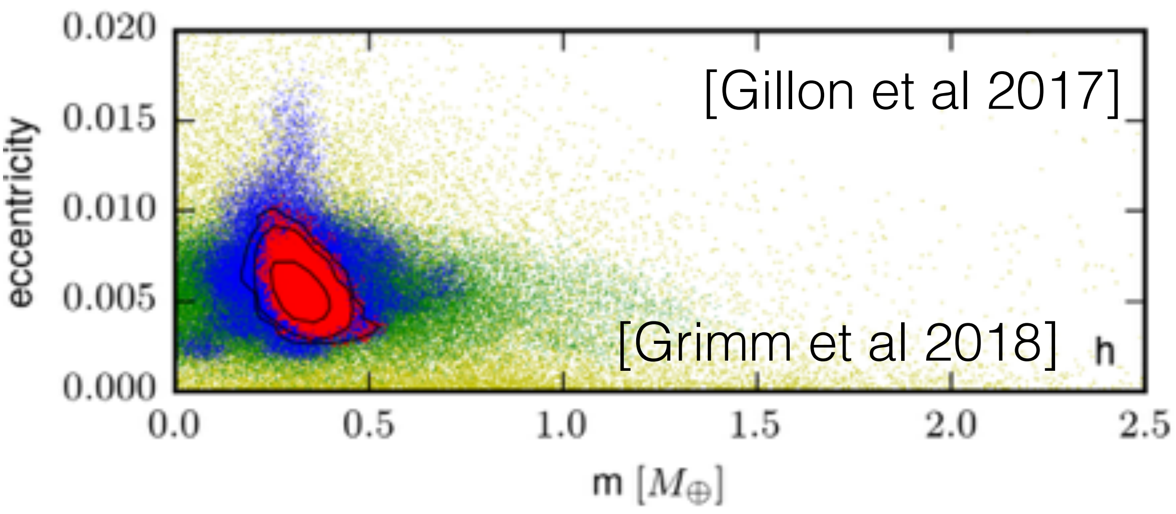
Variational Bayes and beyond: Bayesian inference for big data

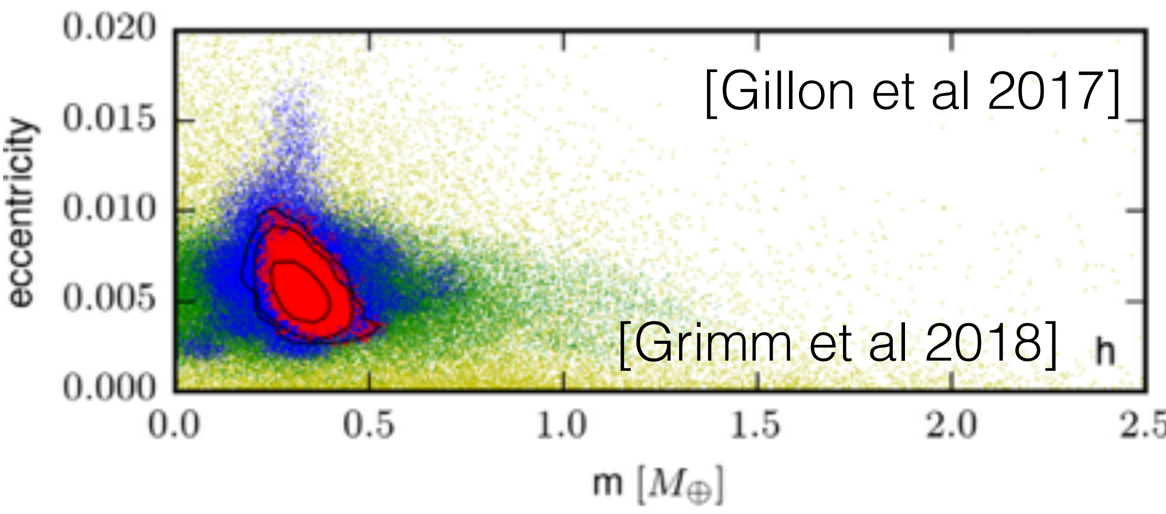
Tamara Broderick
ITT Career Development
Assistant Professor,
MIT

