Statistical and computational trade-offs in Bayesian learning

Tamara Broderick
ITT Career Development
Assistant Professor, MIT

With: Nick Boyd, Ryan Giordano, Joseph Gonzalez, Stefanie Jegelka, Brian Kulis, Michael I. Jordan, Xinghao Pan, Andre Wibisono, Ashia C. Wilson

- Bayesian inference

- Bayesian inference
  - modular, complex models

- Bayesian inference
  - modular, complex models
  - all information about the parameter in the posterior

- Bayesian inference
  - modular, complex models
  - all information about the parameter in the posterior
- Approximating the posterior can be computationally expensive

Statistical/computational trade-offs

- Bayesian inference
  - modular, complex models
  - all information about the parameter in the posterior
- Approximating the posterior can be computationally expensive

Statistical/computational trade-offs

- Bayesian inference
  - modular, complex models
  - all information about the parameter in the posterior
- Approximating the posterior can be computationally expensive
- Computational/statistical gains for trading off some posterior knowledge

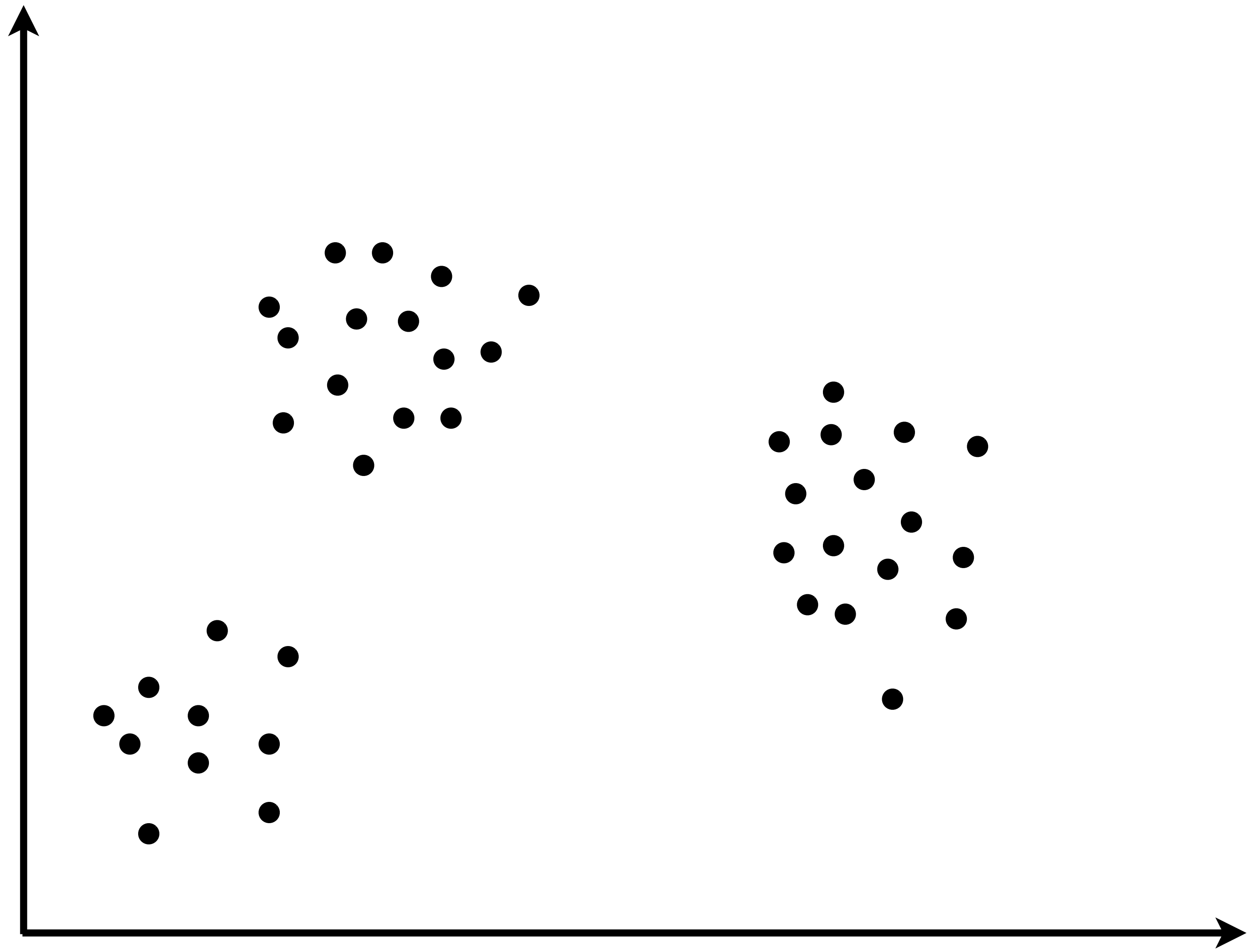
Statistical/computational trade-offs

- Bayesian inference
    - modular, complex models
    - all information about the parameter in the posterior
- Approximating the posterior can be computationally expensive
- Computational/statistical gains for trading off some posterior knowledge
    - point estimates

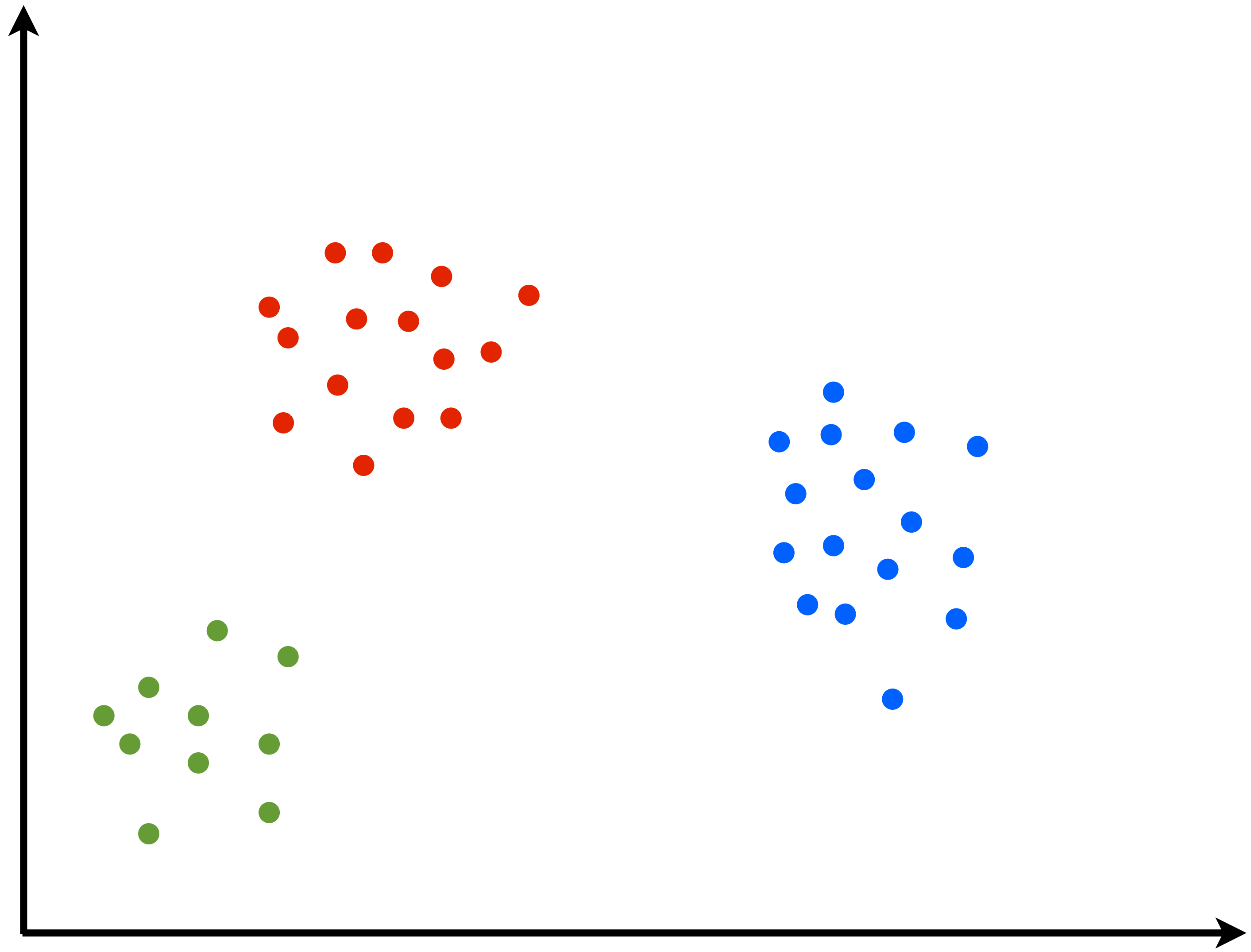
Statistical/computational trade-offs

- Bayesian inference
    - modular, complex models
    - all information about the parameter in the posterior
- Approximating the posterior can be computationally expensive
- Computational/statistical gains for trading off some posterior knowledge
    - point estimates
    - covariances, coherent estimates of uncertainty

Clustering

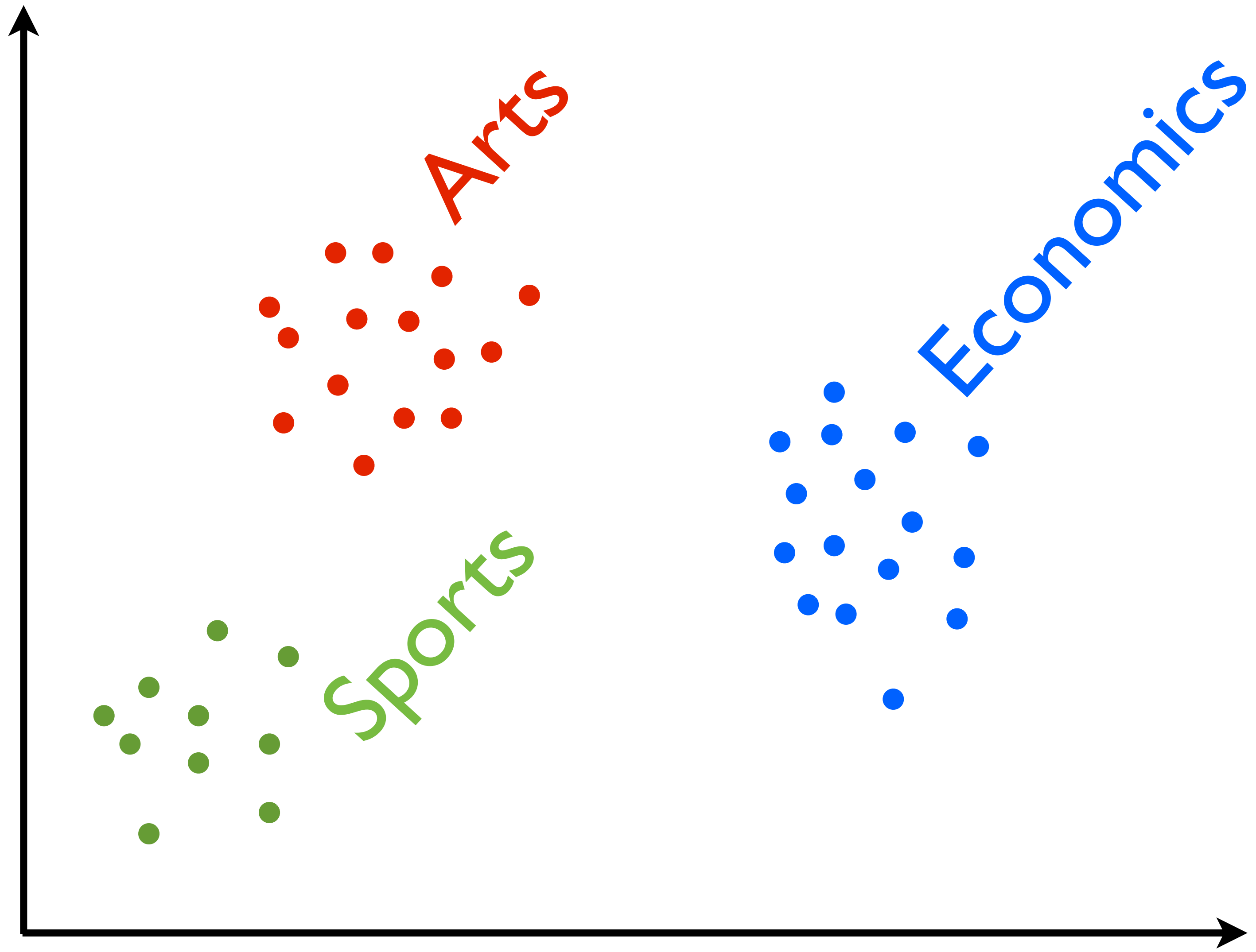


Clustering



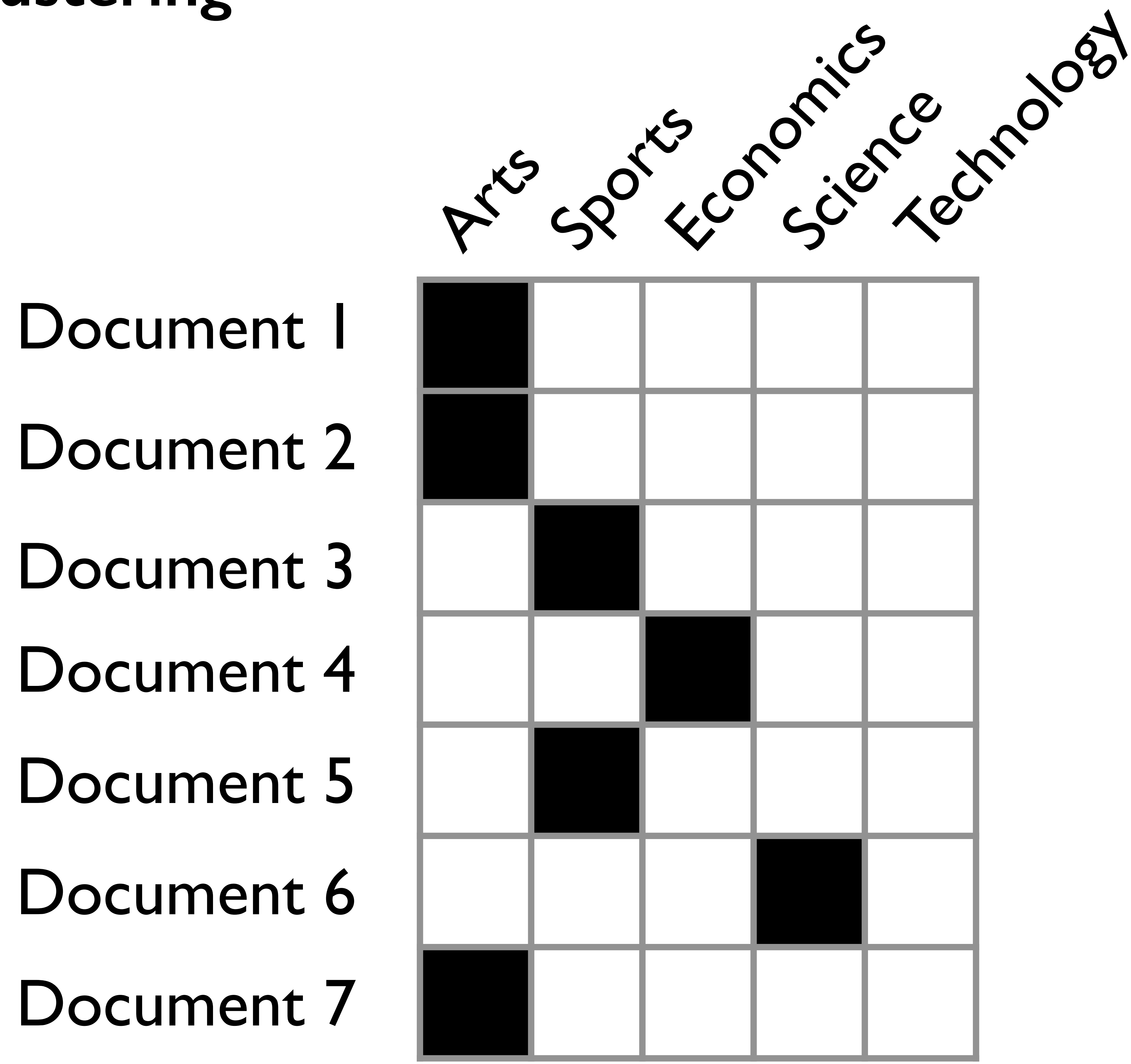
“clusters”

Clustering

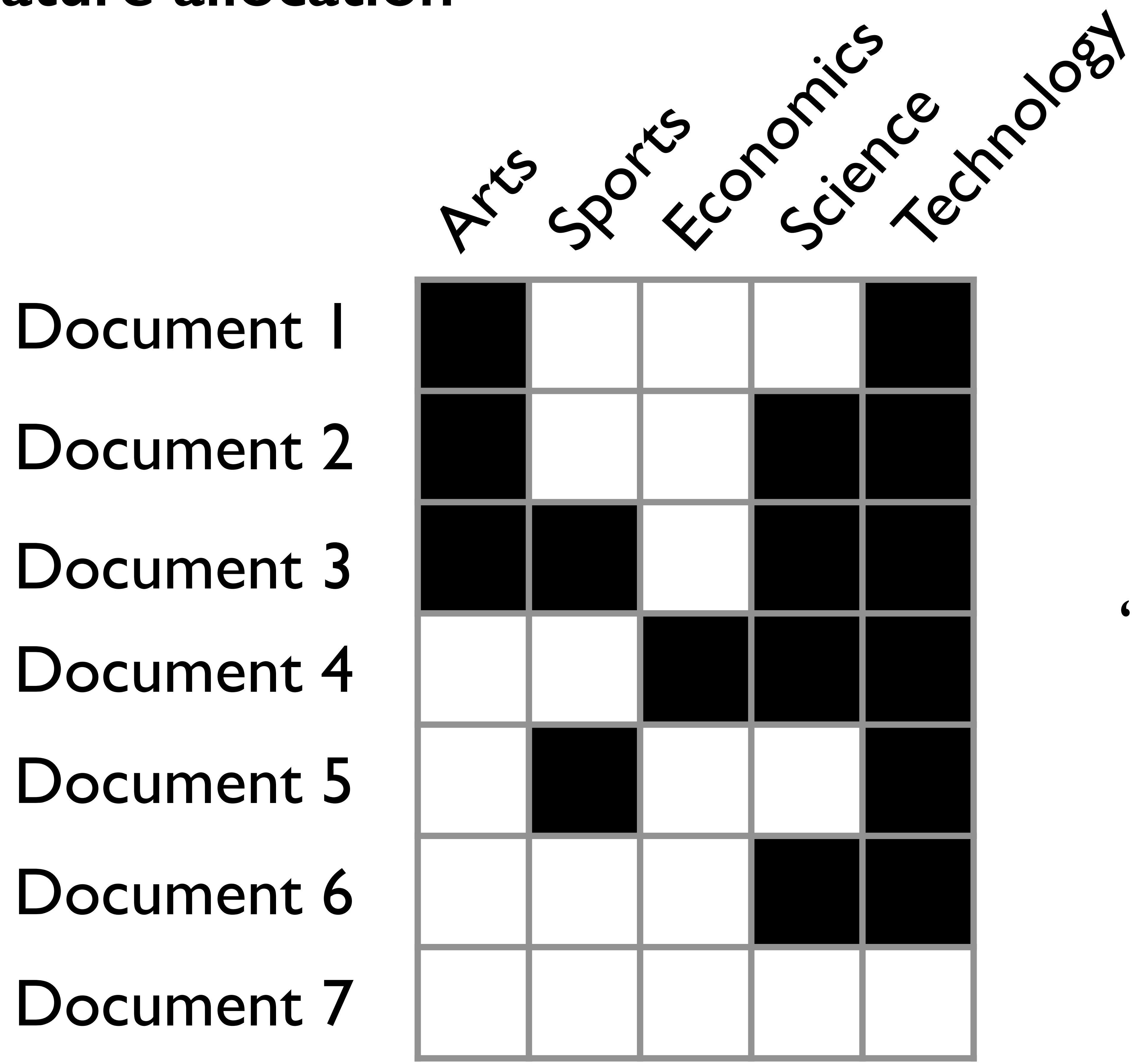


“clusters”

Clustering

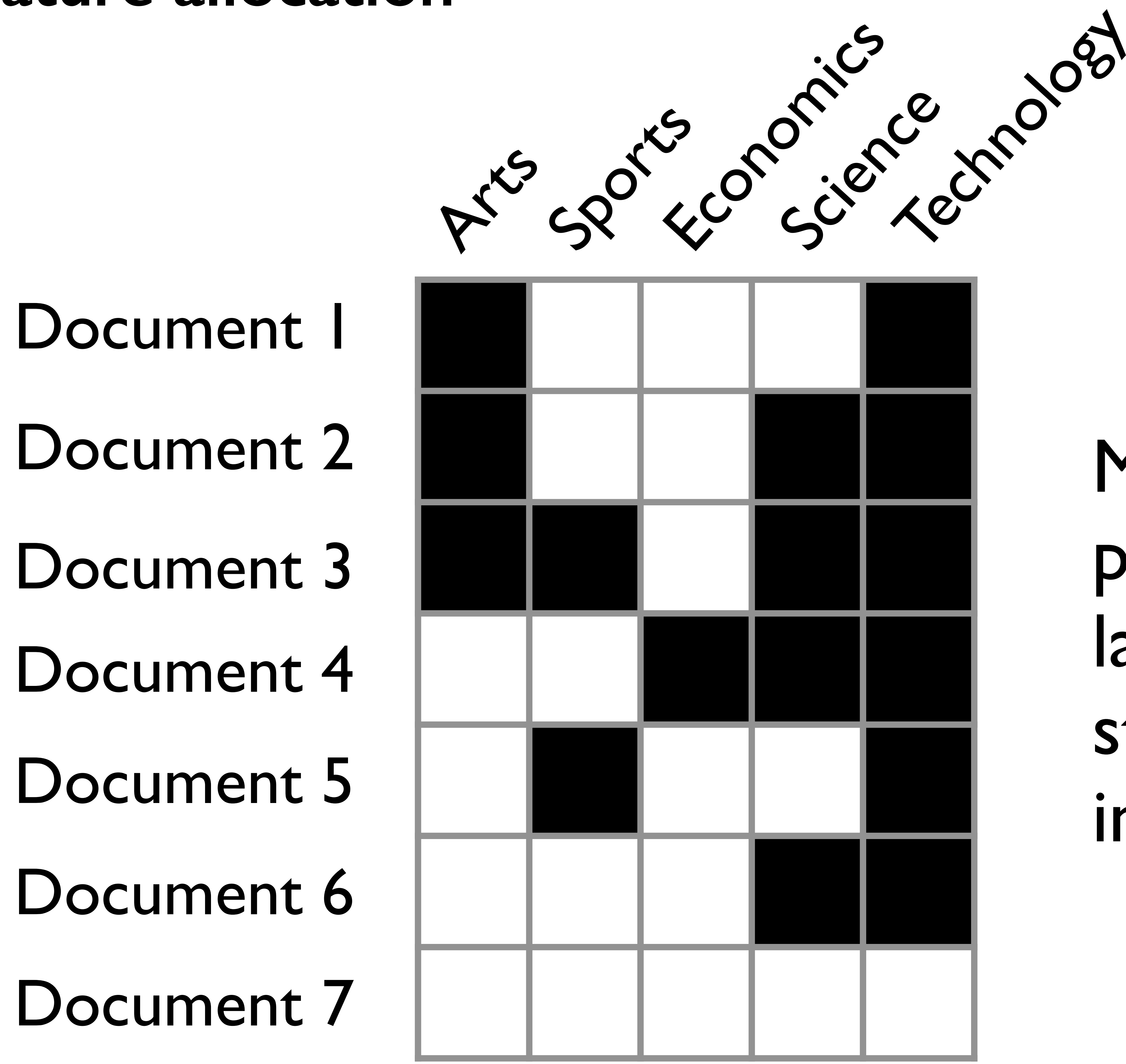

Arts
Sports
Economics
Science
Technology
Document 1
Document 2
Document 3
Document 4
Document 5
Document 6
Document 7

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...

- E.g., Silicon Valley: 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

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

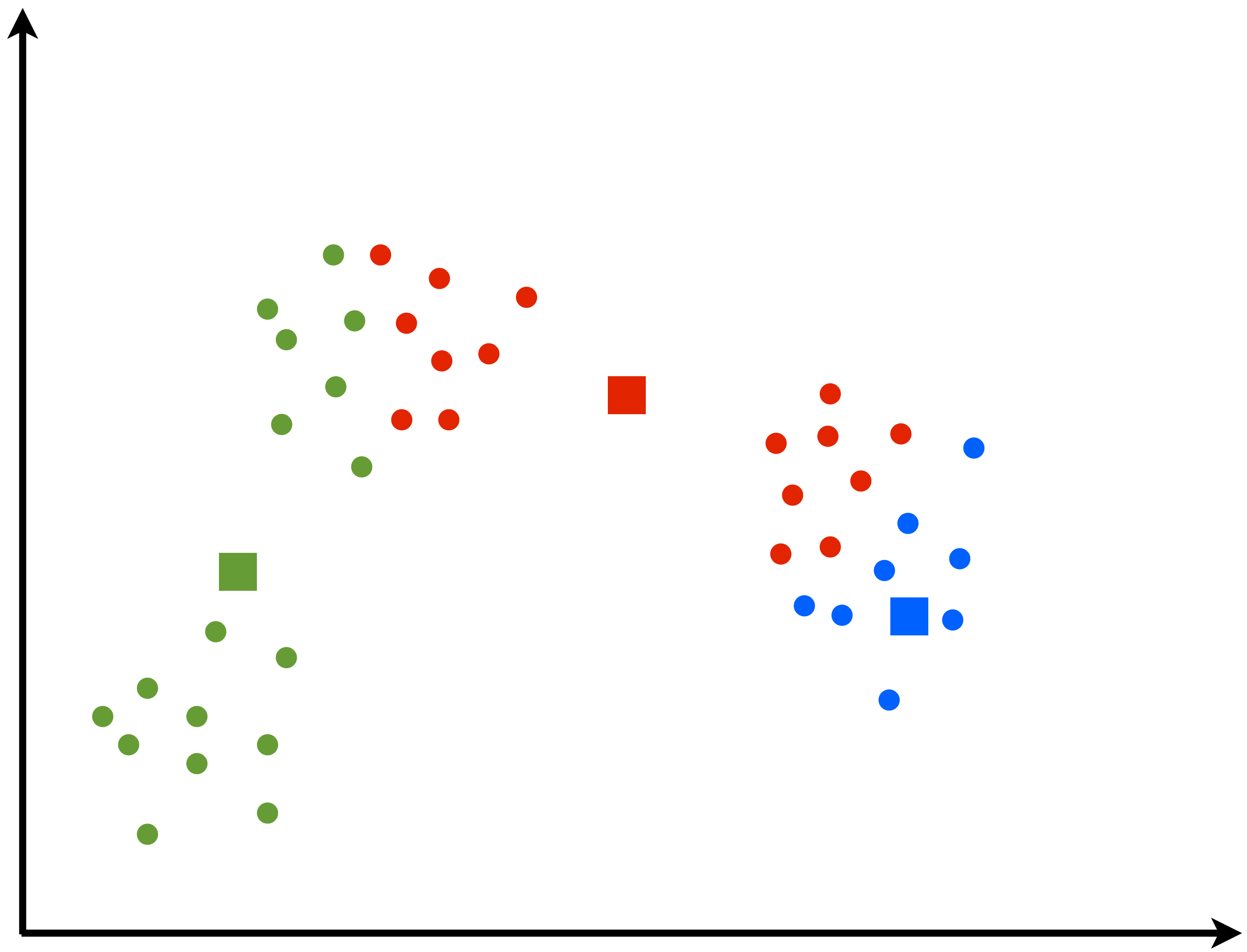
MAD-Bayes

The MAD-Bayes idea

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

K-means

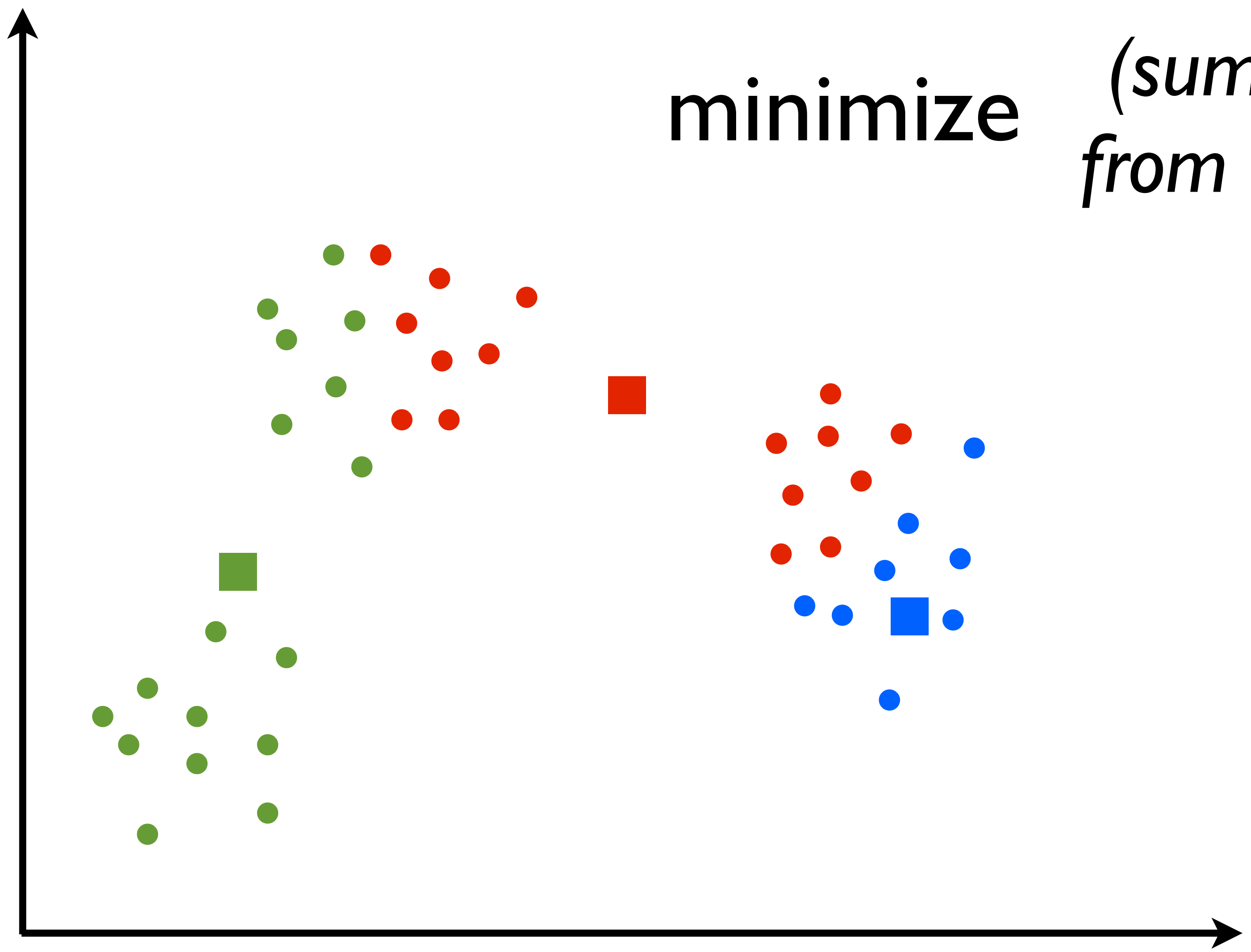
K-means clustering problem



K-means

K-means clustering problem

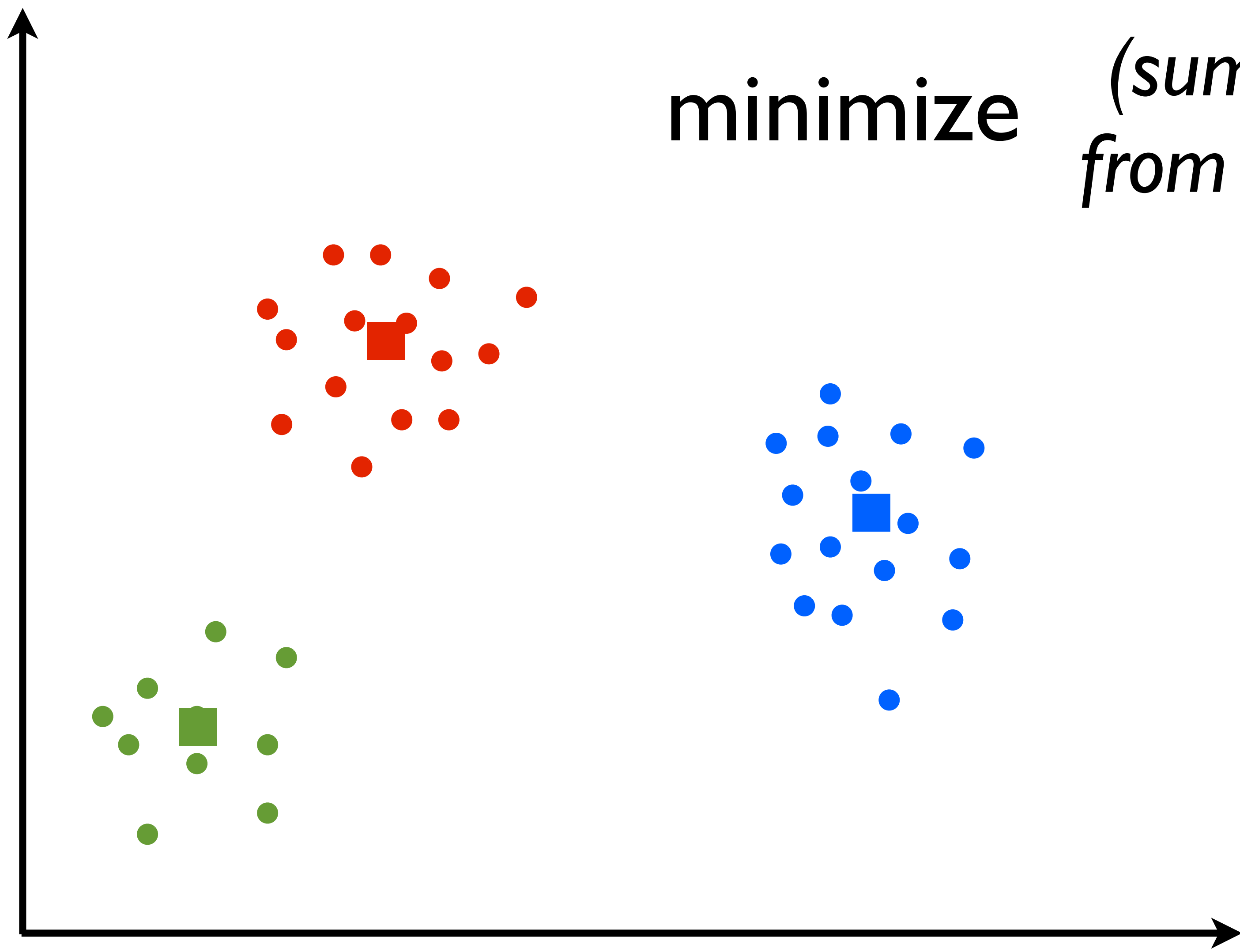
minimize *(sum of square distances from data points to cluster centers)*



