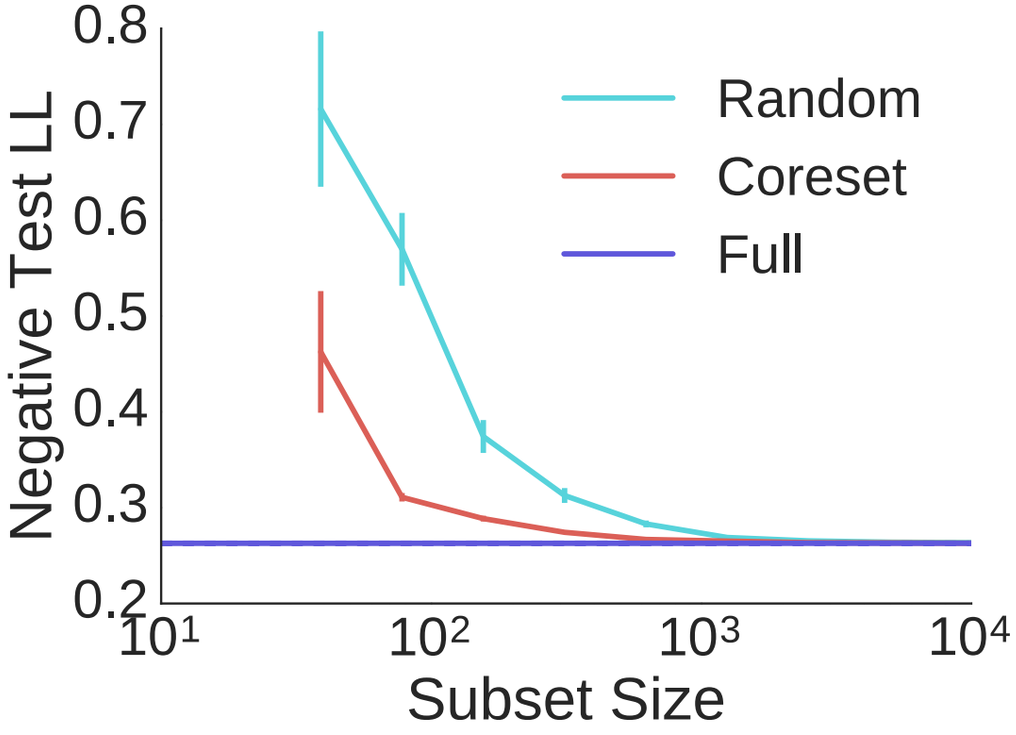
W.p. 1 - δ, it constructs a coreset with $\left|\ln \mathcal{E} - \ln \tilde{\mathcal{E}}\right| \leq \epsilon \left|\ln \mathcal{E}\right|$

Results

Thm sketch (HCB). Choose ε > 0, δ ∈ (0,1). Our algorithm runs in *O*(*N*) time and creates $\text{coreset-size} \sim \text{const} \cdot \epsilon^{-2} + \log(1/\delta)$

W.p. 1 - δ, it constructs a coreset with $\left|\ln \mathcal{E} - \ln \tilde{\mathcal{E}}\right| \leq \epsilon \left|\ln \mathcal{E}\right|$