Bayesian inference

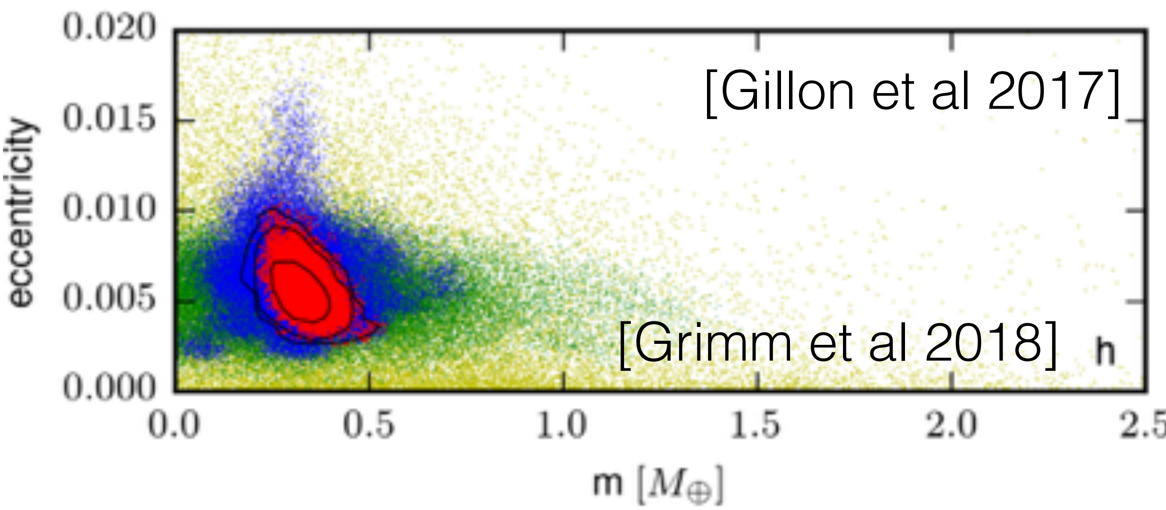
Bayesian inference

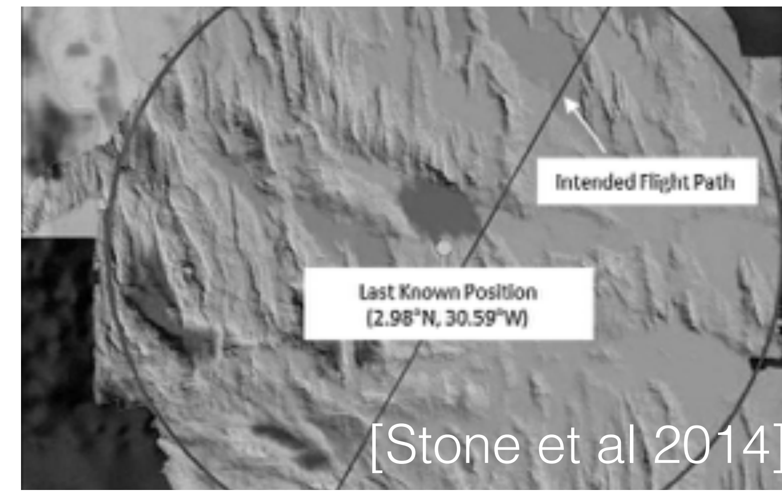


Bayesian inference

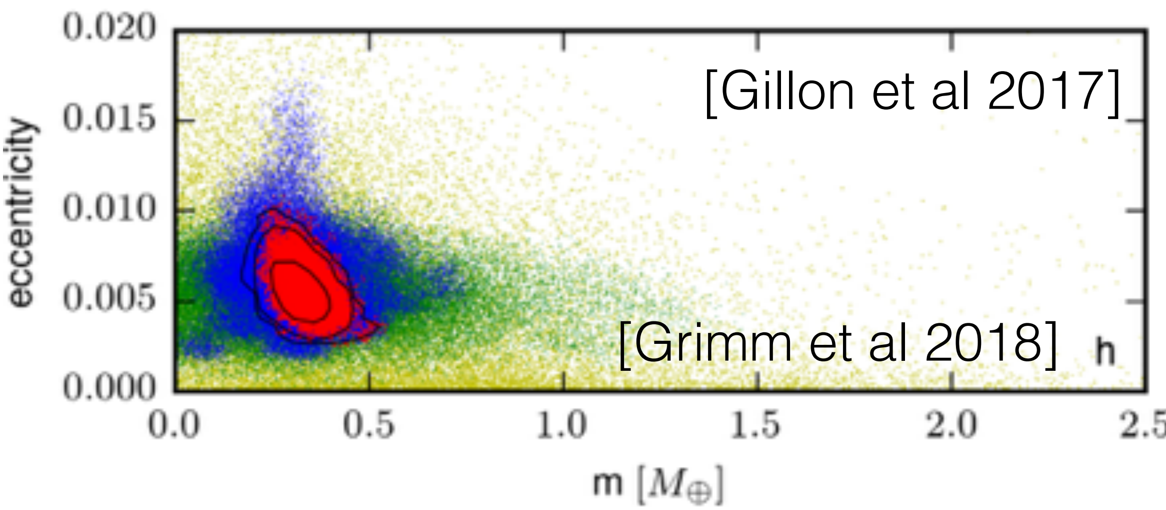


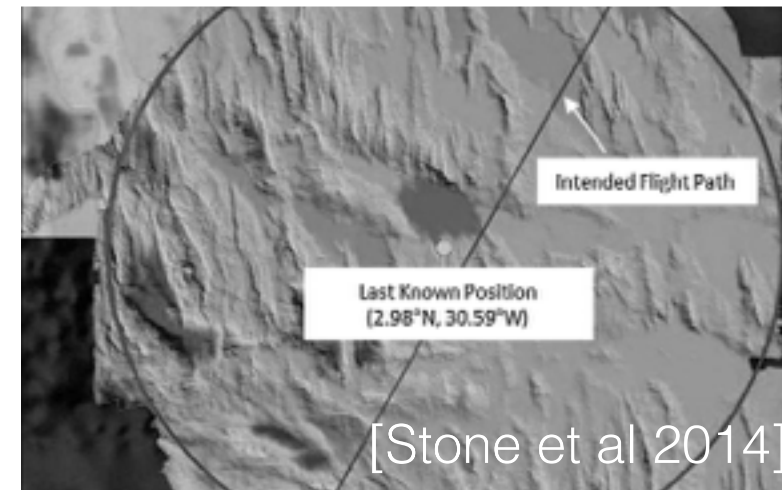
Bayesian inference

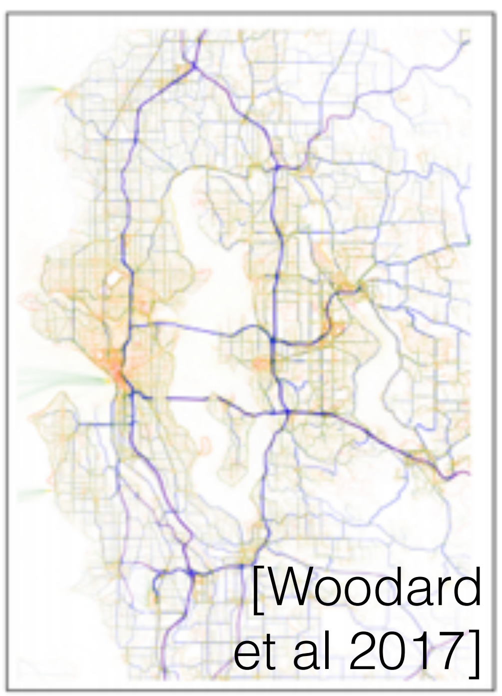




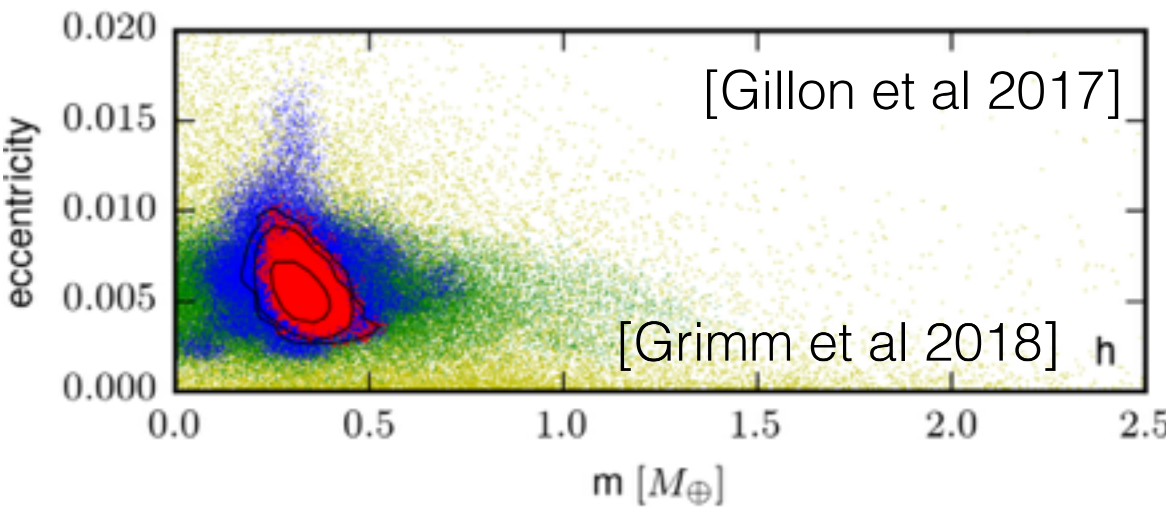
Bayesian inference

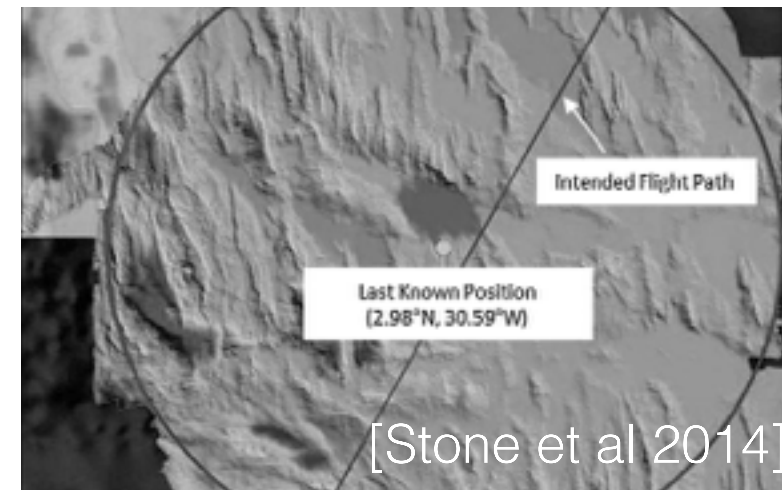


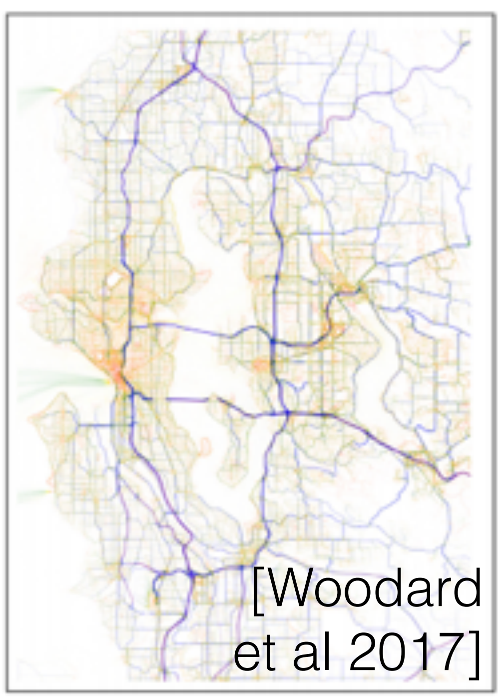


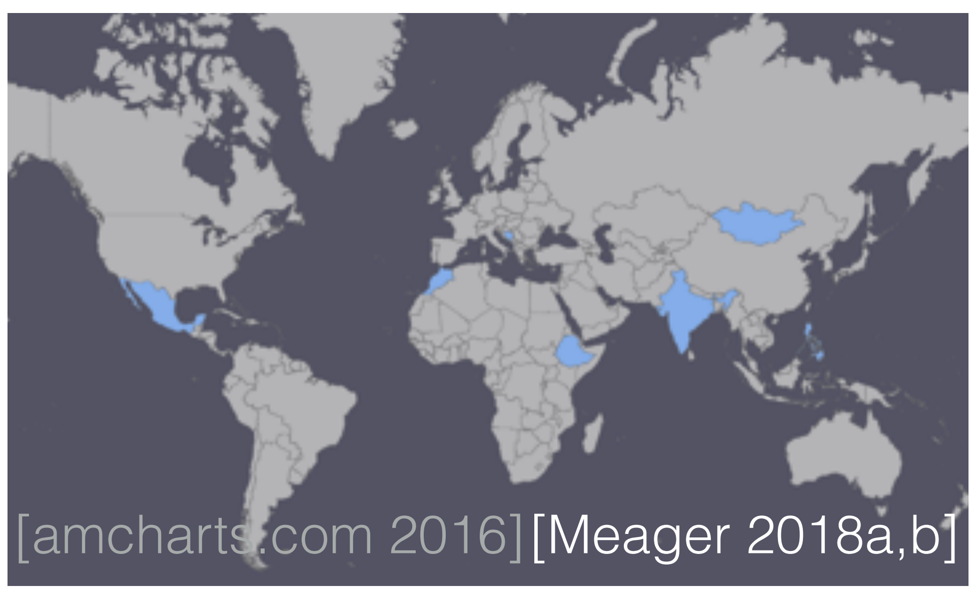


Bayesian inference

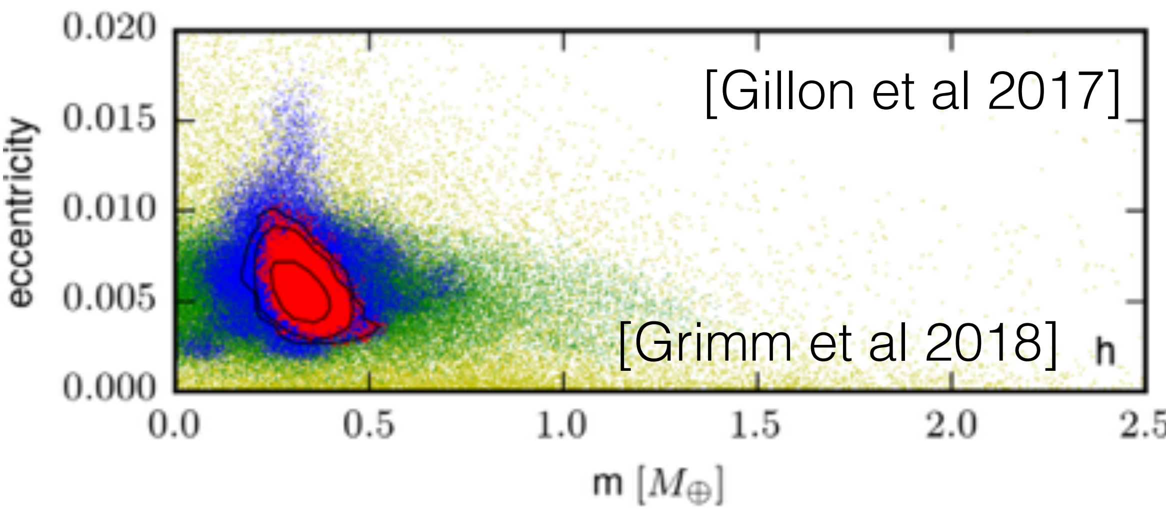


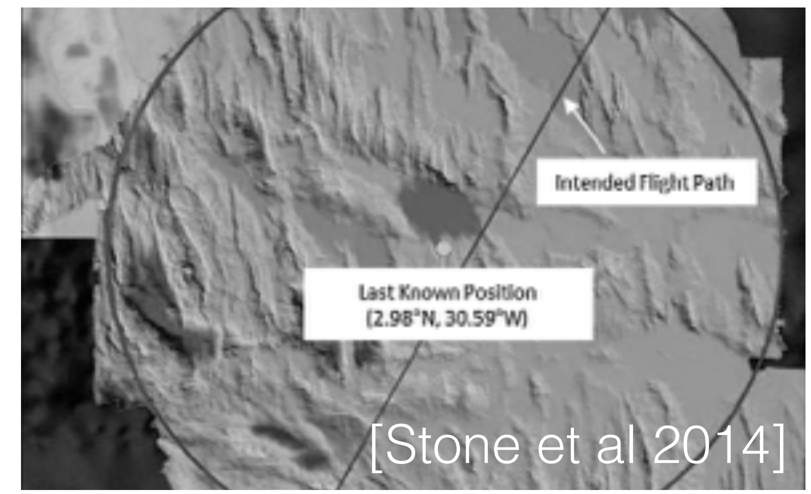


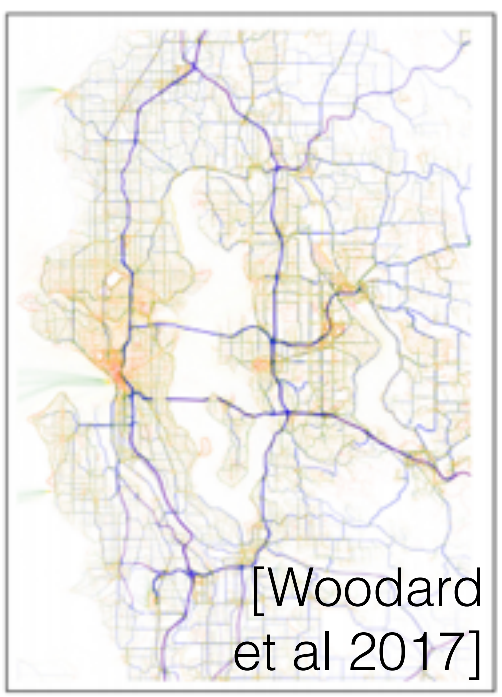


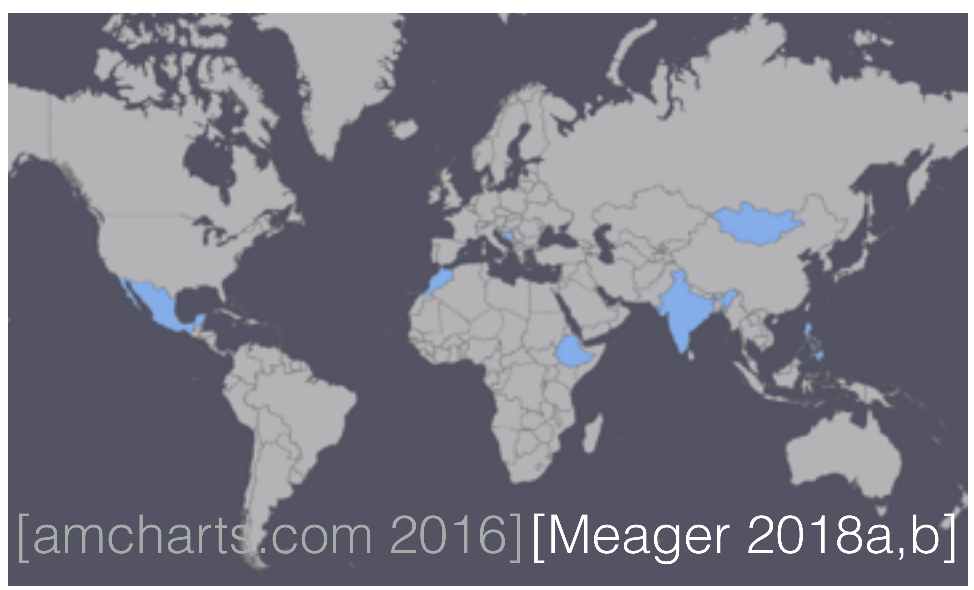


Bayesian inference

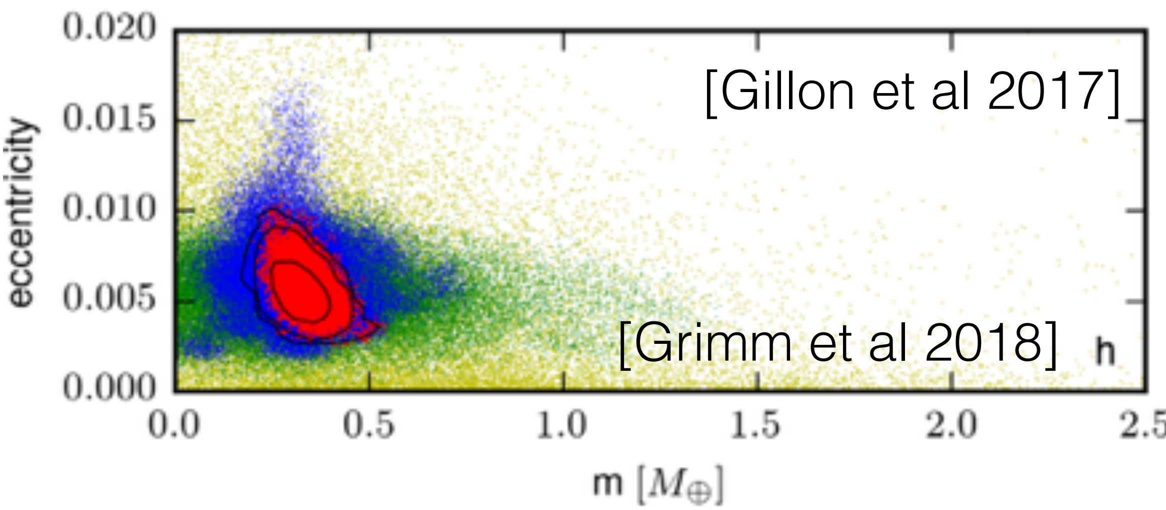


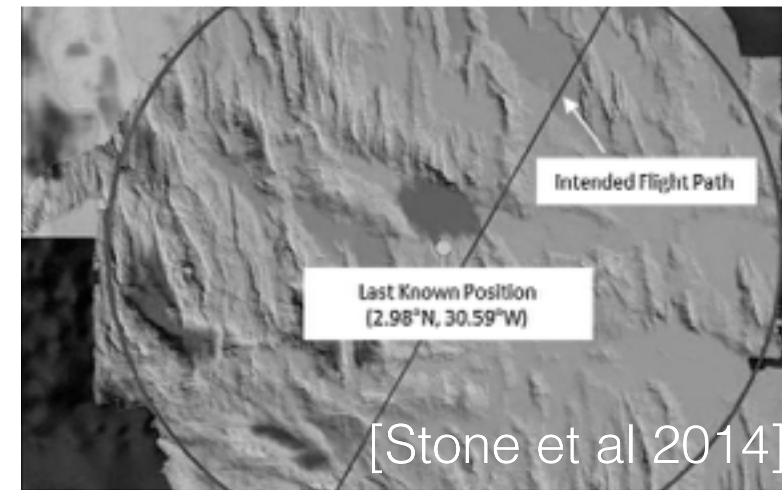


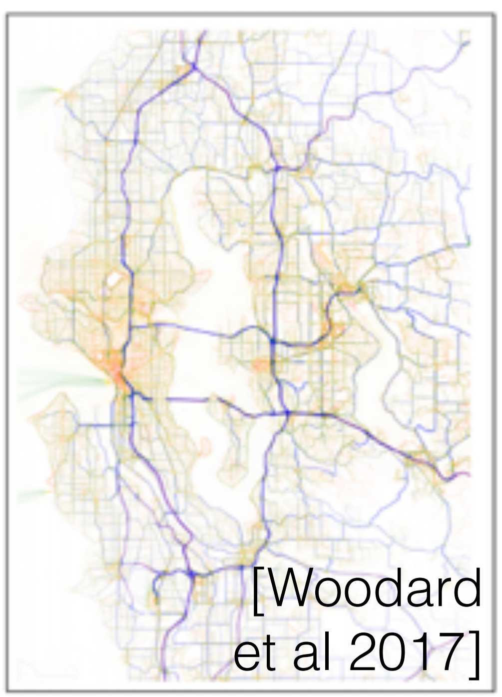


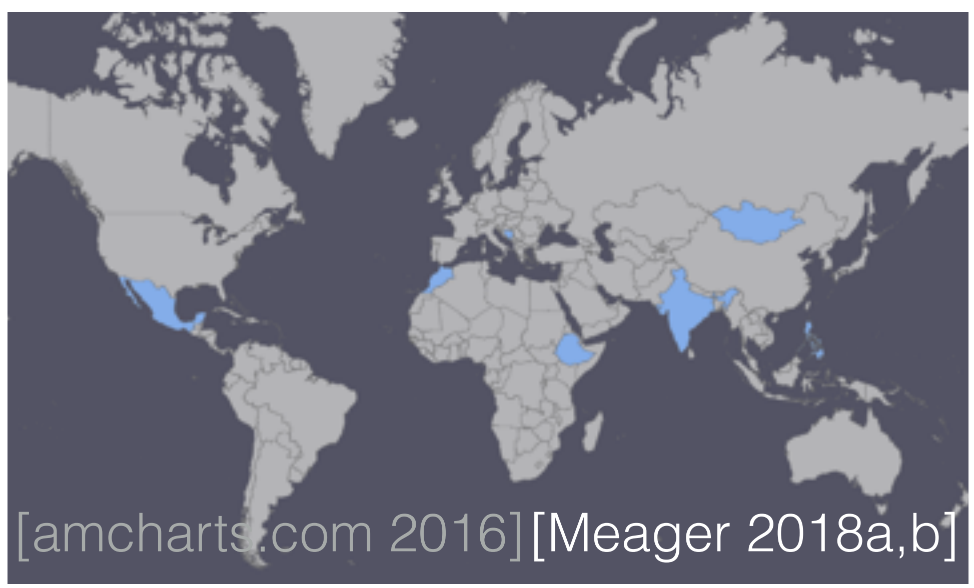


Bayesian inference



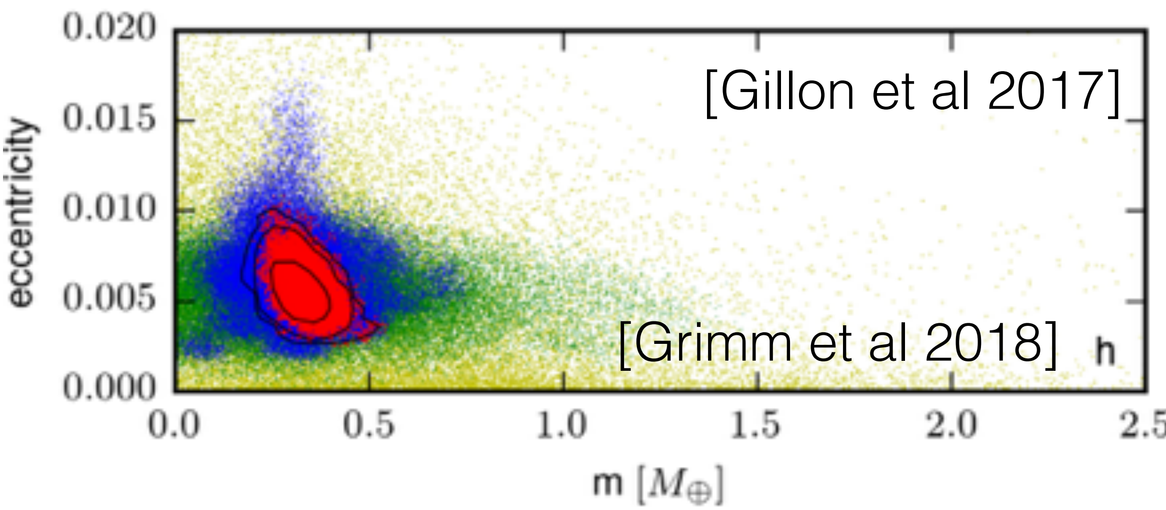


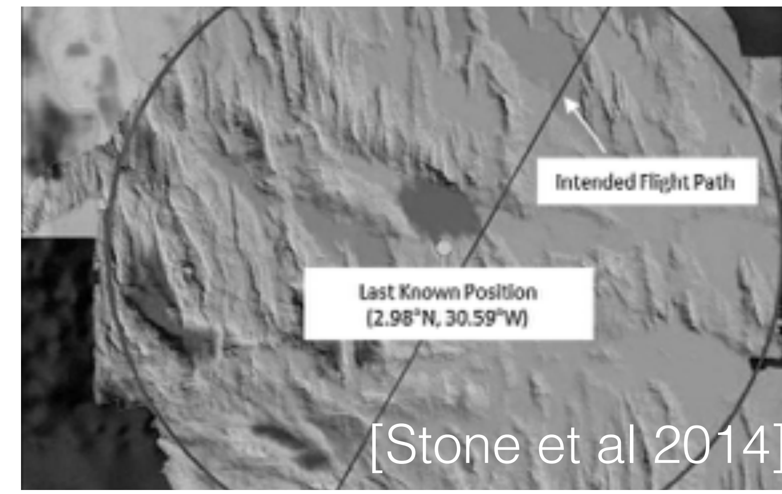


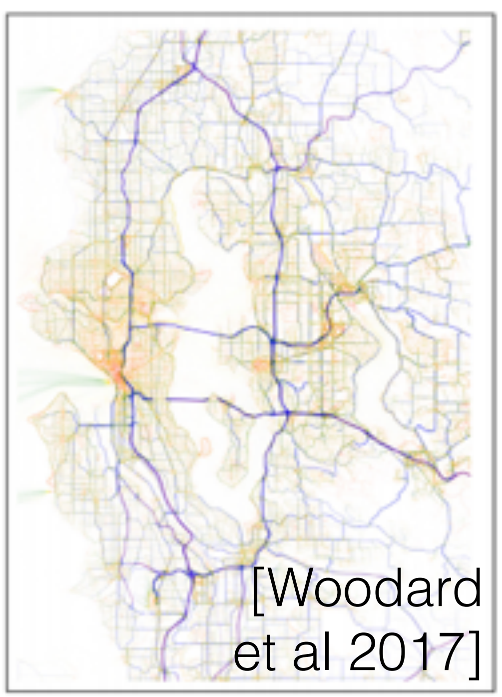


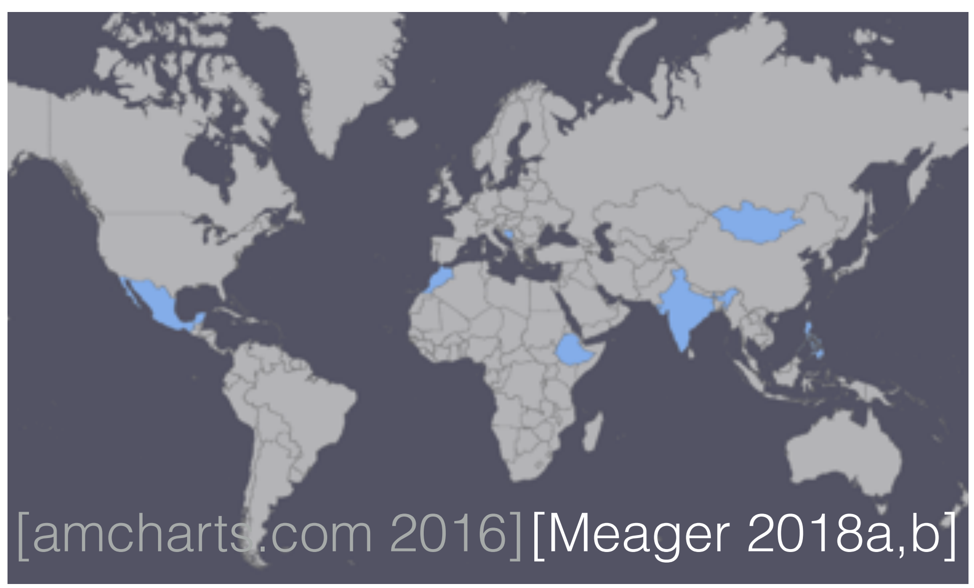
Bayesian inference

- Analysis goals: Point estimates, coherent uncertainties
  - Interpretable, complex, modular; expert information



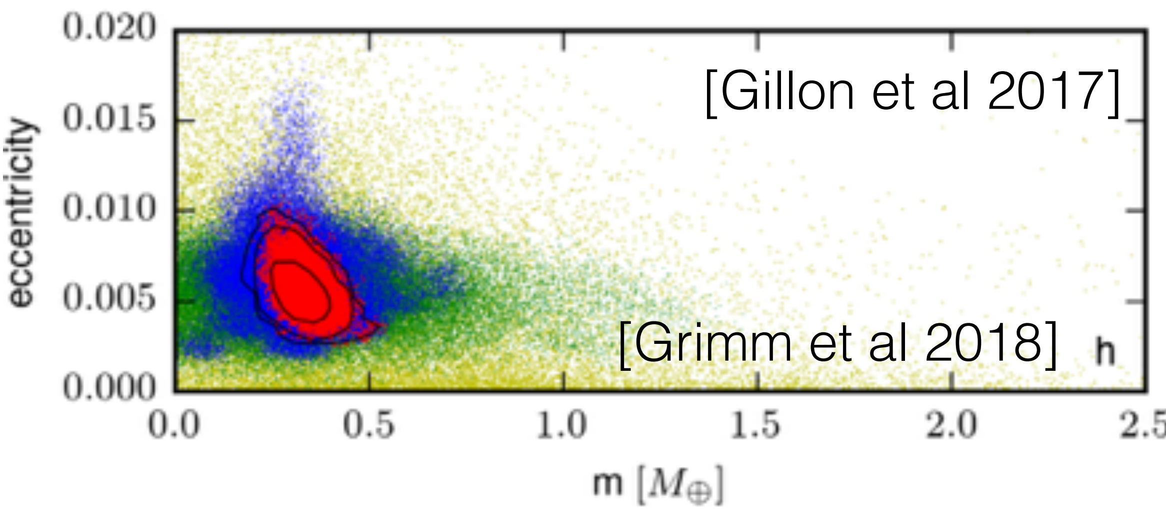


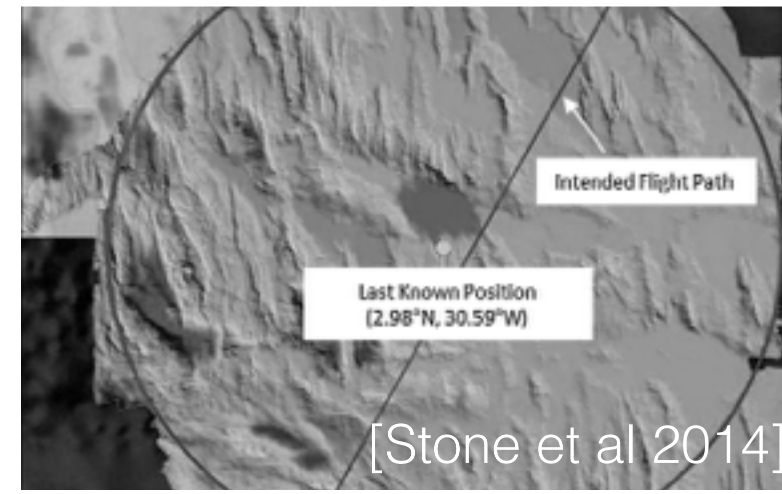


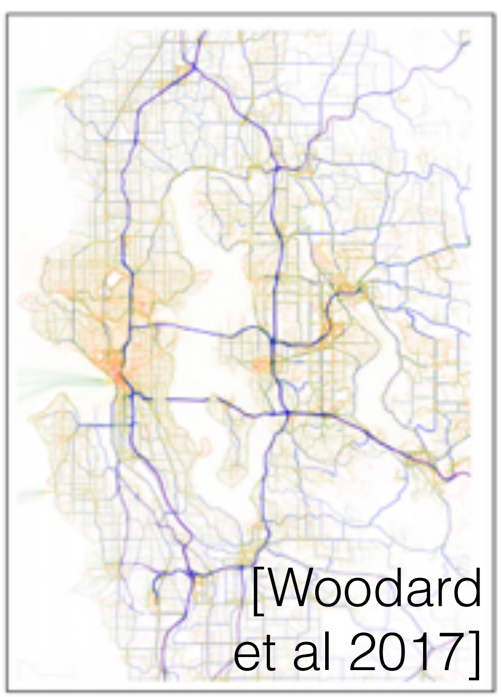


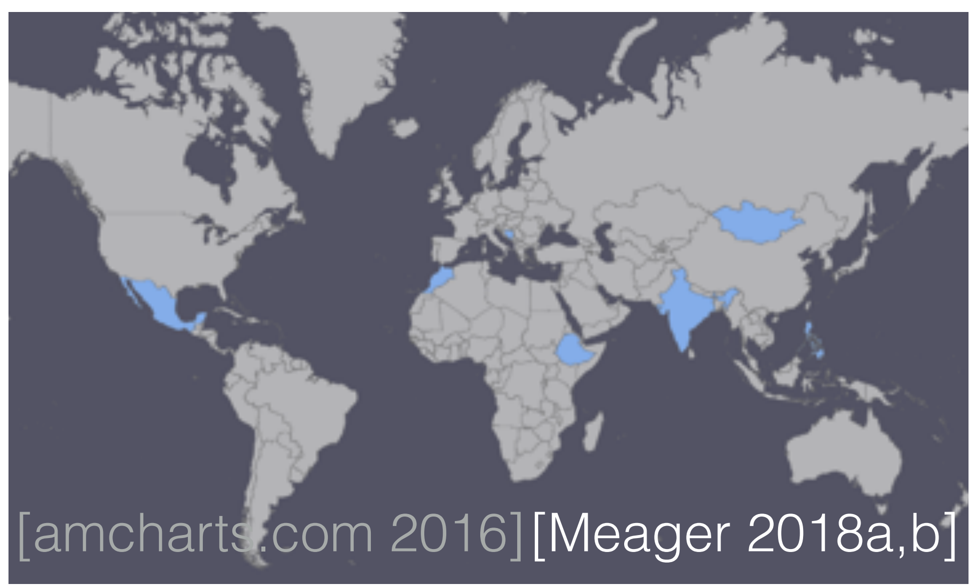
Bayesian inference

- Analysis goals: Point estimates, coherent uncertainties
  - Interpretable, complex, modular; expert information