K-means

K-means clustering problem

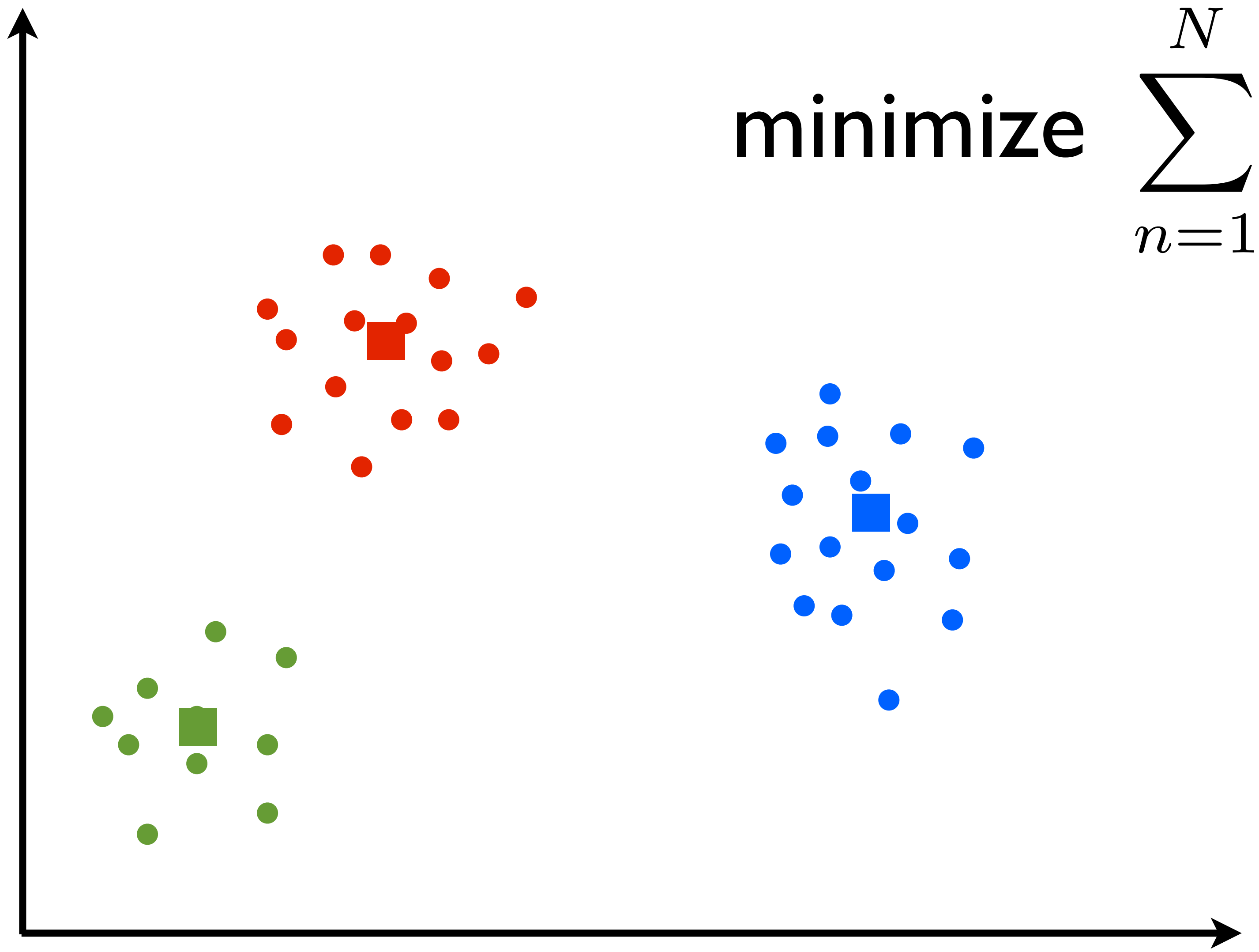
minimize *(sum of square distances from data points to cluster centers)*



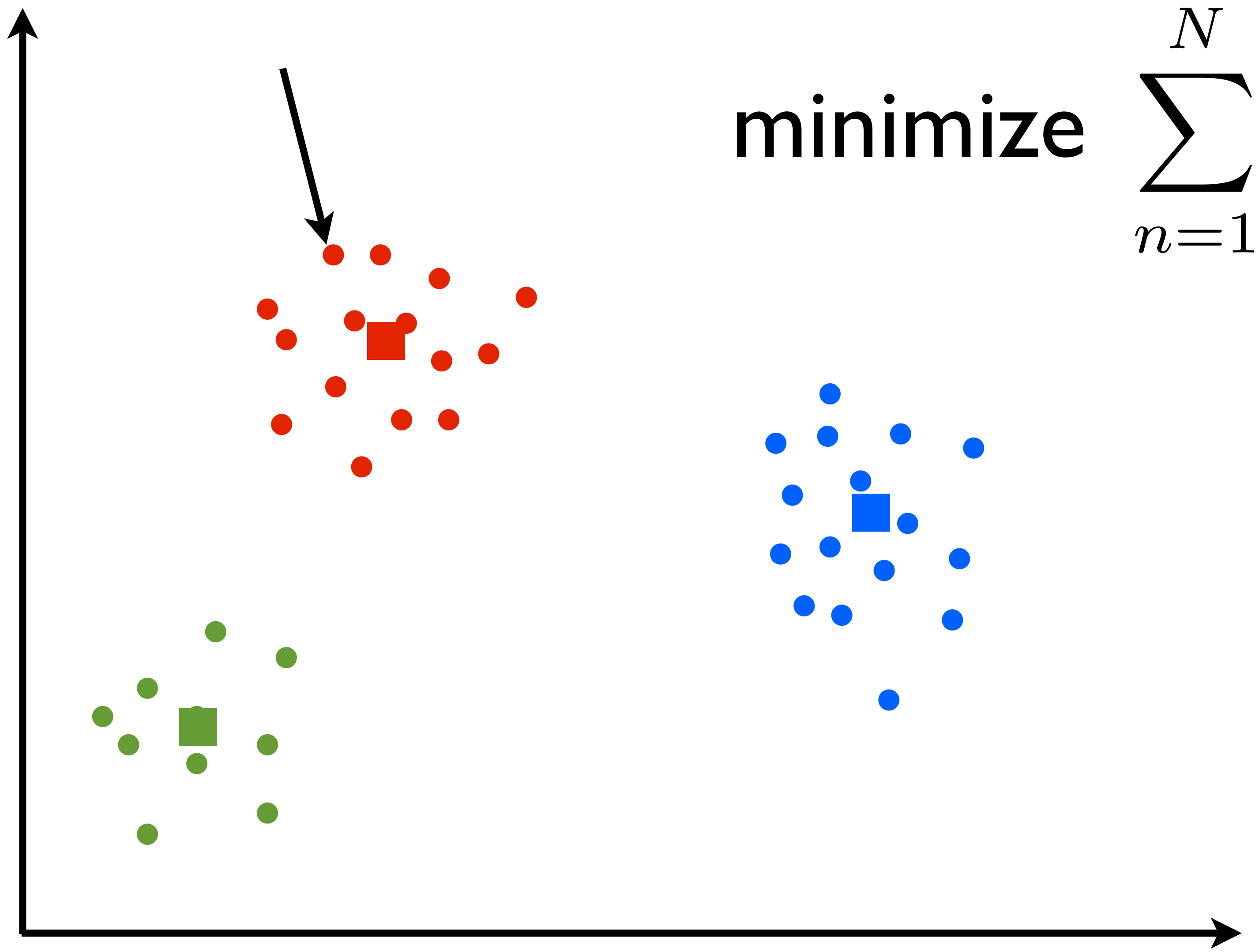
K-means

K-means clustering problem

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



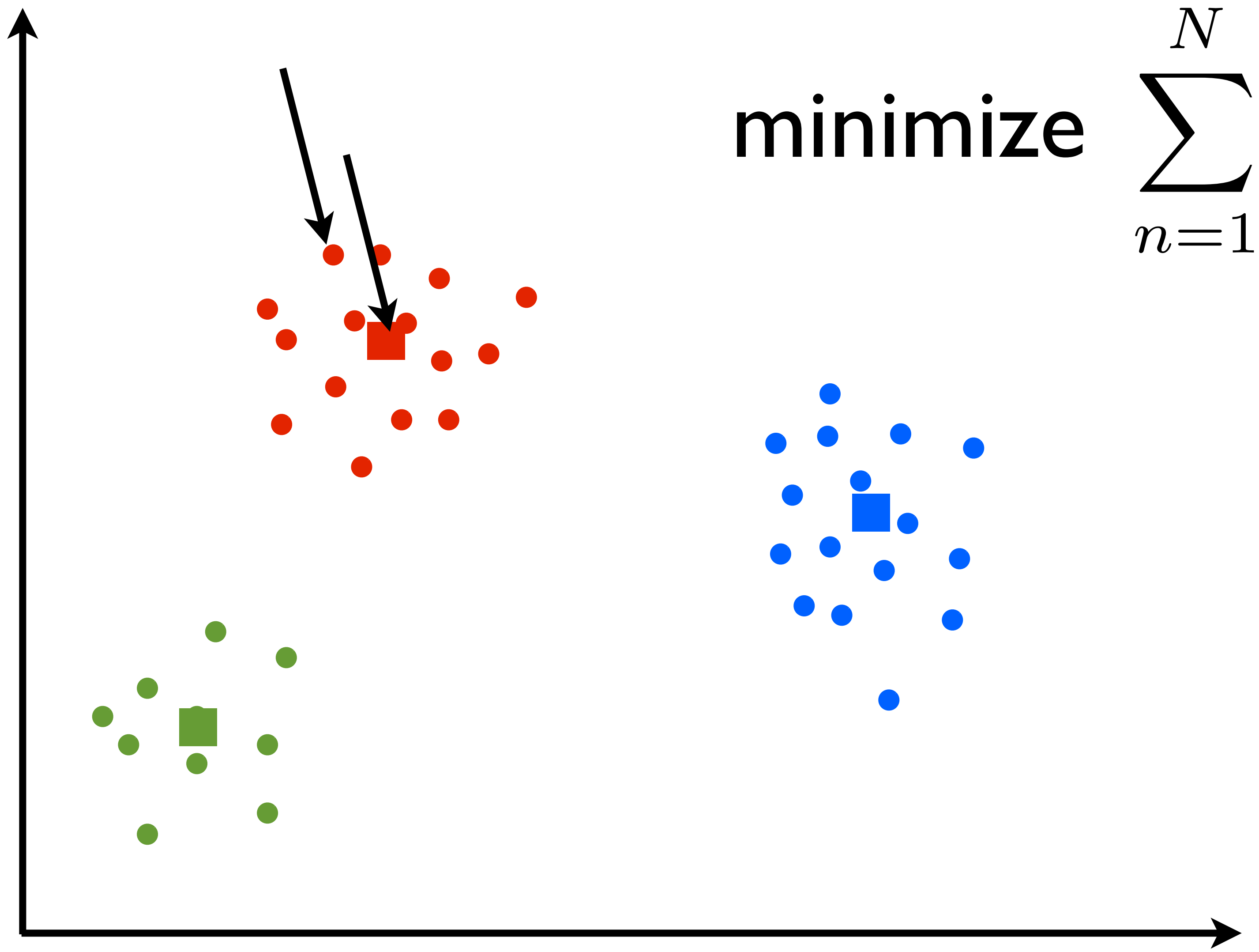
K-means



K-means clustering problem

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

K-means



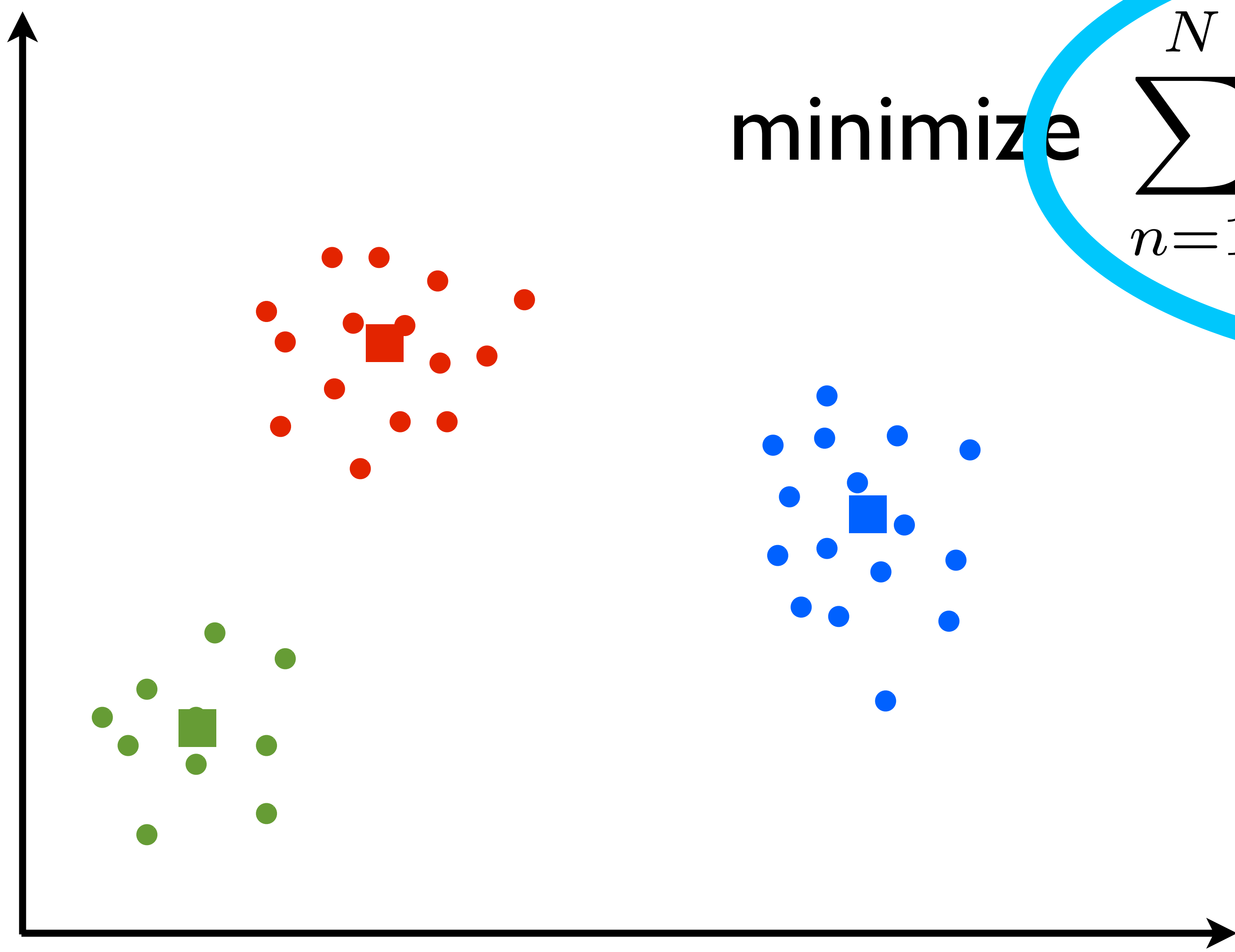
K-means clustering problem

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

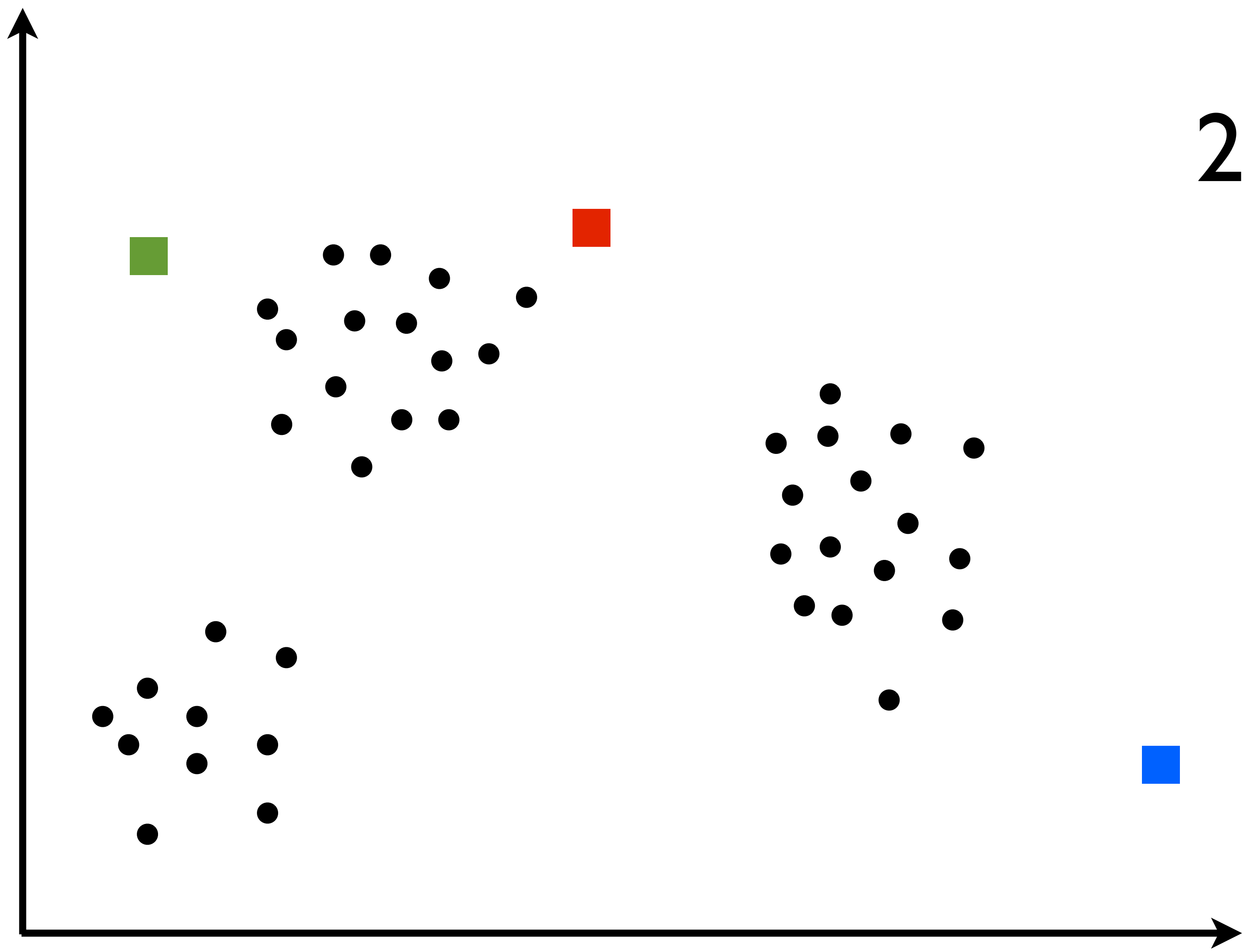
K-means

K-means objective

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



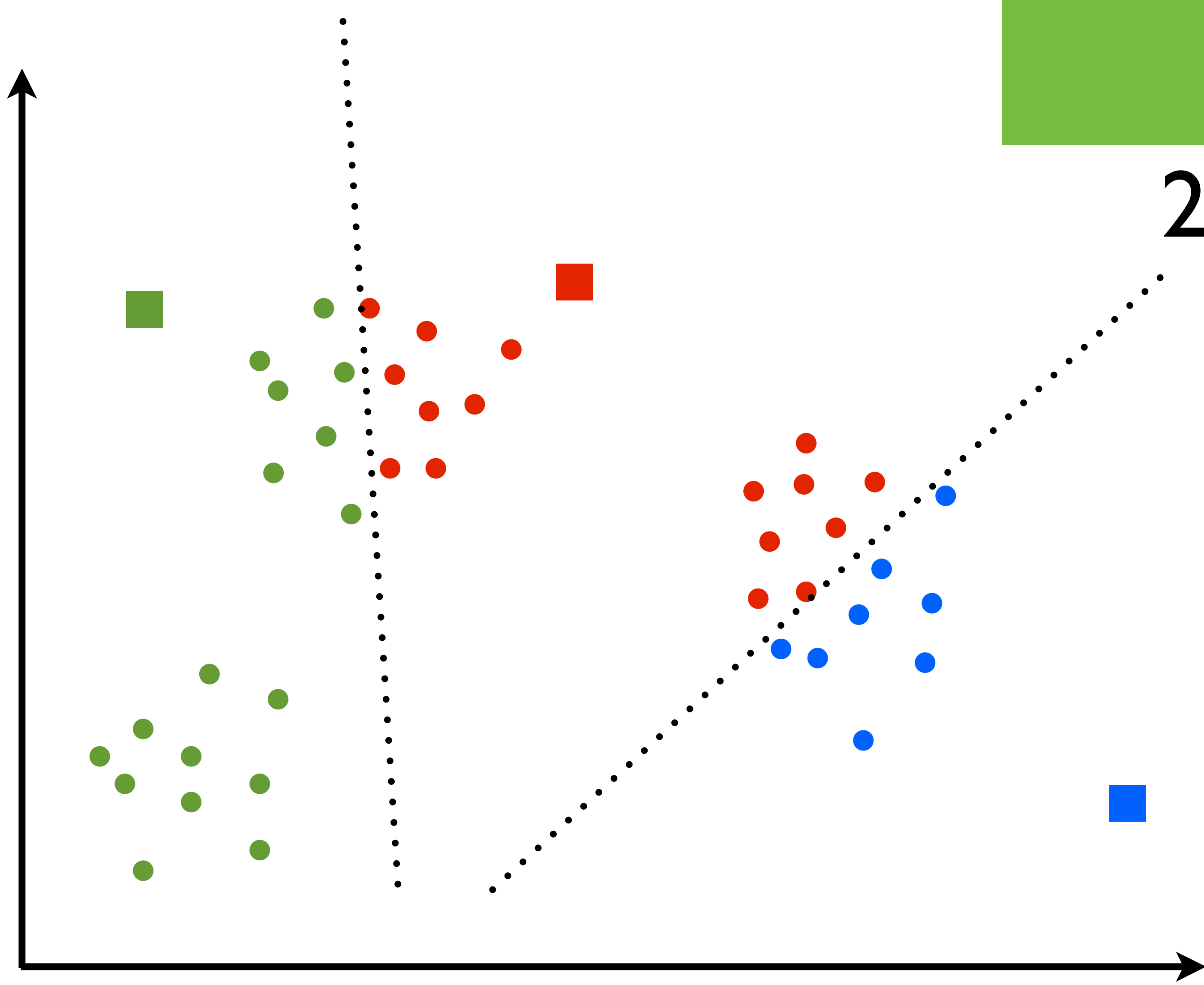
Lloyd's algorithm



Iterate until no changes:

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

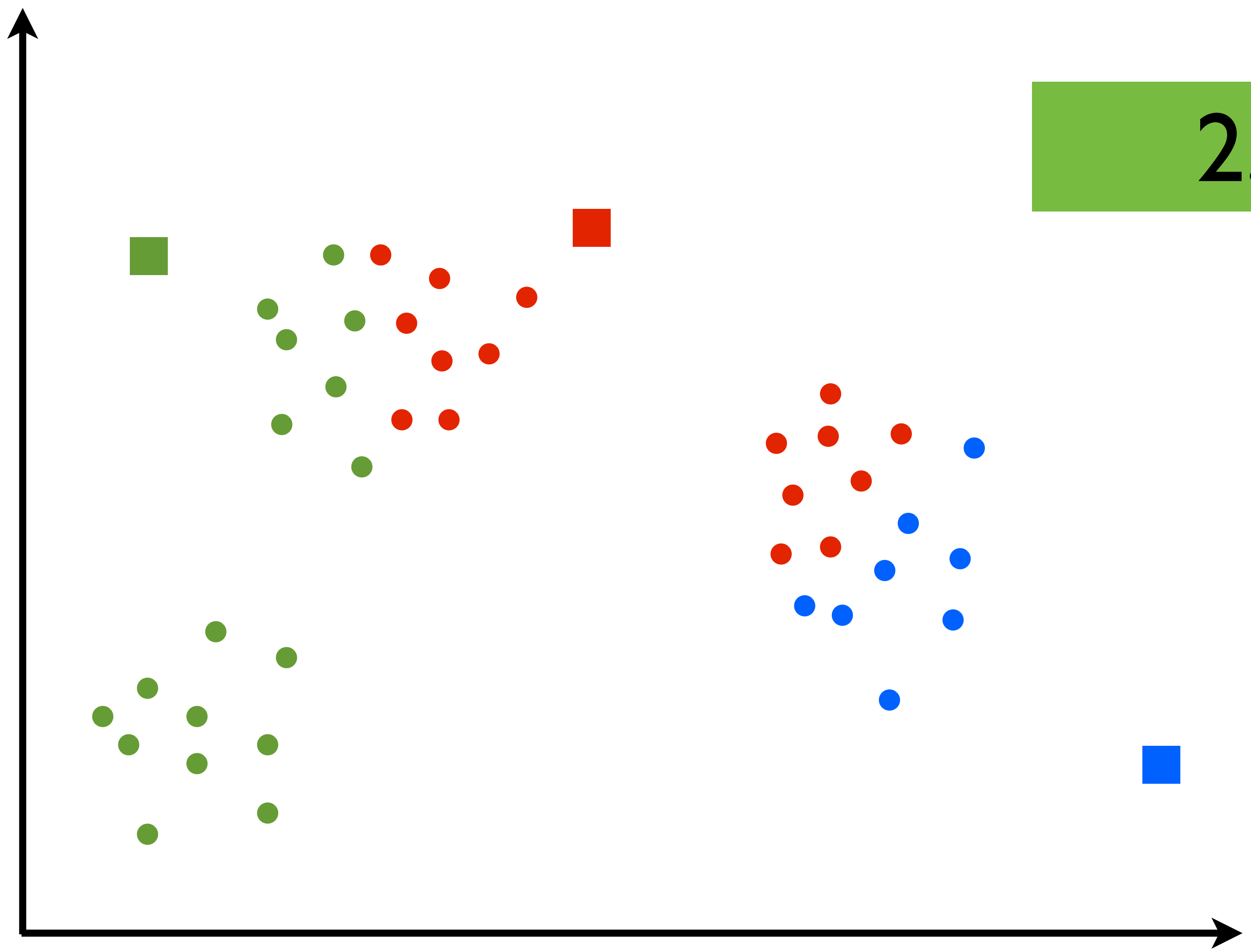
Lloyd's algorithm



Iterate until no changes:

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

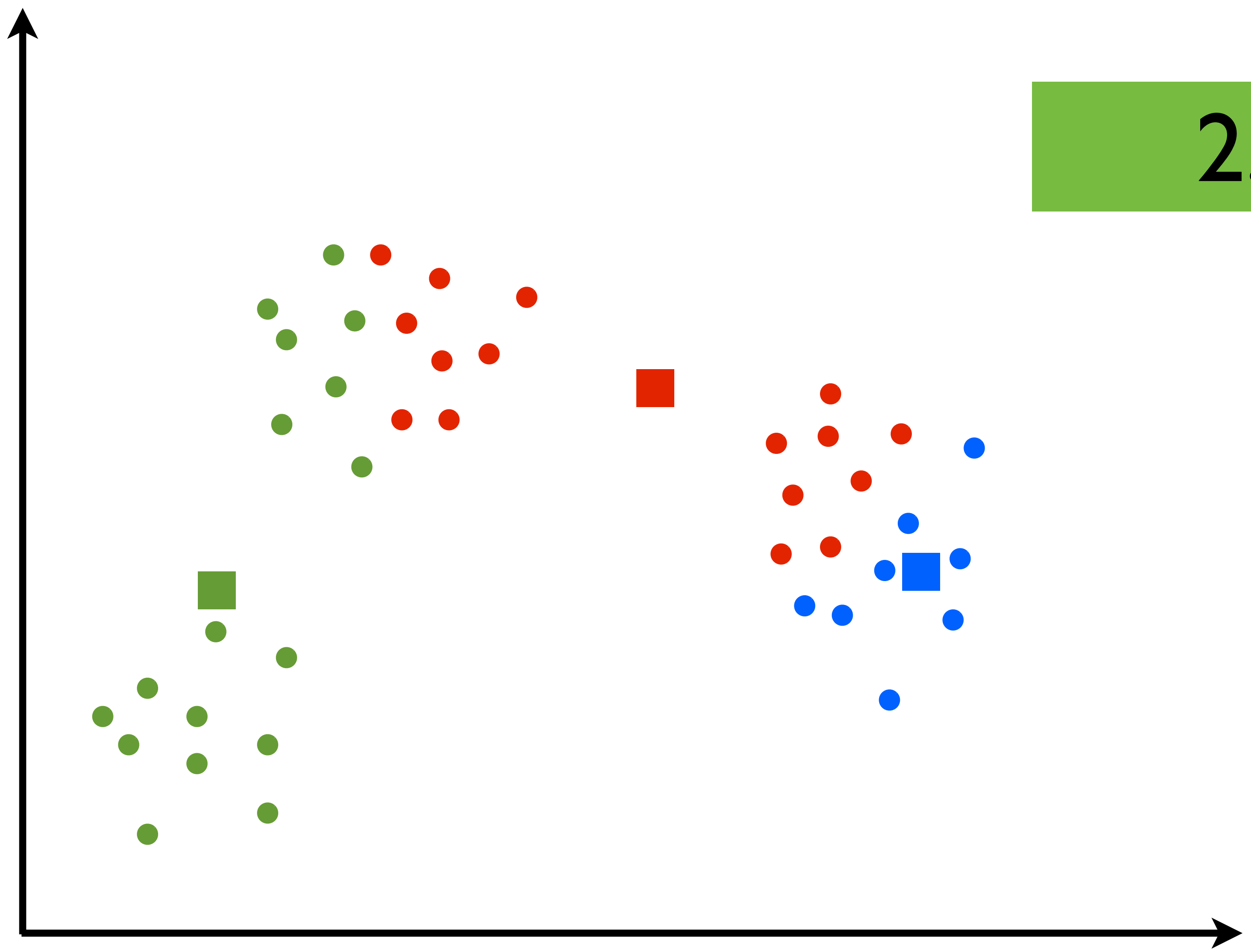
Lloyd's algorithm



Iterate until no changes:

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

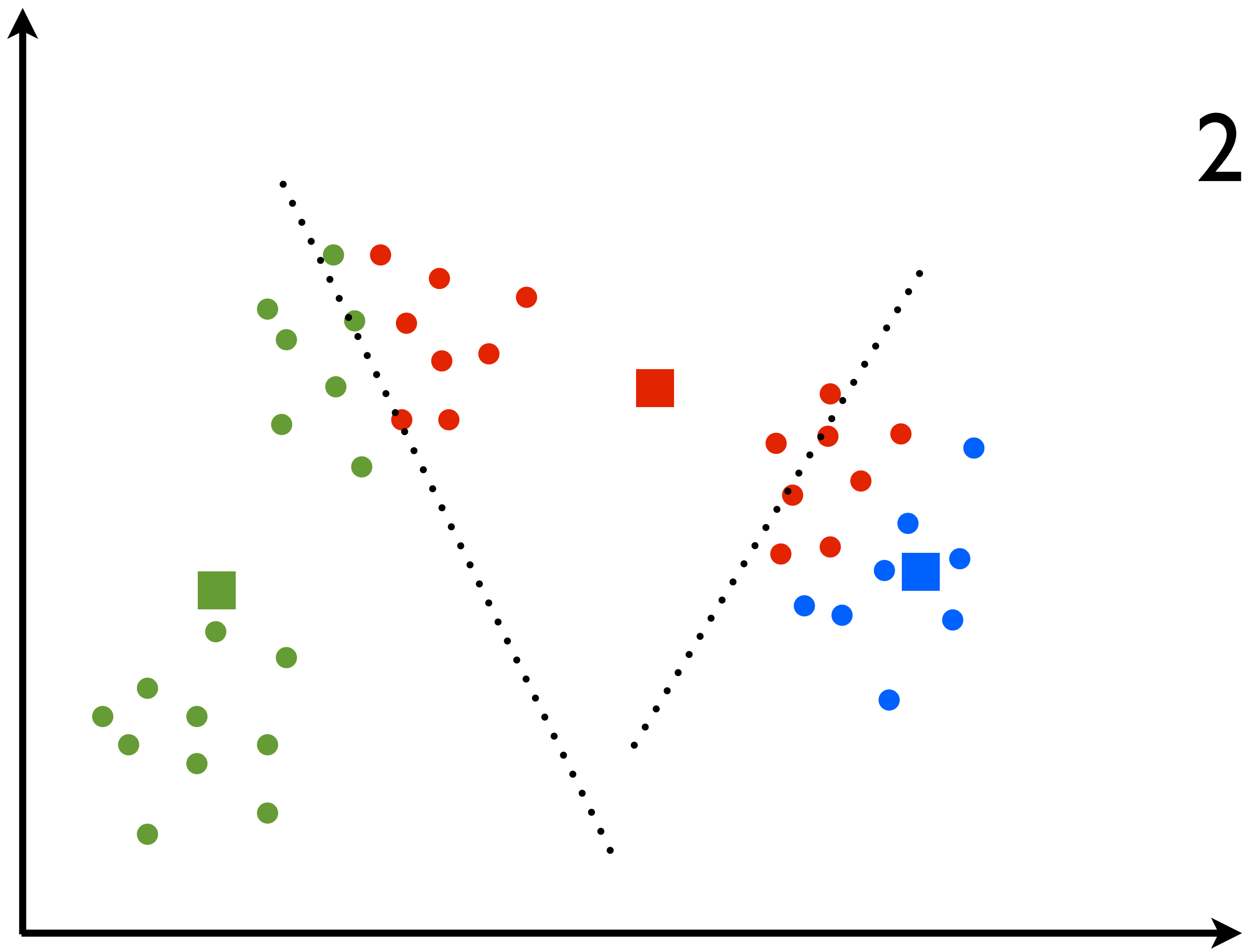
Lloyd's algorithm



Iterate until no changes:

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

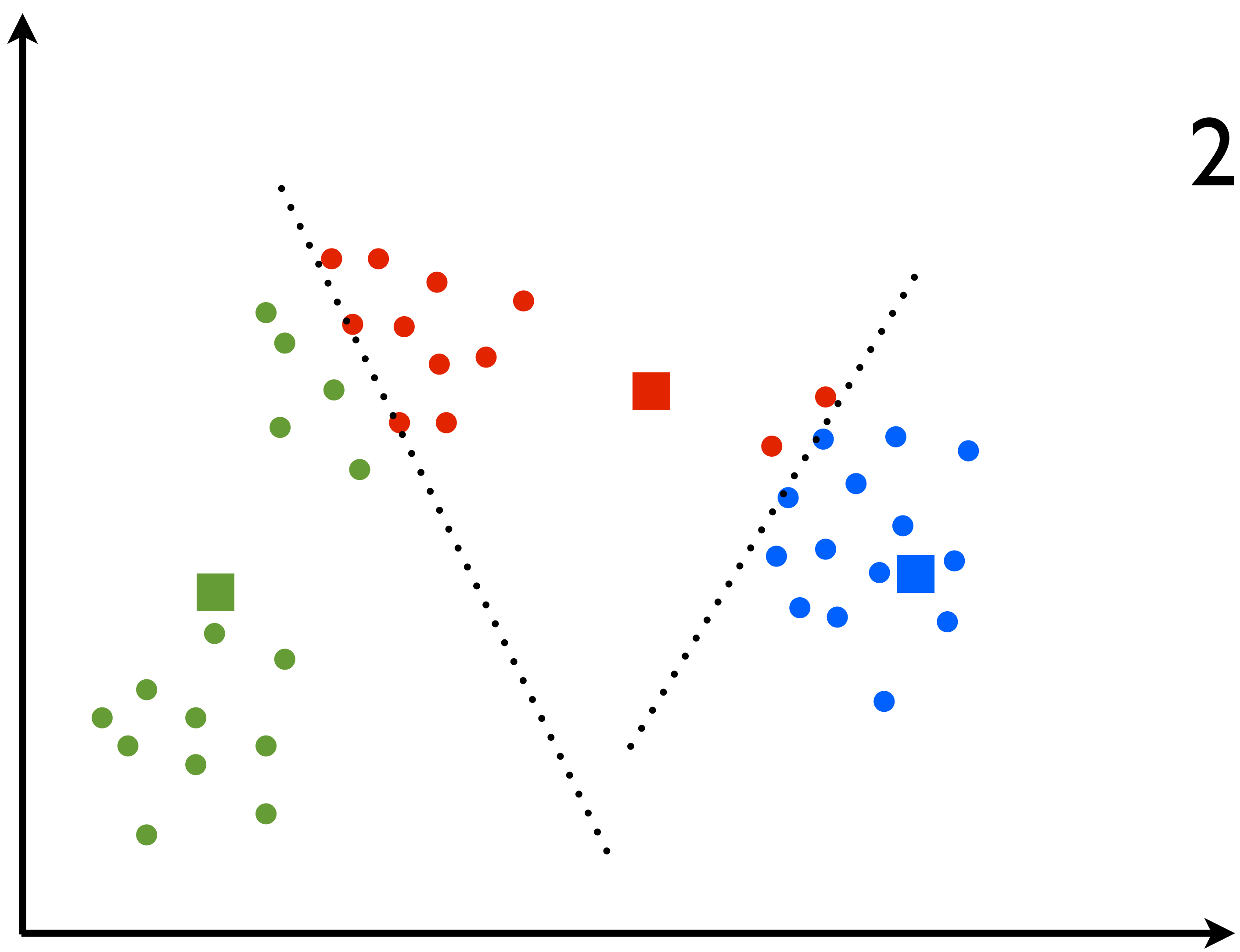
Lloyd's algorithm



Iterate until no changes:

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

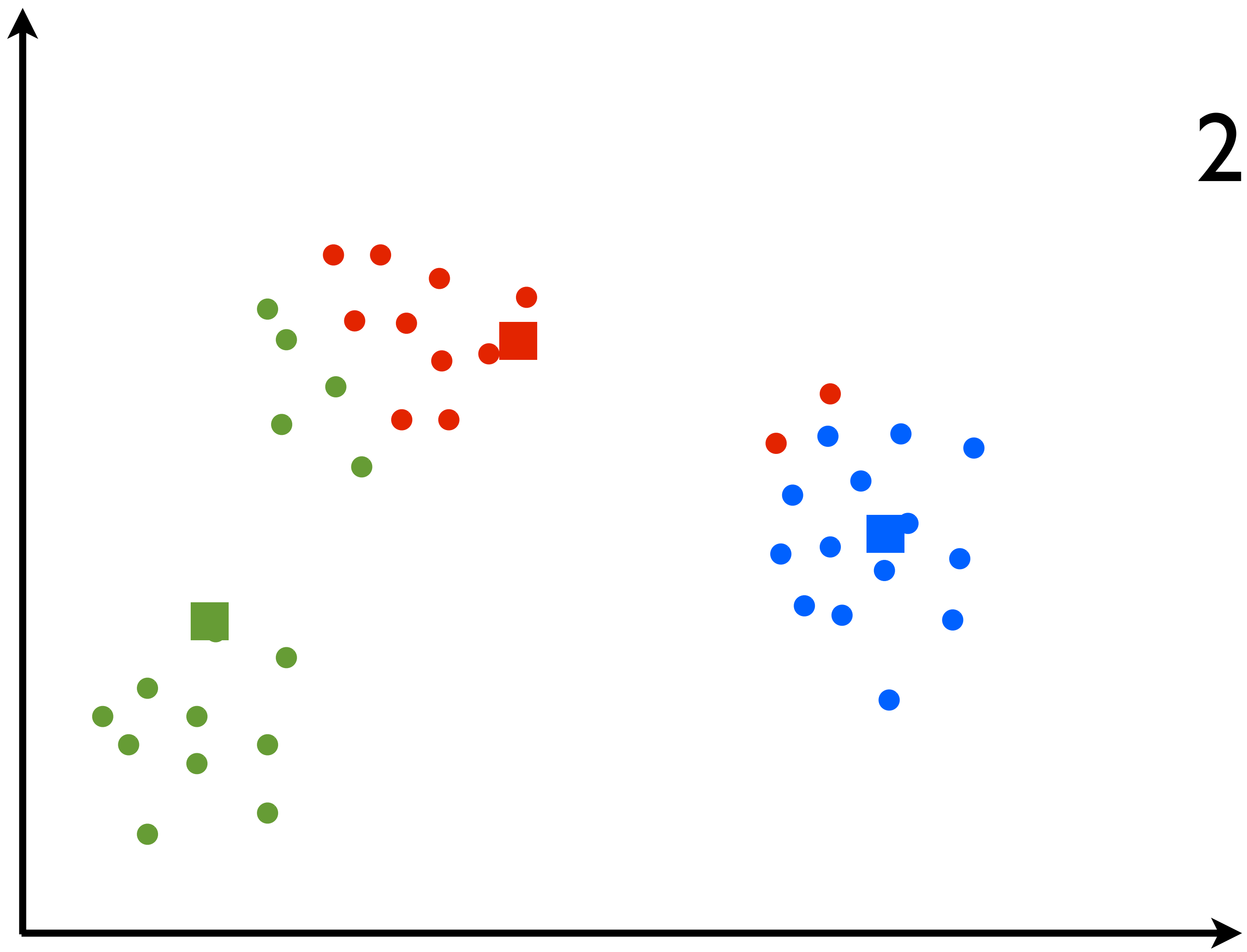
Lloyd's algorithm



Iterate until no changes:

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

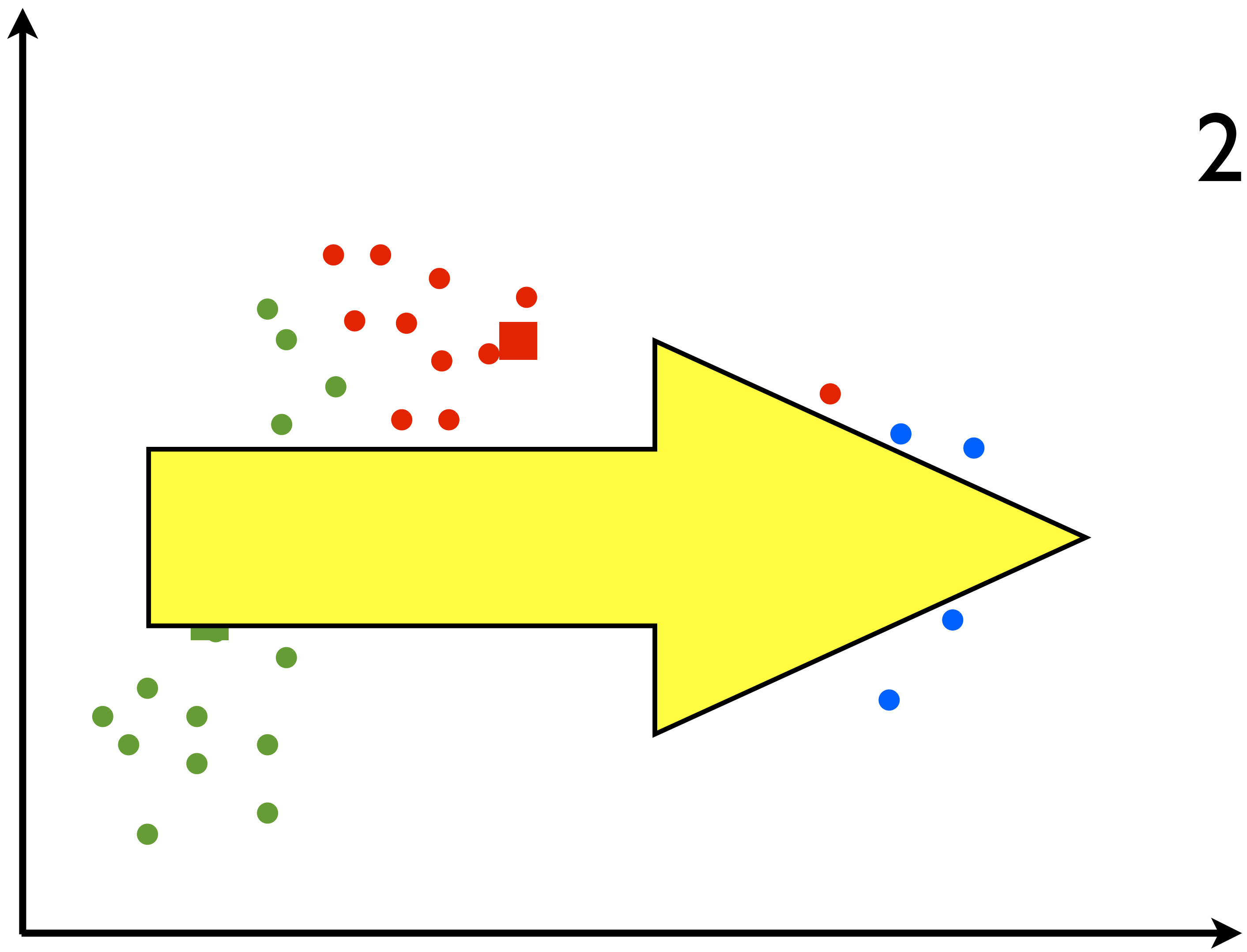
Lloyd's algorithm



Iterate until no changes:

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

Lloyd's algorithm



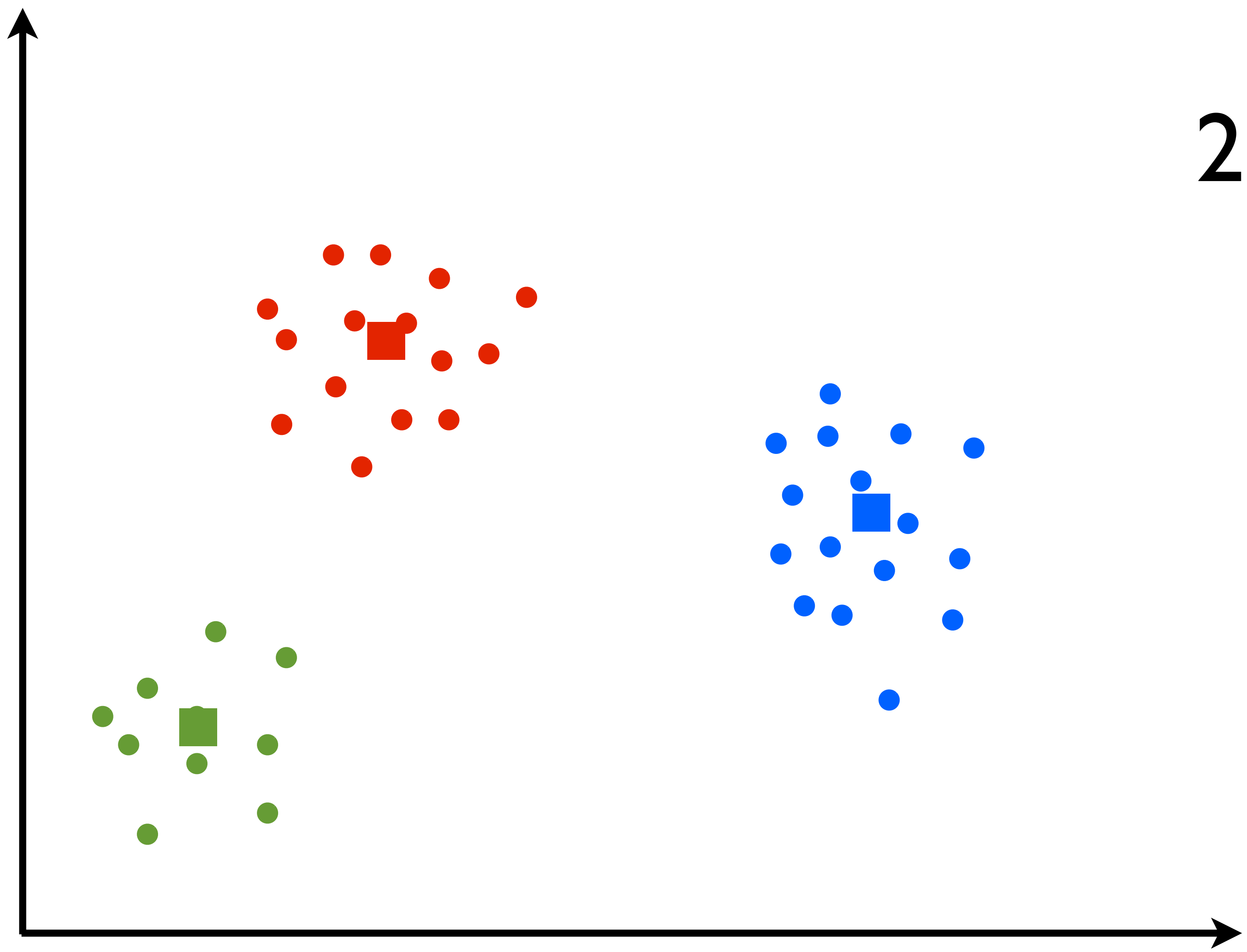
Iterate until no changes:

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

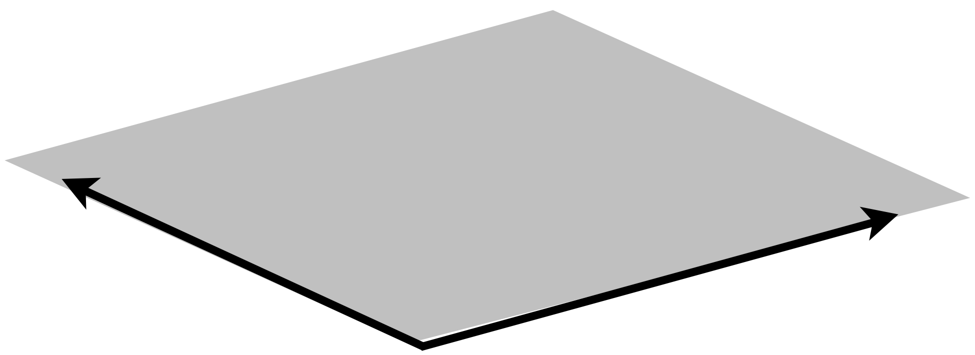
- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective

MAD-Bayes

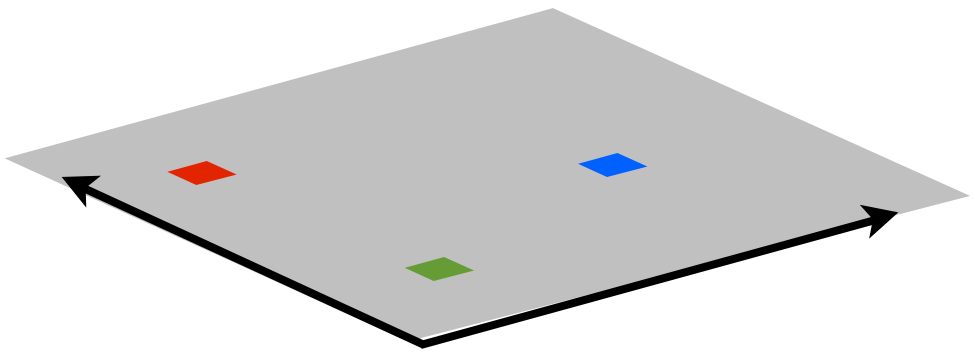
The MAD-Bayes idea

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

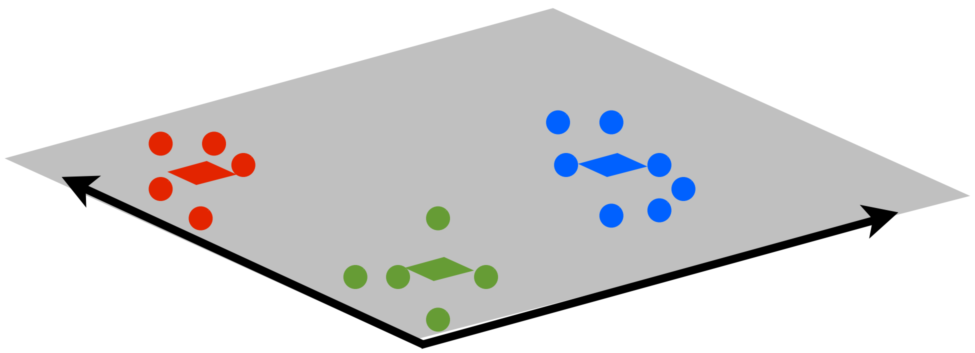
Bayesian model



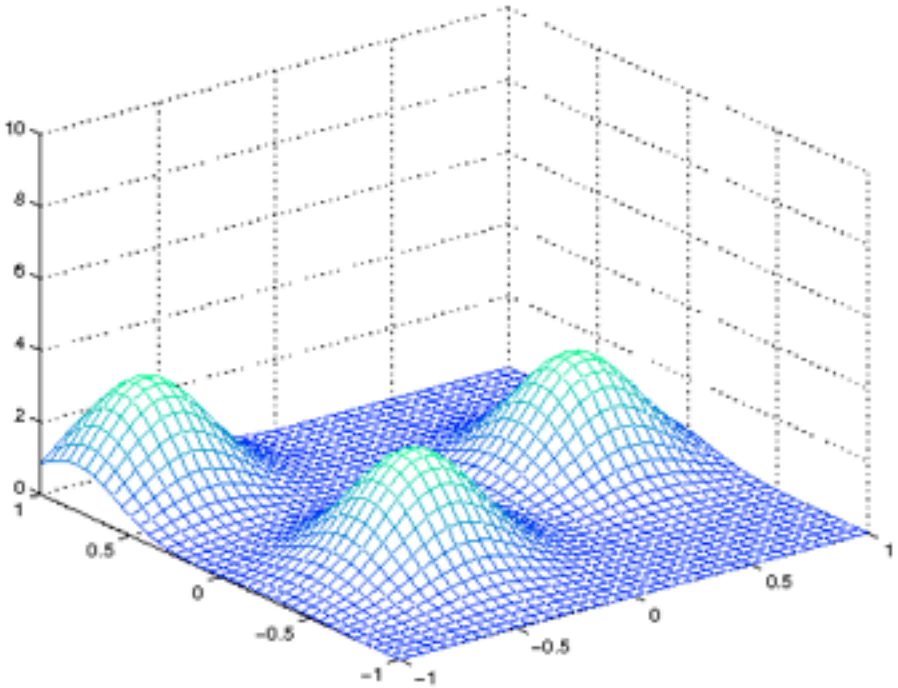
Bayesian model



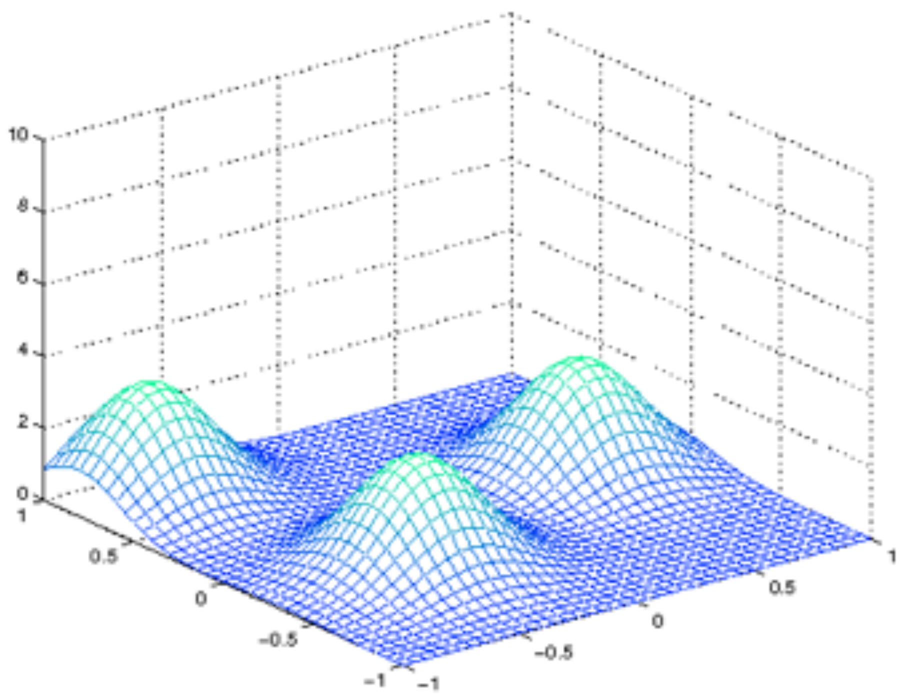
Bayesian model



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

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

MAD-Bayes

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

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

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

MAD-Bayes

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

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

- 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 $\Rightarrow$ K-means

MAD-Bayes

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

Mixture of K Gaussians $\Rightarrow$ K-means

Dirichlet process mixture $\Rightarrow$ 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

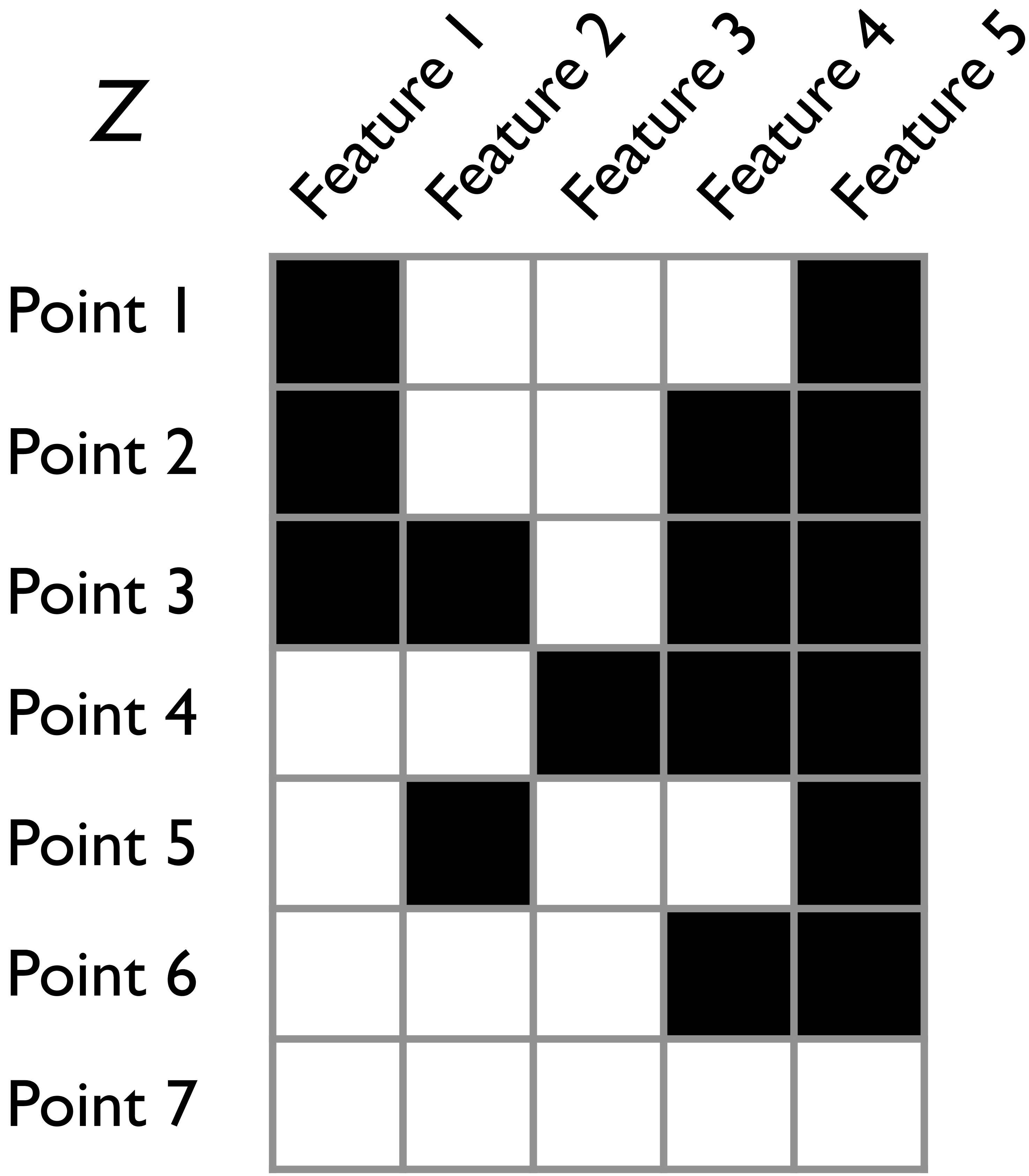
MAD-Bayes

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

MAD-Bayes

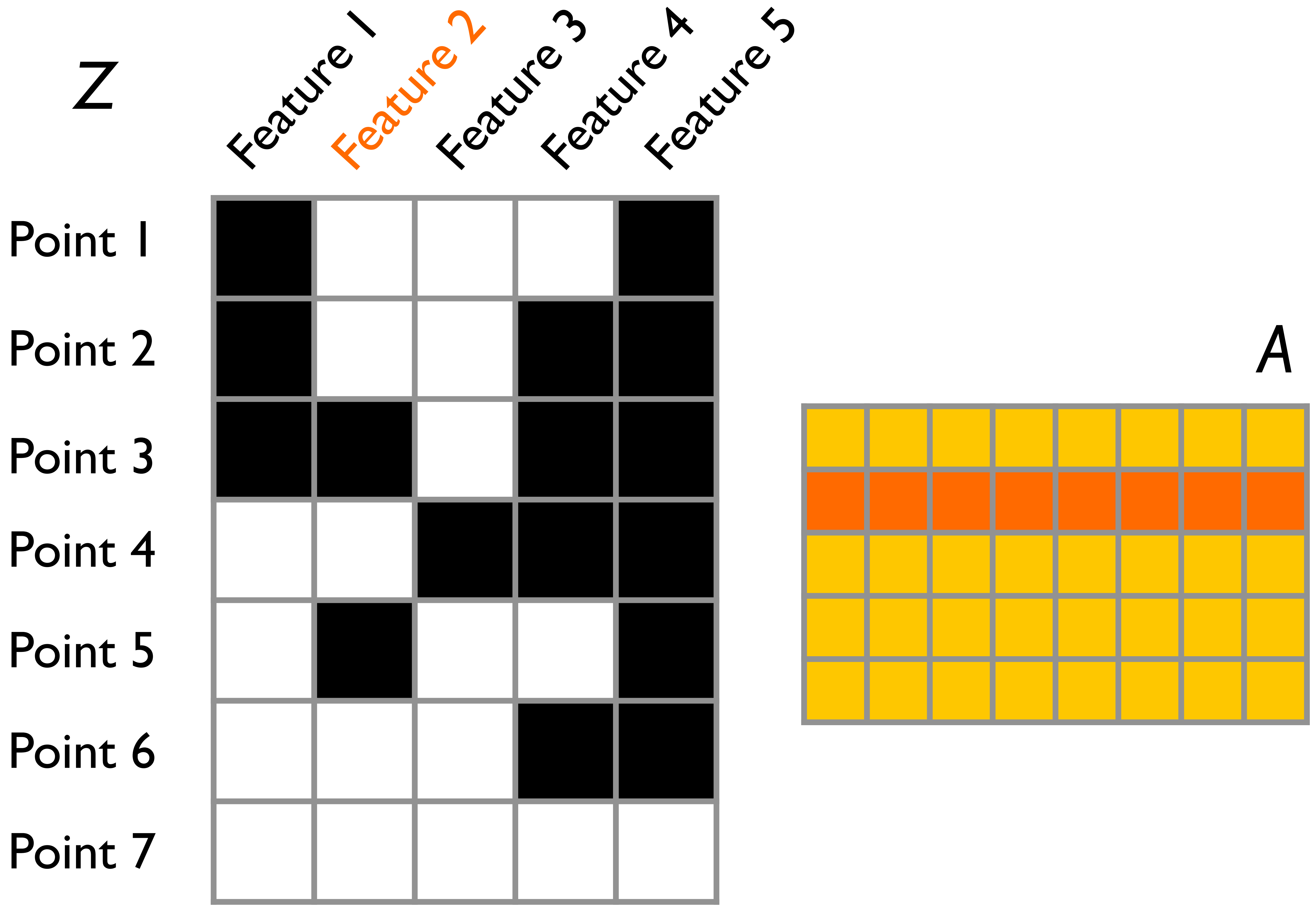
Bayesian posterior		K-means-like objectives
Mixture of K Gaussians	→	K-means
Dirichlet process mixture	→	Unbounded number of clusters
Hierarchical Dirichlet process	→	Multiple data sets share cluster centers
⋮		⋮
Beta process	→	Features