Synthetic Binary10 Data

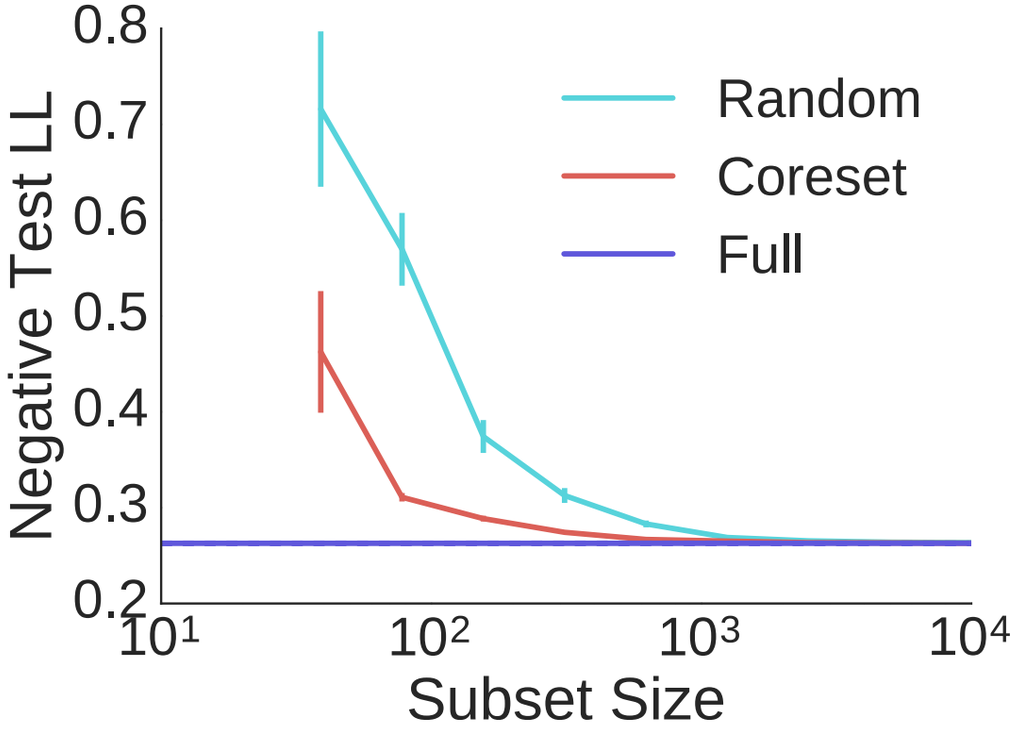


Results

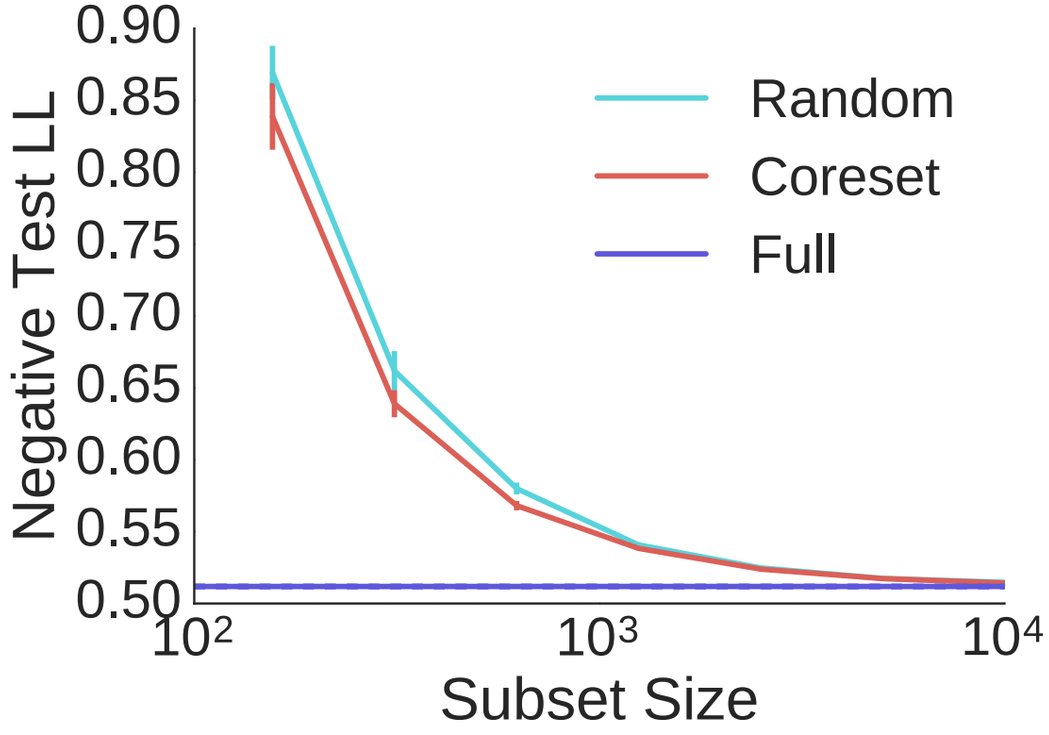
Thm sketch (HCB). Choose ε > 0, δ ∈ (0,1). Our algorithm runs in *O*(*N*) time and creates $\text{coreset-size} \sim \text{const} \cdot \epsilon^{-2} + \log(1/\delta)$

W.p. 1 - δ, it constructs a coreset with $\left|\ln \mathcal{E} - \ln \tilde{\mathcal{E}}\right| \leq \epsilon \left|\ln \mathcal{E}\right|$