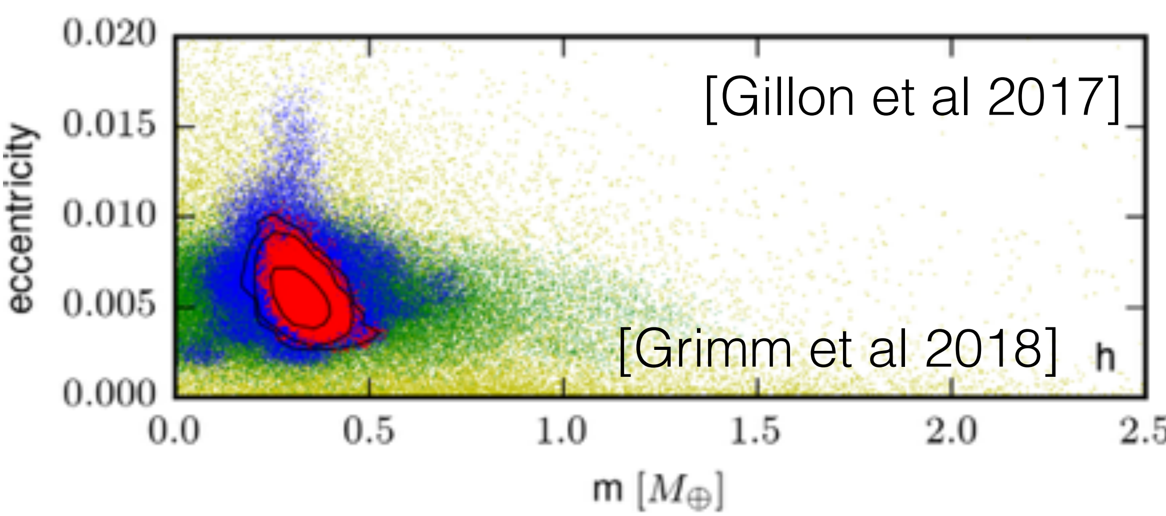


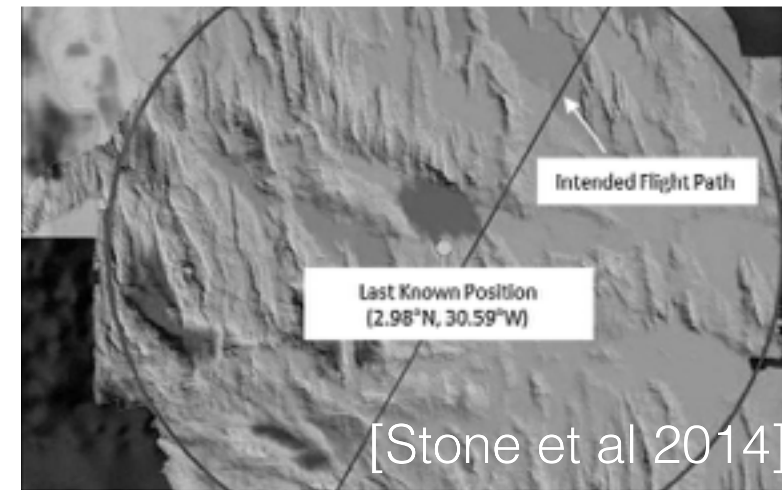


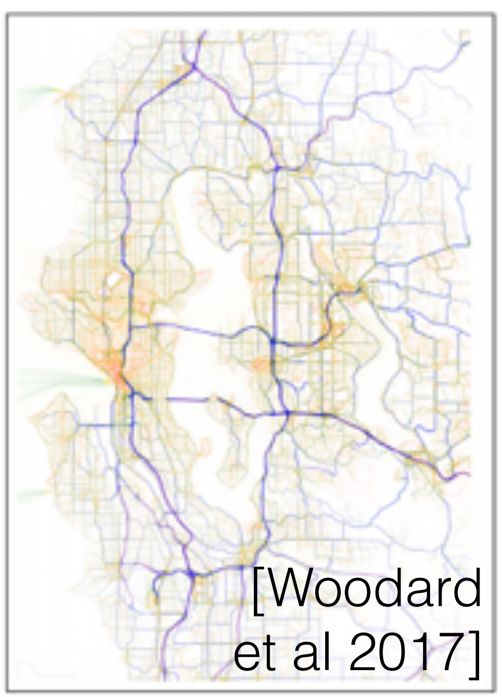
[mc-stan.org]

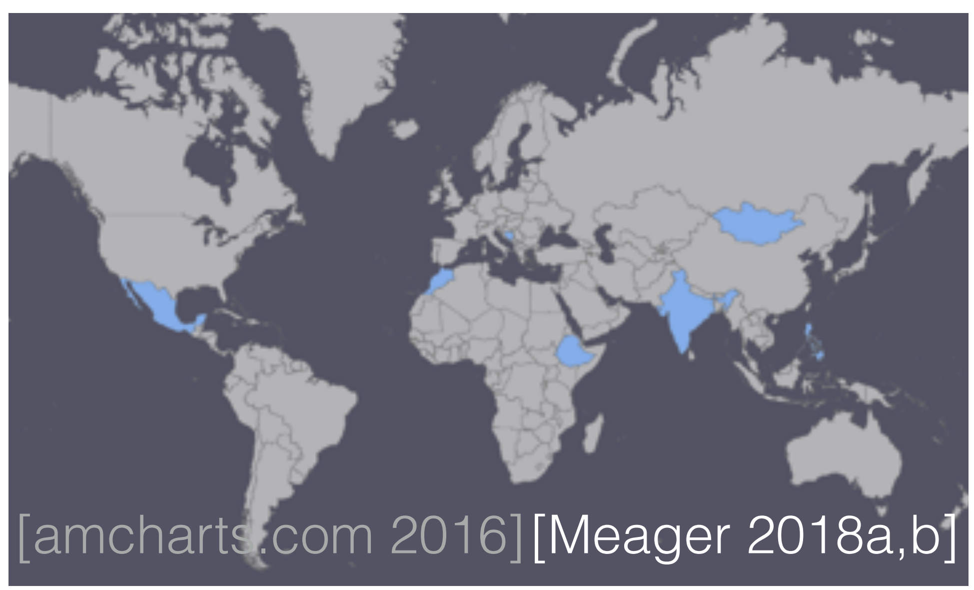
Bayesian inference

- Analysis goals: Point estimates, coherent uncertainties
  - Interpretable, complex, modular; expert information







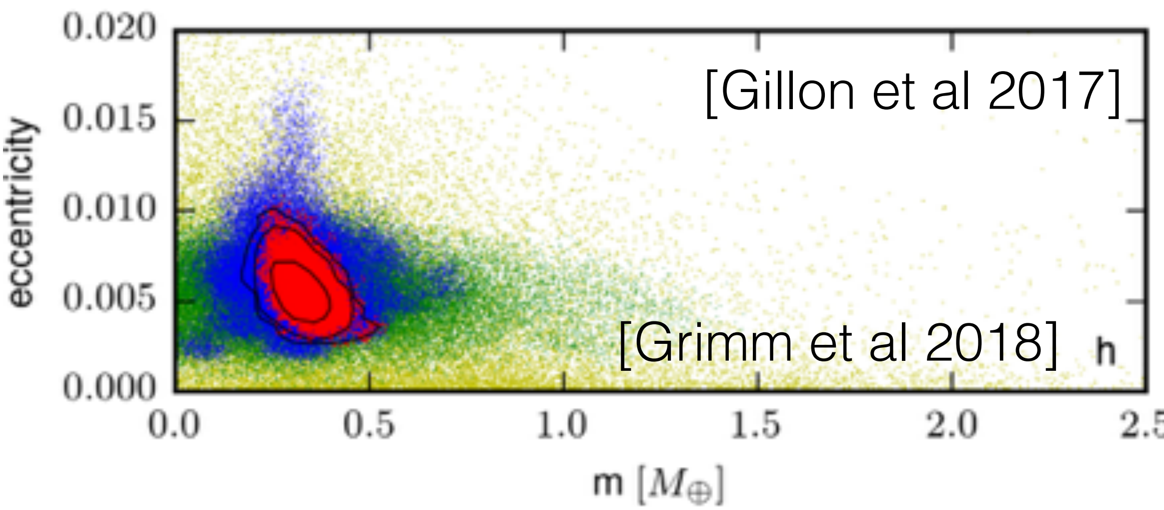


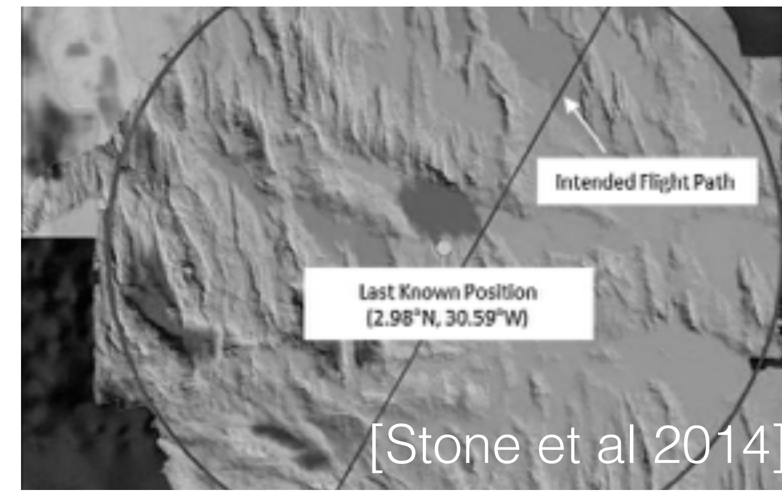
[mc-stan.org]

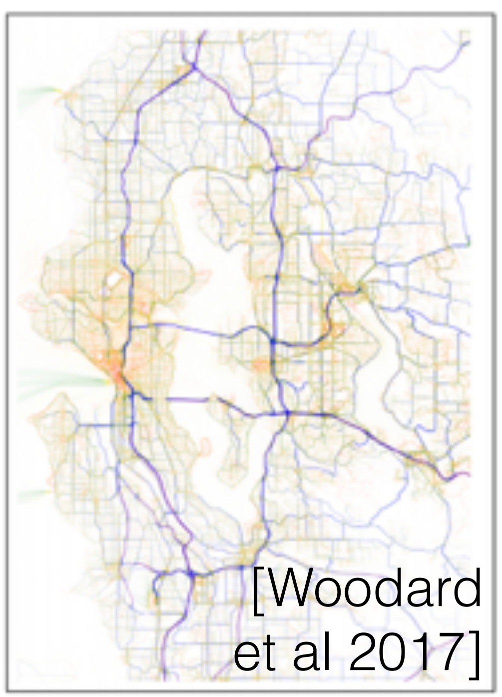
- Challenge: fast (compute, user), reliable inference

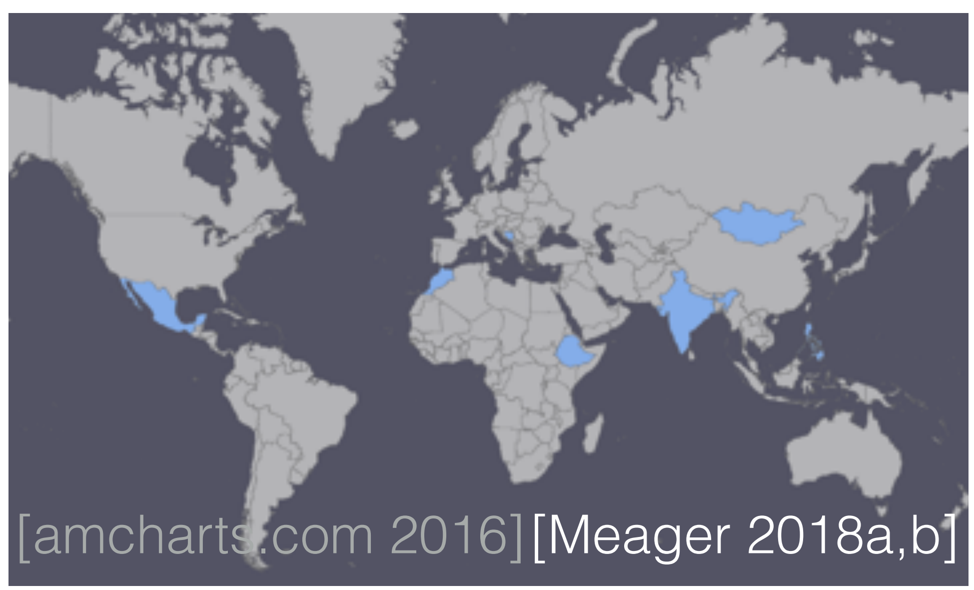
Bayesian inference

- Analysis goals: Point estimates, coherent uncertainties
  - Interpretable, complex, modular; expert information









[mc-stan.org]

- Challenge: fast (compute, user), reliable inference
- Uncertainty doesn't have to disappear in large data sets

Variational Bayes

Variational Bayes

- Modern problems: often large data, large dimensions

Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

[Blei et al 2003]

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

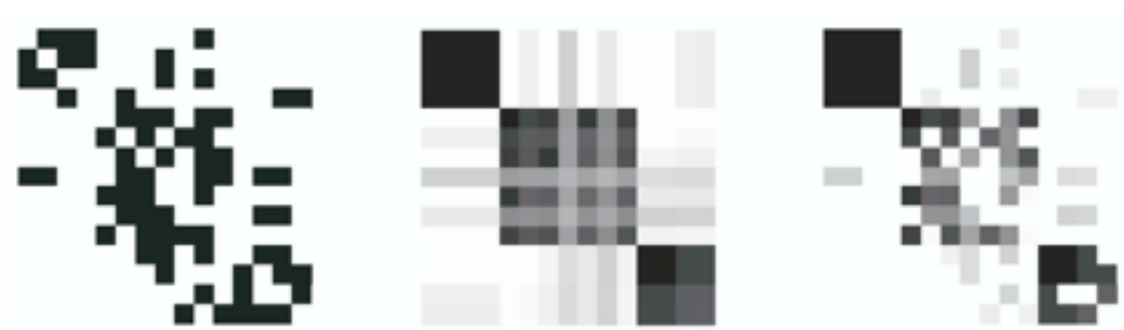
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

[Blei et al 2003]

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



[Airoldi et al 2008]

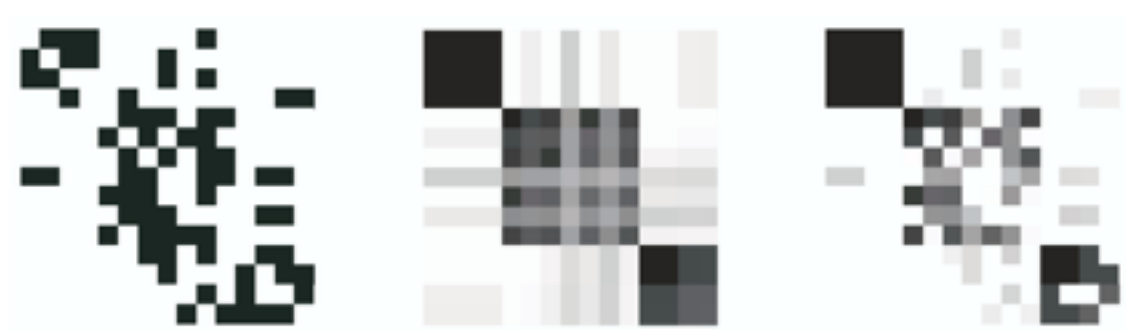
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

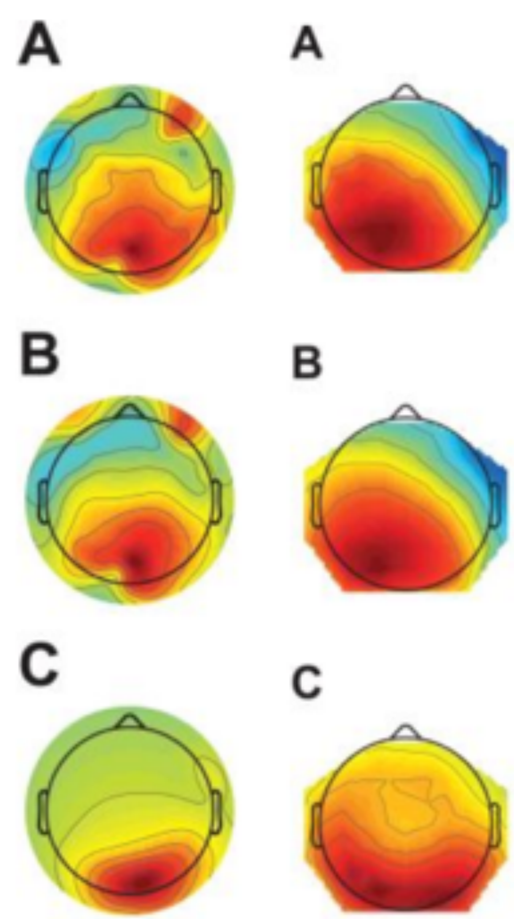
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

[Blei et al 2003]

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



[Airoldi et al 2008]



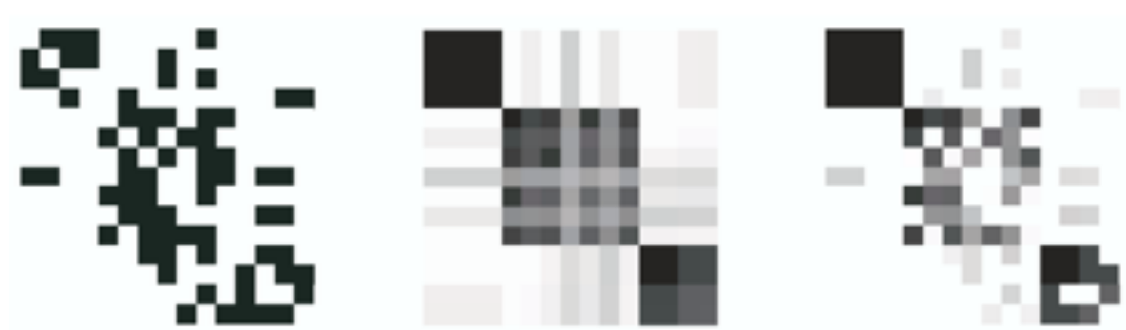
[Gershman et al 2014]

[Blei et al 2018]

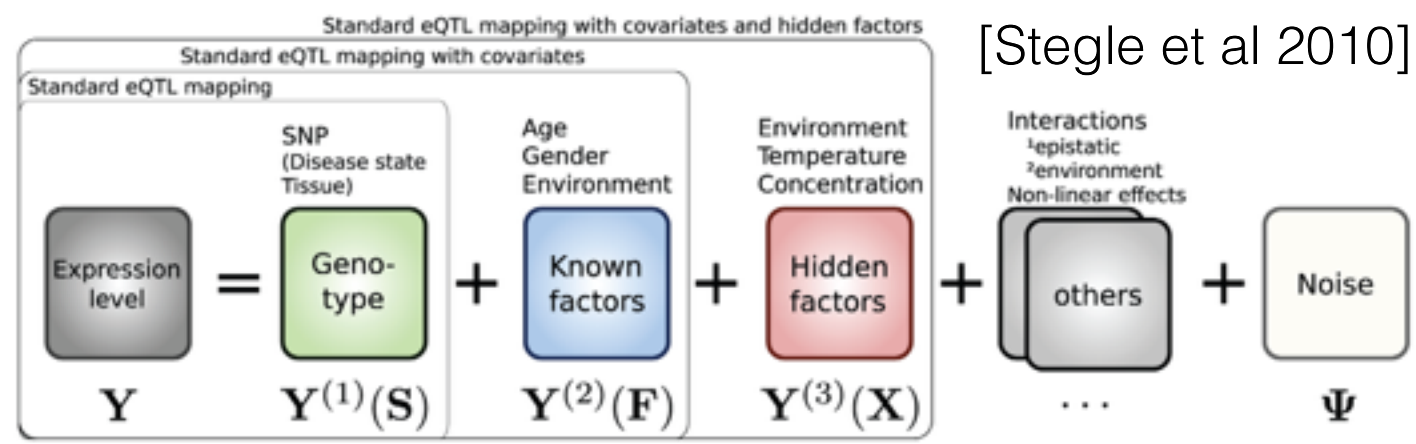
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

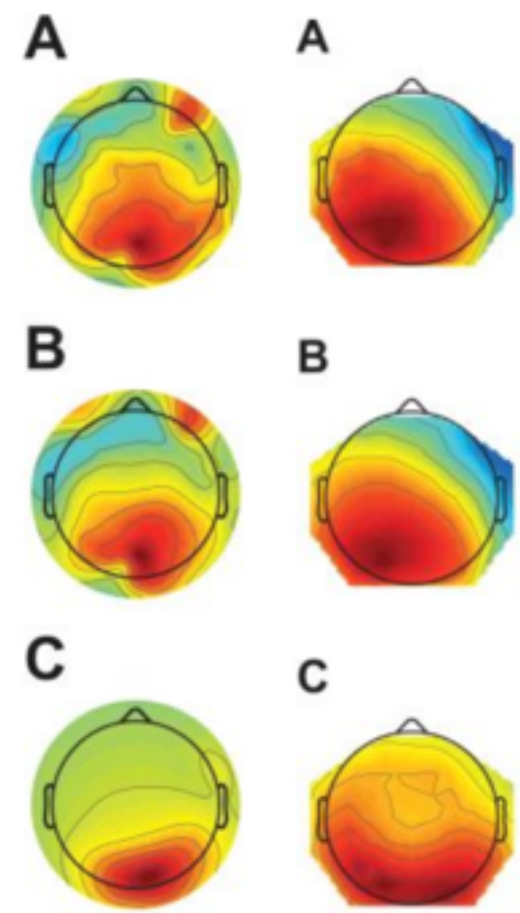
[Blei et al 2003]