Features

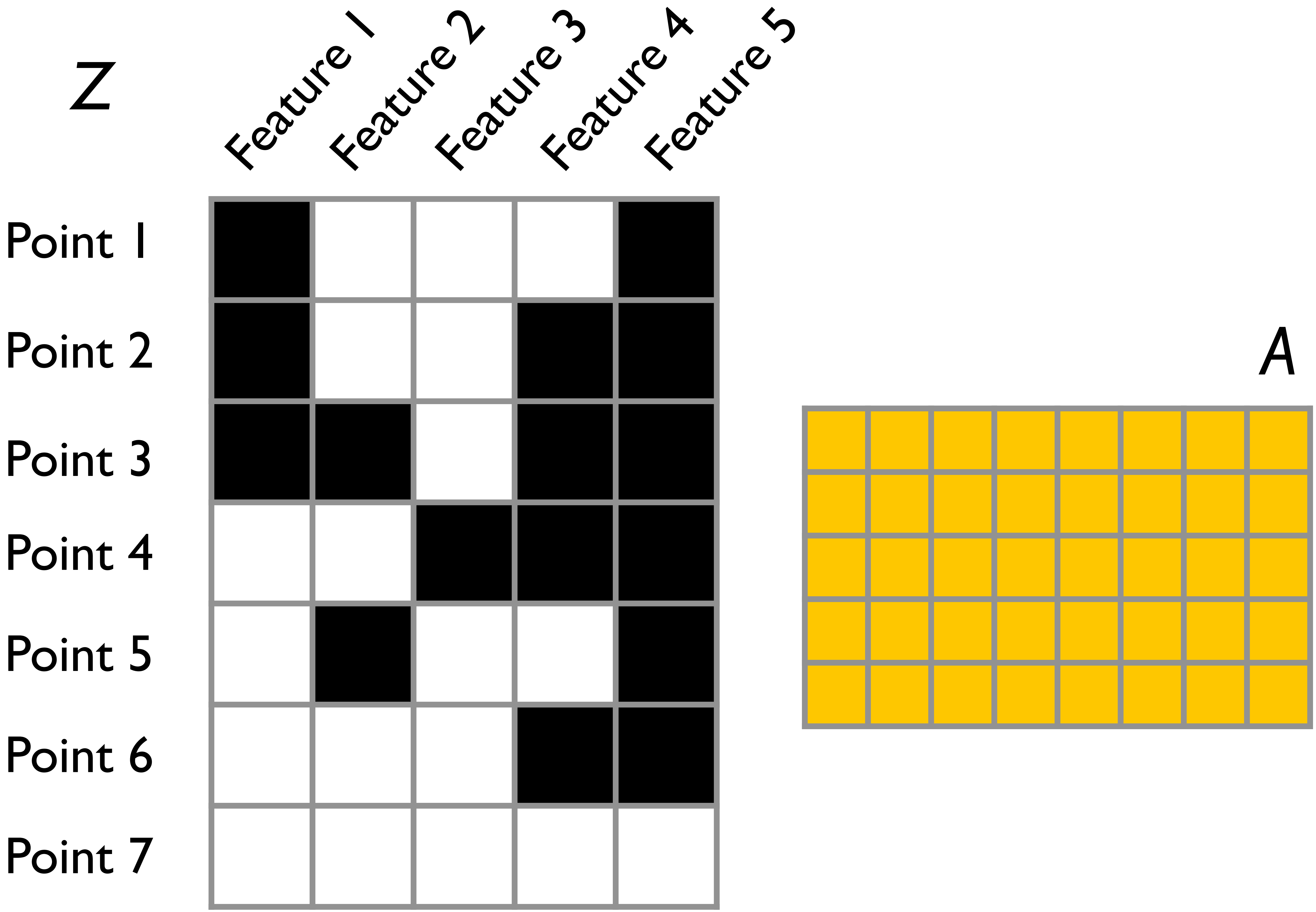


Z
Feature 1
Feature 2
Feature 3
Feature 4
Feature 5
Point 1
Point 2
Point 3
Point 4
Point 5
Point 6
Point 7

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

$$
\begin{aligned}
&\mathbb{P}(Z, A|X) \\
&\qquad \propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{ -\frac{1}{2\sigma^2} \mathbf{tr}((X - ZA)'(X - ZA)) \right\} \\
&\qquad \cdot \frac{\gamma^{K^+} \exp\left\{ -\sum_{n=1}^{N} \frac{\gamma}{n} \right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!} \\
&\qquad \cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{ -\frac{1}{2\rho^2} \mathbf{tr}(A'A) \right\}.
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{P}(Z, A|X) \\
&\propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{ -\frac{1}{2\sigma^2} \mathbf{tr}((X - ZA)'(X - ZA)) \right\} \\
&\cdot \frac{\gamma^{K^+} \exp\left\{ -\sum_{n=1}^{N} \frac{\gamma}{n} \right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!} \\
&\cdot \frac{1}{(2\pi\rho^2)^{K^+D/2}} \exp\left\{ -\frac{1}{2\rho^2} \mathbf{tr}(A'A) \right\}.
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathbb{P}(Z, A|X) \\
&\quad \propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X - ZA)'(X - ZA))\right\} \\
&\quad \cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!} \\
&\quad \cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.
\end{aligned}
$$

MAD-Bayes

BP-means objective

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

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

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

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

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

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 $A \leftarrow (Z'Z)^{-1}Z'X$

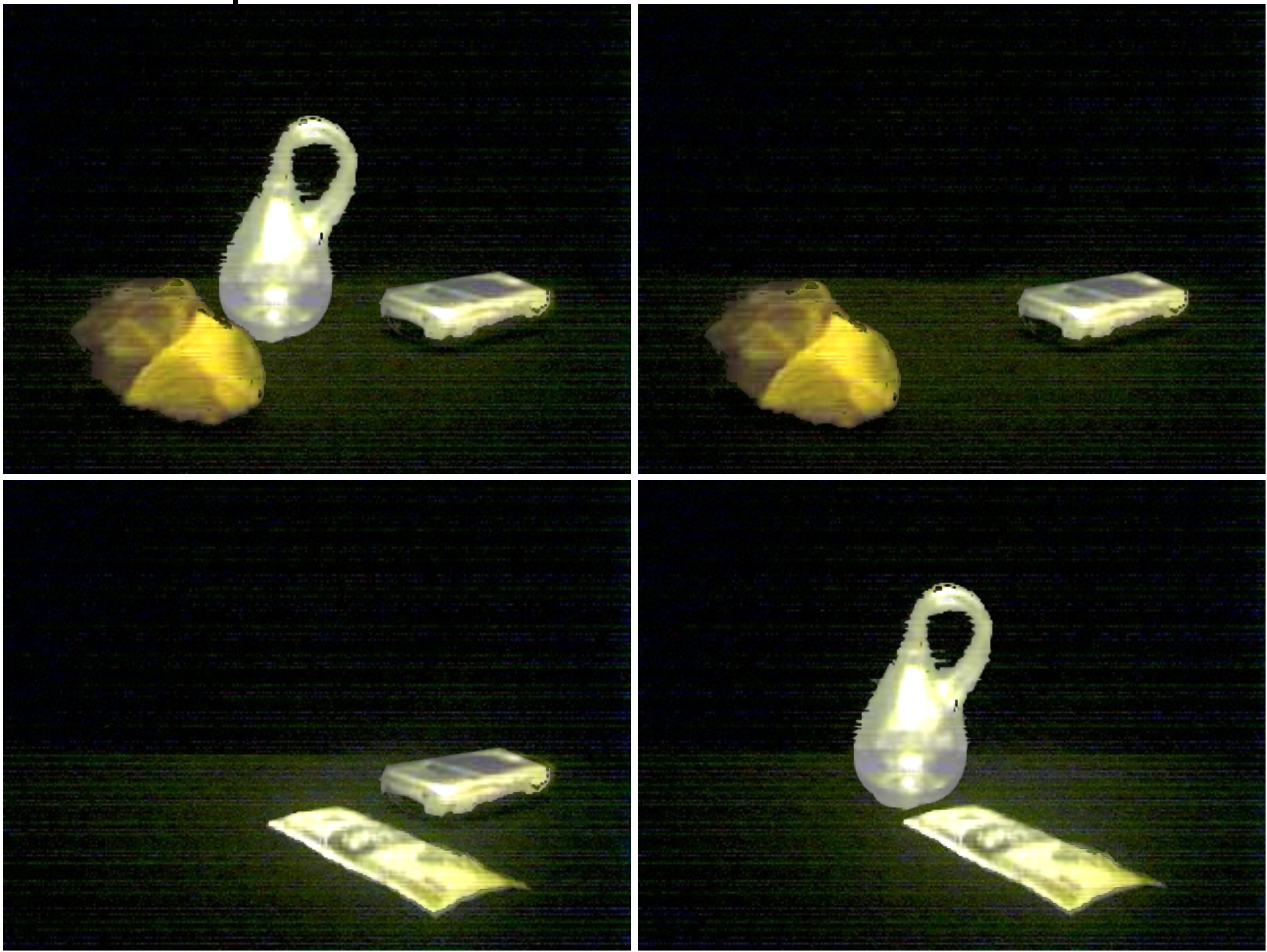
MAD-Bayes

Griffiths & Ghahramani (2006) computer vision problem “tabletop data”

[Griffiths, Ghahramani 2006]

MAD-Bayes

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

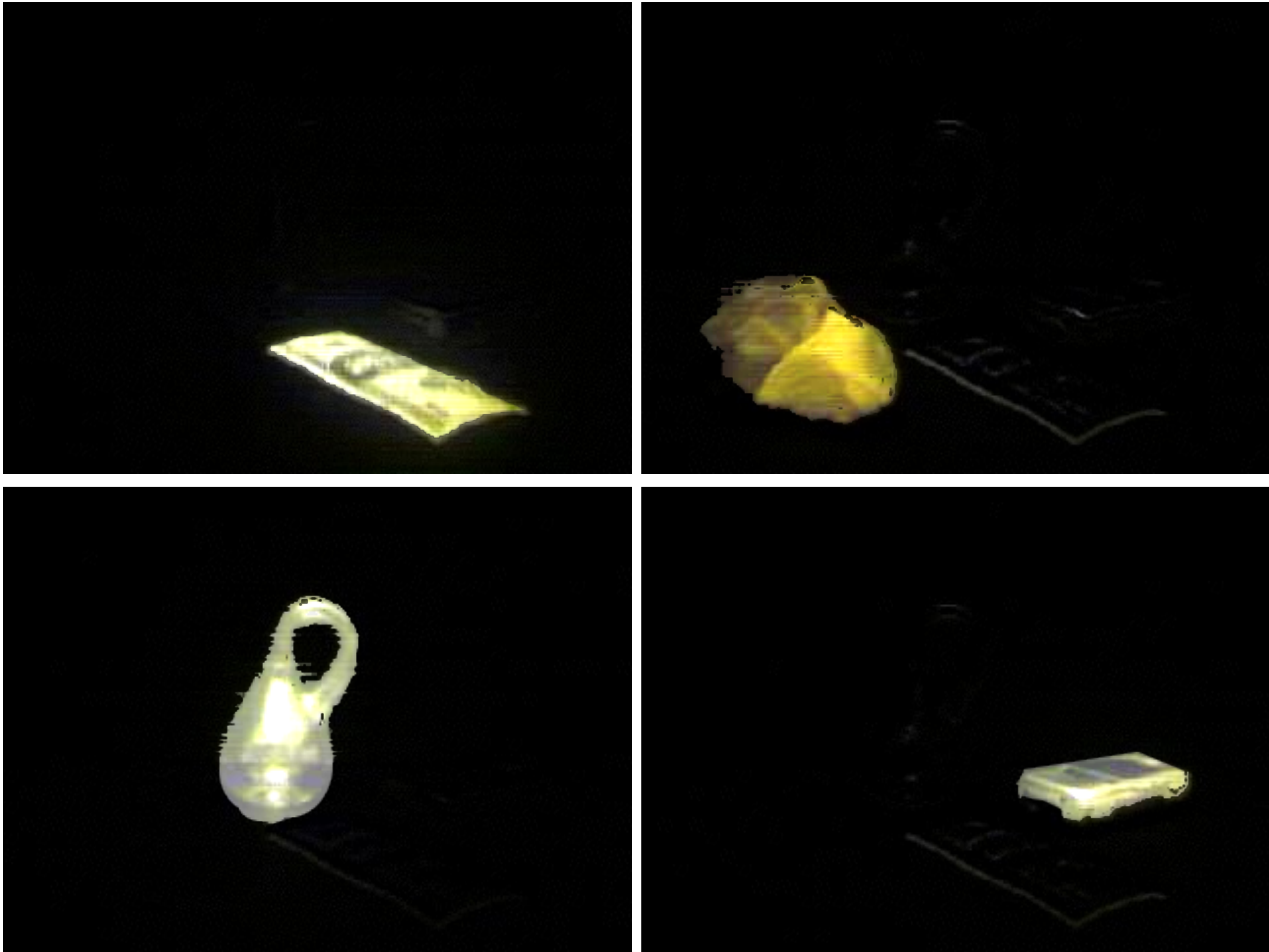


MAD-Bayes

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

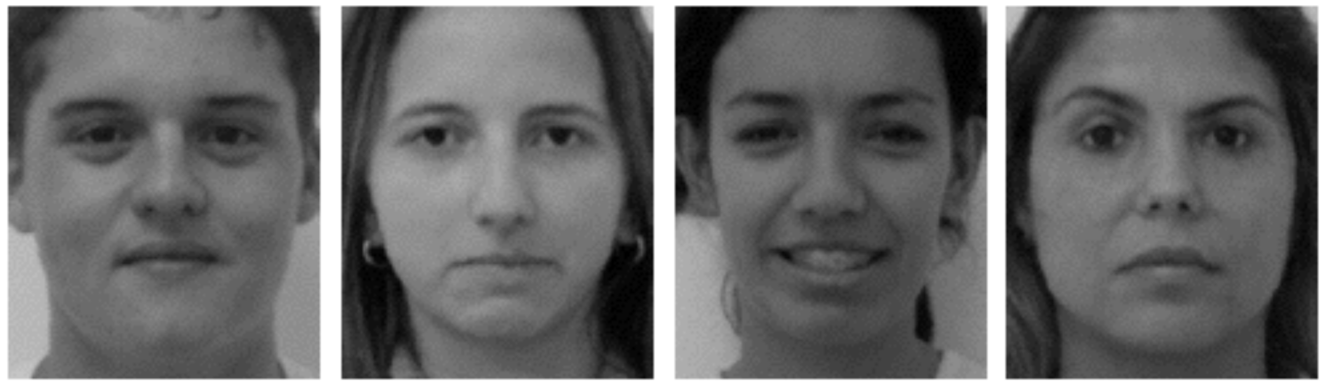
0.36 sec

Still faster by order of magnitude if restart 1000 times

Face data

Pre-aligned faces

Samples

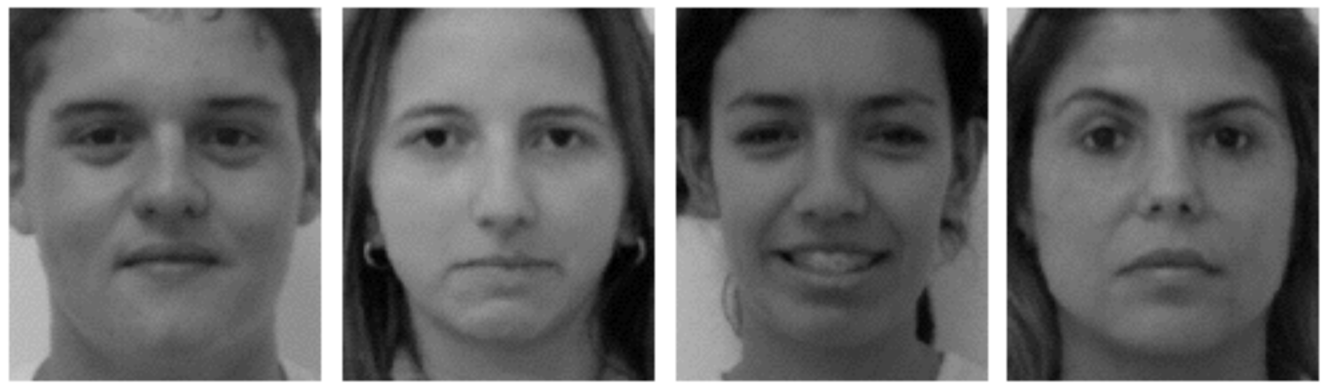


[Thomaz, Giraldi 2010]

Face data

Pre-aligned faces

Samples

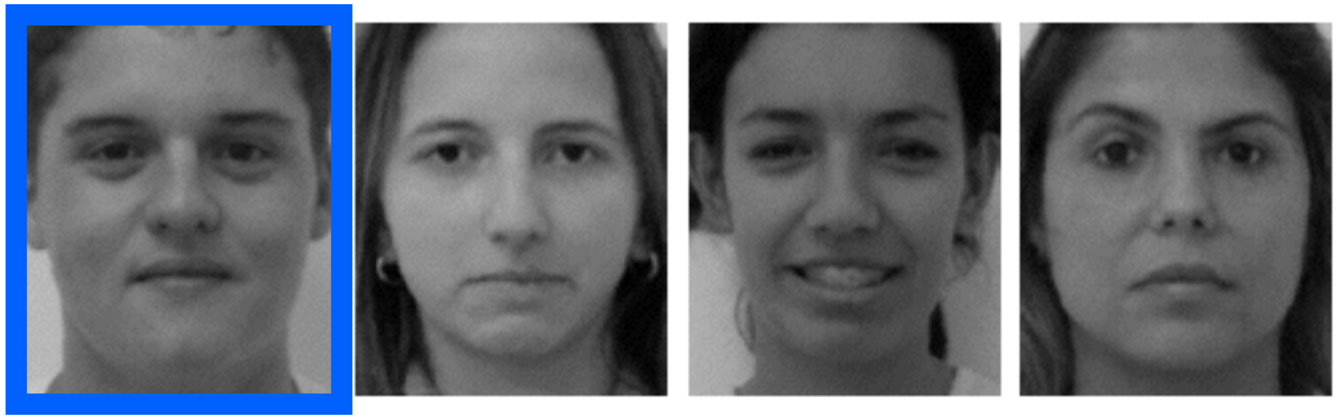


3 **<u>features</u>**
(BP-means)

Face data

Pre-aligned faces

Samples



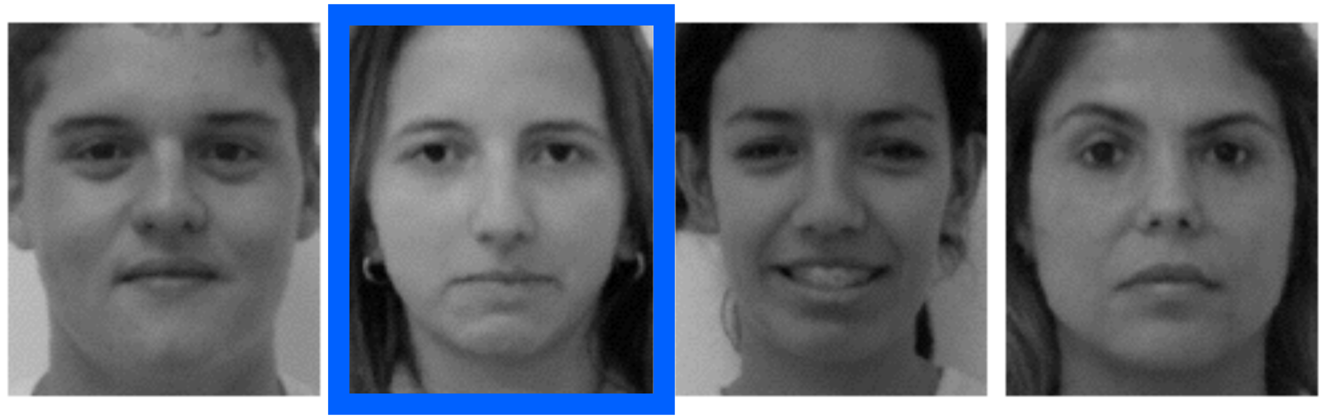
3 <u>features</u>

(BP-means)

Face data

Pre-aligned faces

Samples

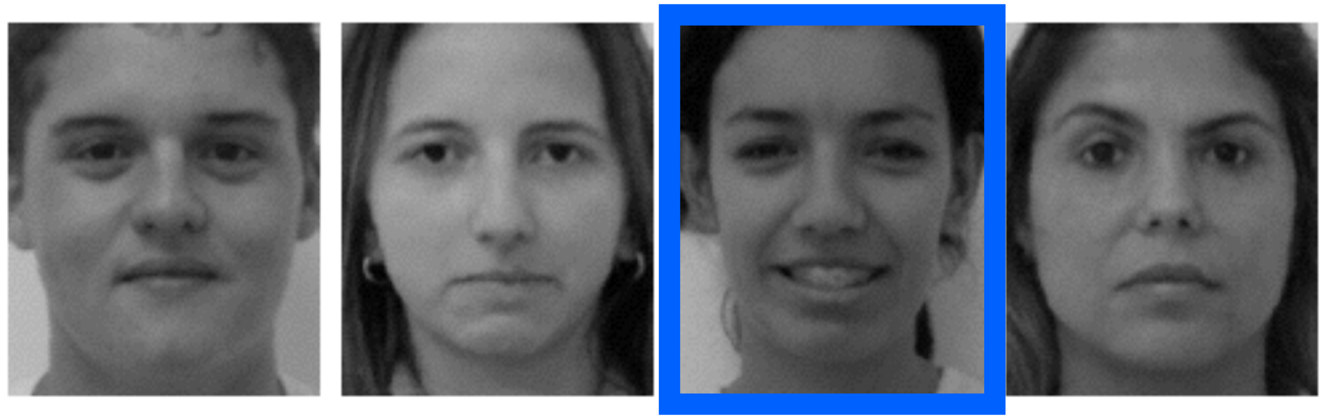


3 **features**
(BP-means)

Face data

Pre-aligned faces

Samples



3 **features**

(BP-means)

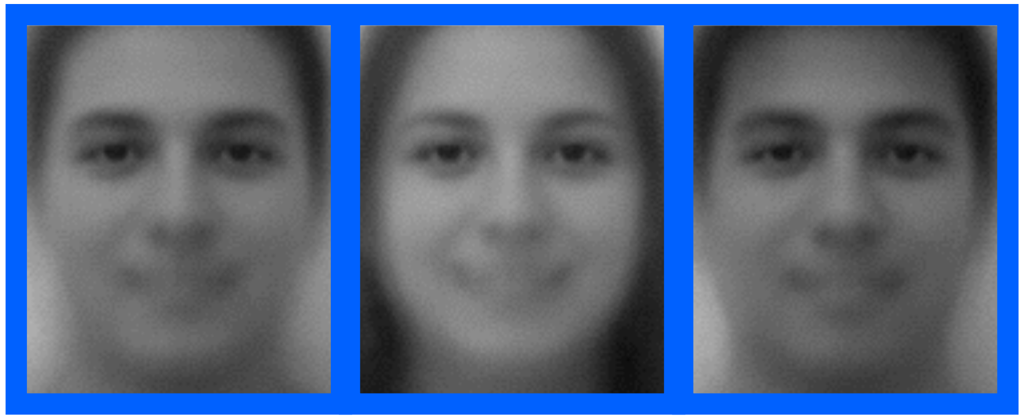
Face data

Pre-aligned faces

Samples

3 **<u>features</u>**

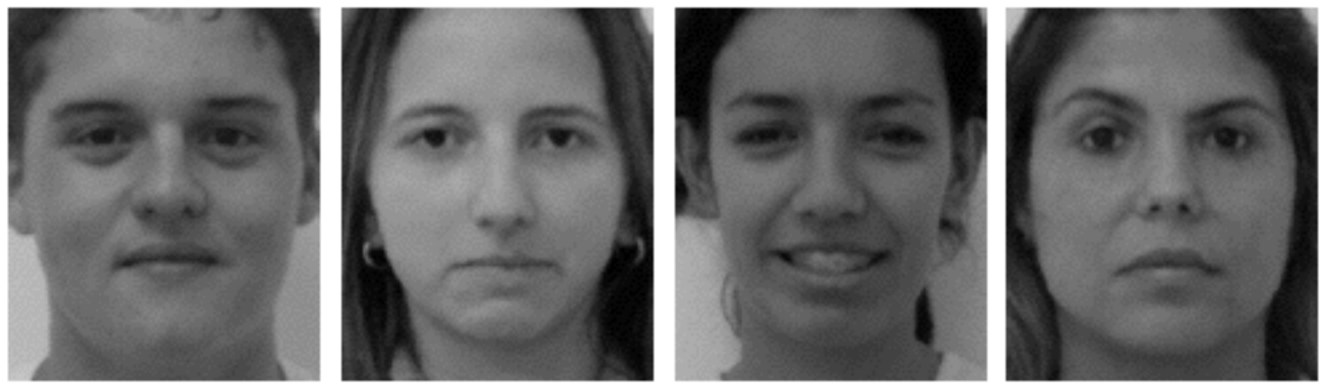
(BP-means)



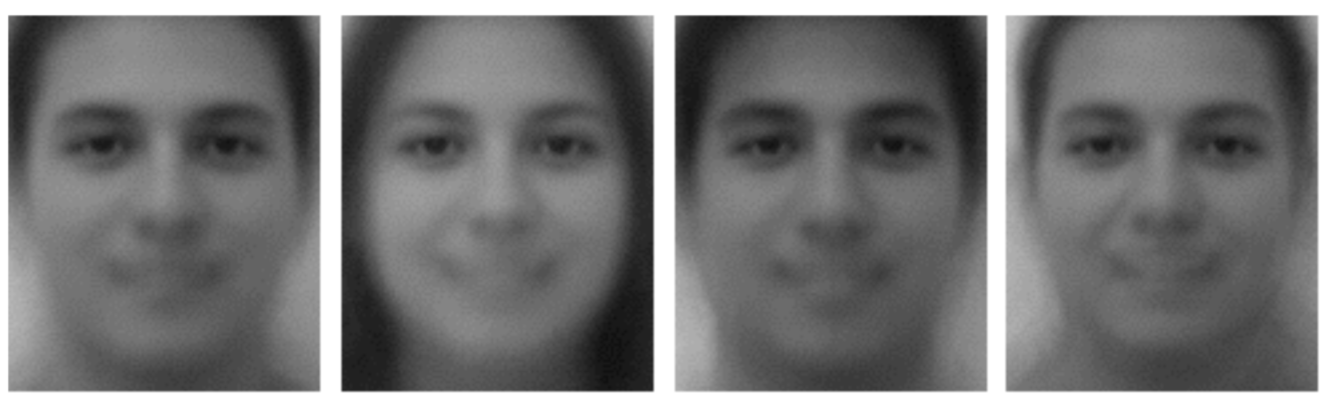
Face data

Pre-aligned faces

Samples



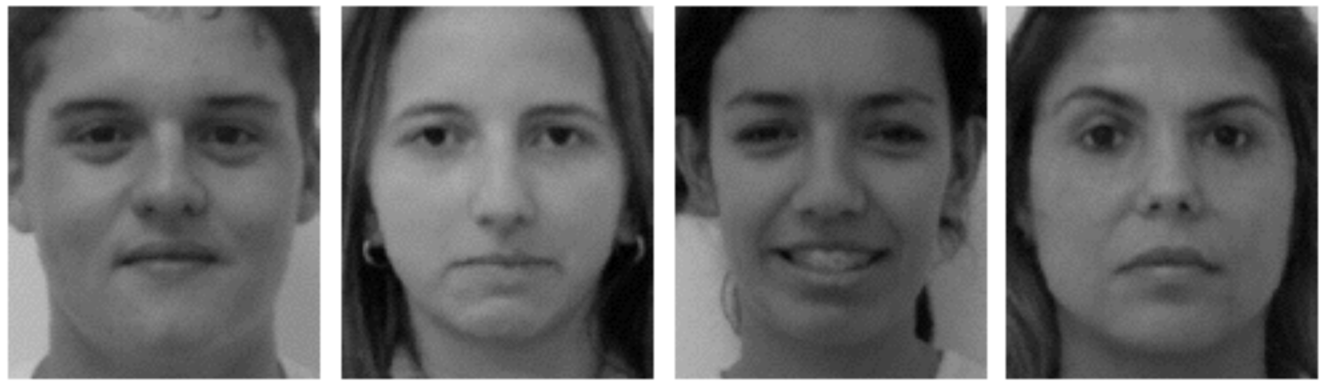
4 clusters
(K-means, K=4)



Face data

Pre-aligned faces

Samples

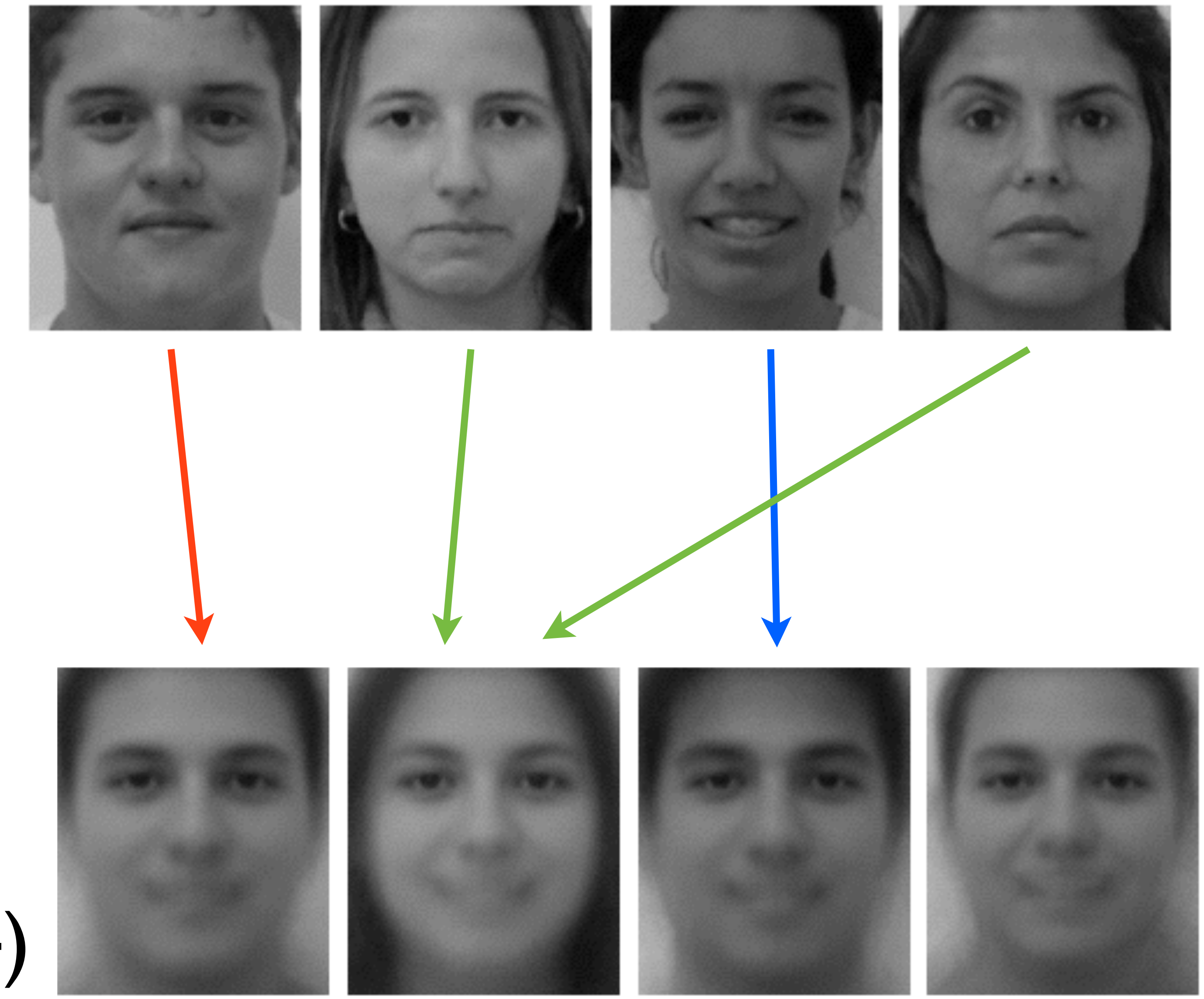


4 clusters
(K-means, K=4)

Face data

Pre-aligned faces

Samples



4 clusters
(K-means, K=4)

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

MAD-Bayes conclusions

- We provide new optimization objectives and regularizers

MAD-Bayes conclusions

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

MAD-Bayes conclusions

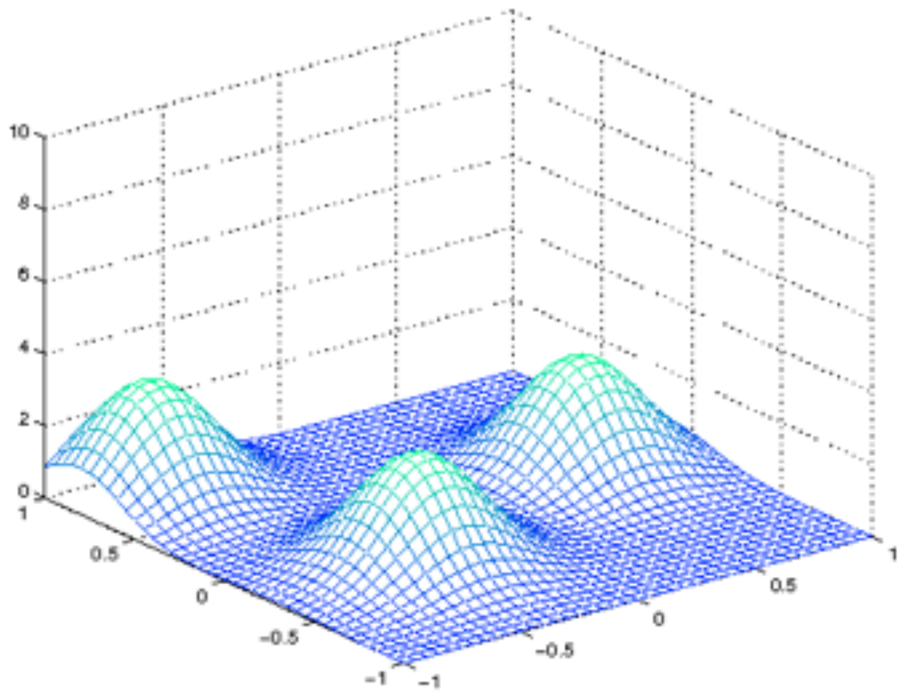
- We provide new optimization objectives and regularizers
  - In fact, general means of obtaining more
  - Straightforward, fast algorithms

What about uncertainty?