Synthetic Binary10 Data



Covertype Data

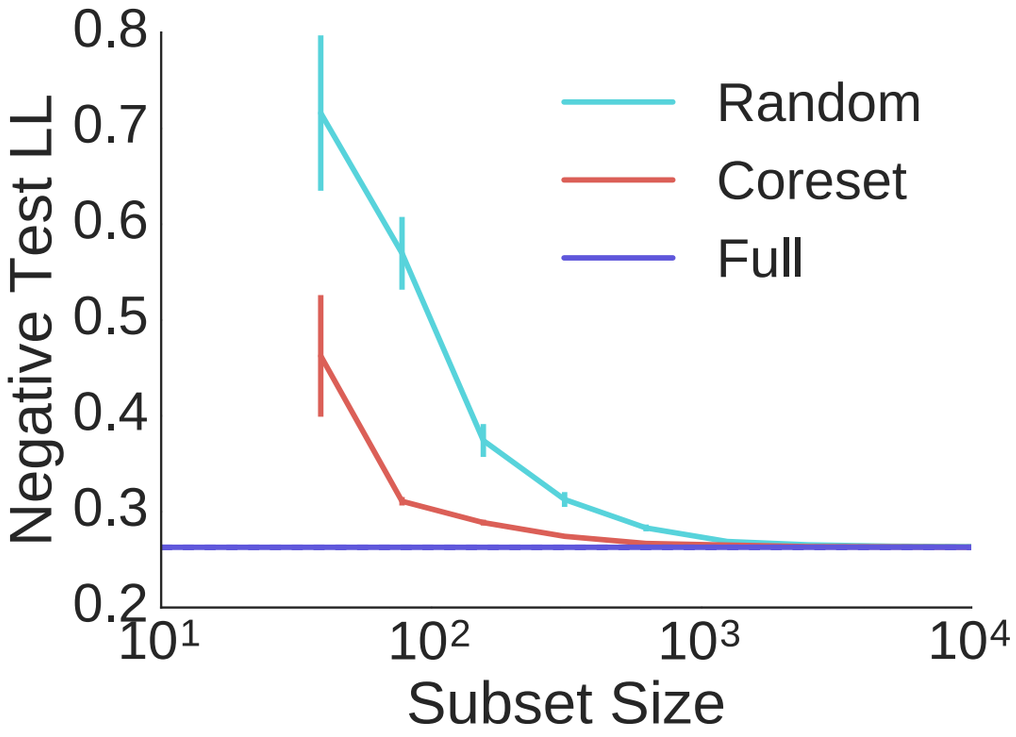


Results

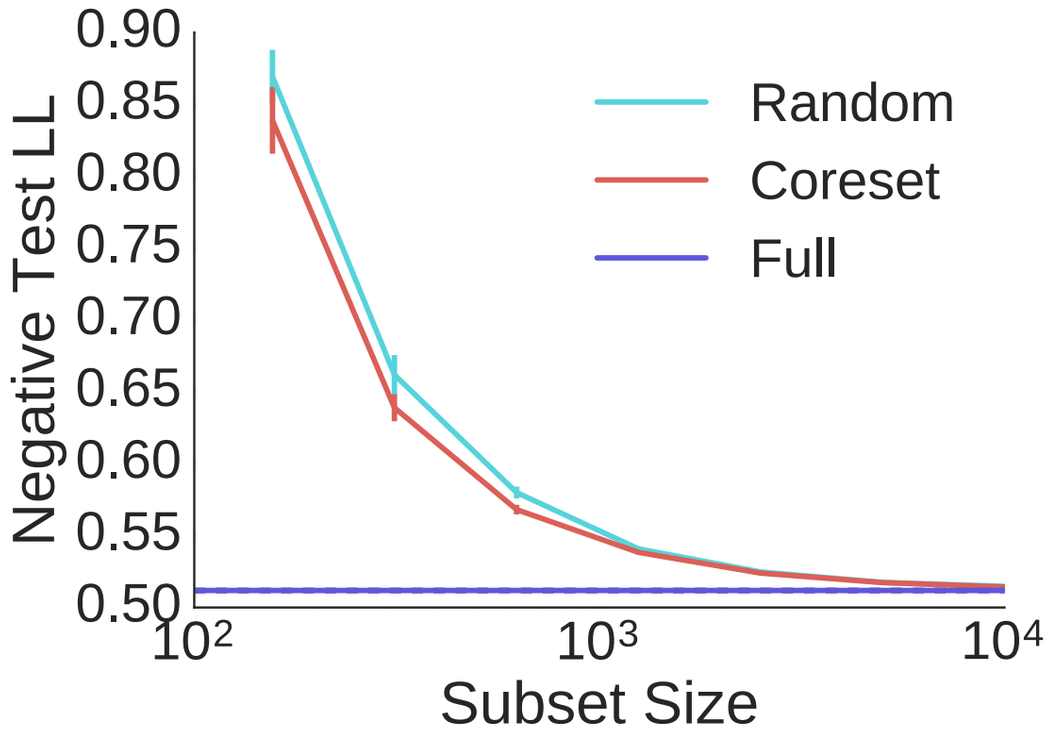
Thm sketch (HCB). Choose ε > 0, δ ∈ (0,1). Our algorithm runs in *O*(*N*) time and creates $\text{coreset-size} \sim \text{const} \cdot \epsilon^{-2} + \log(1/\delta)$

W.p. 1 - δ, it constructs a coreset with $\left|\ln \mathcal{E} - \ln \tilde{\mathcal{E}}\right| \leq \epsilon \left|\ln \mathcal{E}\right|$

Synthetic Binary10 Data



Covertype Data



Can quantify the propagation of error in streaming and parallel settings.