[Airoldi et al 2008]



[Stegle et al 2010]

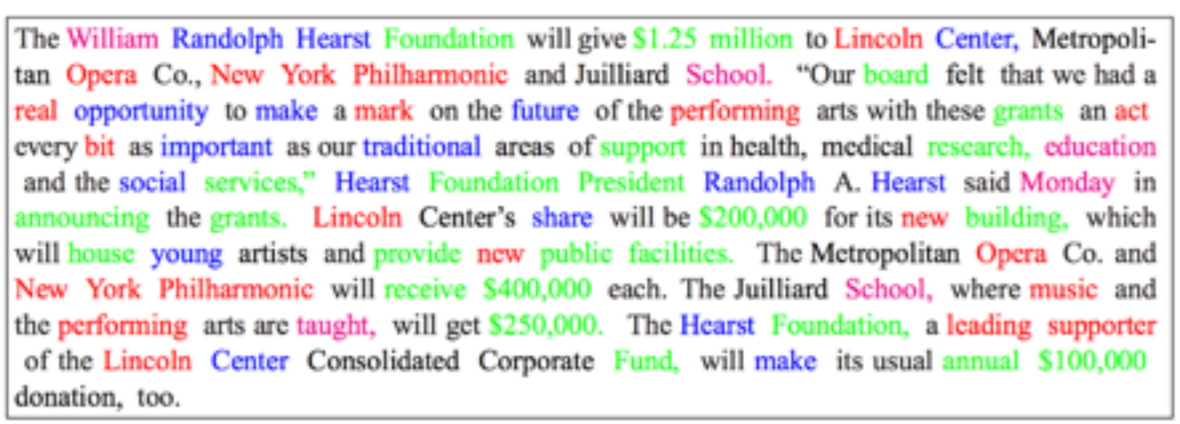


[Gershman et al 2014]

[Blei et al 2018]

Variational Bayes

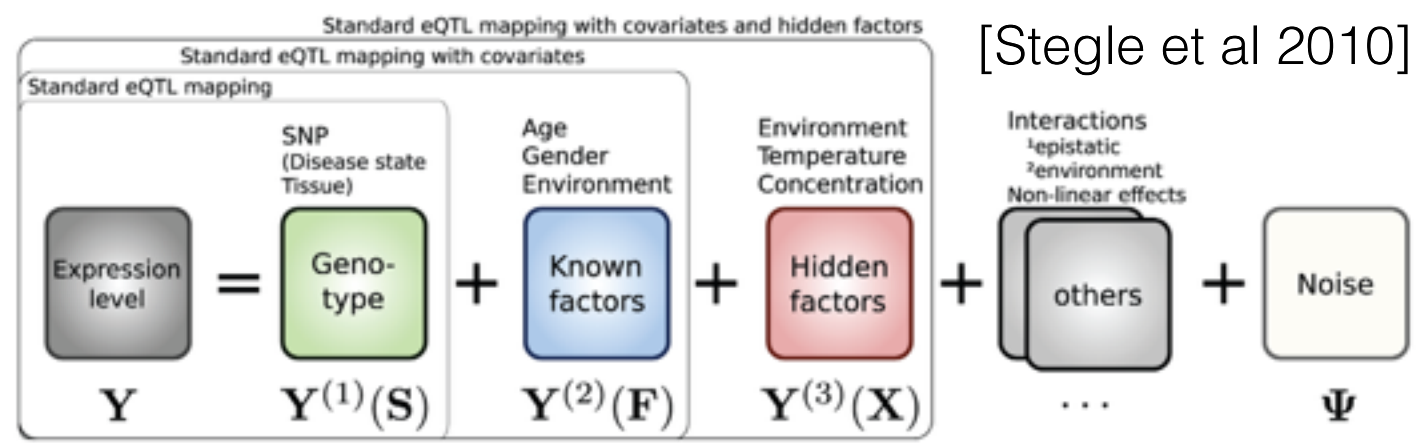
- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast



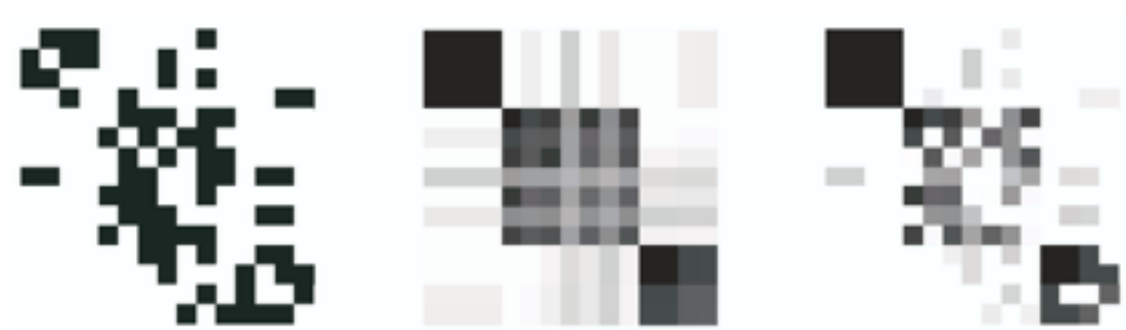
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

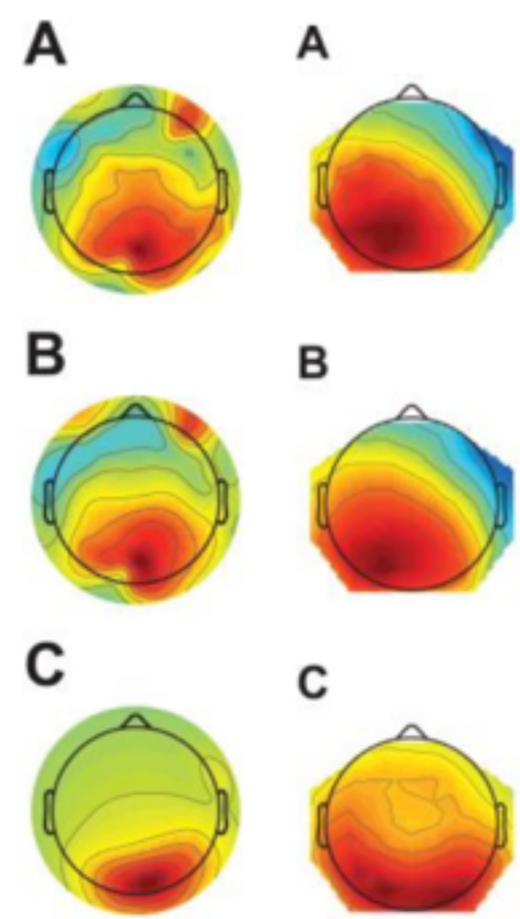
[Blei et al 2003]



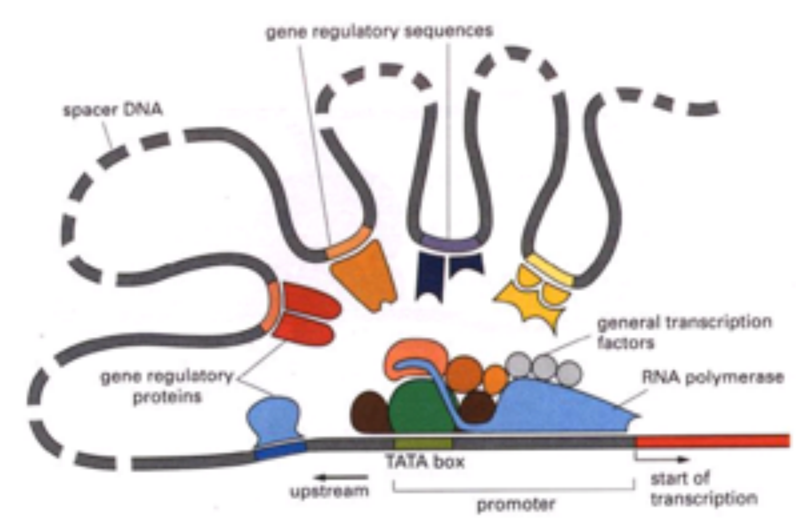
[Stegle et al 2010]



[Airoldi et al 2008]



[Gershman et al 2014]



[Xing et al 2004]
[Xing 2003]

[Blei et al 2018]