What about uncertainty?

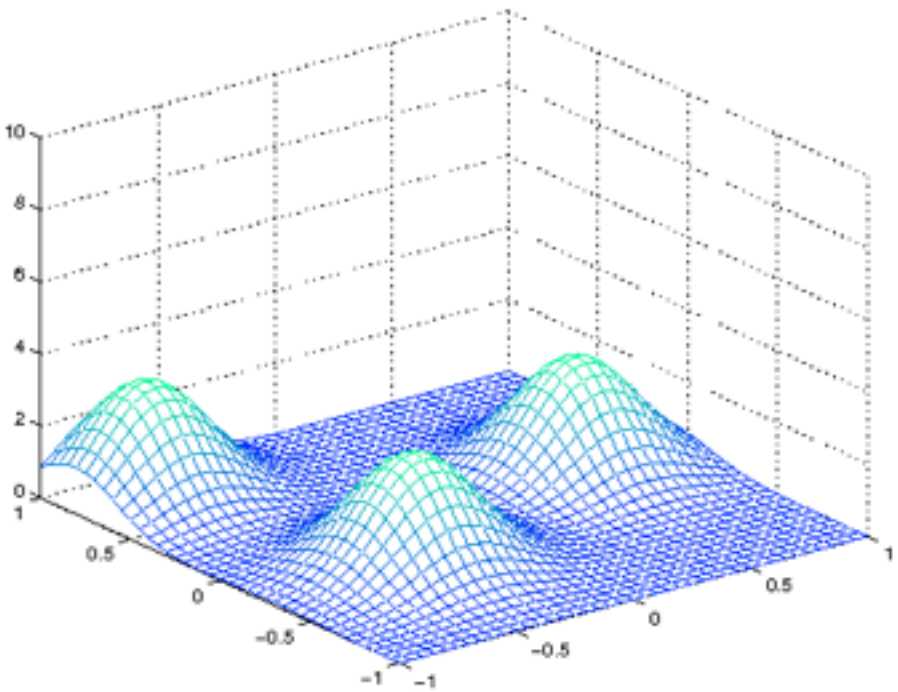
- Variational Bayes (VB)

What about uncertainty?

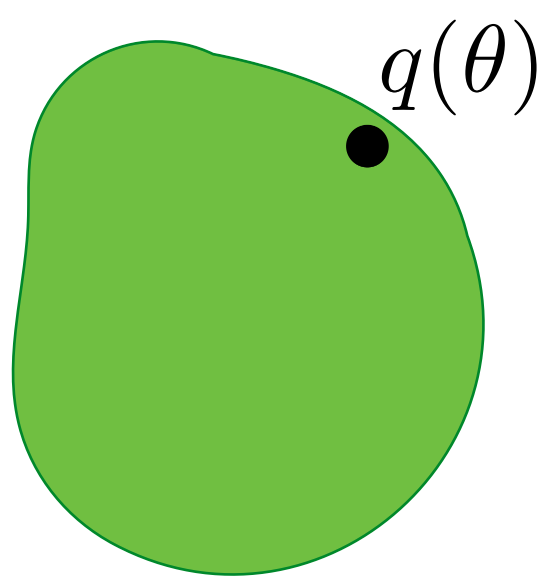


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$

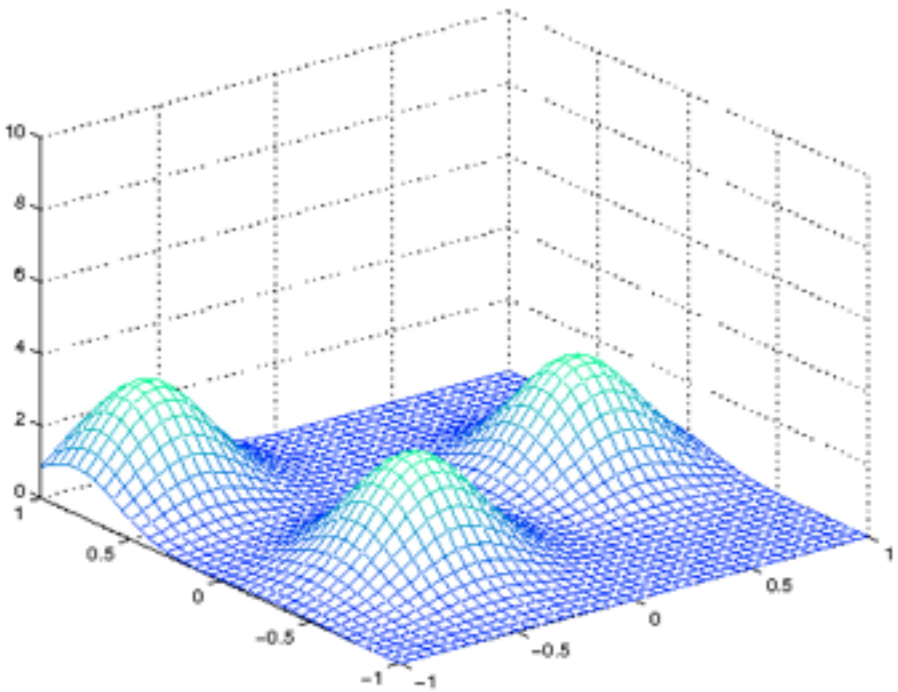
What about uncertainty?



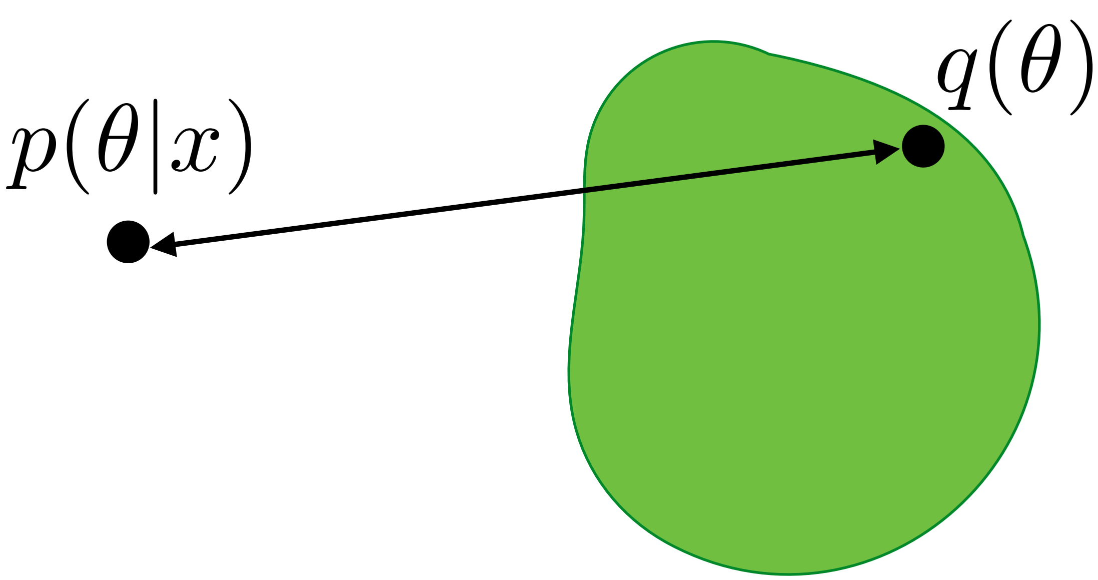
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$



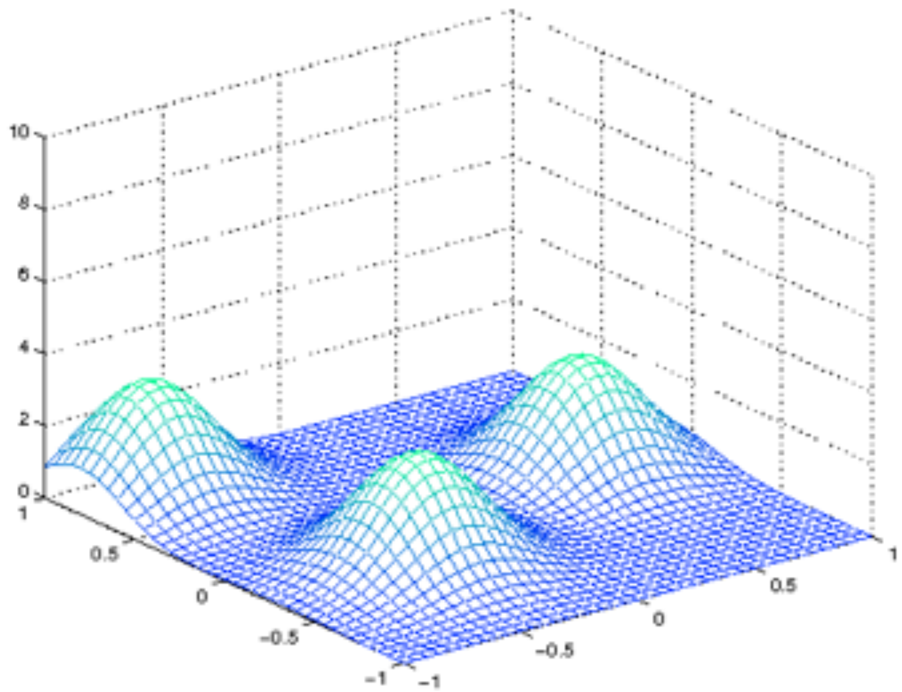
What about uncertainty?



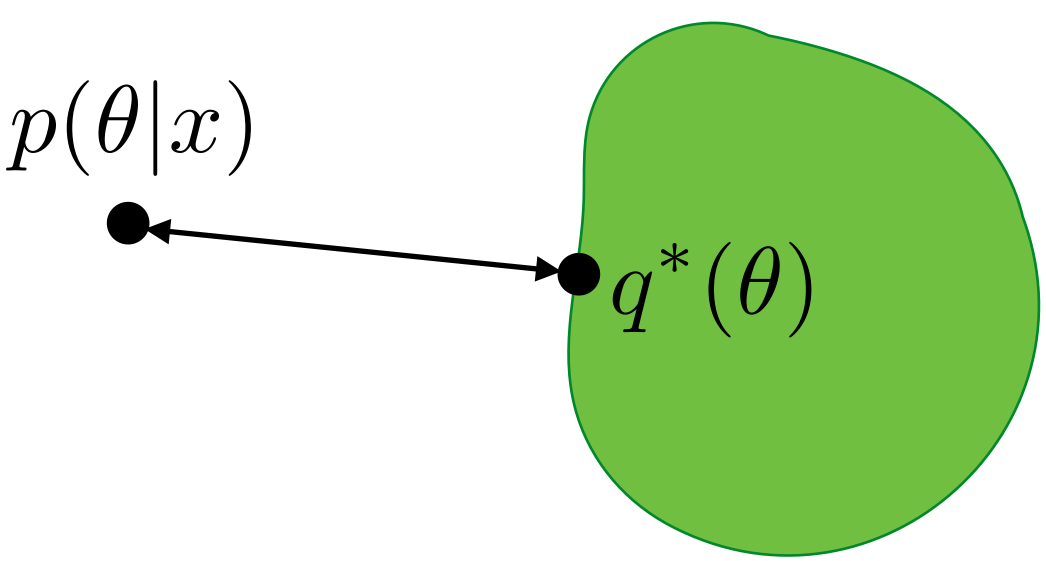
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$



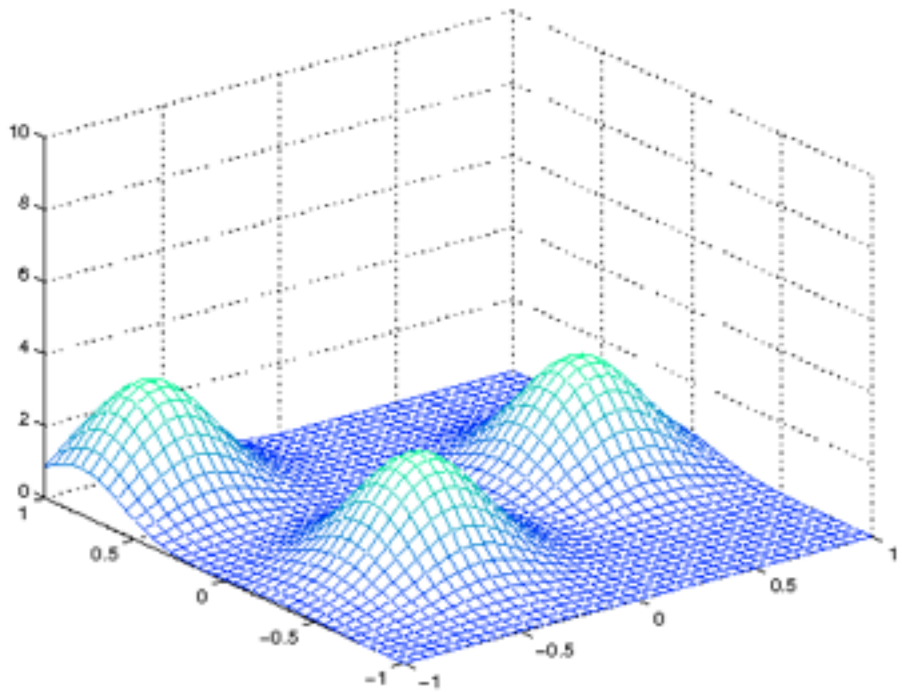
What about uncertainty?



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

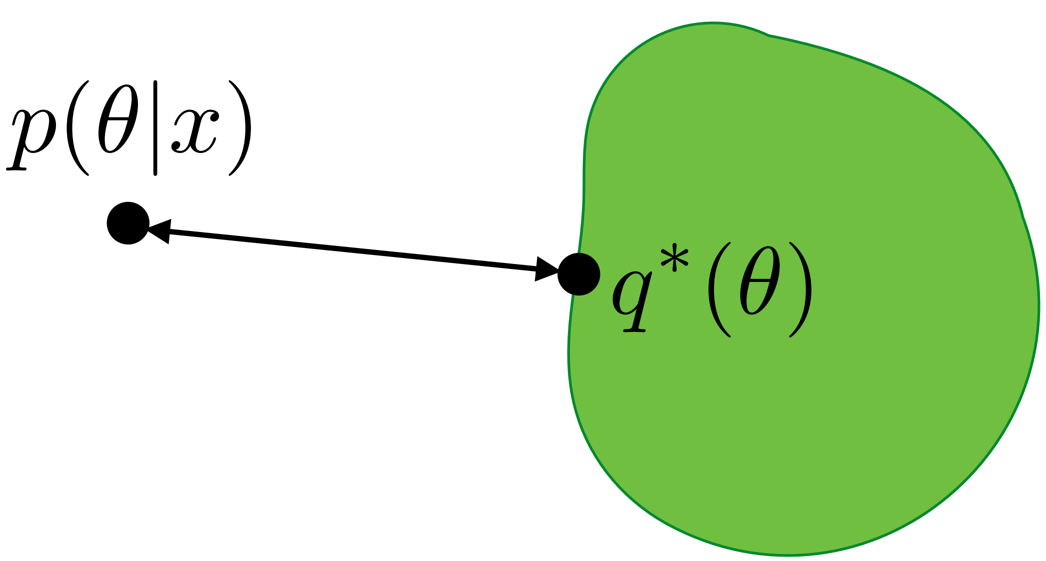


What about uncertainty?

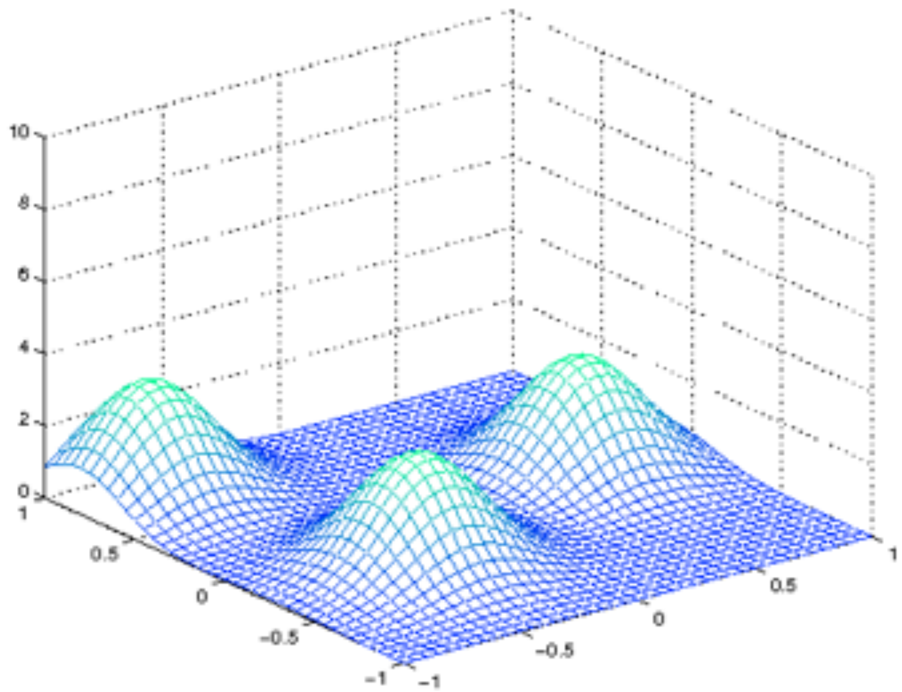


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

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

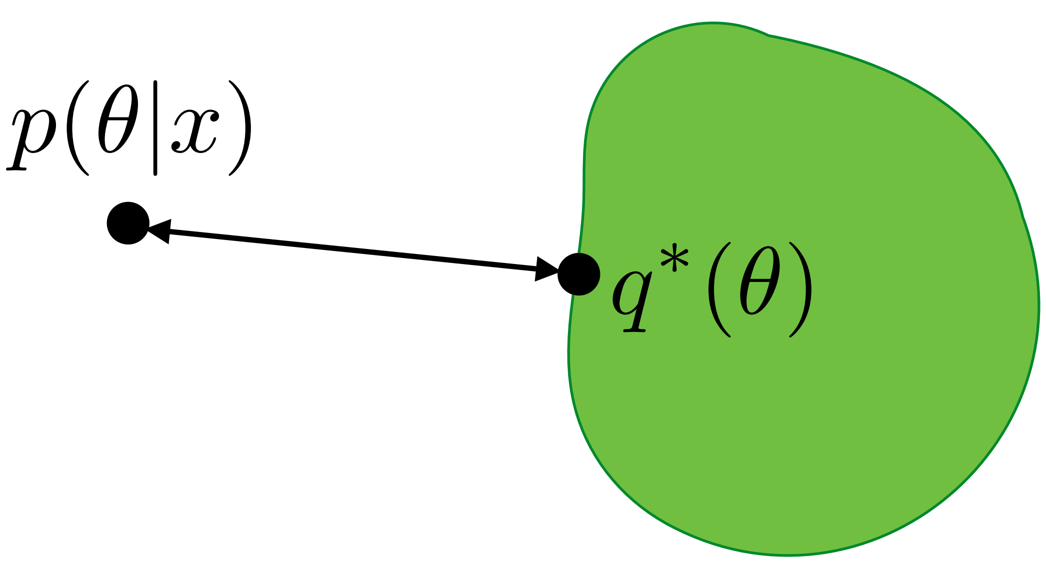


What about uncertainty?



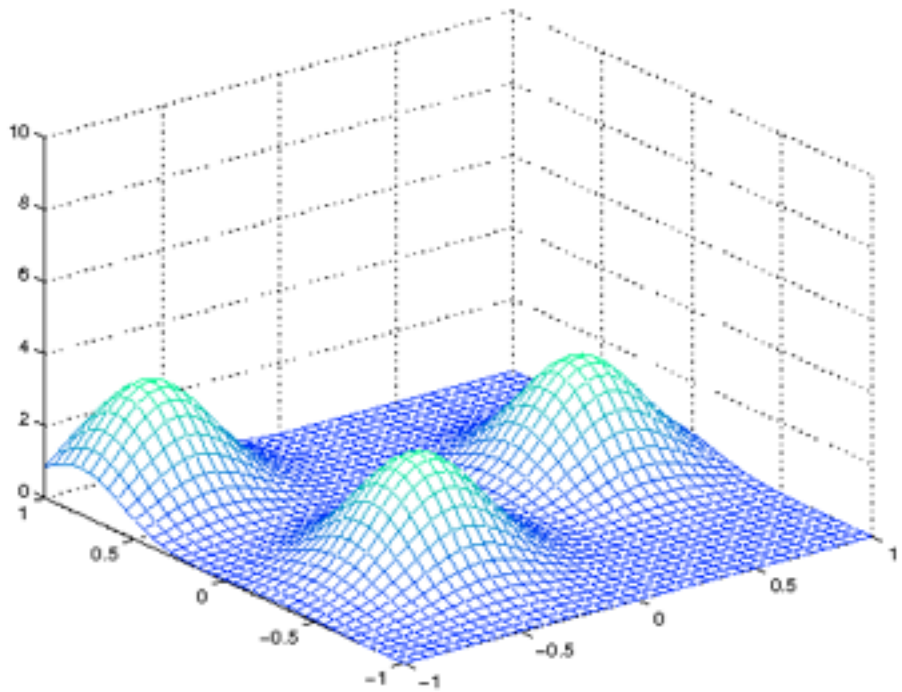
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

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



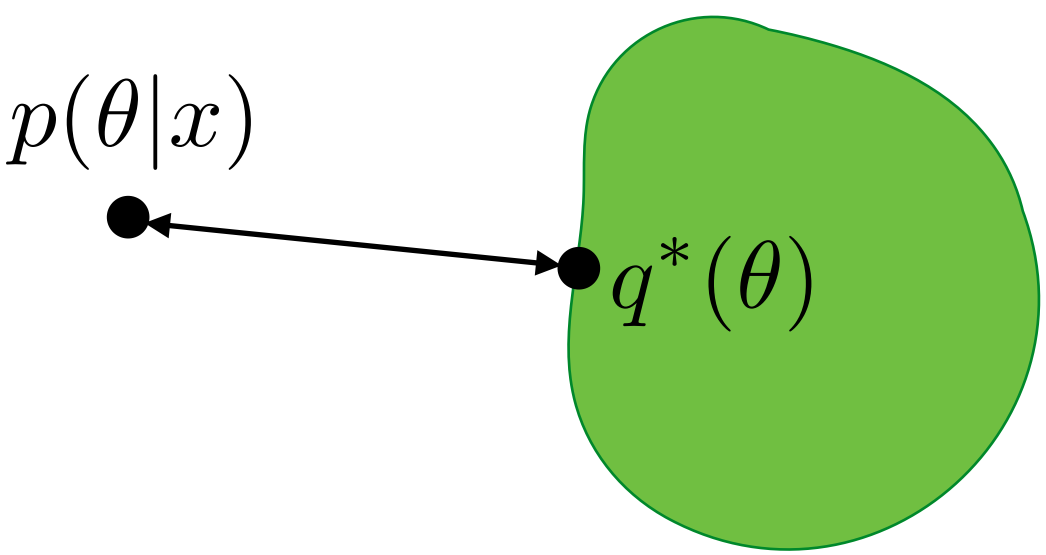
- VB practical success

What about uncertainty?



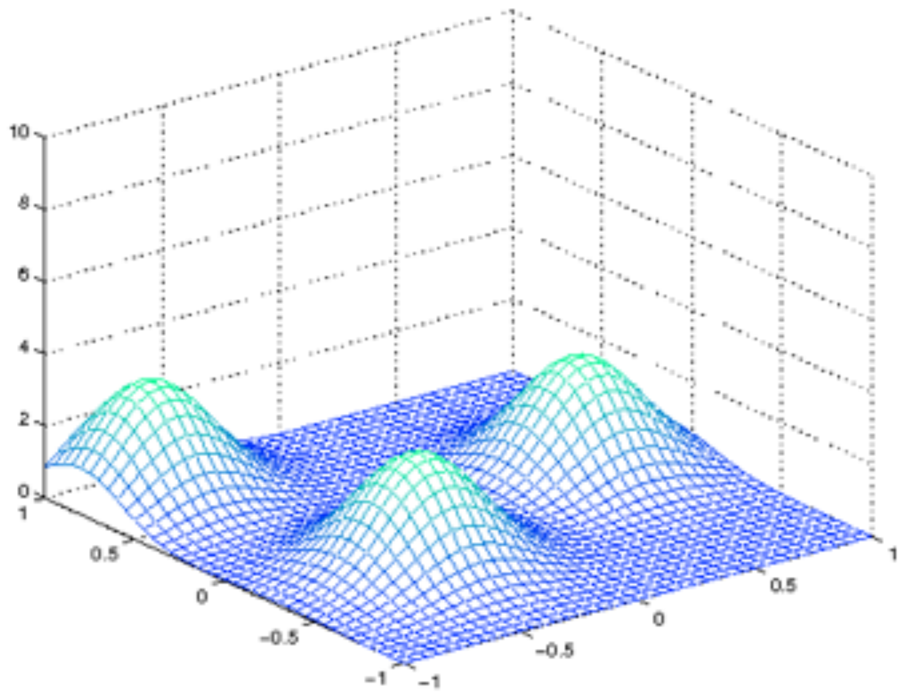
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

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



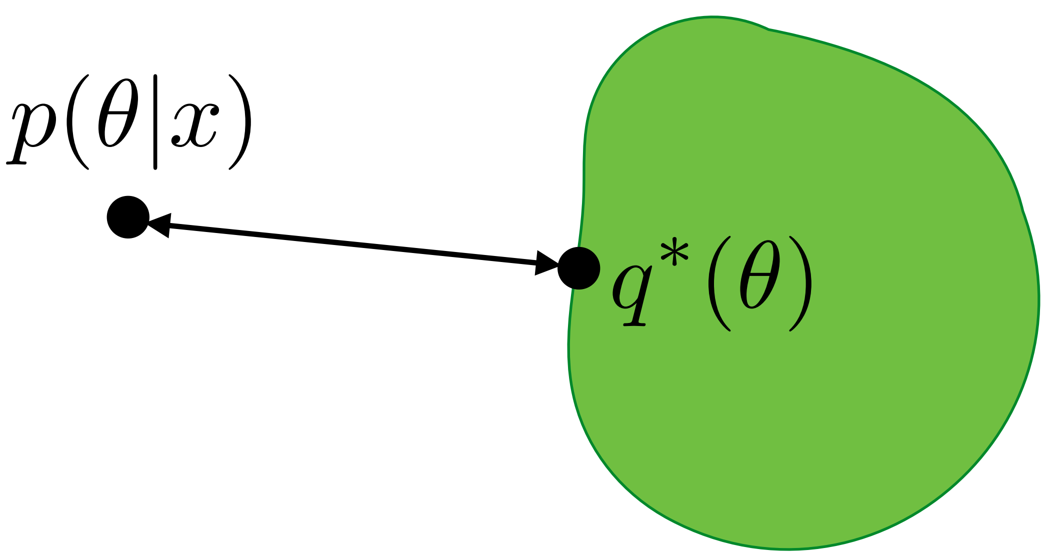
- VB practical success
  - point estimates and prediction

What about uncertainty?



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

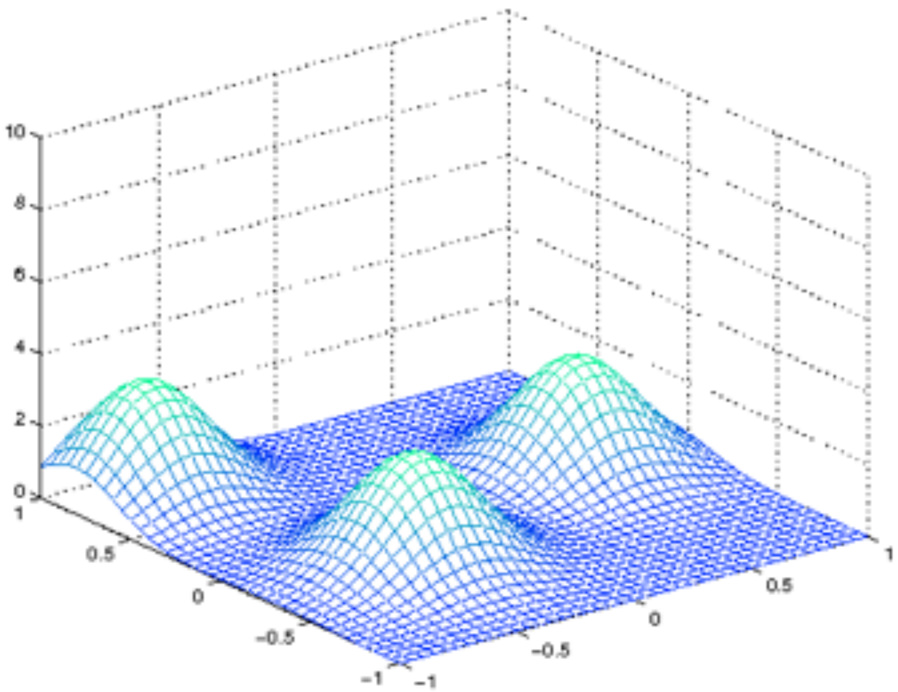
$$KL(q||p(\cdot|x))$$



- VB practical success
  - point estimates and prediction
  - fast

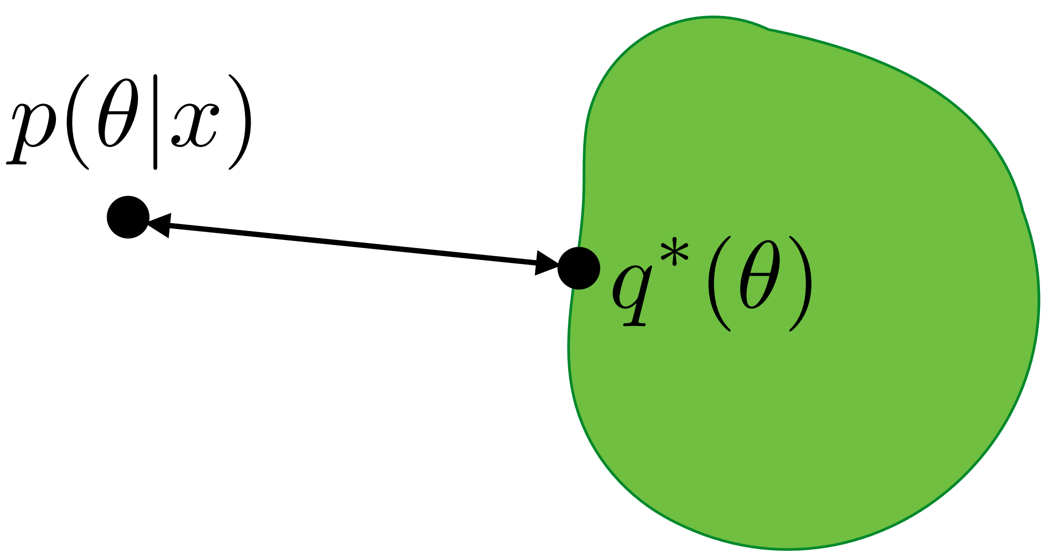
[Broderick, Boyd, Wibisono, Wilson, Jordan 2013]

What about uncertainty?