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes, variational Bayes)
  - Uncertainties (Linear response)
- Robustness trade-offs
  - Local robustness quantification (Linear response)
- Theoretical guarantees on quality

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes, variational Bayes)
  - Uncertainties (Linear response)
- Robustness trade-offs
  - Local robustness quantification (Linear response)
- Theoretical guarantees on quality (coresets)

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes, variational Bayes)
  - Uncertainties (Linear response)
- Robustness trade-offs
  - Local robustness quantification (Linear response)
  - Theoretical guarantees on quality (coresets)

Roadmap

- Posterior approximation trade-offs
  - Point estimates (MAD-Bayes, variational Bayes)
  - Uncertainties (Linear response)
- Robustness trade-offs
  - Local robustness quantification (Linear response)
- Theoretical guarantees on quality (coresets)

- Interested in your data and models
  - Point estimates, uncertainty quantification, robustness quantification
  - Bayesian logistic regression; up next: high-dimensional, other models

References (page 1 of 3)

T Broderick, B Kulis, and MI Jordan. MAD-Bayes: MAP-based asymptotic derivations from Bayes. In *ICML*, 2013.

T Broderick, N Boyd, A Wibisono, AC Wilson, and MI Jordan. Streaming variational Bayes. In *NIPS*, 2013.

T Campbell*, J H Huggins, J How, and T Broderick. Truncated random measures. *Under review*. ArXiv:1603.00861.

R Giordano, T Broderick, and MI Jordan. Linear response methods for accurate covariance estimates from mean field variational Bayes. In *NIPS*, 2015.

R Giordano, T Broderick, R Meager, JH Huggins, and MI Jordan. Fast robustness quantification with variational Bayes. *ICML Workshop on #Data4Good: Machine Learning in Social Good Applications*, 2016. ArXiv:1606.07153.

F Guo, X Wang, K Fan, T Broderick, and D Dunson. Boosting variational inference. *NIPS Workshop on Advances in Approximate Bayesian Inference*, 2016. ArXiv:1611.05559.

JH Huggins, T Campbell, and T Broderick. Coresets for scalable Bayesian logistic regression. In *NIPS*, 2016.

X Pan, JE Gonzalez, S Jegelka, T Broderick, and MI Jordan. Optimistic concurrency control for distributed unsupervised learning. In *NIPS*, 2013.

Tutorials: www.tamarabroderick.com/tutorials.html

References (page 2 of 3)

PK Agarwal, S Har-Peled, and KR Varadarajan. Geometric approximation via coresets. *Combinatorial and computational geometry*, 2005.

R Bardenet, A Doucet, and C Holmes. On Markov chain Monte Carlo methods for tall data. arXiv, 2015.

CM Bishop. *Pattern Recognition and Machine Learning*, 2006.

RA Burrell, N McGranahan, J Bartek, and C Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 2013.

D Dunson. Robust and scalable approach to Bayesian inference. Talk at *ISBA* 2014.

D Feldman and M Langberg. A unified framework for approximating and clustering data. In *Symposium on Theory of Computing*, 2011.

B Fosdick. *Modeling Heterogeneity within and between Matrices and Arrays*, Chapter 4.7. PhD Thesis, University of Washington, 2013.

PJ Green. MAD-Bayes matching and alignment for labelled and unlabelled configurations. In IL Dryden and JT Kent, editors, *Geometry Driven Statistics*, 2015.

TL Griffiths and Z Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2005.

JH Huggins, K Narasimhan, A Saeedi, VK Mansinghka. JUMP-means: Small-variance asymptotics for Markov jump processes. In *ICML*, 2015.

DJC MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

D Madigan, N Raghavan, W Dumouchel, M Nason, C Posse, and G Ridgeway. Likelihood-based data squashing: A modeling approach to instance construction. *Data Mining and Knowledge Discovery*, 2002.

References (page 3 of 3)

R Meager. Understanding the impact of microcredit expansions: A Bayesian hierarchical analysis of 7 randomised experiments. ArXiv:1506.06669, 2015.

J Straub, T Campbell, JP How, and JW Fisher. Small-variance nonparametric clustering on the hypersphere. In *CVPR*, 2015.

RE Turner and M Sahani. Two problems with variational expectation maximisation for time-series models. In D Barber, AT Cemgil, and S Chiappa, editors, *Bayesian Time Series Models*, 2011.

B Wang and M Titterington. Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *AISTATS*, 2004.

Y Xu, P Müller, Y Yuan, K Gulukota, and Y Ji. MAD Bayes for Tumor Heterogeneity—Feature Allocation With Exponential Family Sampling. *Journal of the American Statistical Association*, 2015.