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

Bayesian inference

Bayesian inference

parameters

θ

Bayesian inference

parameters

$$p(\theta)$$

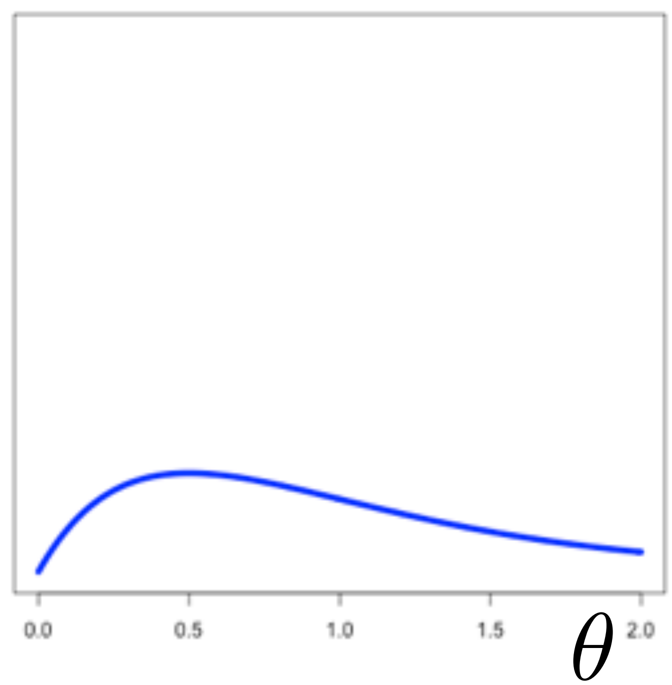
prior

Bayesian inference

parameters

$p(\theta)$

prior

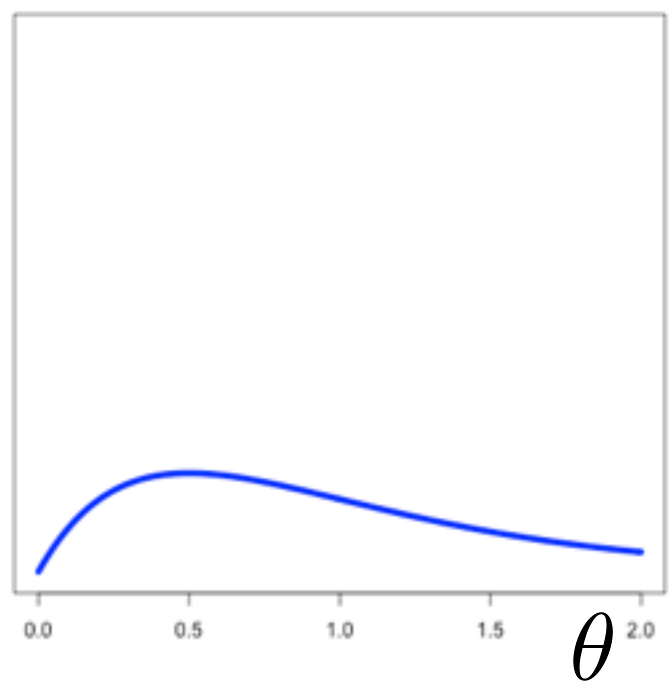


Bayesian inference

parameters

$$p(y_{1:N}|\theta)p(\theta)$$

likelihood prior

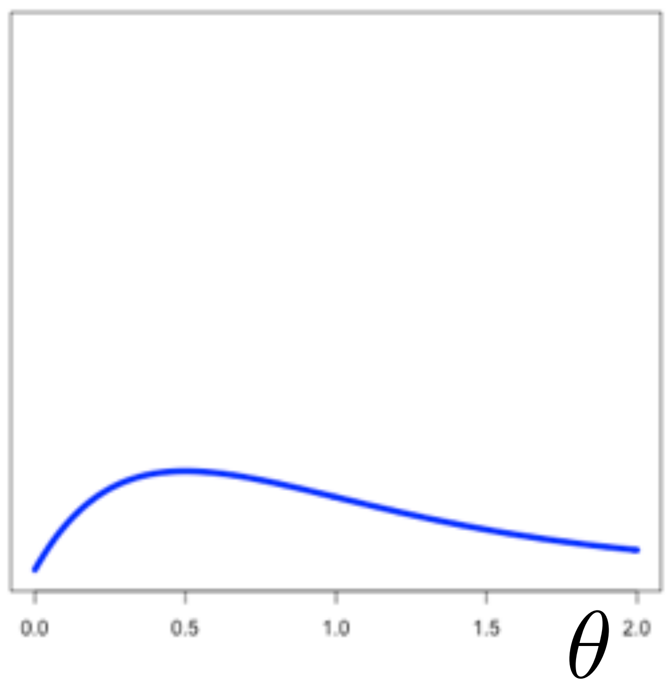


Bayesian inference

data parameters

$$p(y_{1:N}|\theta)p(\theta)$$

likelihood prior

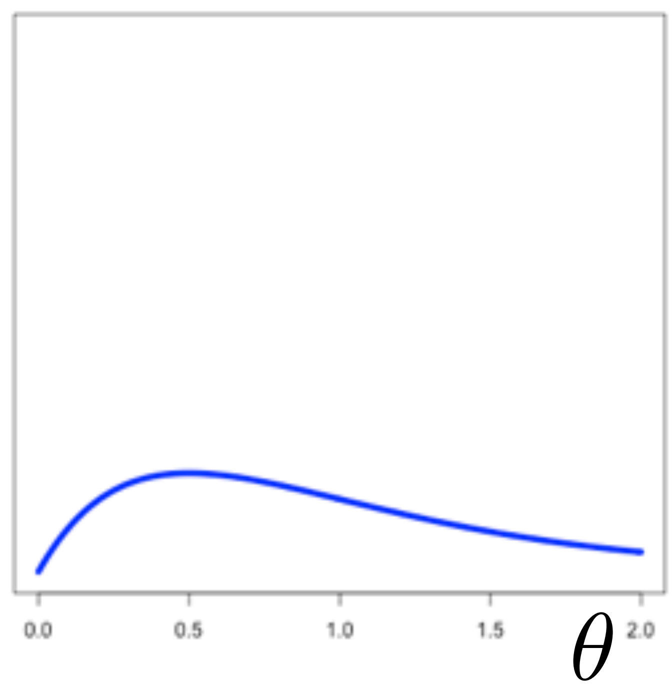


Bayesian inference

data parameters

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior

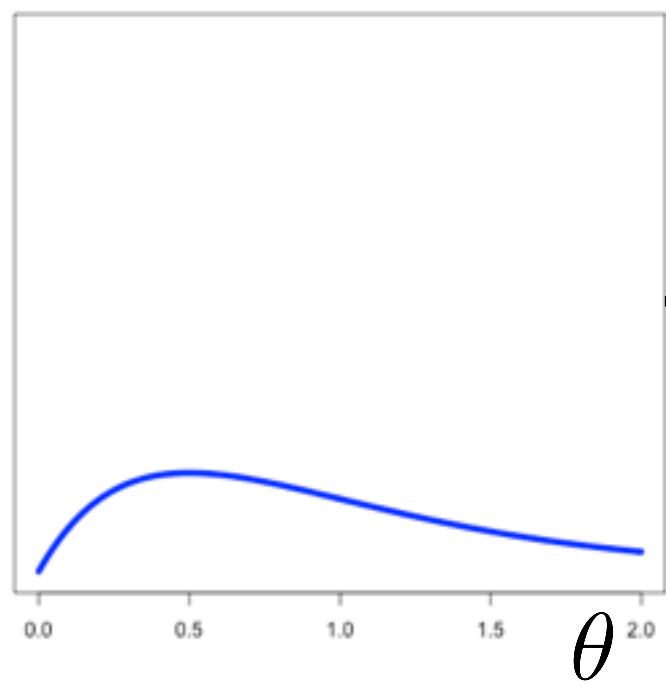


Bayesian inference

data → parameters →

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



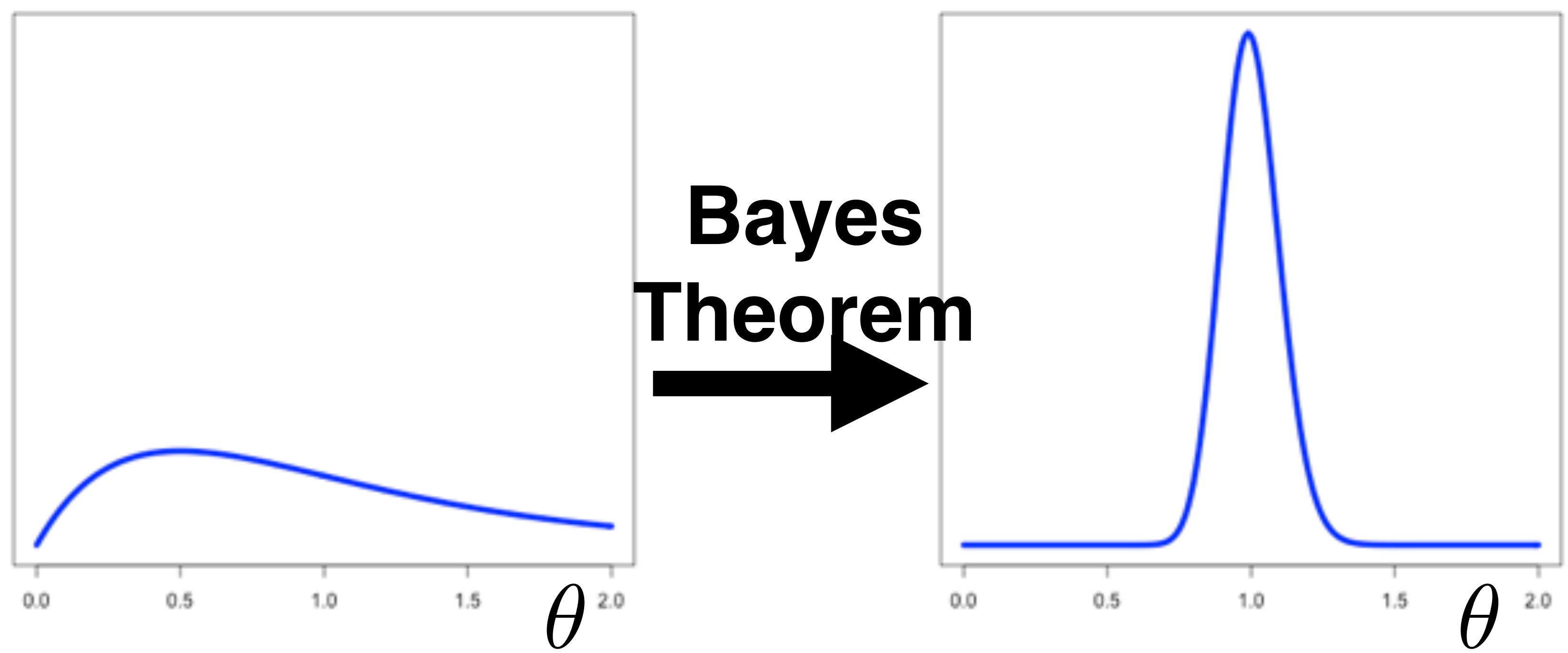
Bayes Theorem

Bayesian inference

data parameters

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior

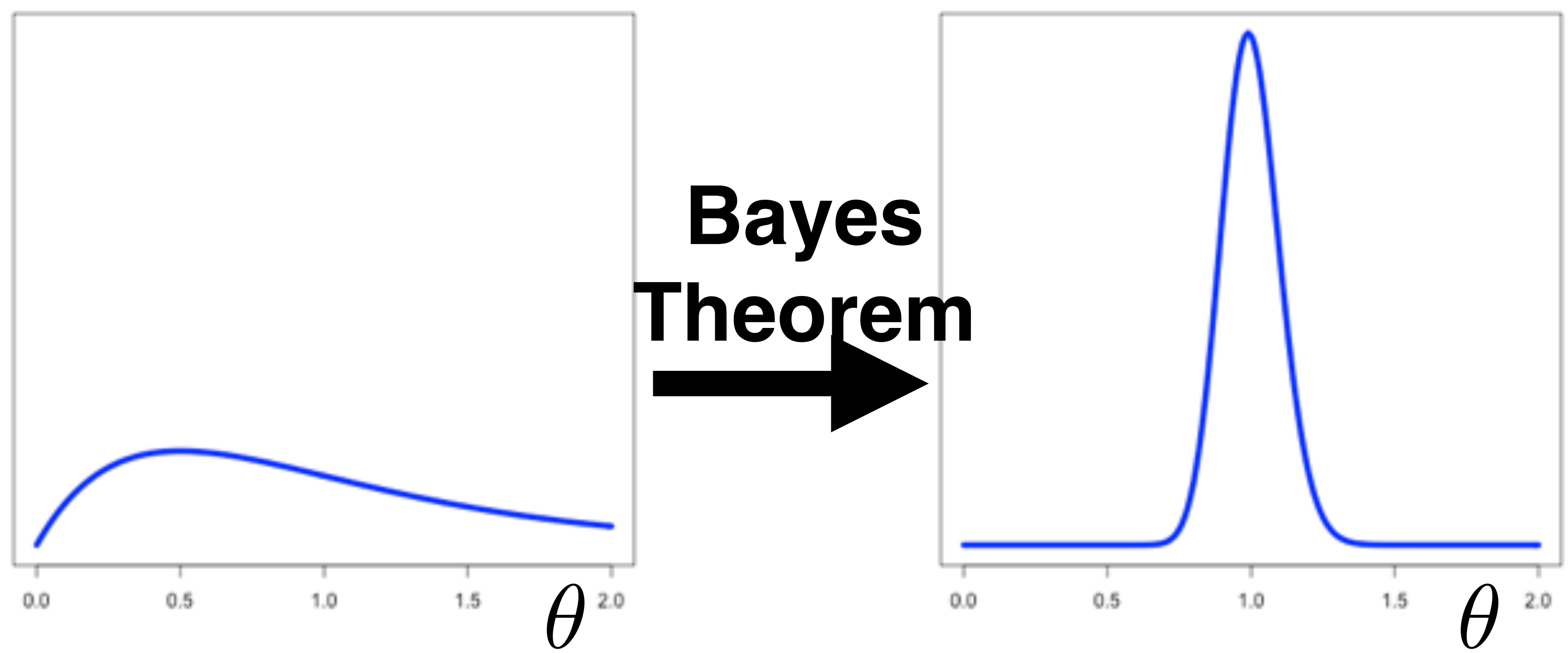


Bayesian inference

data parameters

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



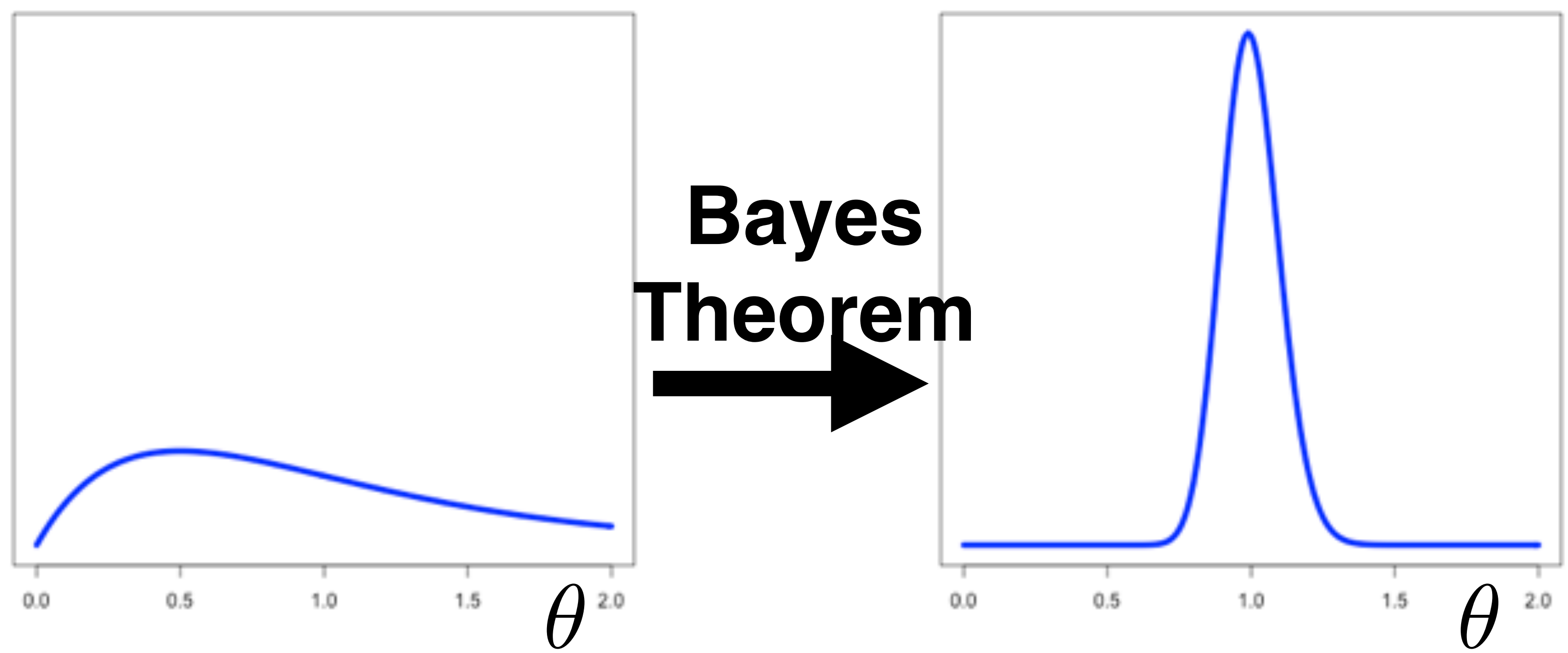
1. Build a model: choose prior, likelihood

Bayesian inference

data parameters

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



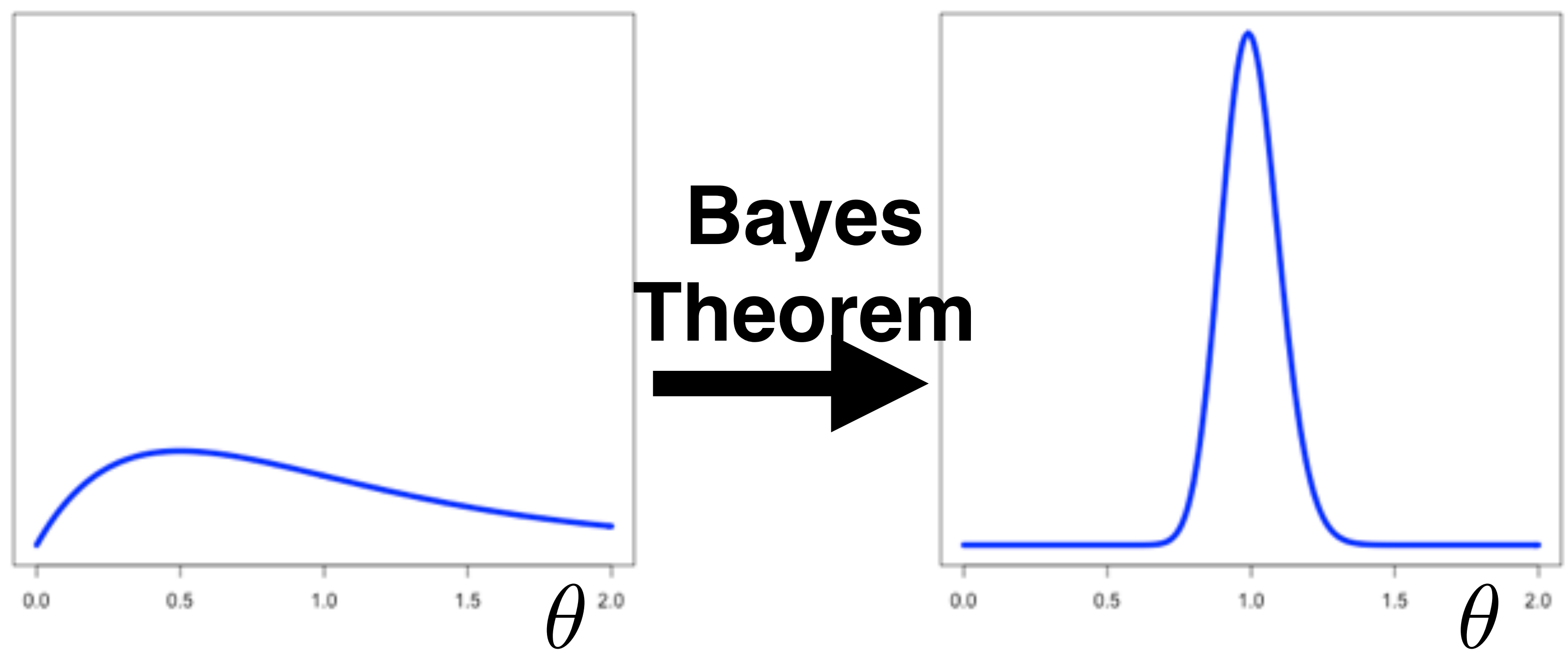
1. Build a model: choose prior, likelihood
2. Compute the posterior

Bayesian inference

data → $y_{1:N}$; parameters → θ

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



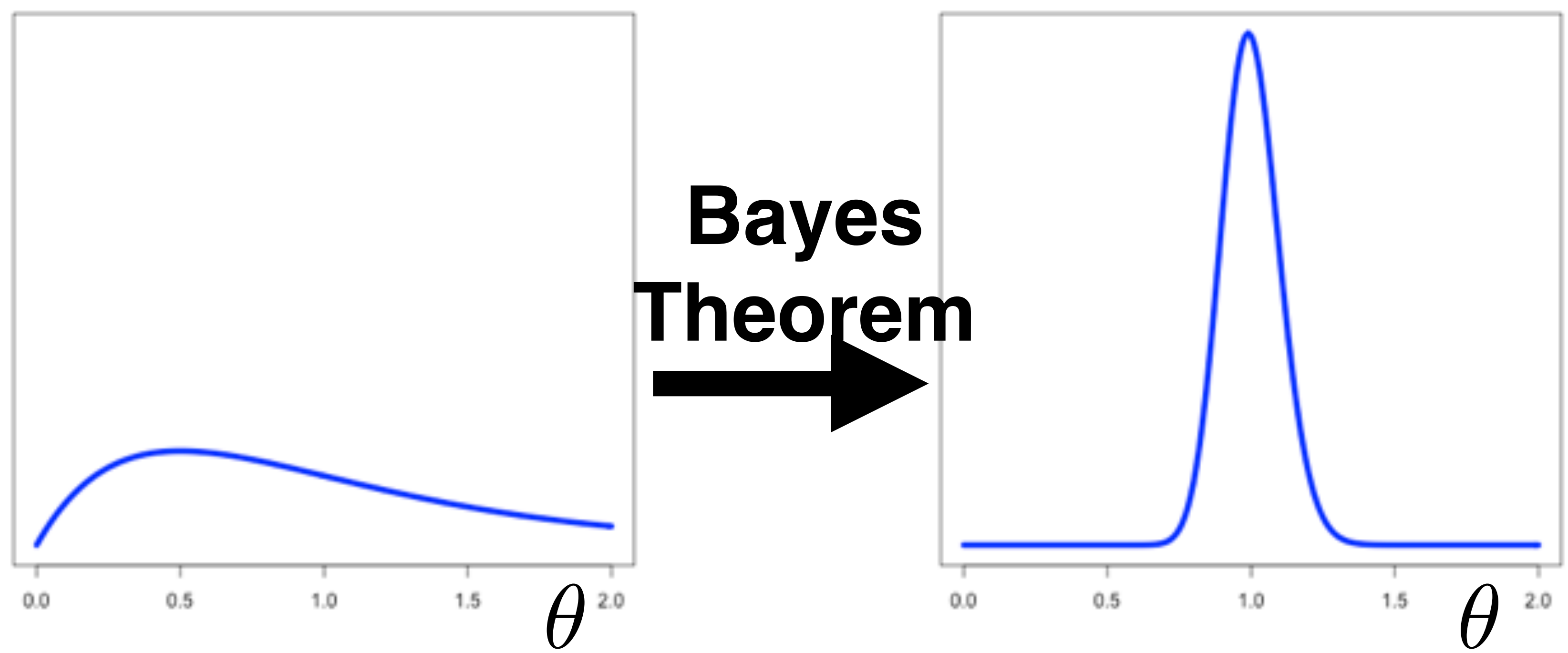
1. Build a model: choose prior, likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

Bayesian inference

data parameters

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



1. Build a model: choose prior, likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

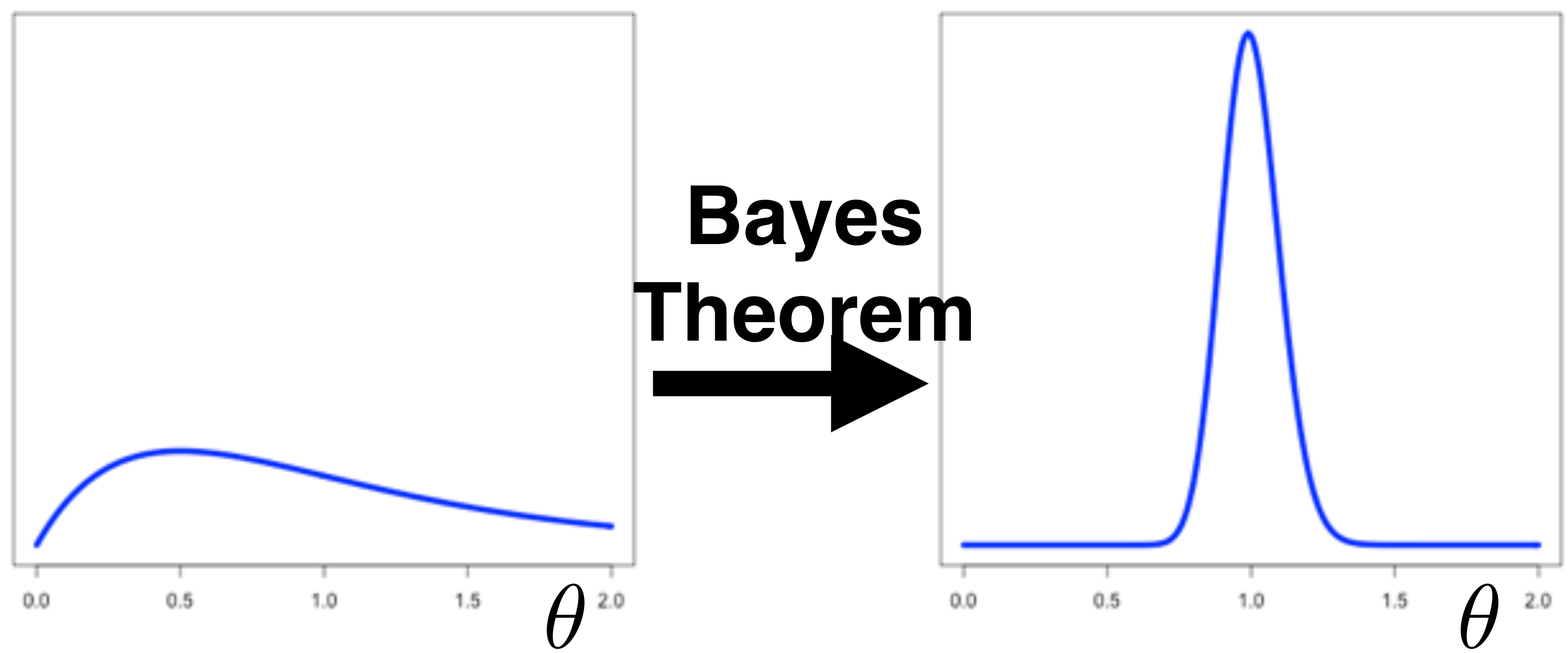
- Why are steps 2 and 3 hard?

Bayesian inference

data parameters

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



1. Build a model: choose prior, likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

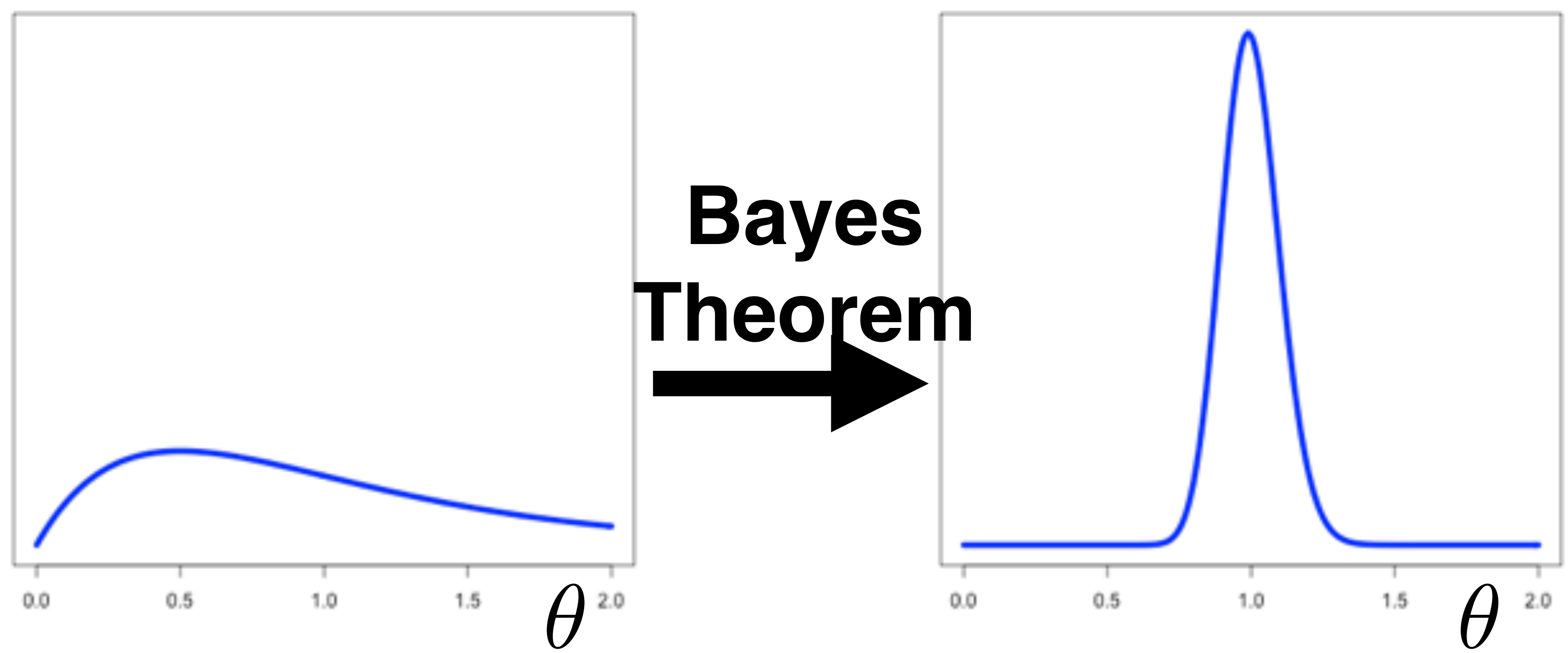
- Why are steps 2 and 3 hard? High-dimensional integration

Bayesian inference

data parameters

$$p(\theta|y_{1:N}) = p(y_{1:N}|\theta)p(\theta)/p(y_{1:N})$$

posterior likelihood prior



1. Build a model: choose prior, likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

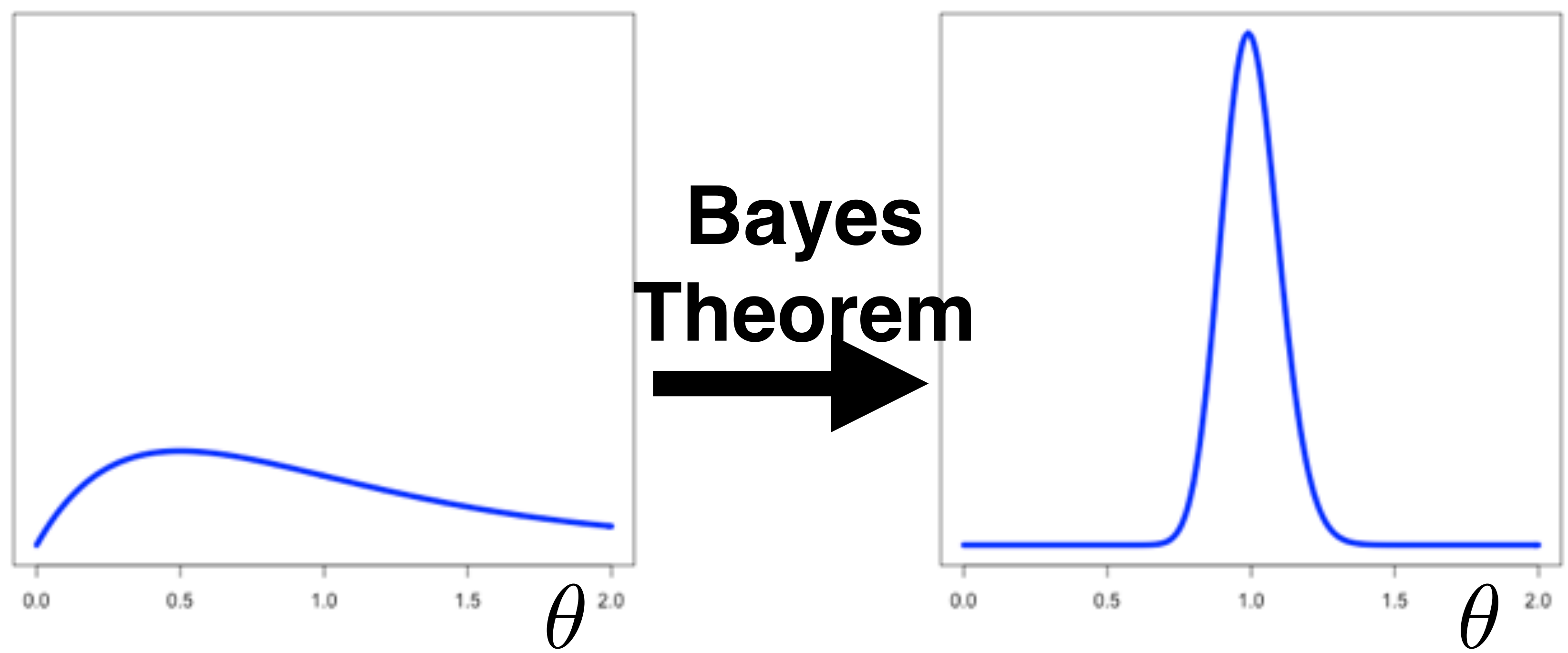
- Why are steps 2 and 3 hard? High-dimensional integration