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

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



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

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

What about uncertainty?

What about uncertainty?

- Variational Bayes

What about uncertainty?

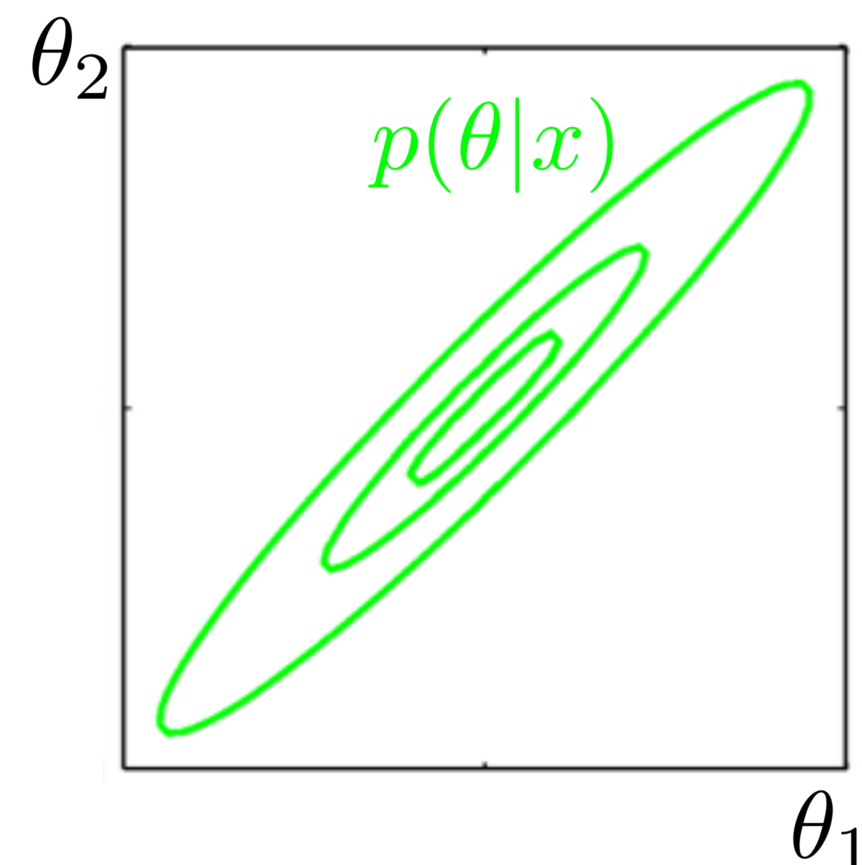
- Variational Bayes

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

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]

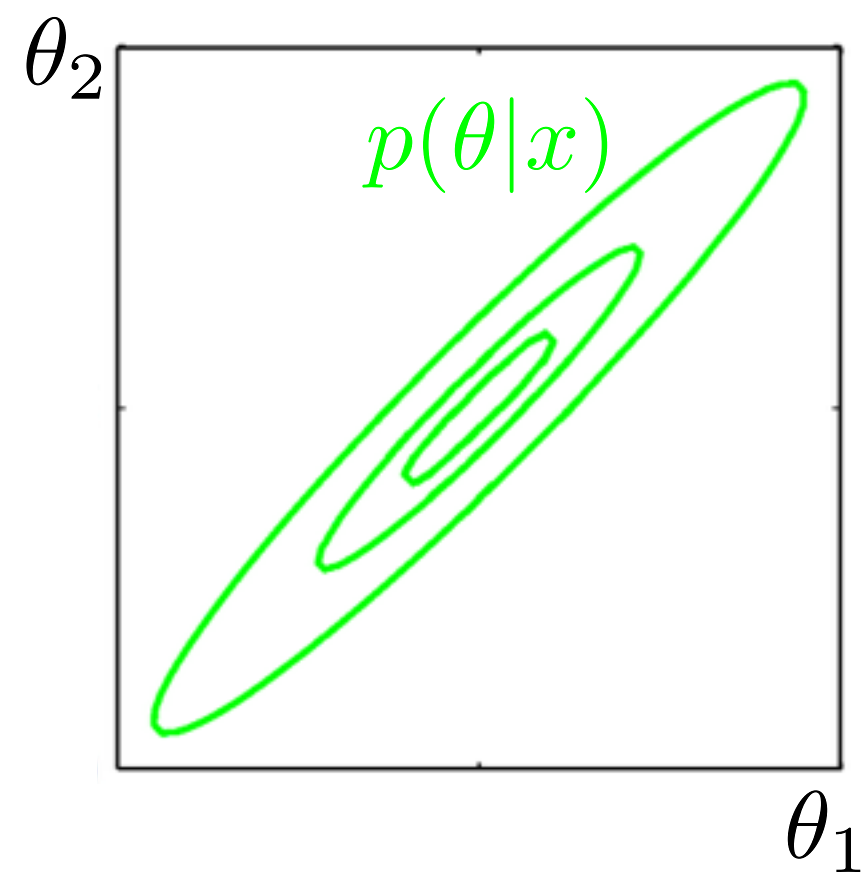
What about uncertainty?

- 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]

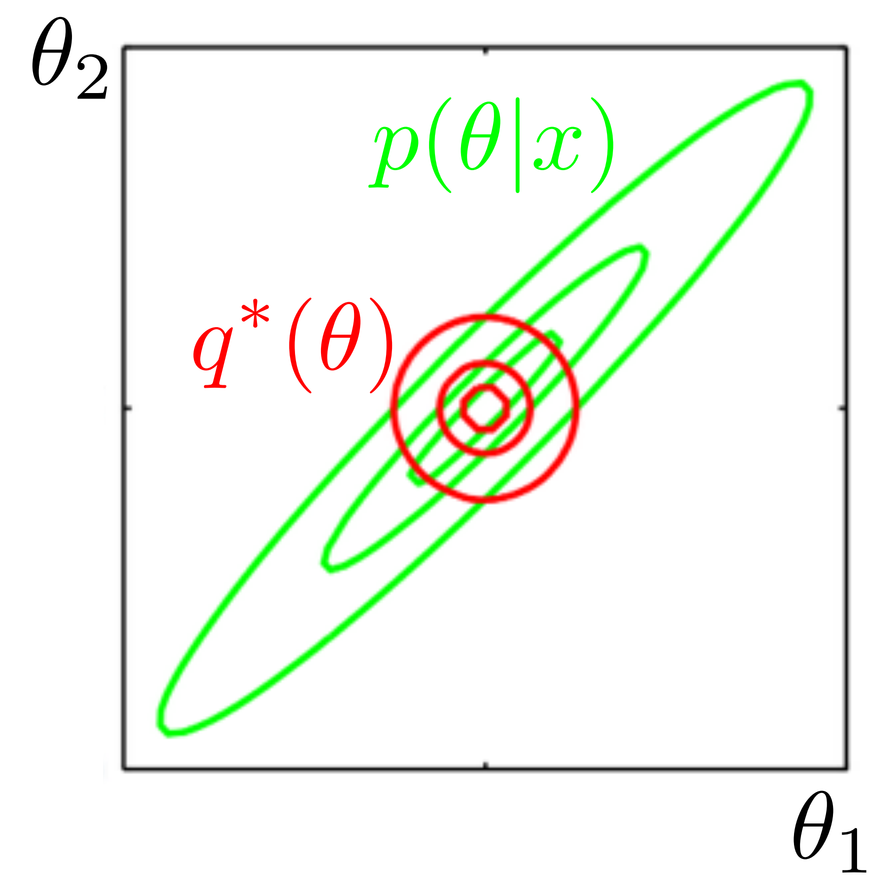
What about uncertainty?

- 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?

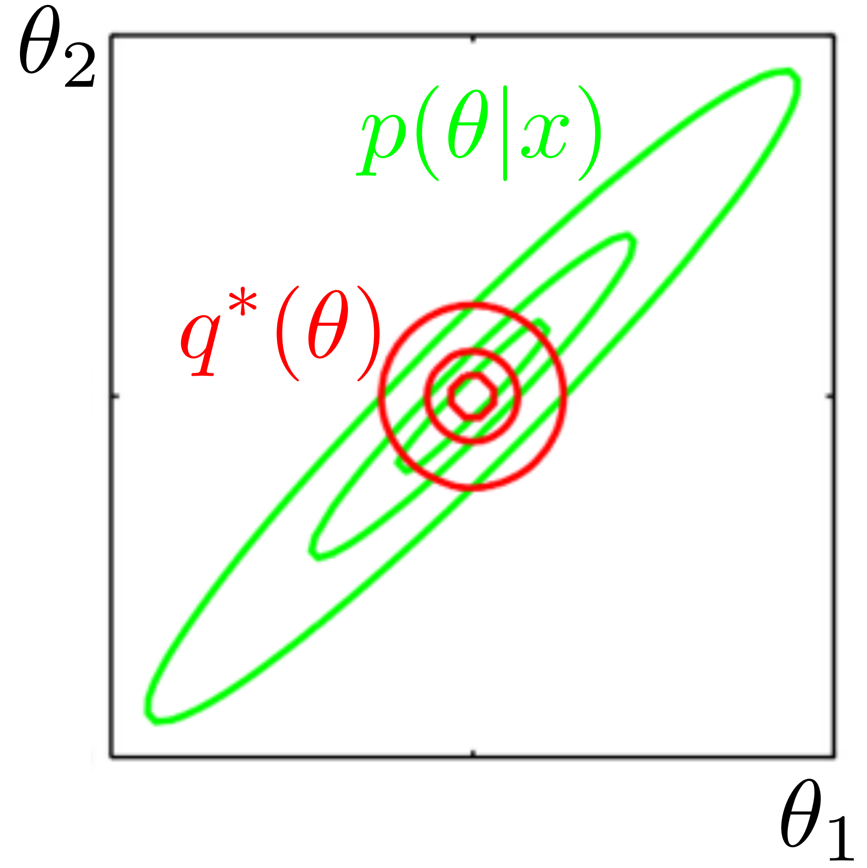
- Variational Bayes

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

- 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

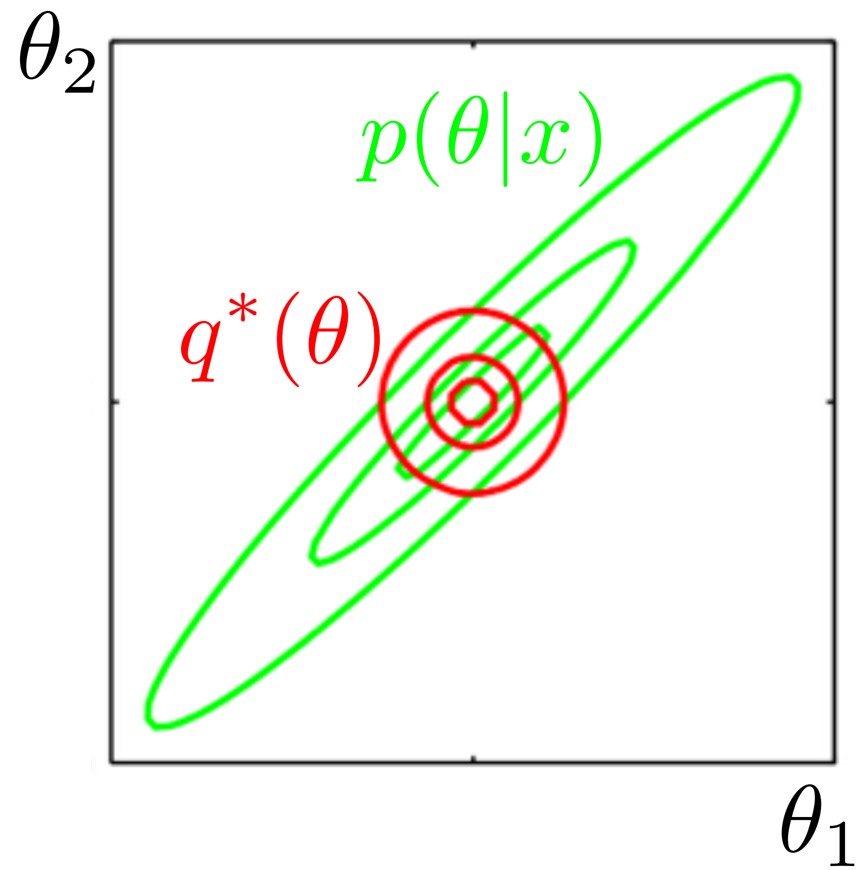
$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- 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

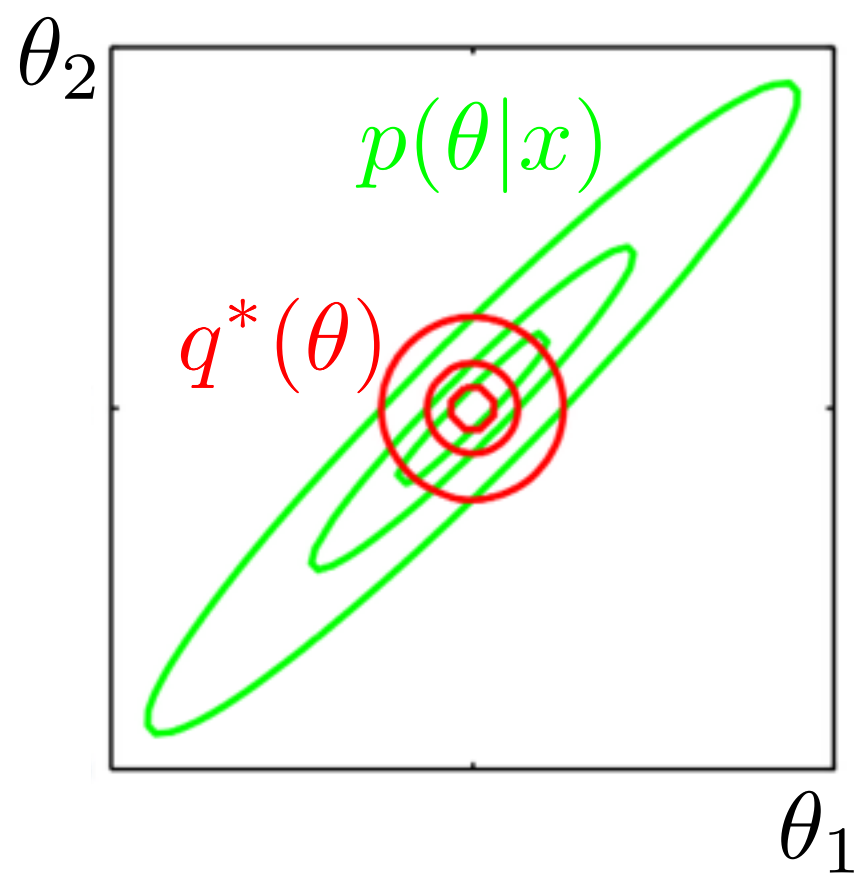
$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- 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

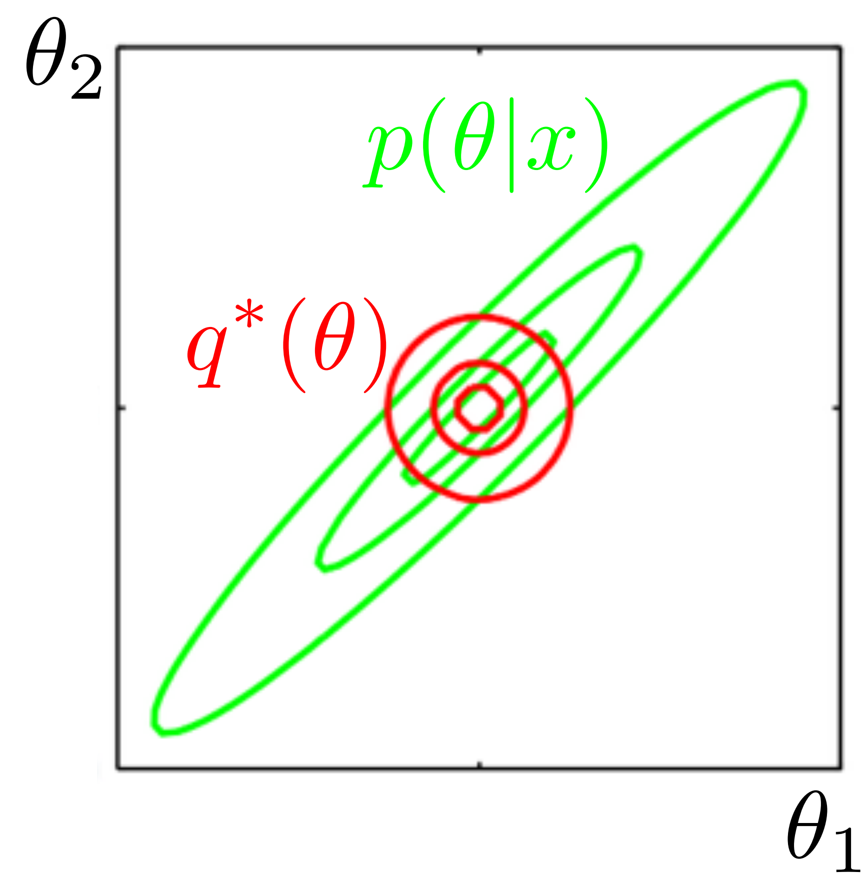
$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- 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]
[Dunson 2014; Bardenet, Doucet, Holmes 2015]

1. Derive *Linear Response Variational Bayes* (LRVB) variance/covariance correction

2. Accuracy experiments

3. Scalability

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

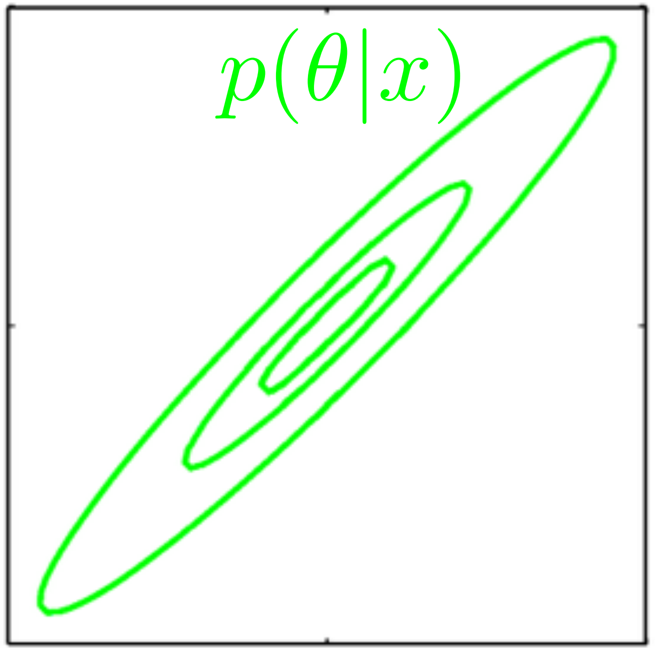
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad\qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

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]

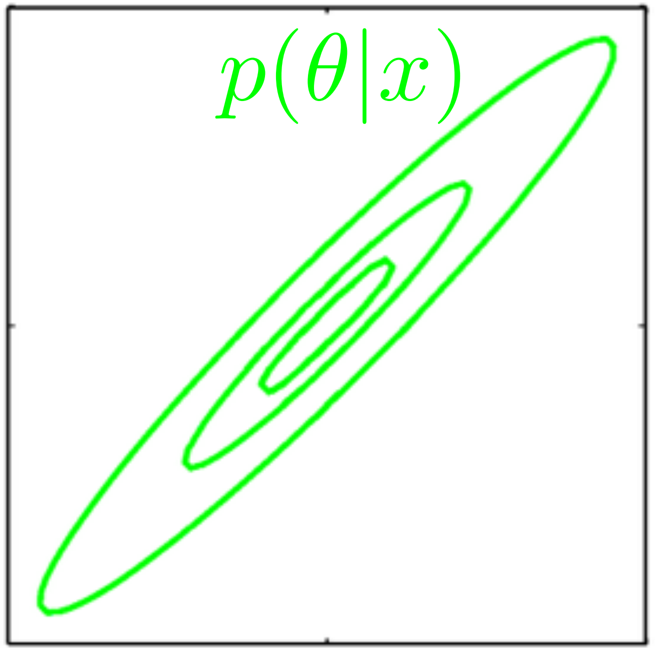
Linear response

- Cumulant-generating function

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

- True posterior covariance

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



[Bishop 2006]

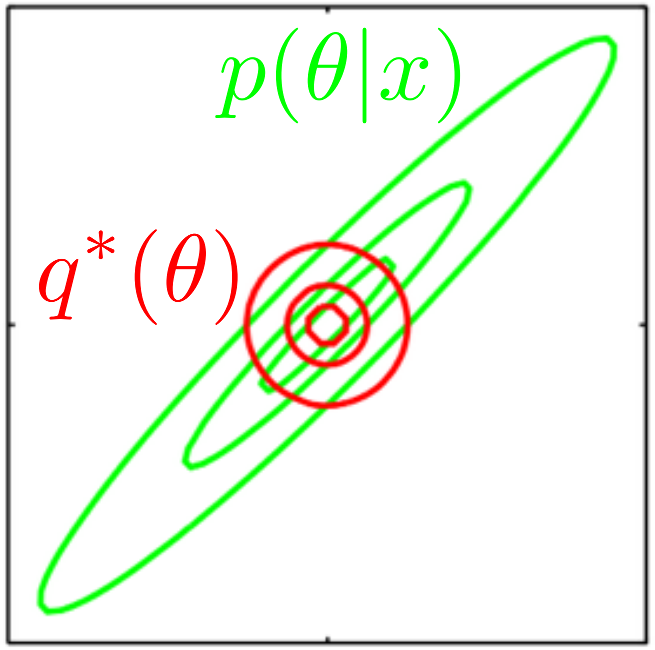
Linear response

- 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]

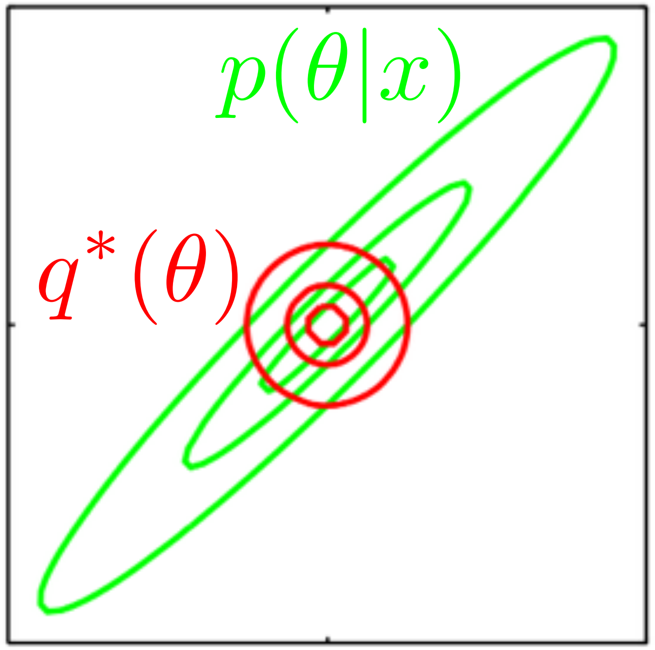
Linear response

- 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

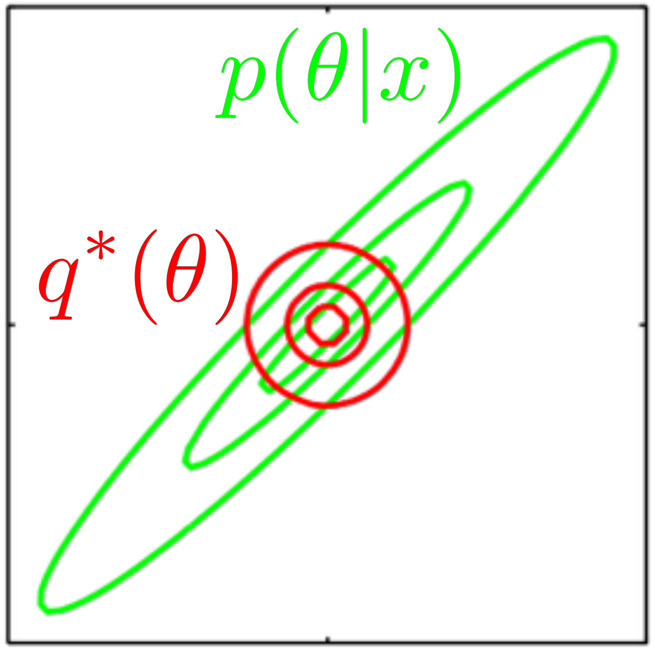
- Cumulant-generating function

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

- 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

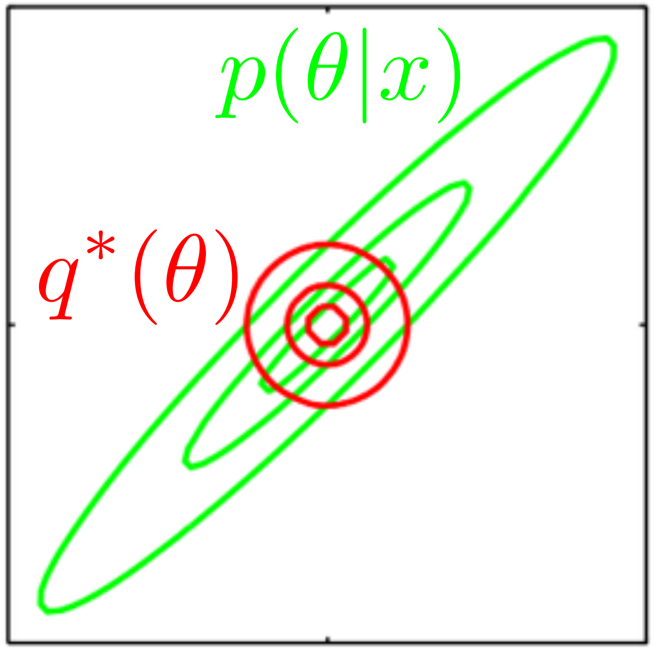
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- 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

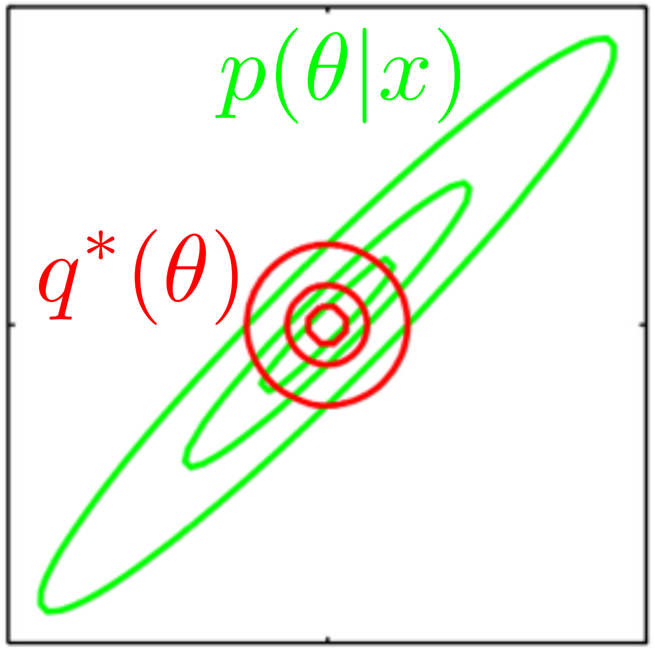
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- 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

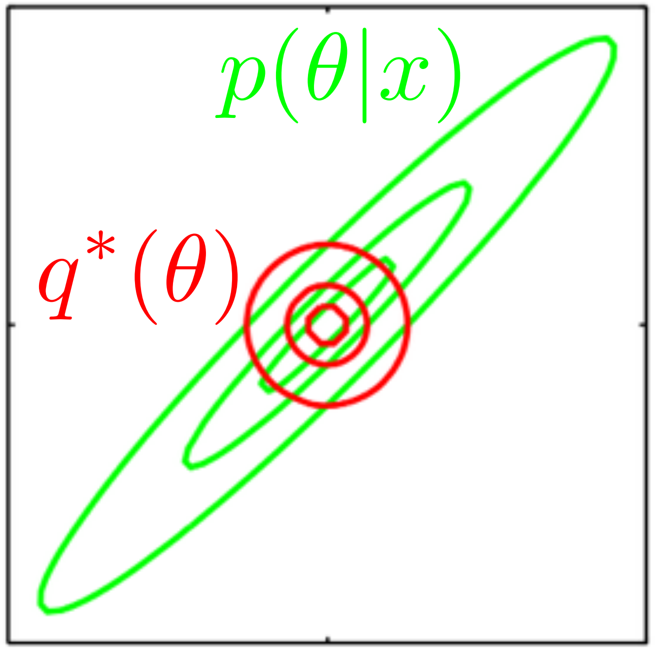
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- 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"

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



[Bishop 2006]

Linear response

- Cumulant-generating function

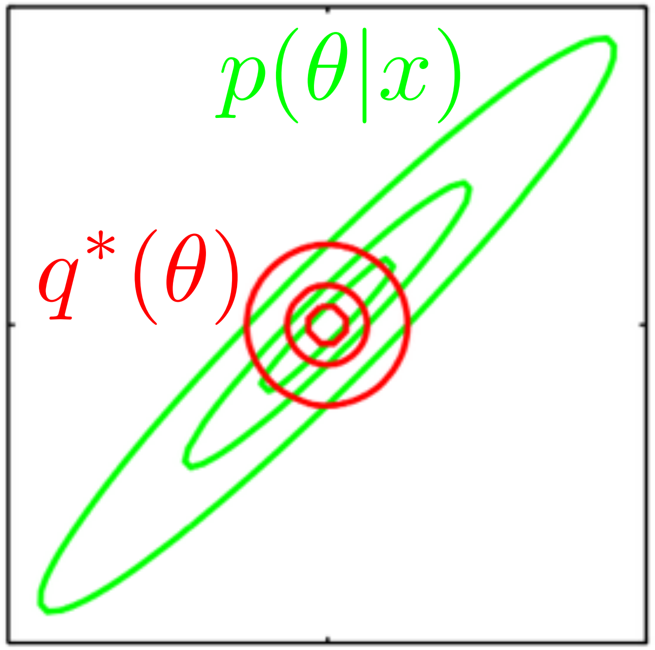
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- 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

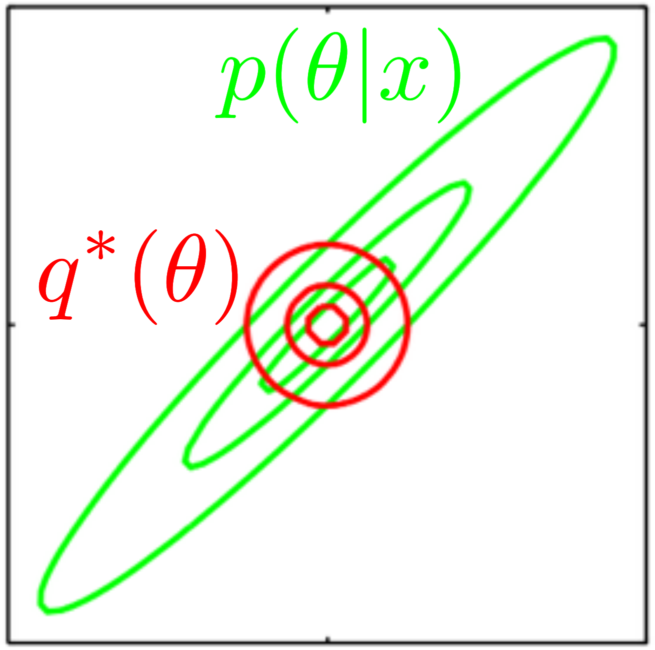
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- 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"

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



[Bishop 2006]

Linear response

- Cumulant-generating function

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

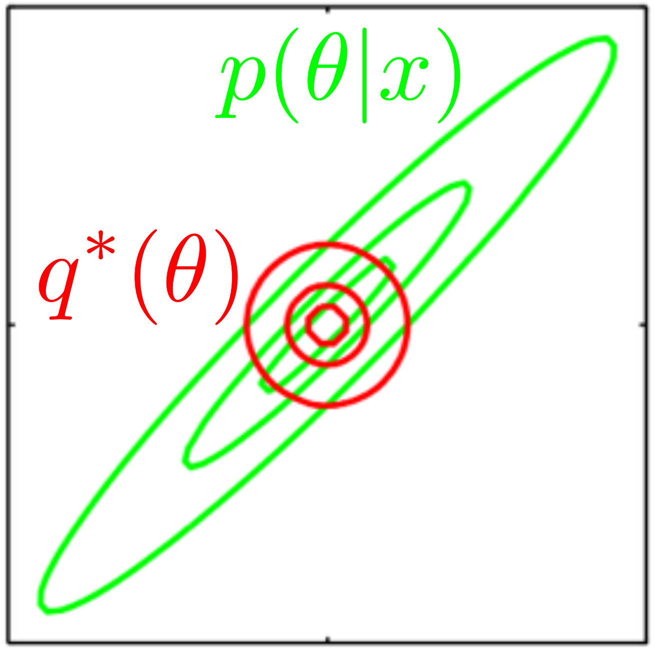
- True posterior covariance vs MFVB covariance