Bayesian inference

data → $y_{1:N}$, parameters → θ

$$p(\theta|y_{1:N}) = p(y_{1:N}|\theta)p(\theta)/p(y_{1:N})$$

posterior likelihood prior evidence



1. Build a model: choose prior, likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

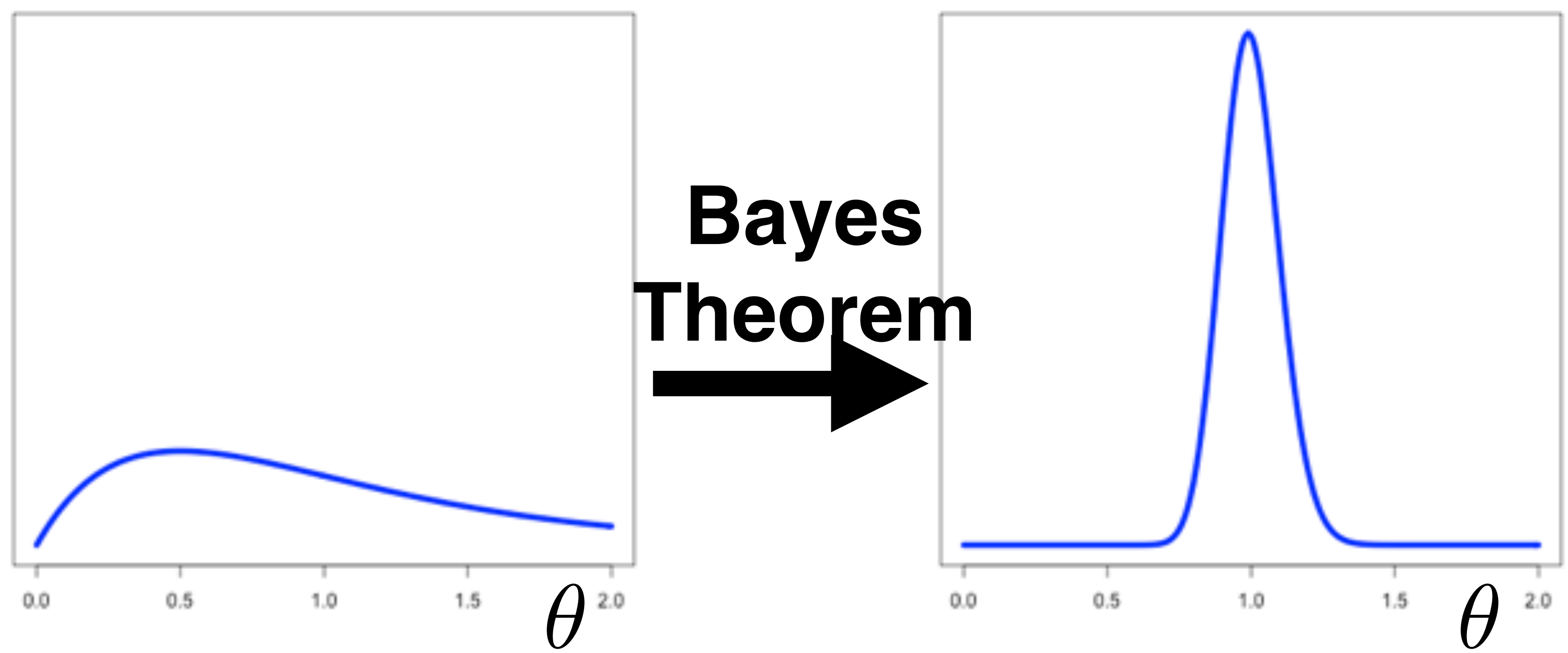
- Why are steps 2 and 3 hard? High-dimensional integration

Bayesian inference

data → $y_{1:N}$ parameters → θ

$$p(\theta|y_{1:N}) = p(y_{1:N}|\theta)p(\theta) / \int p(y_{1:N},\theta)d\theta$$

posterior likelihood prior evidence



1. Build a model: choose prior, likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

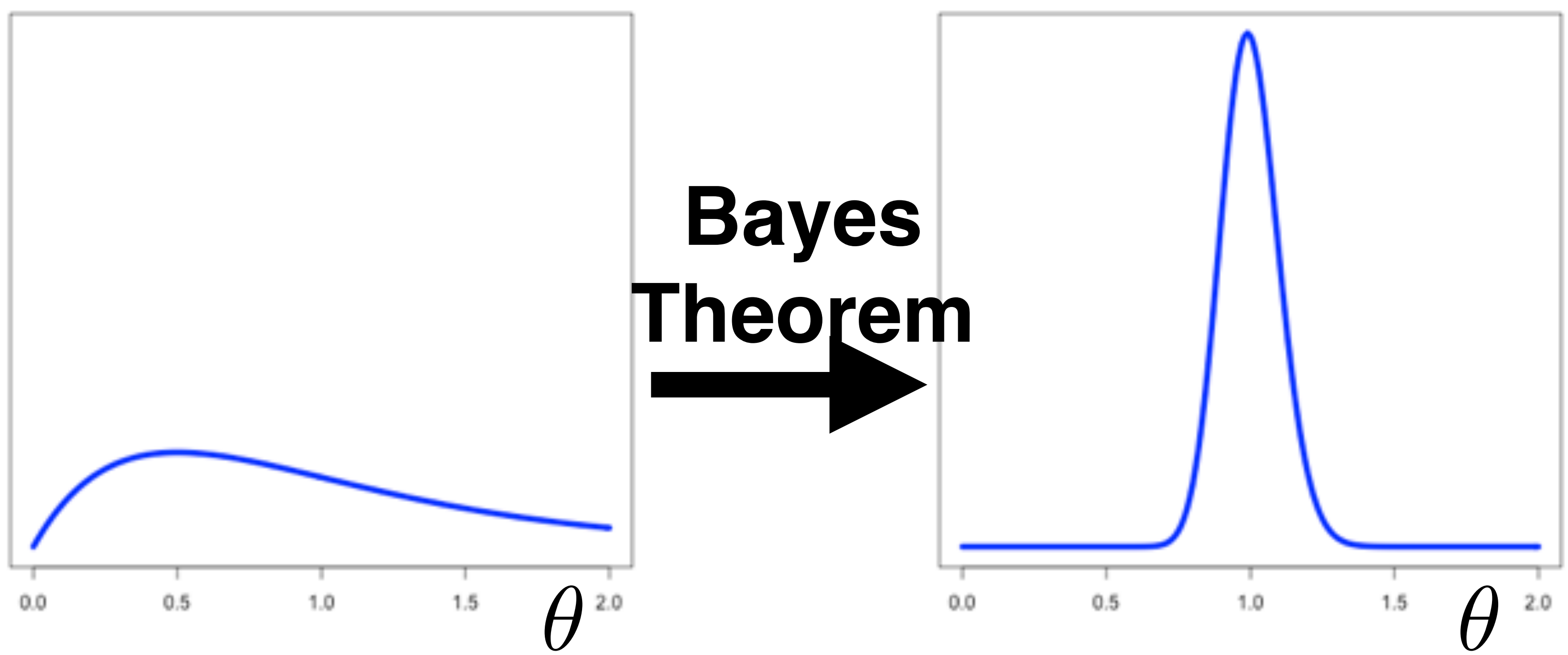
- Why are steps 2 and 3 hard? High-dimensional integration

Bayesian inference

data → $y_{1:N}$, parameters → θ

$$p(\theta|y_{1:N}) = p(y_{1:N}|\theta)p(\theta) / \int p(y_{1:N}, \theta) d\theta$$

posterior likelihood prior evidence



1. Build a model: choose prior, likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

- Why are steps 2 and 3 hard? High-dimensional integration
- Turn to approximation

Approximate Bayesian Inference

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

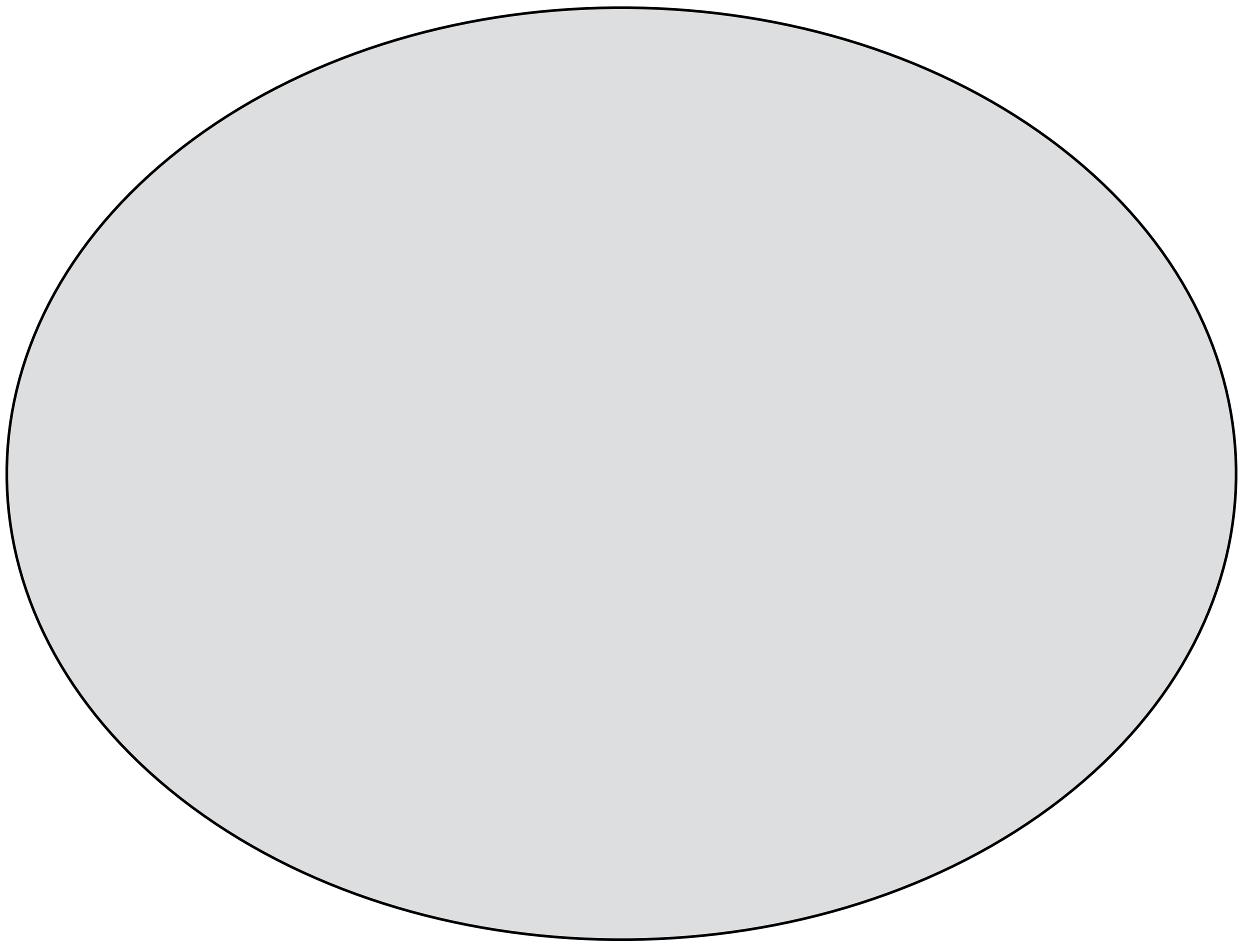
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



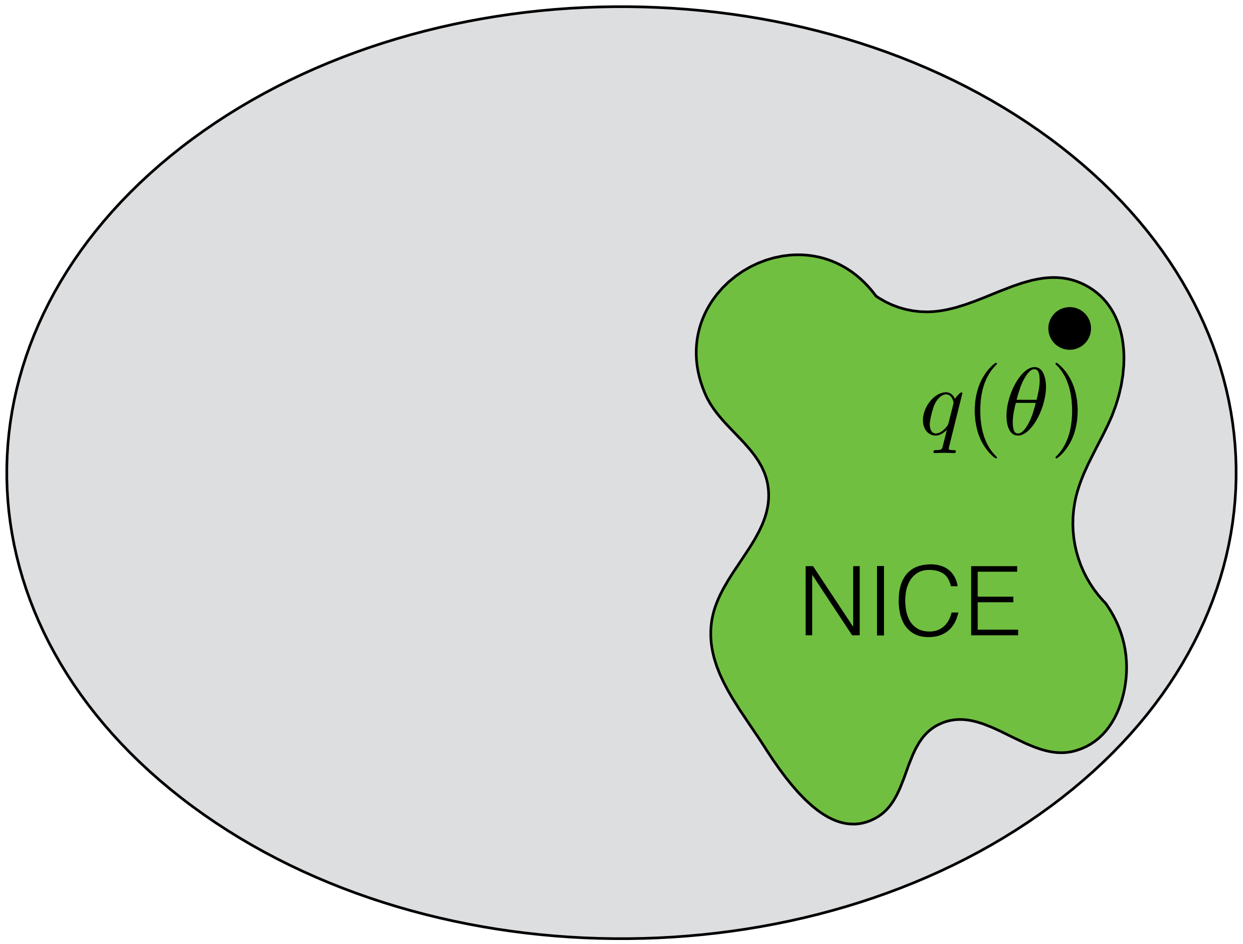
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



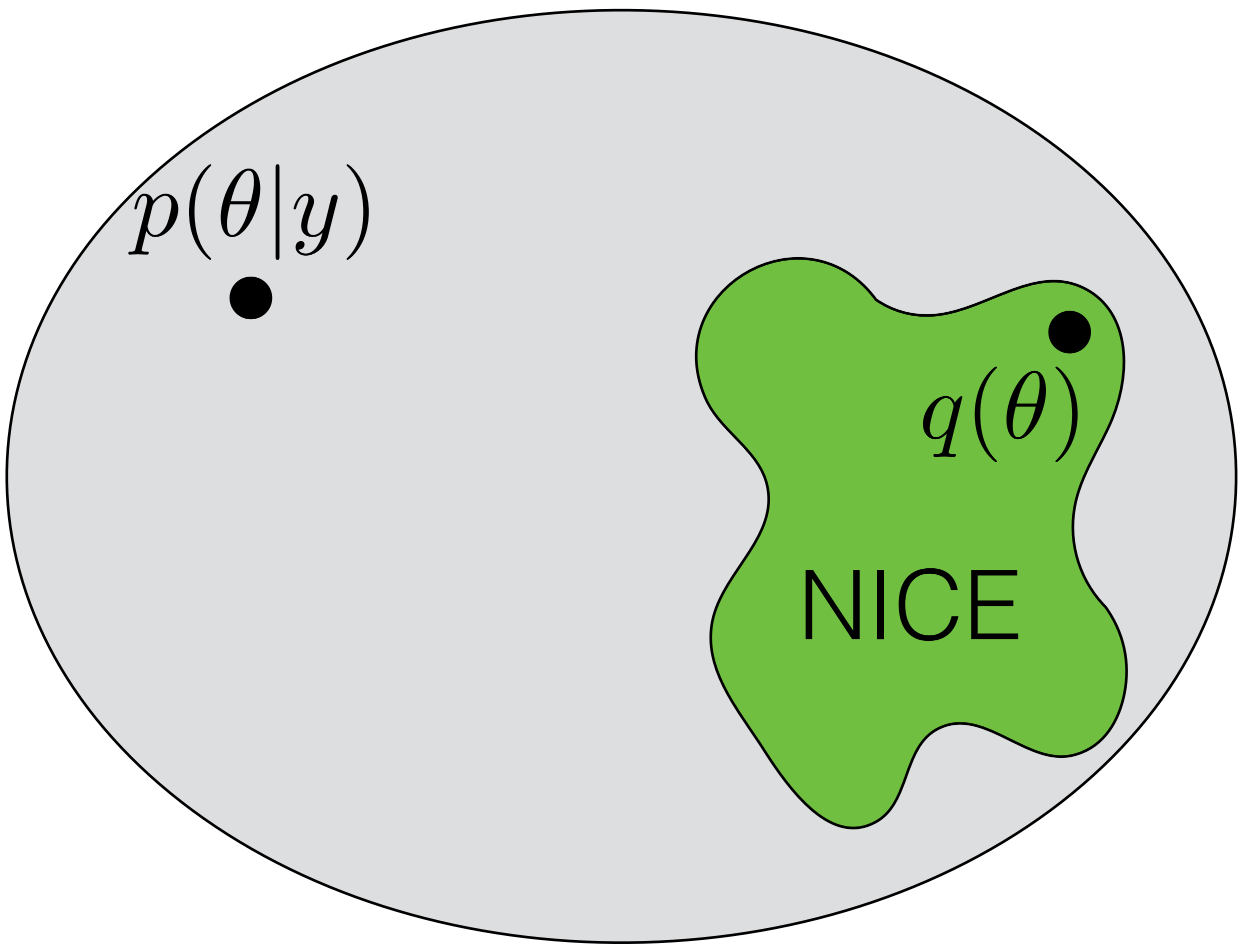
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



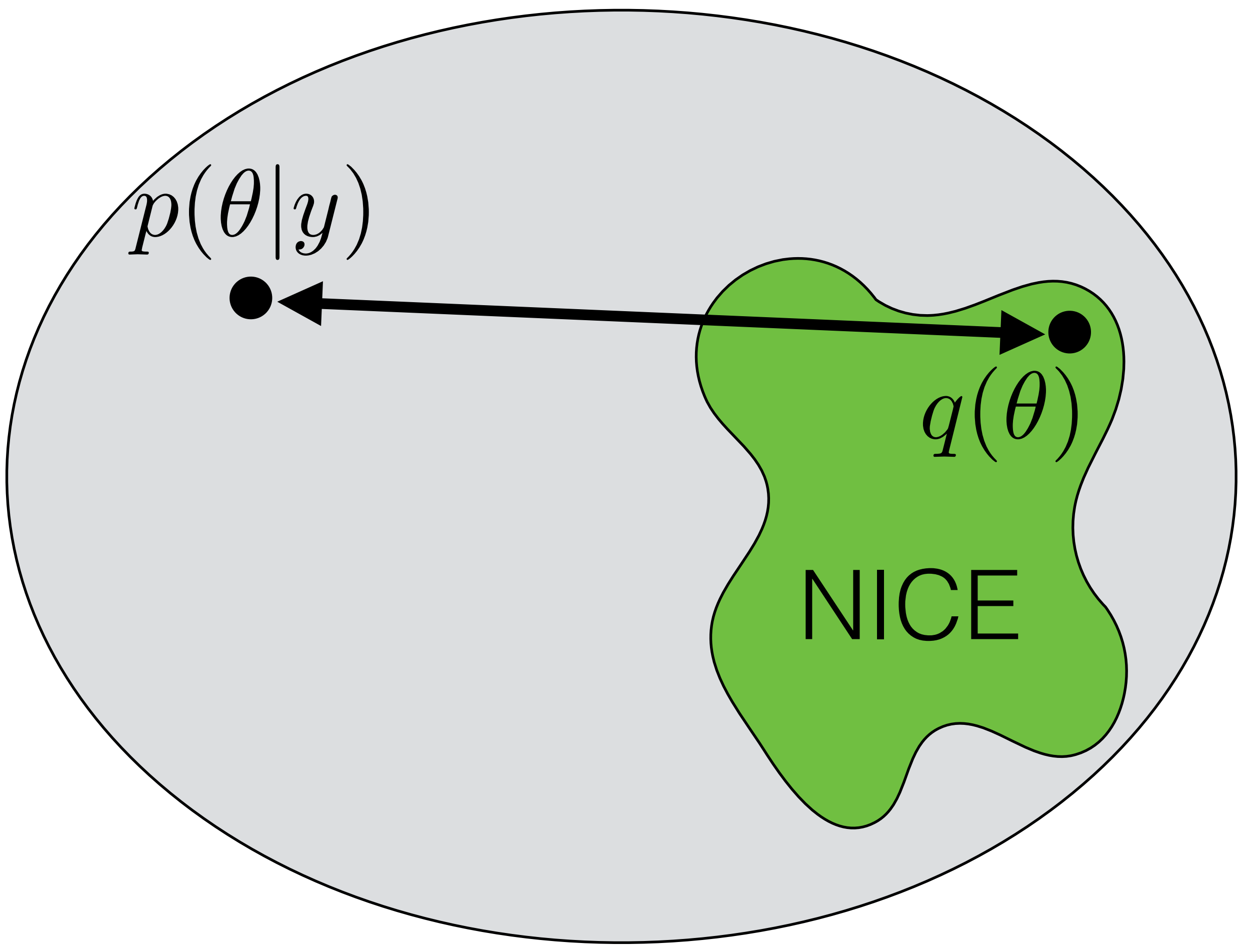
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