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

- "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

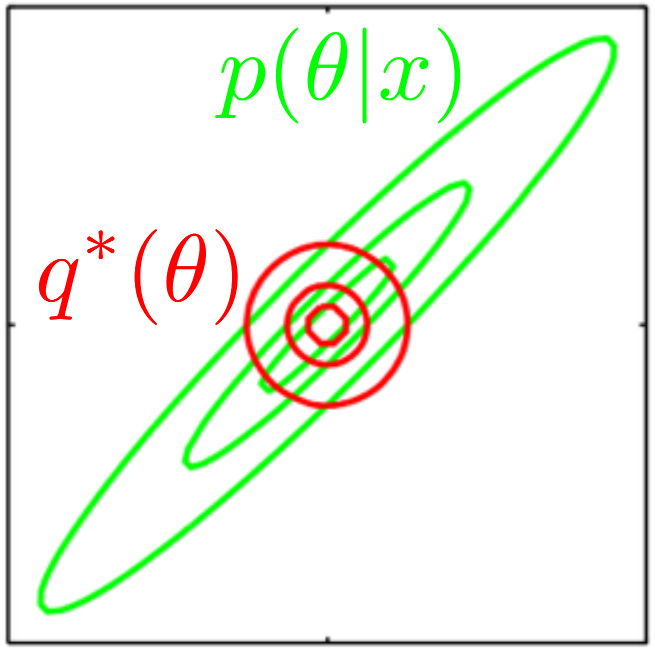
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

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

- The LRVB approximation

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



[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

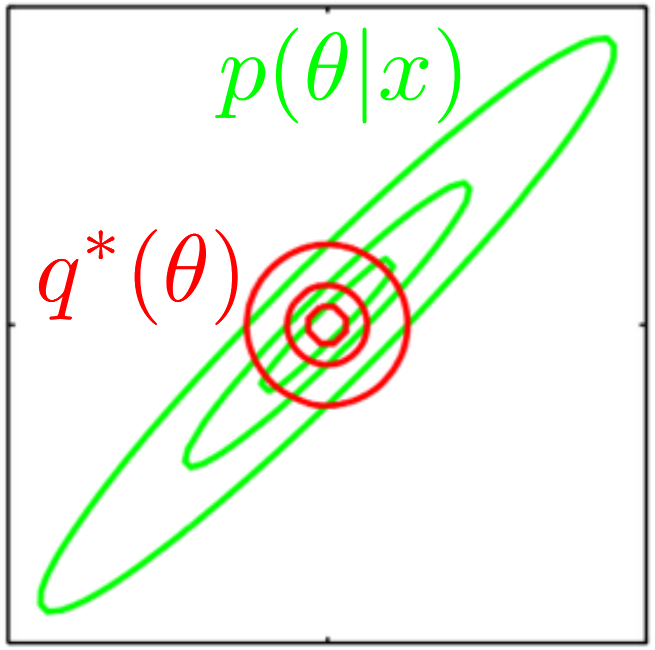
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "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

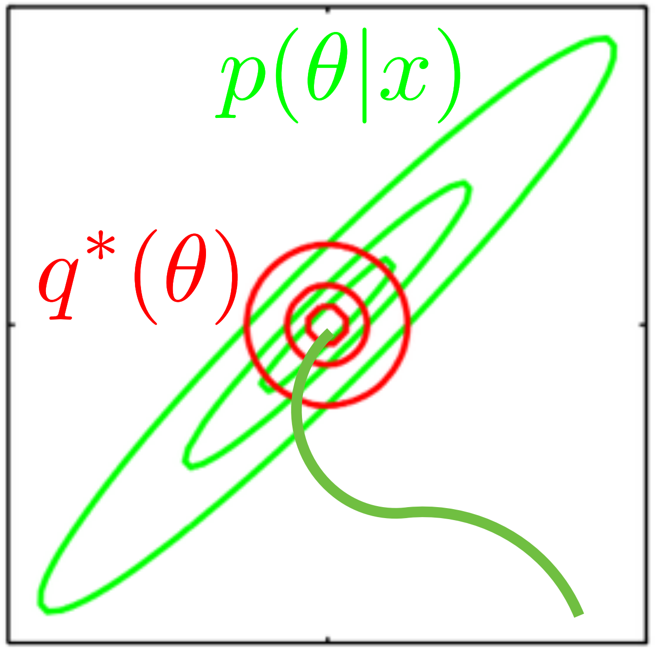
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "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

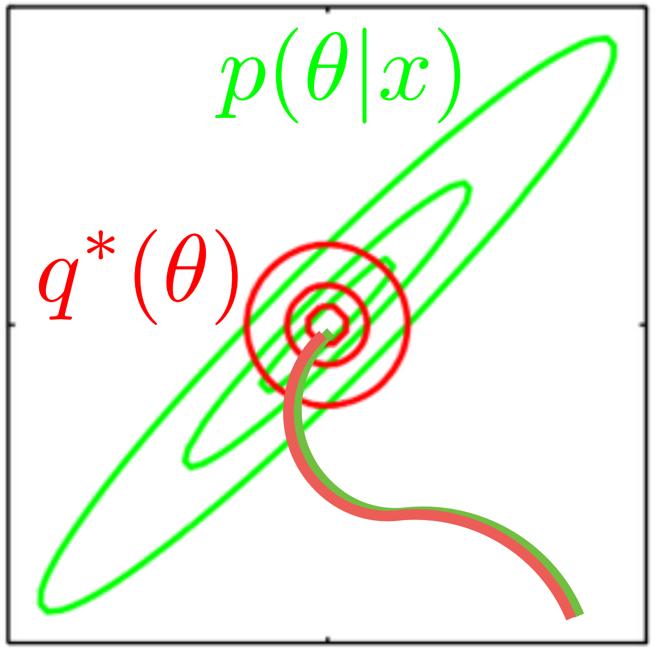
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

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

- The LRVB approximation

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



[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

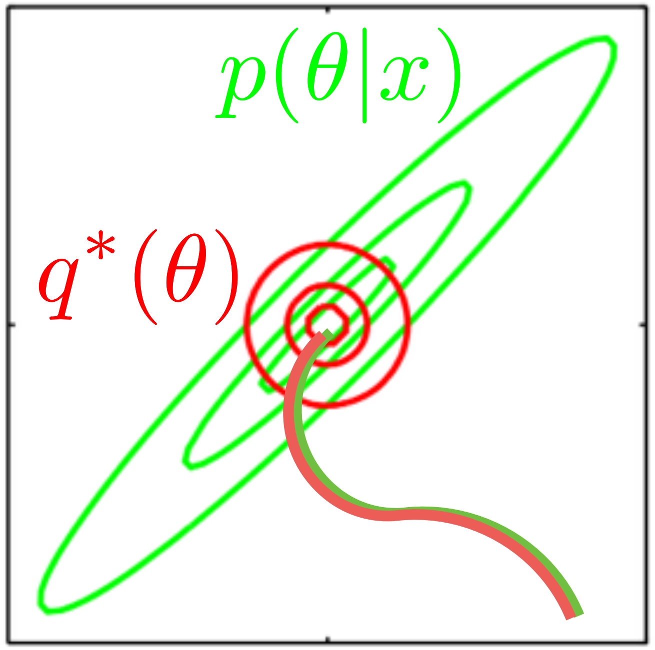
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "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

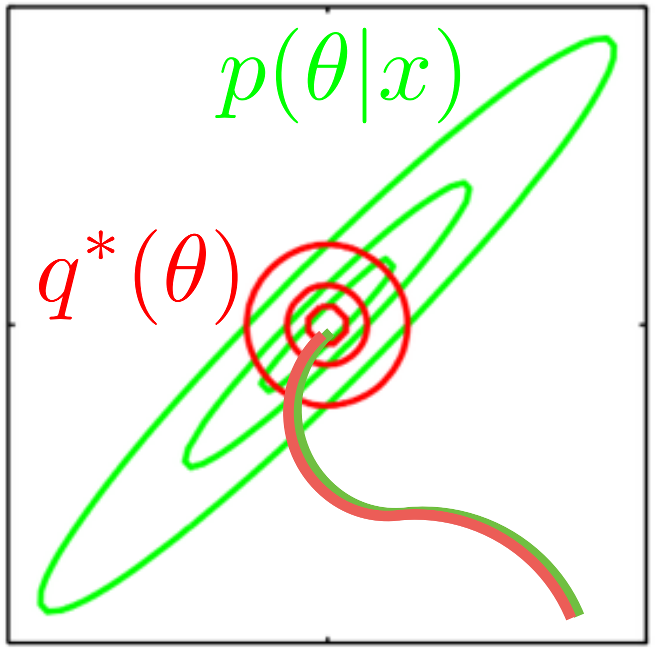
$$\Sigma := \left.\frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t)\right|_{t=0} \qquad V := \left.\frac{d^2}{dt^T dt} C_{q^*}(t)\right|_{t=0}$$

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

Getting rid of *t*

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

Getting rid of t

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

Getting rid of t

- 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

Getting rid of *t*

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

Getting rid of *t*

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

Getting rid of t

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} m_t^* \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t
- KL optimization: fixed point equation in the mean params

$$0 = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t = m_t^*}$$

Getting rid of *t*

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} m_t^* \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t
- KL optimization: fixed point equation in the mean params

$$m_t^* = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t = m_t^*} + m_t^*$$

Getting rid of t

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} m_t^* \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t
- KL optimization: fixed point equation in the mean params

$$m_t^* = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t = m_t^*} + m_t^*$$

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

Getting rid of t

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} m_t^* \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t
- KL optimization: fixed point equation in the mean params

$$m_t^* = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t = m_t^*} + m_t^*$$

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

- KL decomposition: $KL = \mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta|x) =: S - L$

Getting rid of t

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} m_t^* \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t
- KL optimization: fixed point equation in the mean params

$$m_t^* = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t = m_t^*} + m_t^*$$

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

- KL decomposition: $KL = \mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta|x) =: S - L$

$$\hat{\Sigma} = \left(V^{-1} - H \right)^{-1}$$

Getting rid of t

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} m_t^* \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t
- KL optimization: fixed point equation in the mean params

$$m_t^* = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t = m_t^*} + m_t^*$$

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

- KL decomposition: $KL = \mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta|x) =: S - L$

$$\hat{\Sigma} = \left(V^{-1} - H \right)^{-1} \quad \text{for} \quad H := \left. \frac{\partial^2 L}{\partial m \partial m^T} \right|_{m=m^*}$$

Getting rid of t

- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} m_t^* \right|_{t=0}$
- Suppose q_t exponential family with mean parametrization m_t
- KL optimization: fixed point equation in the mean params

$$m_t^* = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t = m_t^*} + m_t^*$$

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

- KL decomposition: $KL = \mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta|x) =: S - L$

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

LRVB estimator

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

$$m_t^* = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t = m_t^*} + m_t^*$$

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

- KL decomposition: $KL = \mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta|x) =: S - L$

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

LRVB estimator

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

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

- KL decomposition: $KL = \mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta|x) =: S - L$

$$\hat{\Sigma} = \left(V^{-1} - H\right)^{-1} = (I - VH)^{-1} V \quad \text{for} \quad H := \frac{\partial^2 L}{\partial m \partial m^T} \Big|_{m=m^*}$$

LRVB estimator

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

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

- KL decomposition: $KL = \mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta|x) =: S - L$

$\hat{\Sigma} = \left(V^{-1} - H\right)^{-1} = (I - VH)^{-1} V$ for $H := \frac{\partial^2 L}{\partial m \partial m^T} \Big|_{m=m^*}$

LRVB estimator

- LRVB covariance estimate

$$\hat{\Sigma} := \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \Big|_{t=0}$$

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

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

LRVB estimator

- LRVB covariance estimate

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

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

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

LRVB estimator

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

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

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

LRVB estimator

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

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

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

LRVB estimator

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

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

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

- Symmetric and positive definite at local min of KL

LRVB estimator

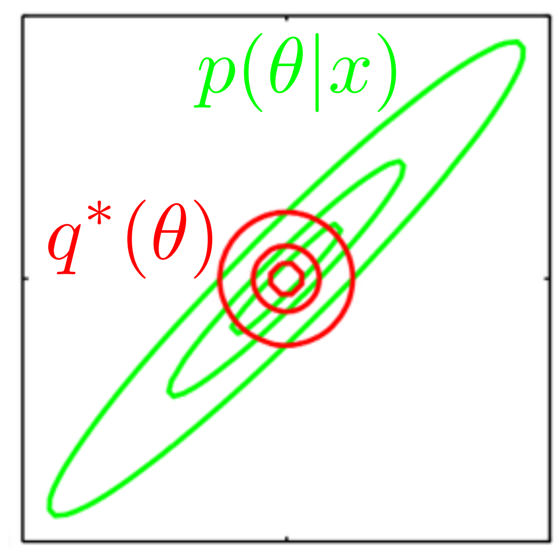
- LRVB covariance estimate $\hat{\Sigma} := \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0}$

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

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

- 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]

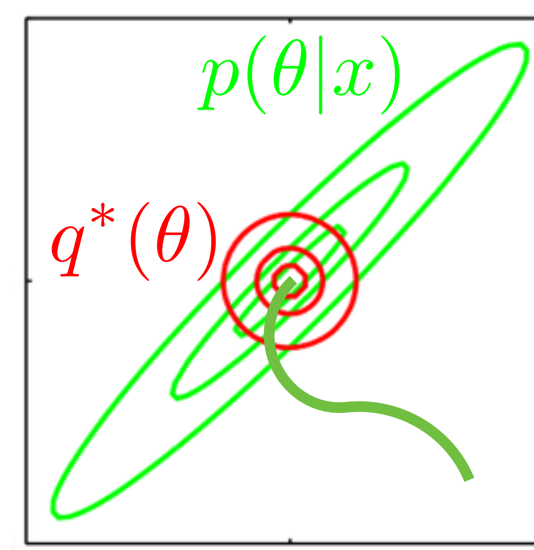
LRVB estimator

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

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

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

- 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]

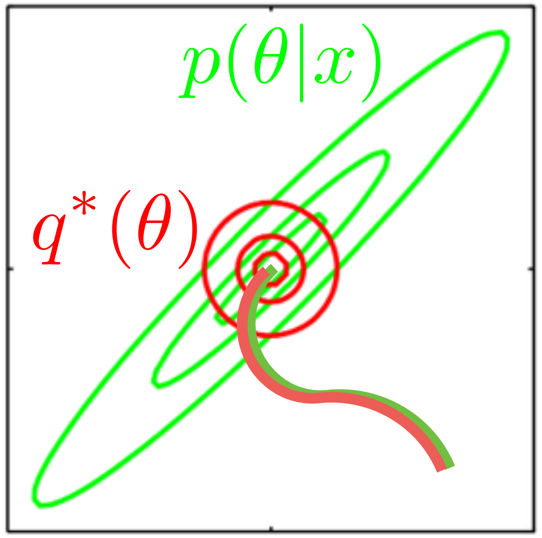
LRVB estimator

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

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

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

- 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]

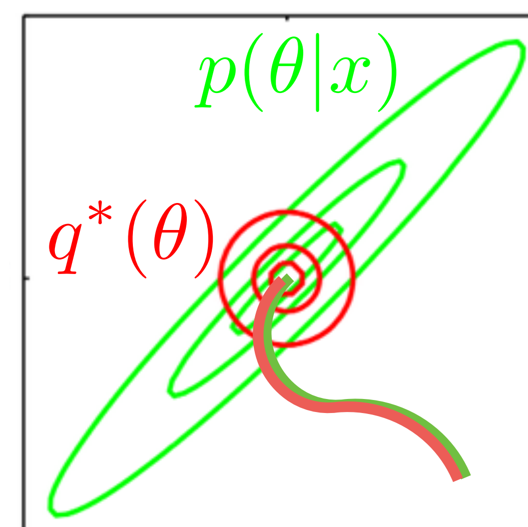
LRVB estimator

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

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

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

- 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 VB gives exact mean (e.g. multivariate normal)



[Bishop 2006]

1. Derive *Linear Response Variational Bayes* (LRVB) variance/covariance correction

2. Accuracy experiments

3. Scalability experiments

Experiments

Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

$$z_n|\beta,\tau \overset{indep}{\sim} \mathcal{N}\left(z_n|\beta x_n, \tau^{-1}\right), \quad y_n|z_n \overset{indep}{\sim} \mathrm{Poisson}\left(y_n|\exp(z_n)\right),$$

$$\beta \sim \mathcal{N}(\beta|0, \sigma_\beta^2), \quad \tau \sim \mathrm{Gamma}(\tau|\alpha_\tau, \beta_\tau)$$

Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

$$z_n|\beta, \tau \overset{indep}{\sim} \mathcal{N}\left(z_n|\beta x_n, \tau^{-1}\right), \quad y_n|z_n \overset{indep}{\sim} \mathrm{Poisson}\left(y_n|\exp(z_n)\right),$$

$$\beta \sim \mathcal{N}(\beta|0, \sigma_\beta^2), \quad \tau \sim \mathrm{Gamma}(\tau|\alpha_\tau, \beta_\tau)$$

- MFVB assumption:

$$q(\beta, \tau, z) = q(\beta)q(\tau)\prod_{n=1}^{N} q(z_n)$$

Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

$$z_n|\beta,\tau \overset{indep}{\sim} \mathcal{N}\left(z_n|\beta x_n, \tau^{-1}\right), \quad y_n|z_n \overset{indep}{\sim} \mathrm{Poisson}\left(y_n|\exp(z_n)\right),$$

$$\beta \sim \mathcal{N}(\beta|0,\sigma_\beta^2), \quad \tau \sim \mathrm{Gamma}(\tau|\alpha_\tau,\beta_\tau)$$

- MFVB assumption:

$$q(\beta,\tau,z) = q(\beta)q(\tau)\prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathcal{N}(z_n)$$

Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

$$z_n|\beta,\tau \stackrel{indep}{\sim} \mathcal{N}\left(z_n|\beta x_n, \tau^{-1}\right), \quad y_n|z_n \stackrel{indep}{\sim} \mathrm{Poisson}\left(y_n|\exp(z_n)\right),$$

$$\beta \sim \mathcal{N}(\beta|0, \sigma_\beta^2), \quad \tau \sim \mathrm{Gamma}(\tau|\alpha_\tau, \beta_\tau)$$

- MFVB assumption:

$$q(\beta, \tau, z) = q(\beta)q(\tau)\prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathcal{N}(z_n)$$

- 100 simulated data sets, 500 data points each, R `MCMCglmm` package (20,000 samples)

Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

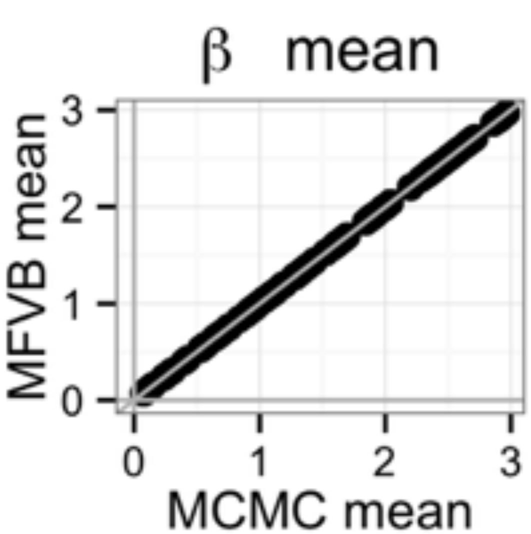
$$z_n|\beta,\tau \overset{indep}{\sim} \mathcal{N}\left(z_n|\beta x_n, \tau^{-1}\right), \quad y_n|z_n \overset{indep}{\sim} \mathrm{Poisson}\left(y_n|\exp(z_n)\right),$$

$$\beta \sim \mathcal{N}(\beta|0,\sigma_\beta^2), \quad \tau \sim \mathrm{Gamma}(\tau|\alpha_\tau,\beta_\tau)$$

- MFVB assumption:

$$q(\beta,\tau,z) = q(\beta)q(\tau)\prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathcal{N}(z_n)$$

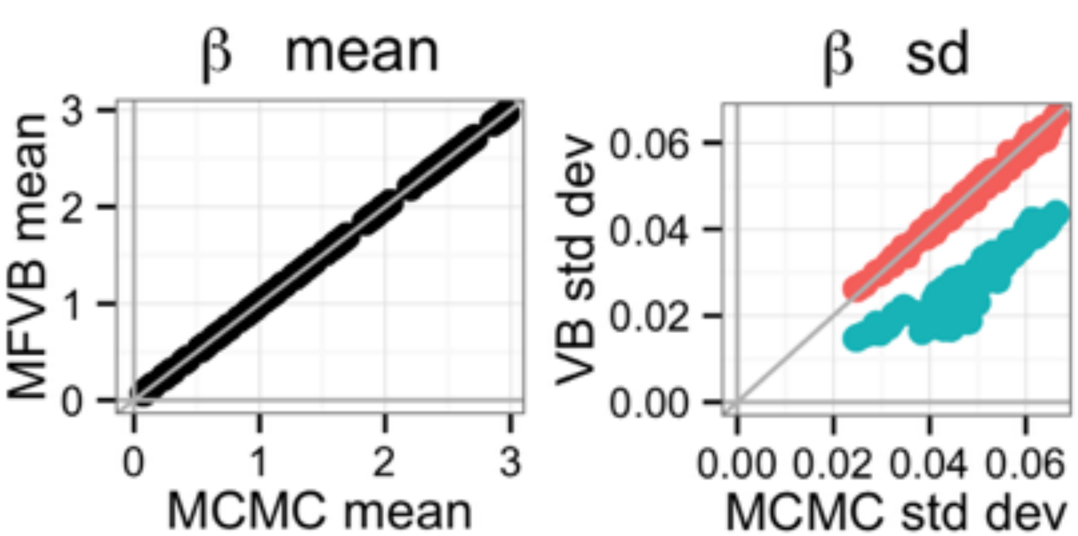
- 100 simulated data sets, 500 data points each, R `MCMCglmm` package (20,000 samples)



Experiments

- Non-conjugate normal-Poisson generalized linear mixed model
$$z_n|\beta,\tau \overset{indep}{\sim} \mathcal{N}\left(z_n|\beta x_n,\tau^{-1}\right), \quad y_n|z_n \overset{indep}{\sim} \text{Poisson}\left(y_n|\exp(z_n)\right),$$
$$\beta \sim \mathcal{N}(\beta|0,\sigma_\beta^2), \quad \tau \sim \text{Gamma}(\tau|\alpha_\tau,\beta_\tau)$$
- MFVB assumption:
$$q(\beta,\tau,z) = q(\beta)q(\tau)\prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathcal{N}(z_n)$$
- 100 simulated data sets, 500 data points each, R `MCMCglmm` package (20,000 samples)

LRVB, **MFVB**



Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

$$z_n|\beta,\tau \overset{indep}{\sim} \mathcal{N}\left(z_n|\beta x_n, \tau^{-1}\right), \quad y_n|z_n \overset{indep}{\sim} \text{Poisson}\left(y_n|\exp(z_n)\right),$$

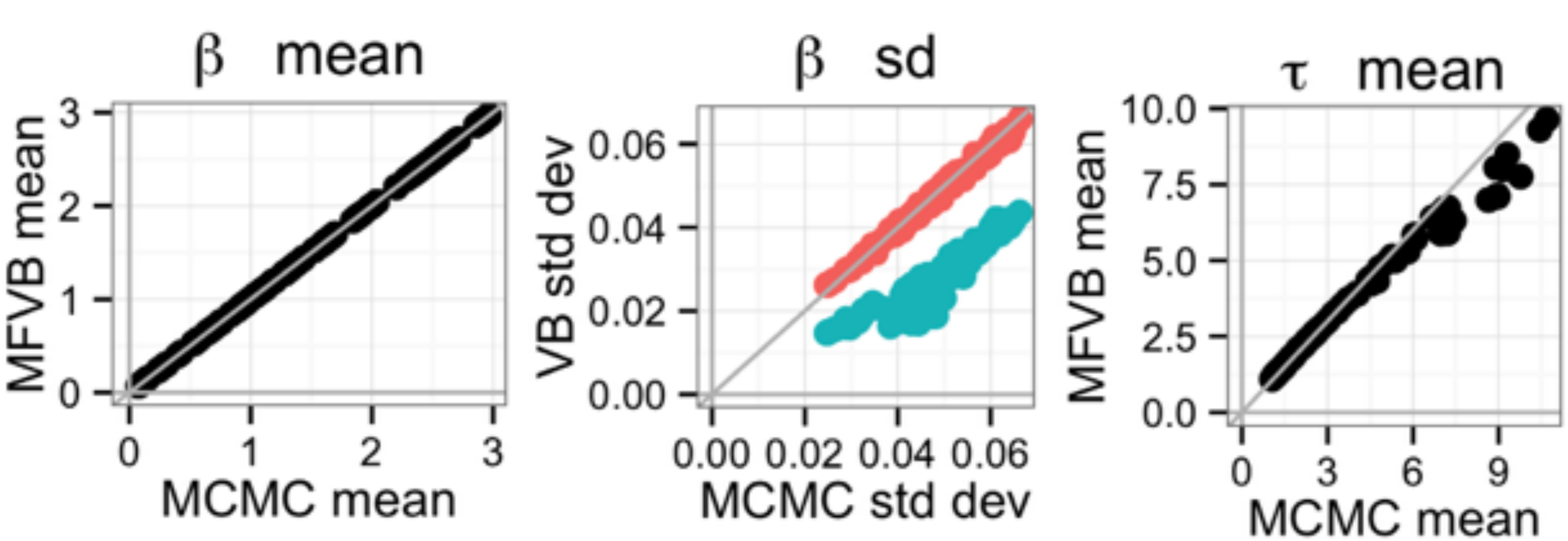
$$\beta \sim \mathcal{N}(\beta|0,\sigma_\beta^2), \quad \tau \sim \text{Gamma}(\tau|\alpha_\tau,\beta_\tau)$$

- MFVB assumption:

$$q(\beta,\tau,z) = q(\beta)q(\tau)\prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathcal{N}(z_n)$$

- 100 simulated data sets, 500 data points each, R `MCMCglmm` package (20,000 samples)

LRVB, **MFVB**



Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

$$z_n|\beta,\tau \overset{indep}{\sim} \mathcal{N}\left(z_n|\beta x_n, \tau^{-1}\right), \quad y_n|z_n \overset{indep}{\sim} \text{Poisson}\left(y_n|\exp(z_n)\right),$$

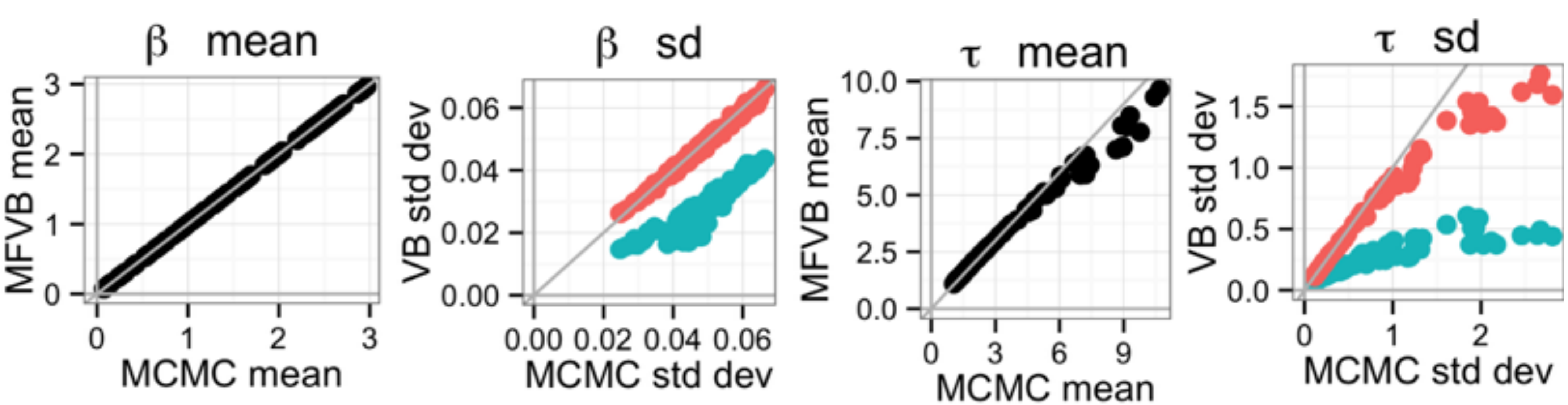
$$\beta \sim \mathcal{N}(\beta|0,\sigma_\beta^2), \quad \tau \sim \text{Gamma}(\tau|\alpha_\tau,\beta_\tau)$$

- MFVB assumption:

$$q(\beta,\tau,z) = q(\beta)q(\tau)\prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathcal{N}(z_n)$$

- 100 simulated data sets, 500 data points each, R `MCMCglmm` package (20,000 samples)

LRVB, MFVB



Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

$$z_n|\beta,\tau \overset{indep}{\sim} \mathcal{N}\left(z_n|\beta x_n, \tau^{-1}\right), \quad y_n|z_n \overset{indep}{\sim} \text{Poisson}\left(y_n|\exp(z_n)\right),$$

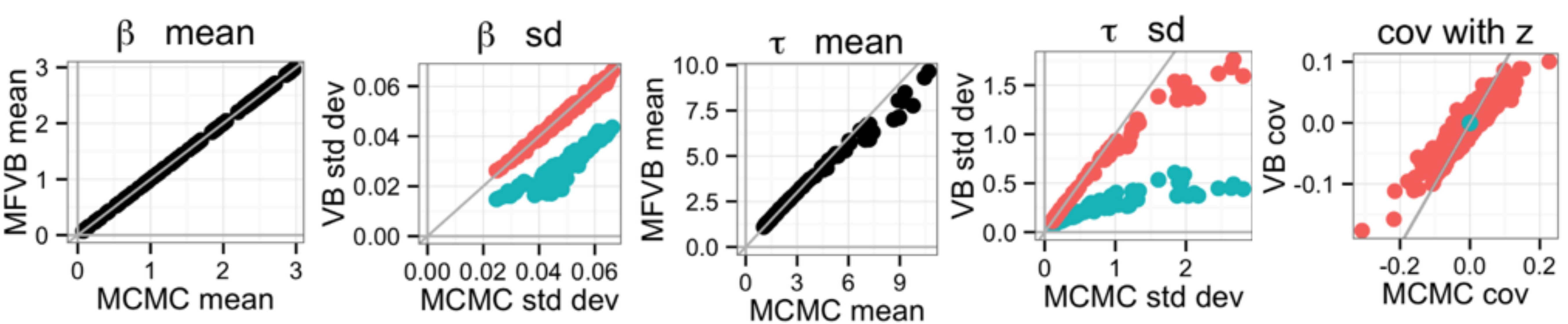
$$\beta \sim \mathcal{N}(\beta|0, \sigma_\beta^2), \quad \tau \sim \text{Gamma}(\tau|\alpha_\tau, \beta_\tau)$$

- MFVB assumption:

$$q(\beta, \tau, z) = q(\beta)q(\tau)\prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathcal{N}(z_n)$$

- 100 simulated data sets, 500 data points each, R `MCMCglmm` package (20,000 samples)

LRVB, **MFVB**



Experiments

Experiments

- Linear model with random effects

Experiments

- Linear model with random effects

$$y_n|\beta, z, \tau \overset{indep}{\sim} \mathcal{N}\left(y_n|\beta^T x_n + r_n z_{k(n)}, \tau^{-1}\right), \quad z_k|\nu \overset{iid}{\sim} \mathcal{N}\left(z_k|0, \nu^{-1}\right)$$

$$\beta \sim \mathcal{N}(\beta|0, \Sigma_\beta), \quad \nu \sim \Gamma(\nu|\alpha_\nu, \beta_\nu), \quad \tau \sim \Gamma(\tau|\alpha_\tau, \beta_\tau)$$

Experiments

- Linear model with random effects

$$y_n|\beta, z, \tau \overset{indep}{\sim} \mathcal{N}\left(y_n|\beta^T x_n + r_n z_{k(n)}, \tau^{-1}\right), \quad z_k|\nu \overset{iid}{\sim} \mathcal{N}\left(z_k|0, \nu^{-1}\right)$$

$$\beta \sim \mathcal{N}(\beta|0, \Sigma_\beta), \quad \nu \sim \Gamma(\nu|\alpha_\nu, \beta_\nu), \quad \tau \sim \Gamma(\tau|\alpha_\tau, \beta_\tau)$$

- MFVB assumption: $q(\beta, \nu, \tau, z) = q(\beta)q(\tau)q(\nu) \prod_{k=1}^{K} q(z_n)$