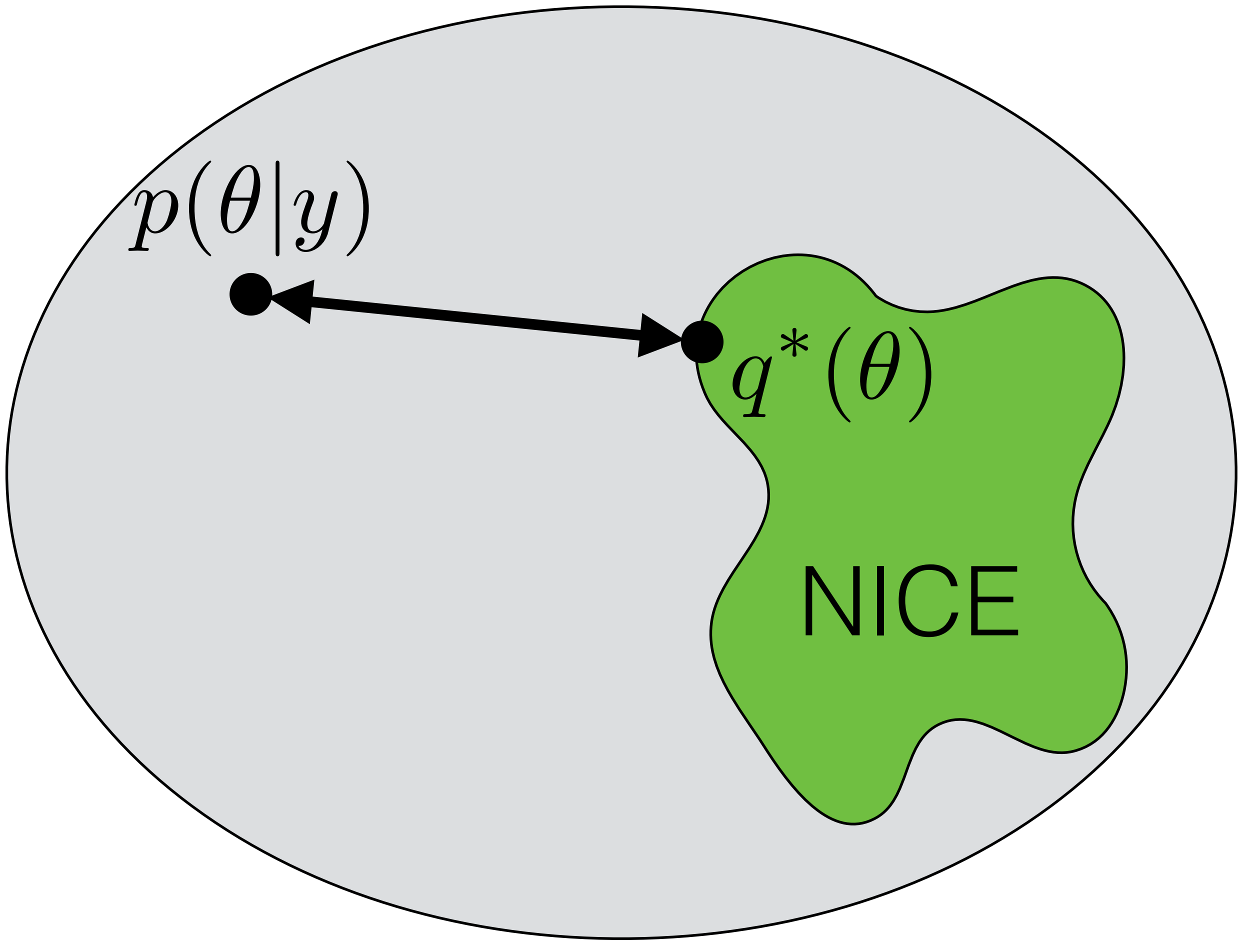
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



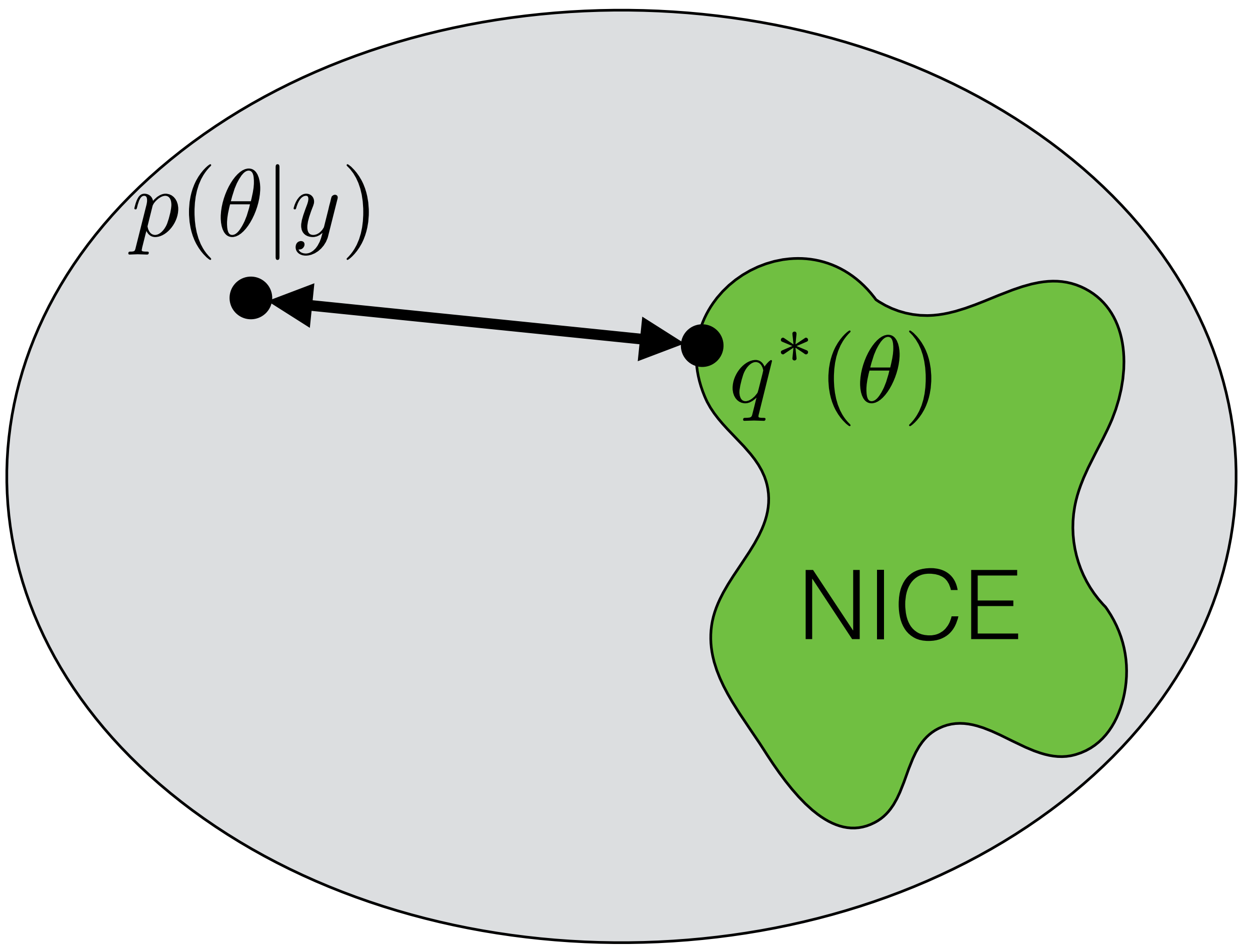
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

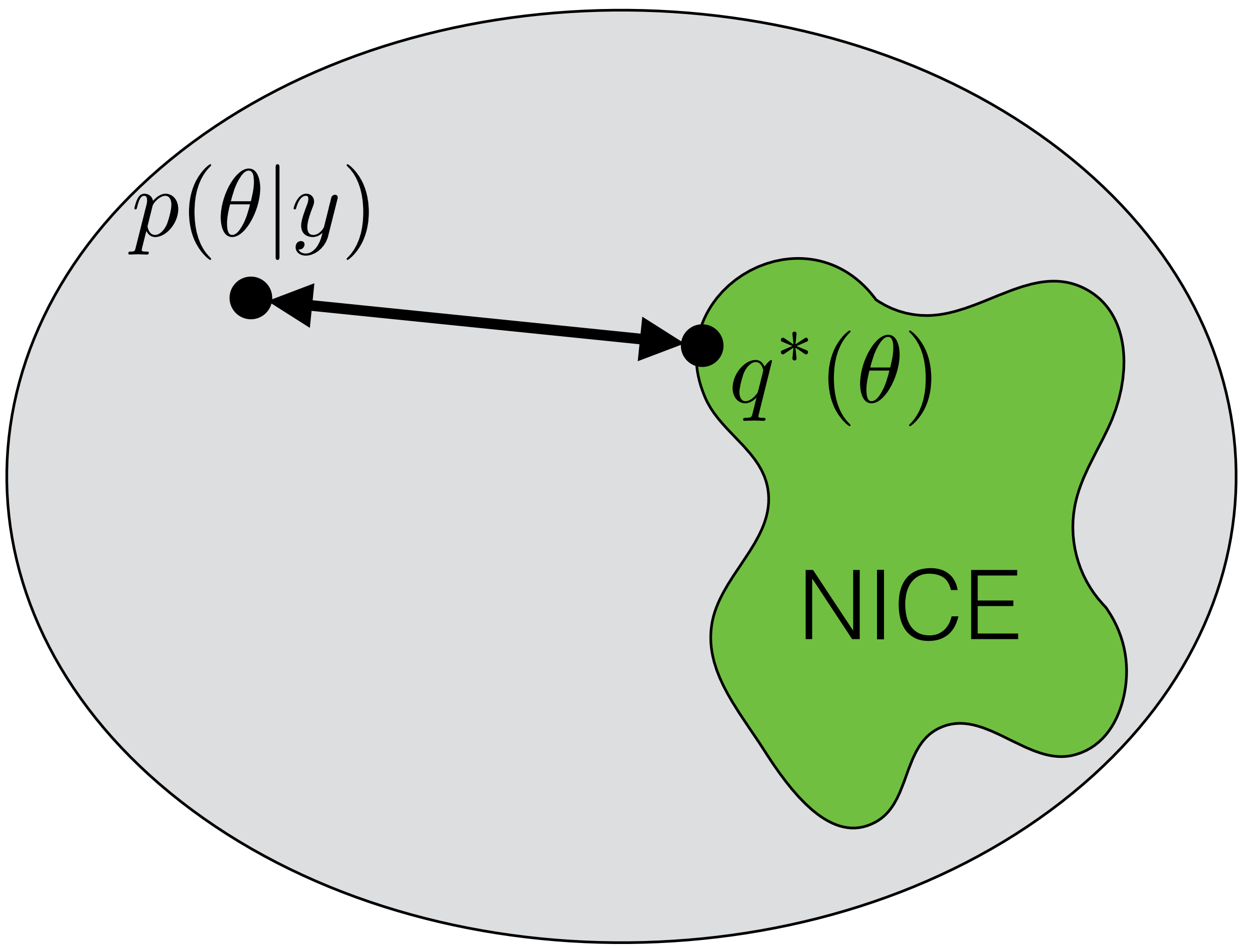
- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

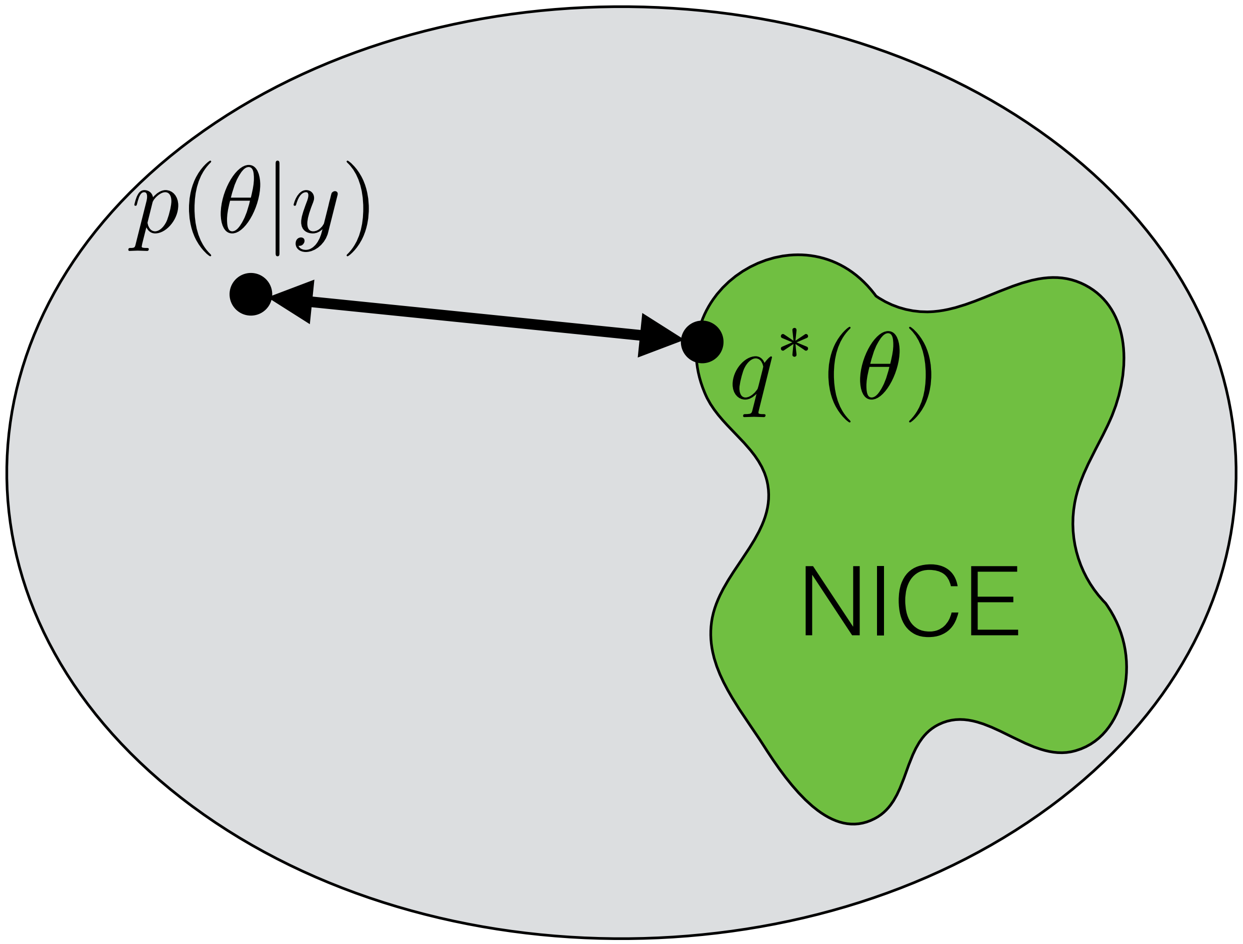
- Approximate posterior with q^*

$$q^* = \mathrm{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

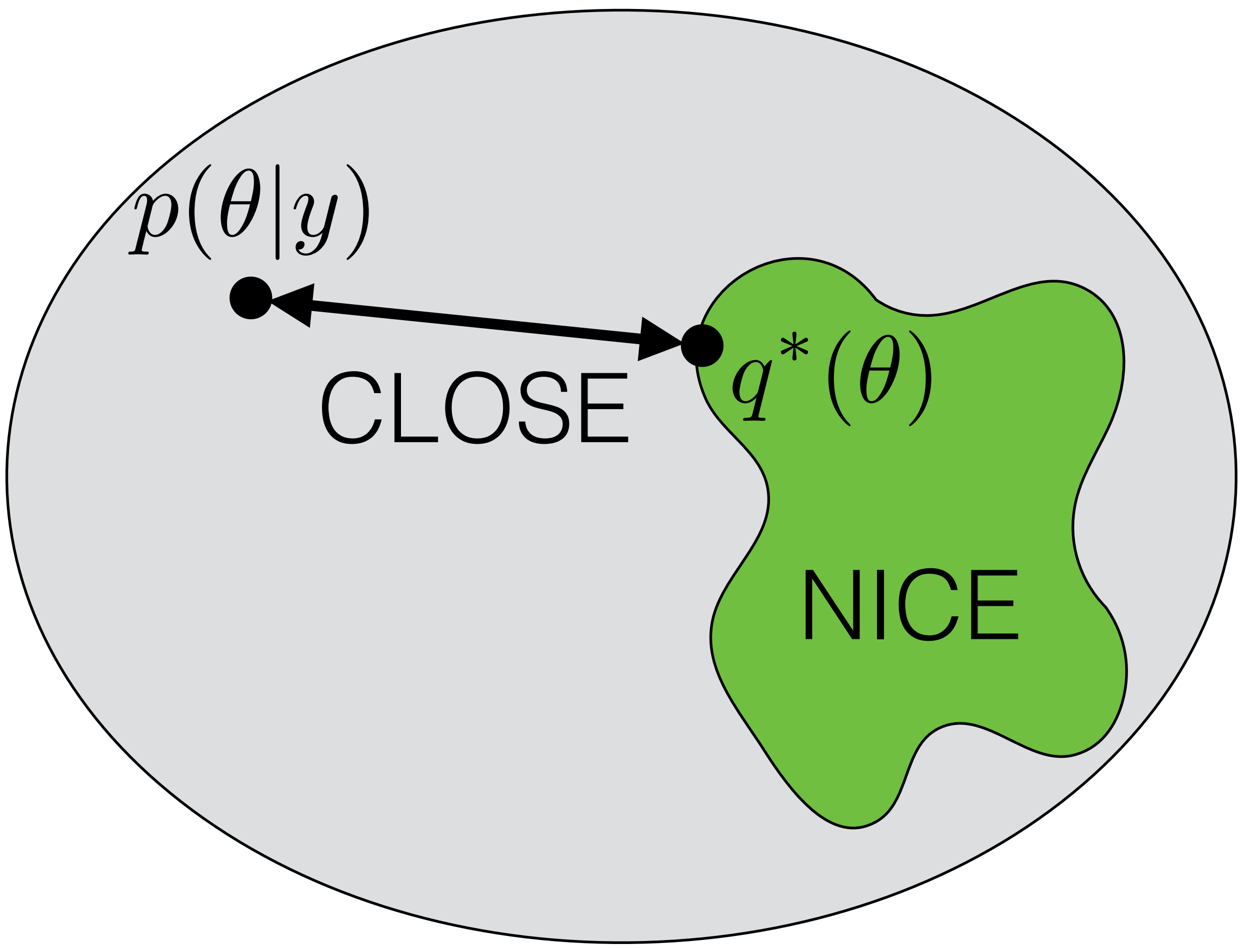
- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