Experiments

- Linear model with random effects

$$y_n|\beta, z, \tau \overset{indep}{\sim} \mathcal{N}\left(y_n|\beta^T x_n + r_n z_{k(n)}, \tau^{-1}\right), \quad z_k|\nu \overset{iid}{\sim} \mathcal{N}\left(z_k|0, \nu^{-1}\right)$$

$$\beta \sim \mathcal{N}(\beta|0, \Sigma_\beta), \quad \nu \sim \Gamma(\nu|\alpha_\nu, \beta_\nu), \quad \tau \sim \Gamma(\tau|\alpha_\tau, \beta_\tau)$$

- MFVB assumption: $q(\beta, \nu, \tau, z) = q(\beta)q(\tau)q(\nu)\prod_{k=1}^{K} q(z_n)$

- 100 simulated data sets, 300 data points each, R `MCMCglmm` package (20,000 samples)

Experiments

- Linear model with random effects

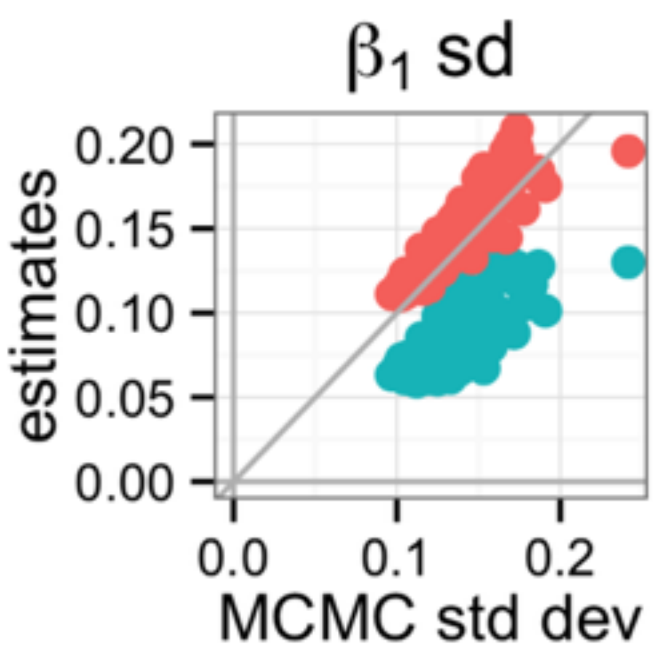
$$y_n | \beta, z, \tau \overset{indep}{\sim} \mathcal{N}\left(y_n | \beta^T x_n + r_n z_{k(n)}, \tau^{-1}\right), \quad z_k | \nu \overset{iid}{\sim} \mathcal{N}\left(z_k | 0, \nu^{-1}\right)$$

$$\beta \sim \mathcal{N}(\beta | 0, \Sigma_\beta), \quad \nu \sim \Gamma(\nu | \alpha_\nu, \beta_\nu), \quad \tau \sim \Gamma(\tau | \alpha_\tau, \beta_\tau)$$

- MFVB assumption: $q(\beta, \nu, \tau, z) = q(\beta) q(\tau) q(\nu) \prod_{k=1}^{K} q(z_n)$

- 100 simulated data sets, 300 data points each, R `MCMCglmm` package (20,000 samples)

LRVB, **MFVB**



Experiments

- Linear model with random effects

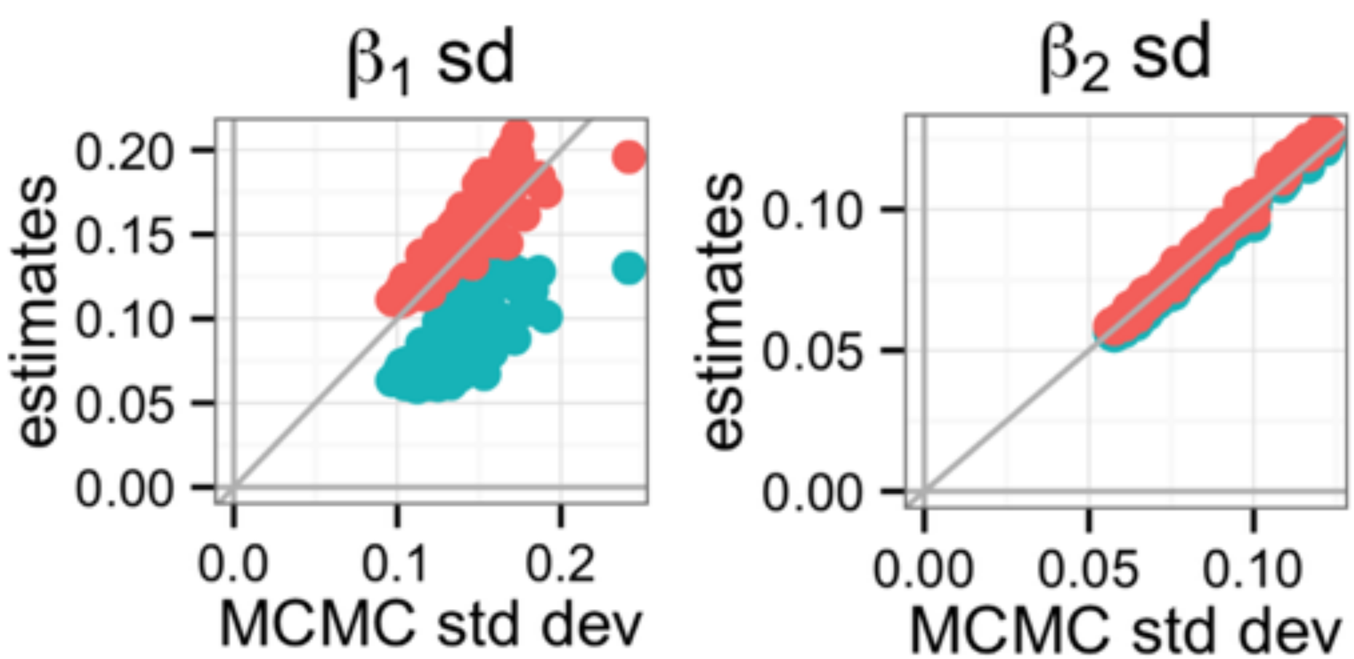
$$y_n|\beta, z, \tau \overset{indep}{\sim} \mathcal{N}\left(y_n|\beta^T x_n + r_n z_{k(n)}, \tau^{-1}\right), \quad z_k|\nu \overset{iid}{\sim} \mathcal{N}\left(z_k|0, \nu^{-1}\right)$$

$$\beta \sim \mathcal{N}(\beta|0, \Sigma_\beta), \quad \nu \sim \Gamma(\nu|\alpha_\nu, \beta_\nu), \quad \tau \sim \Gamma(\tau|\alpha_\tau, \beta_\tau)$$

- MFVB assumption: $$q(\beta, \nu, \tau, z) = q(\beta)q(\tau)q(\nu)\prod_{k=1}^{K} q(z_n)$$

- 100 simulated data sets, 300 data points each, R `MCMCglmm` package (20,000 samples)

LRVB, **MFVB**



Experiments

Experiments

- Gaussian mixture model

Experiments

- Gaussian mixture model

$$P(z_{nk} = 1) = \pi_k, \quad p(x|\pi, \mu, \Lambda, z) = \prod_{n=1:N} \prod_{k=1:K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{nk}}$$

with conjugate priors on π, μ, Λ

Experiments

- Gaussian mixture model
$$P(z_{nk} = 1) = \pi_k, \quad p(x|\pi, \mu, \Lambda, z) = \prod_{n=1:N} \prod_{k=1:K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{nk}}$$
with conjugate priors on π, μ, Λ
- MFVB assumption: $\left[\prod_{k=1}^{K} q(\mu_k)q(\Lambda_k)q(\pi_k)\right] \prod_{n=1}^{N} q(z_n)$

Experiments

- Gaussian mixture model

$$P(z_{nk} = 1) = \pi_k, \quad p(x|\pi, \mu, \Lambda, z) = \prod_{n=1:N} \prod_{k=1:K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{nk}}$$

with conjugate priors on π, μ, Λ

- MFVB assumption: $\left[\prod_{k=1}^{K} q(\mu_k) q(\Lambda_k) q(\pi_k)\right] \prod_{n=1}^{N} q(z_n)$
- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R `bayesm` package (function `rnmixGibbs`; at least 500 effective samples)

Experiments

- Gaussian mixture model

$$P(z_{nk} = 1) = \pi_k, \quad p(x|\pi, \mu, \Lambda, z) = \prod_{n=1:N} \prod_{k=1:K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{nk}}$$

with conjugate priors on π, μ, Λ

- MFVB assumption: $\left[\prod_{k=1}^{K} q(\mu_k) q(\Lambda_k) q(\pi_k)\right] \prod_{n=1}^{N} q(z_n)$
- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R `bayesm` package (function `rnmixGibbs`; at least 500 effective samples)
- MNIST digits: 12,665 0s and 1s; PCA for 25 dimensions

Experiments

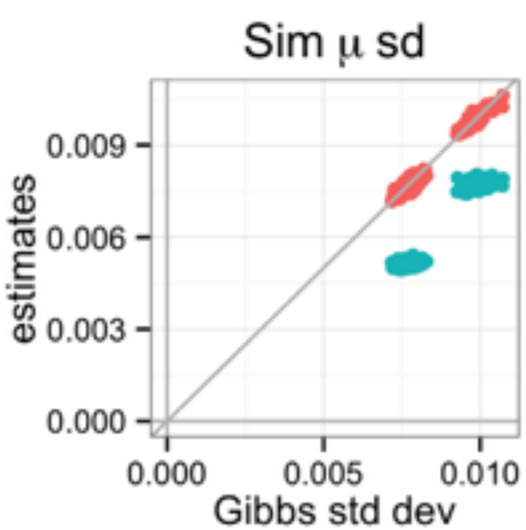
- Gaussian mixture model

$$P(z_{nk}=1)=\pi_k, \quad p(x|\pi,\mu,\Lambda,z)=\prod_{n=1:N}\prod_{k=1:K}\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1})^{z_{nk}}$$

with conjugate priors on π, μ, Λ

- MFVB assumption: $\left[\prod_{k=1}^{K} q(\mu_k)q(\Lambda_k)q(\pi_k)\right]\prod_{n=1}^{N} q(z_n)$
- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R `bayesm` package (function `rnmixGibbs`; at least 500 effective samples)

- MNIST digits: 12,665 0s and 1s; PCA for 25 dimensions



LRVB, **MFVB**

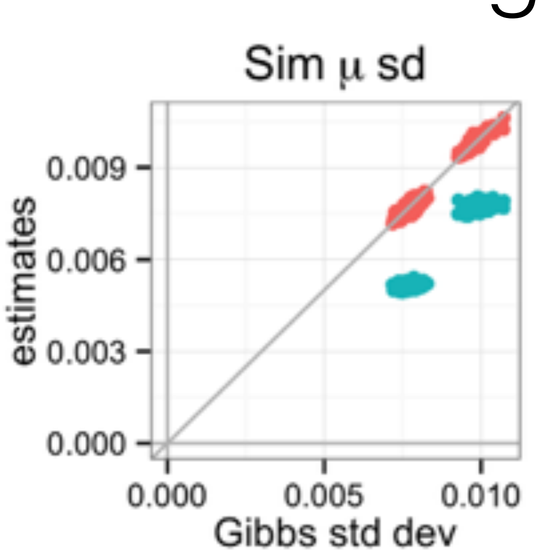
Experiments

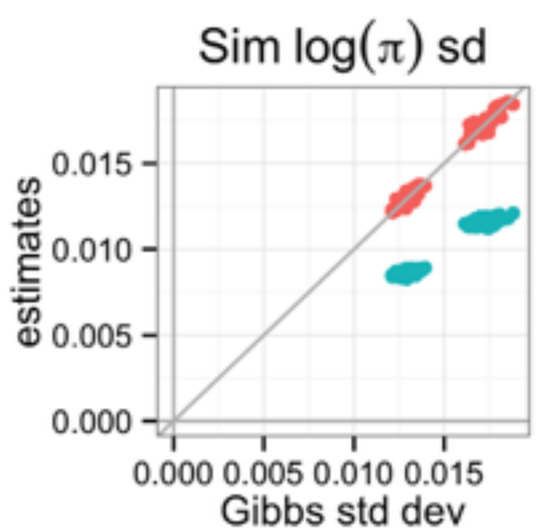
- Gaussian mixture model

$$P(z_{nk}=1)=\pi_k, \quad p(x|\pi,\mu,\Lambda,z)=\prod_{n=1:N}\prod_{k=1:K}\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1})^{z_{nk}}$$

with conjugate priors on π, μ, Λ

- MFVB assumption: $\left[\prod_{k=1}^{K} q(\mu_k)q(\Lambda_k)q(\pi_k)\right]\prod_{n=1}^{N} q(z_n)$
- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R `bayesm` package (function `rnmixGibbs`; at least 500 effective samples)
- MNIST digits: 12,665 0s and 1s; PCA for 25 dimensions





LRVB, **MFVB**

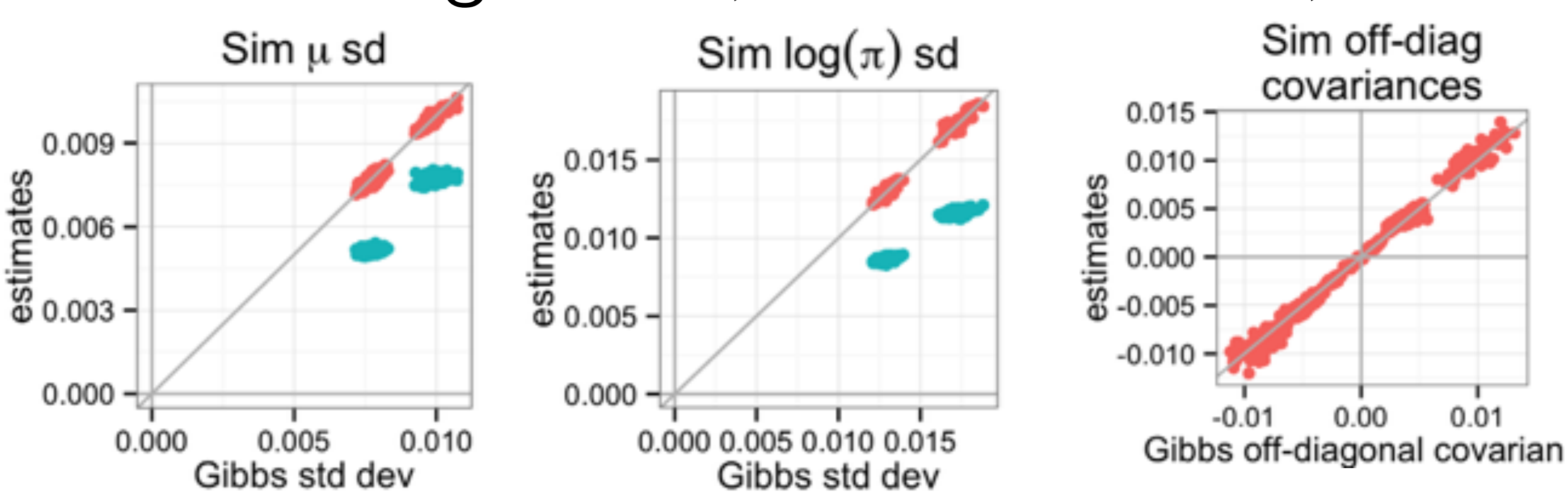
Experiments

- Gaussian mixture model

$$P(z_{nk}=1)=\pi_k, \quad p(x|\pi,\mu,\Lambda,z)=\prod_{n=1:N}\prod_{k=1:K}\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1})^{z_{nk}}$$

with conjugate priors on π, μ, Λ

- MFVB assumption: $\left[\prod_{k=1}^{K} q(\mu_k)q(\Lambda_k)q(\pi_k)\right]\prod_{n=1}^{N} q(z_n)$
- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R `bayesm` package (function `rnmixGibbs`; at least 500 effective samples)
- MNIST digits: 12,665 0s and 1s; PCA for 25 dimensions



LRVB, **MFVB**

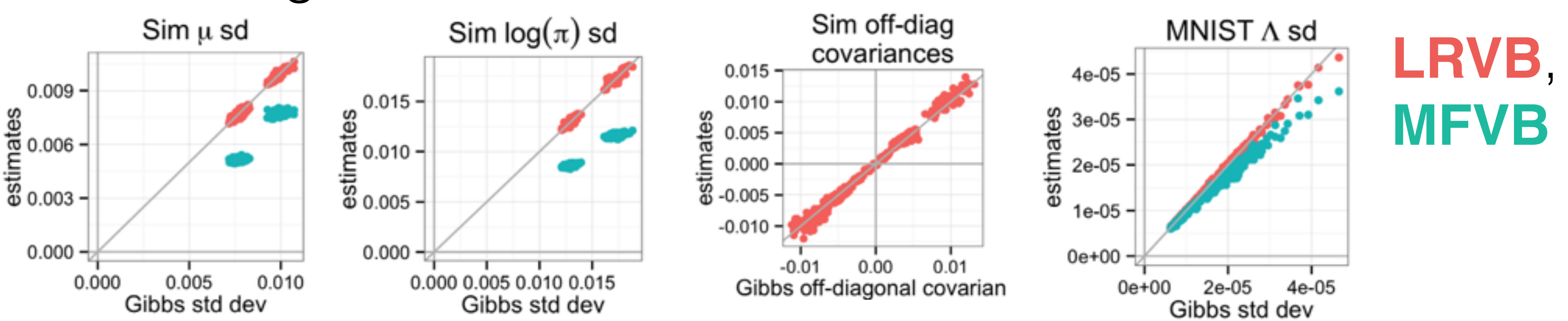
Experiments

- Gaussian mixture model

$$P(z_{nk} = 1) = \pi_k, \quad p(x|\pi, \mu, \Lambda, z) = \prod_{n=1:N} \prod_{k=1:K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{nk}}$$

with conjugate priors on π, μ, Λ

- MFVB assumption: $\left[\prod_{k=1}^{K} q(\mu_k)q(\Lambda_k)q(\pi_k)\right] \prod_{n=1}^{N} q(z_n)$
- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R `bayesm` package (function `rnmixGibbs`; at least 500 effective samples)
- MNIST digits: 12,665 0s and 1s; PCA for 25 dimensions



1. Derive *Linear Response Variational Bayes* (LRVB) variance/covariance correction

2. Accuracy experiments

3. Scalability experiments

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

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

Scaling the matrix inverse

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

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

- Schur complement

Scaling the matrix inverse

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

- Decomposition of parameter vector

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

- 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

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

$$H = \begin{bmatrix} H_\alpha & H_{\alpha z} \\ H_{z\alpha} & H_z \end{bmatrix}$$

- 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

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

- 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

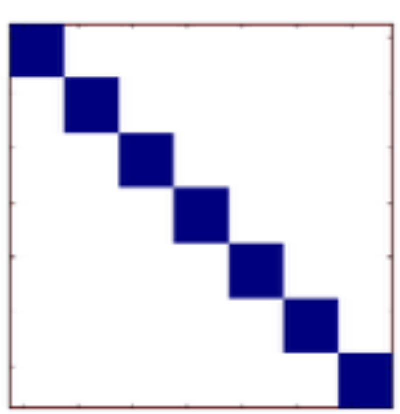
$$\theta = (\alpha^T, z^T)^T \qquad\qquad H = \begin{bmatrix} H_\alpha & H_{\alpha z} \\ H_{z\alpha} & H_z \end{bmatrix}$$

- 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

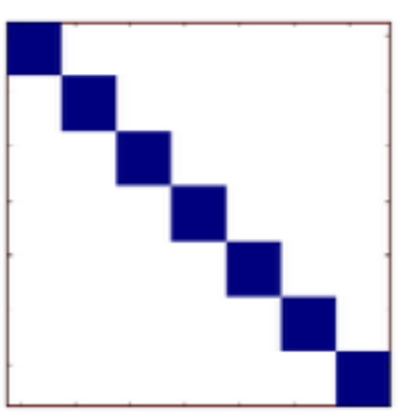
- LRVB estimate $\hat{\Sigma} = (I - VH)^{-1} V$
- Decomposition of parameter vector

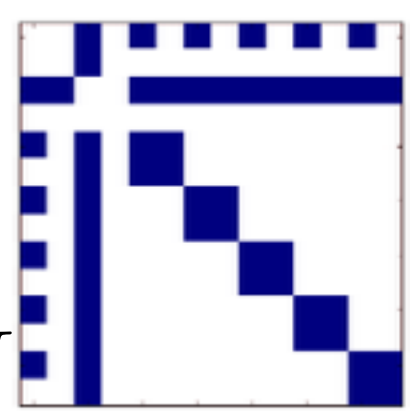
$$\theta = (\alpha^T, z^T)^T \qquad\qquad H = \begin{bmatrix} H_\alpha & H_{\alpha z} \\ H_{z\alpha} & H_z \end{bmatrix}$$

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

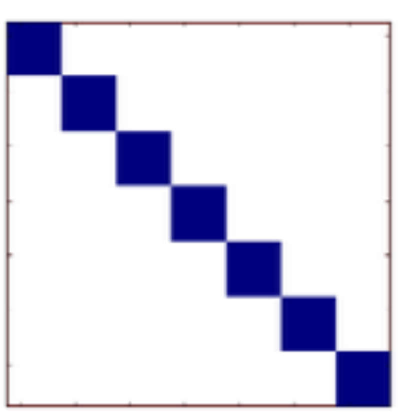
- Decomposition of parameter vector

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

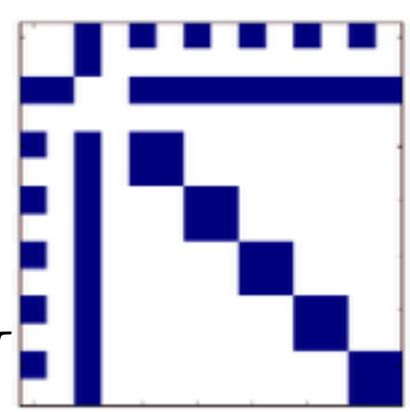
- Schur complement

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

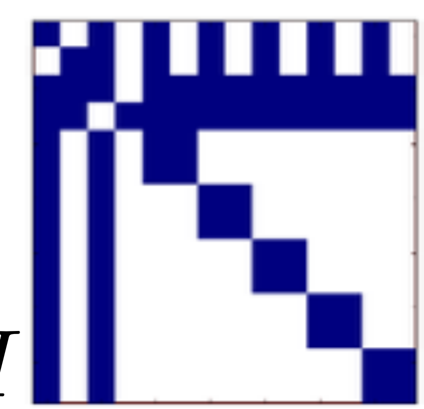
- Sparsity patterns



V



H



$I - VH$

Experiments

Experiments

- Scaling: Gaussian mixture model (*K* components, *P* dimensions, *N* data points)

Experiments

- Scaling: Gaussian mixture model (*K* components, *P* dimensions, *N* data points)
- The number of parameters in μ, π, Λ grows as $O(KP^2)$

Experiments

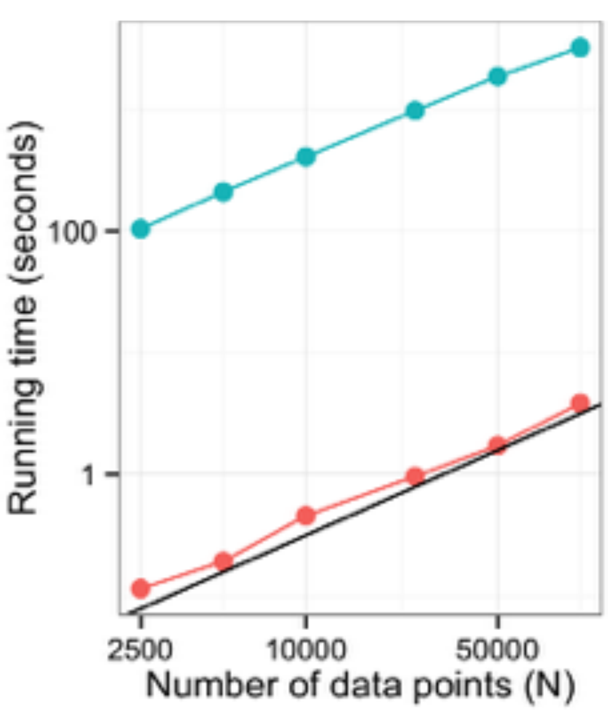
- Scaling: Gaussian mixture model (*K* components, *P* dimensions, *N* data points)
- The number of parameters in μ, π, Λ grows as $O(KP^2)$
- The number of parameters in z grows as $O(KN)$

Experiments

- Scaling: Gaussian mixture model (*K* components, *P* dimensions, *N* data points)
- The number of parameters in μ, π, Λ grows as $O(KP^2)$
- The number of parameters in z grows as $O(KN)$
- Worst case scaling: $O(K^3), O(P^6), O(N)$

Experiments

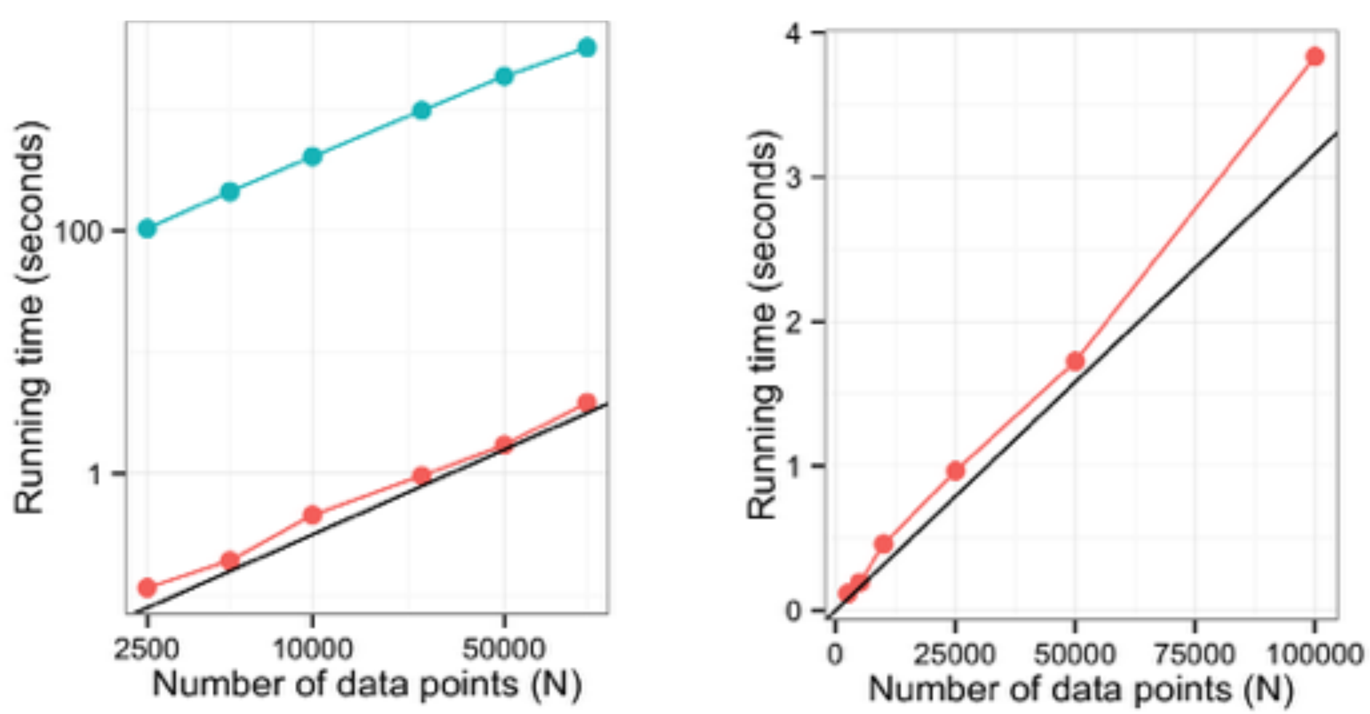
- Scaling: Gaussian mixture model (*K* components, *P* dimensions, *N* data points)
- The number of parameters in μ, π, Λ grows as $O(KP^2)$
- The number of parameters in z grows as $O(KN)$
- Worst case scaling: $O(K^3), O(P^6), O(N)$



LRVB,
Gibbs

Experiments

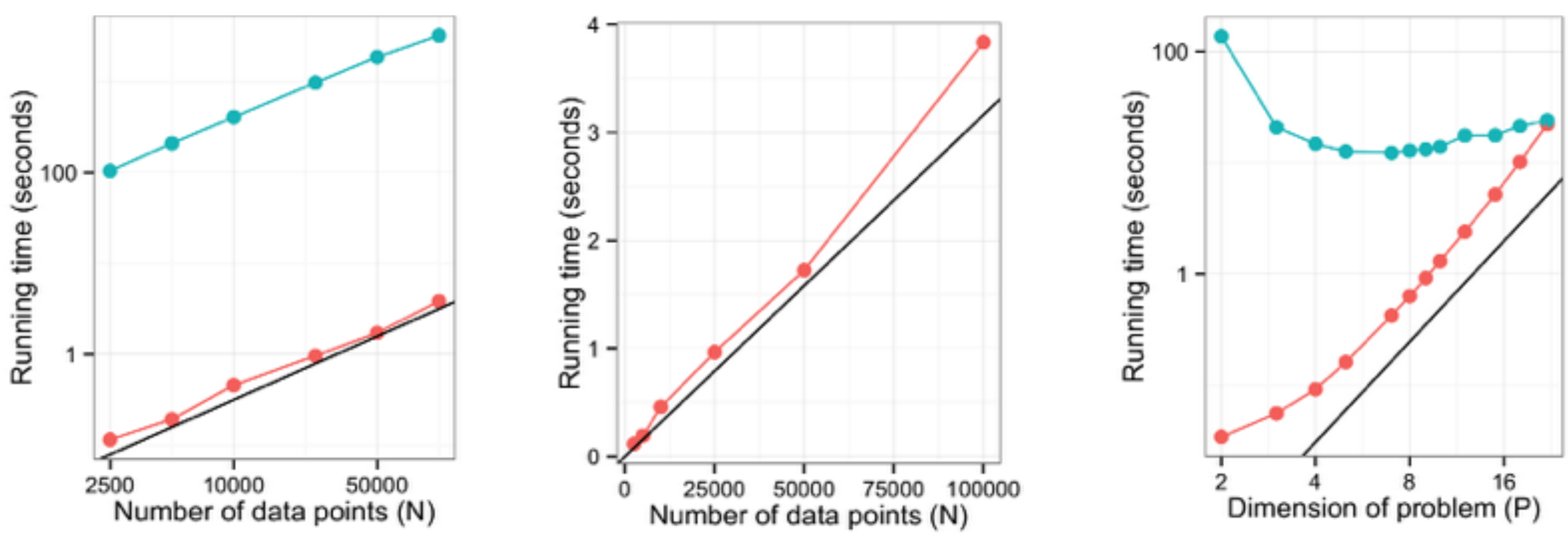
- Scaling: Gaussian mixture model (*K* components, *P* dimensions, *N* data points)
- The number of parameters in μ, π, Λ grows as $O(KP^2)$
- The number of parameters in z grows as $O(KN)$
- Worst case scaling: $O(K^3), O(P^6), O(N)$



LRVB,
Gibbs

Experiments

- Scaling: Gaussian mixture model (*K* components, *P* dimensions, *N* data points)
- The number of parameters in μ, π, Λ grows as $O(KP^2)$
- The number of parameters in z grows as $O(KN)$
- Worst case scaling: $O(K^3), O(P^6), O(N)$



LRVB, **Gibbs**

Conclusions, etc

Conclusions, etc

- MAD-Bayes: fast point estimates

Conclusions, etc

- MAD-Bayes: fast point estimates
- LRVB covariance correction: in many cases, accurate covariance estimates for VB

Conclusions, etc

- MAD-Bayes: fast point estimates
- LRVB covariance correction: in many cases, accurate covariance estimates for VB
- Open questions:

Conclusions, etc

- MAD-Bayes: fast point estimates
- LRVB covariance correction: in many cases, accurate covariance estimates for VB
- Open questions:
    - Mean correction

Conclusions, etc

- MAD-Bayes: fast point estimates
- LRVB covariance correction: in many cases, accurate covariance estimates for VB
- Open questions:
  - Mean correction
  - Global parameter scaling

Conclusions, etc

- MAD-Bayes: fast point estimates
- LRVB covariance correction: in many cases, accurate covariance estimates for VB
- Open questions:
  - Mean correction
  - Global parameter scaling
  - Targeting other posterior statistics besides point estimates and covariance

Conclusions, etc

- MAD-Bayes: fast point estimates
- LRVB covariance correction: in many cases, accurate covariance estimates for VB
- Open questions:
  - Mean correction
  - Global parameter scaling
  - Targeting other posterior statistics besides point estimates and covariance
- More LRVB: Bayesian robustness (work in progress)

References

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*.

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.

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

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, and MI Jordan. Robust inference with variational Bayes. In *NIPS 2015 Workshop in Advances in Approximate Bayesian Inference*.

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

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

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.