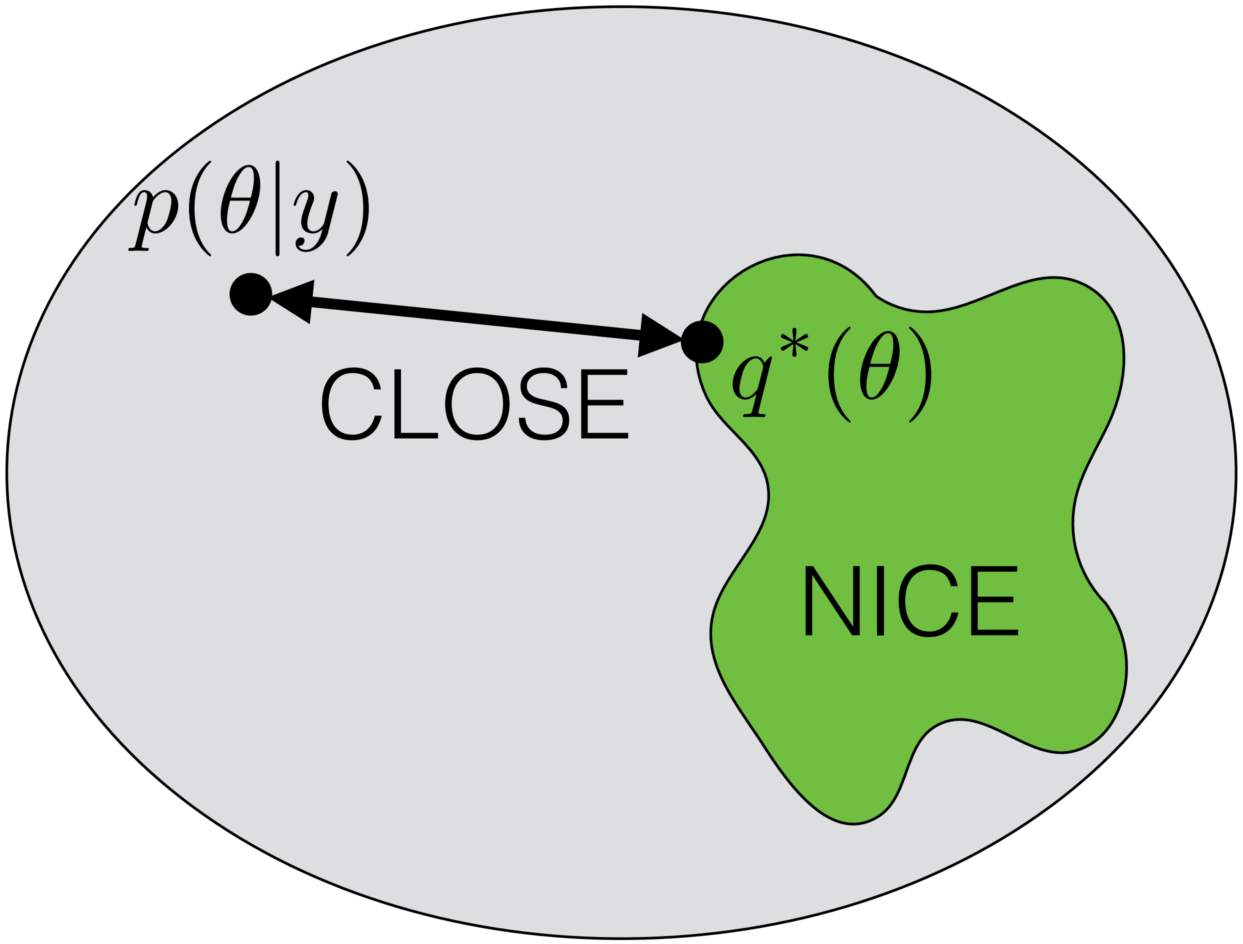
- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

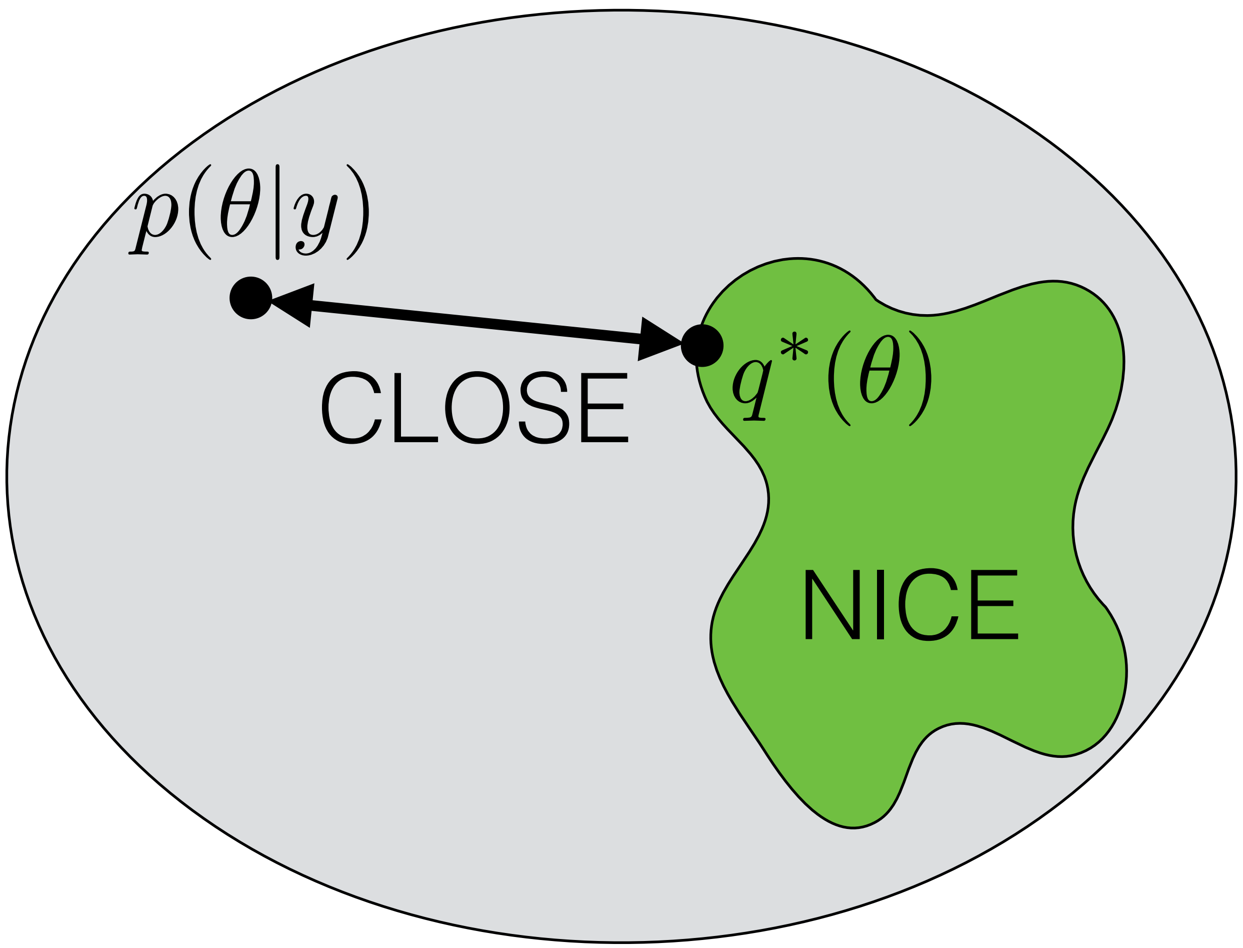
- Approximate posterior with q^*

$$q^* = \mathrm{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

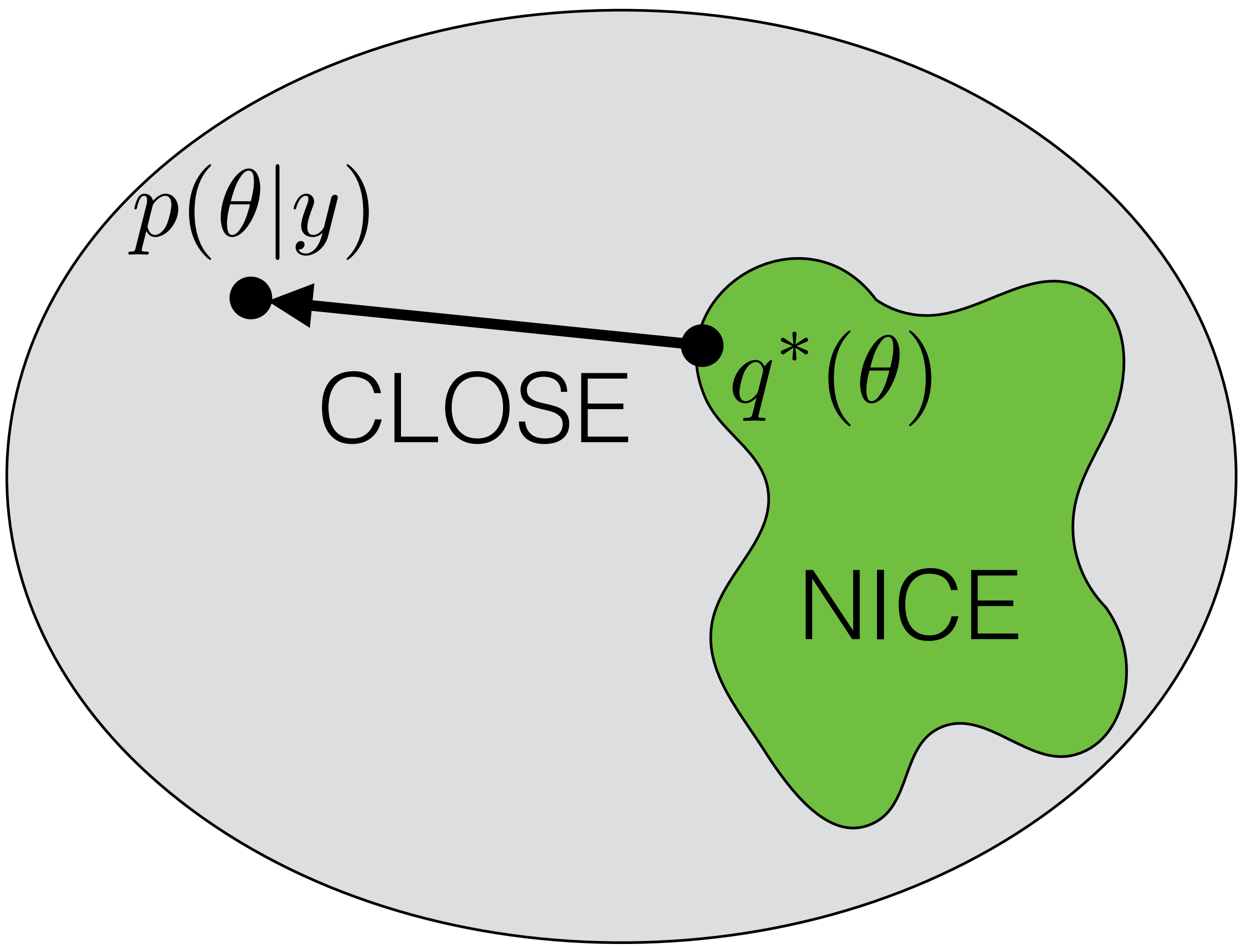
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

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

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \mathrm{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

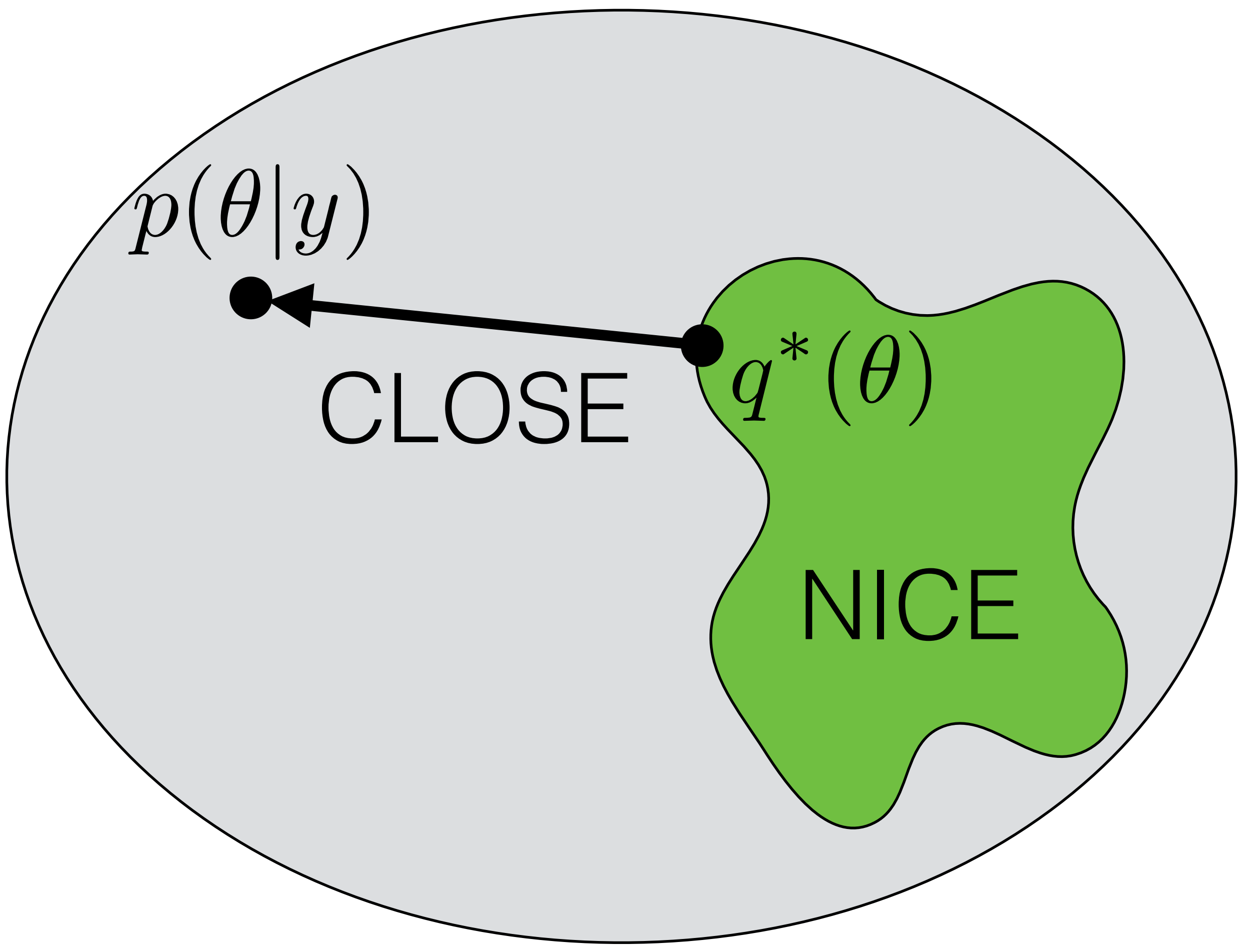
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

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

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

- Variational Bayes (VB): *f* is Kullback-Leibler divergence

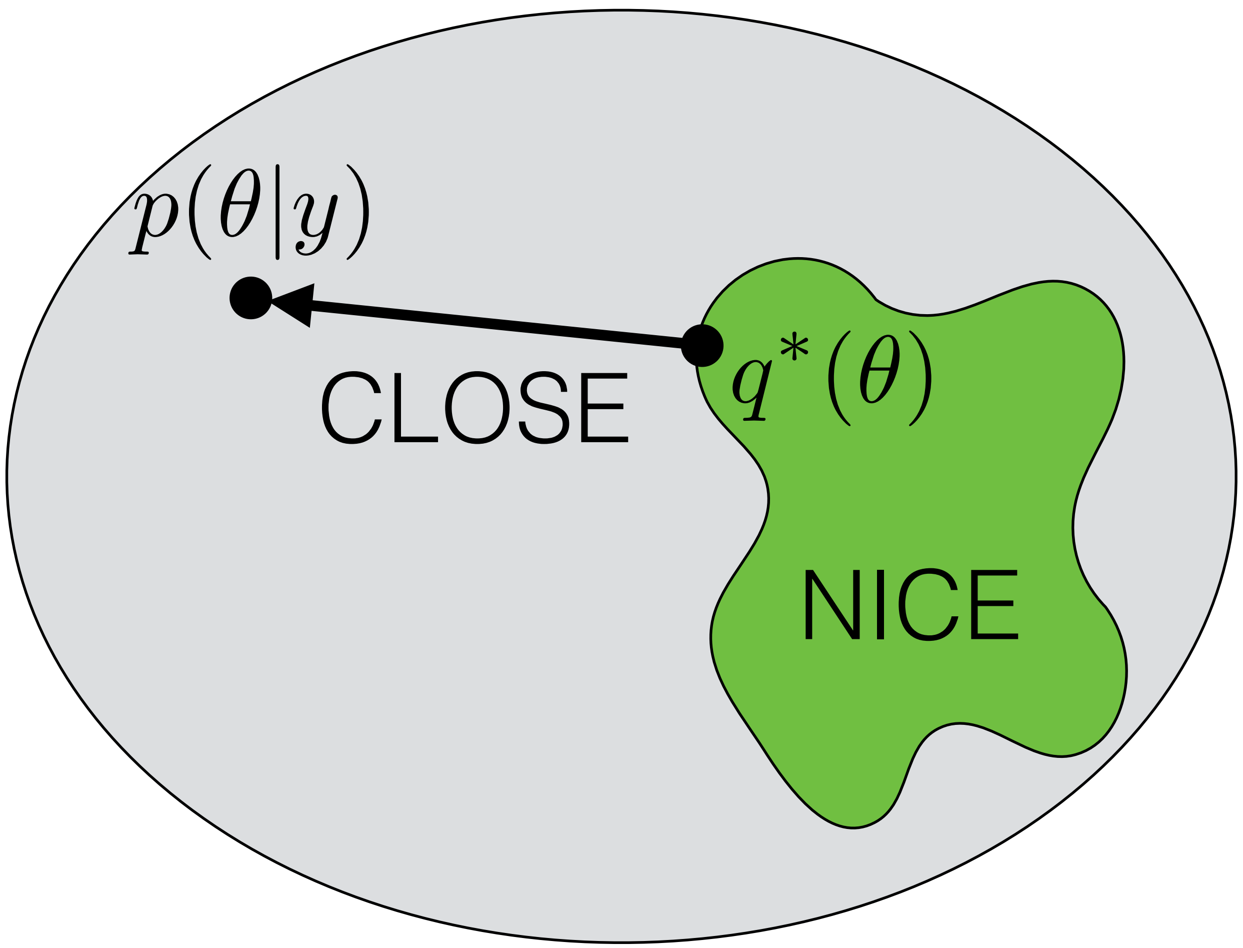
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

- Variational Bayes (VB): *f* is Kullback-Leibler divergence

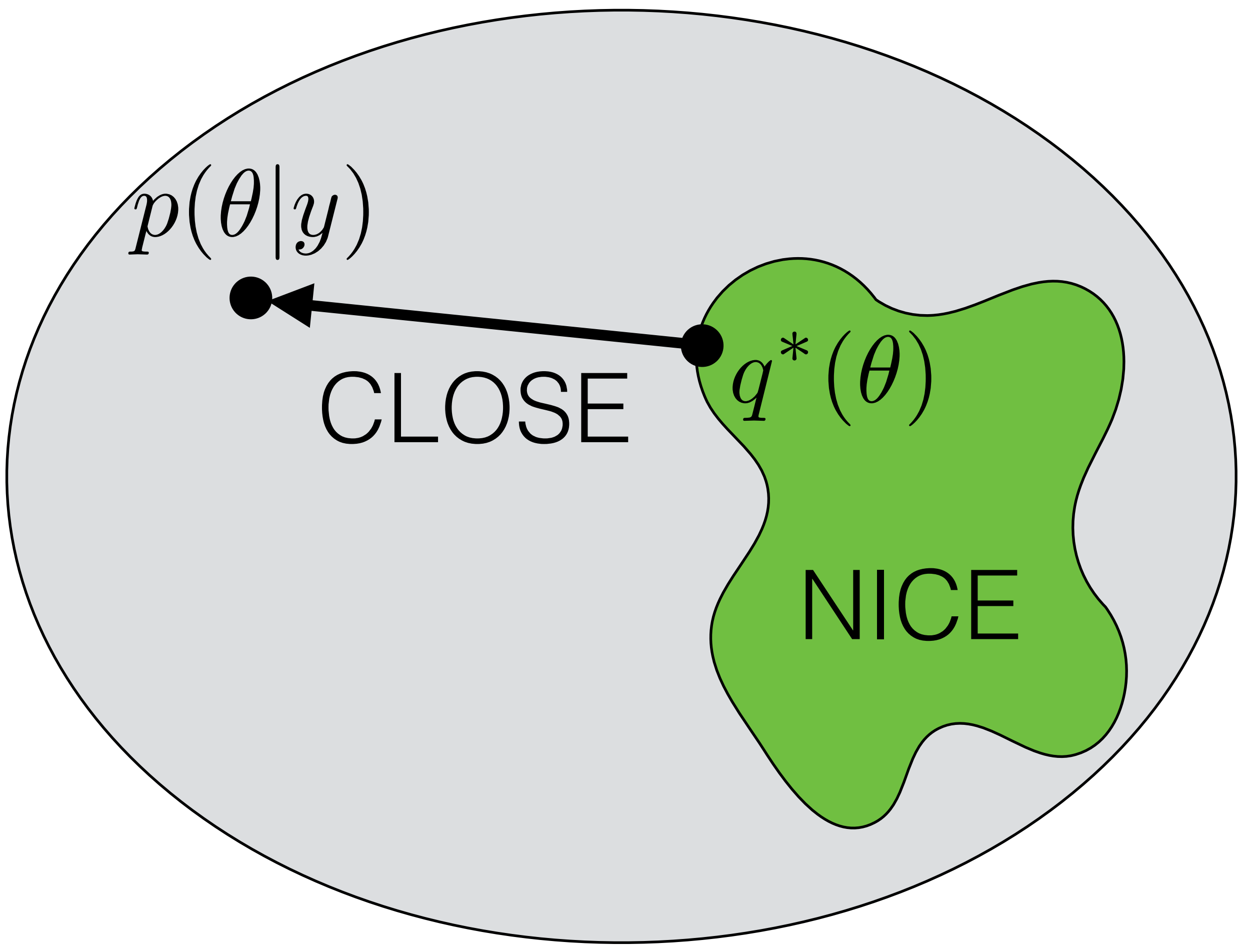
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success: point estimates and prediction

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

- Variational Bayes (VB): *f* is Kullback-Leibler divergence

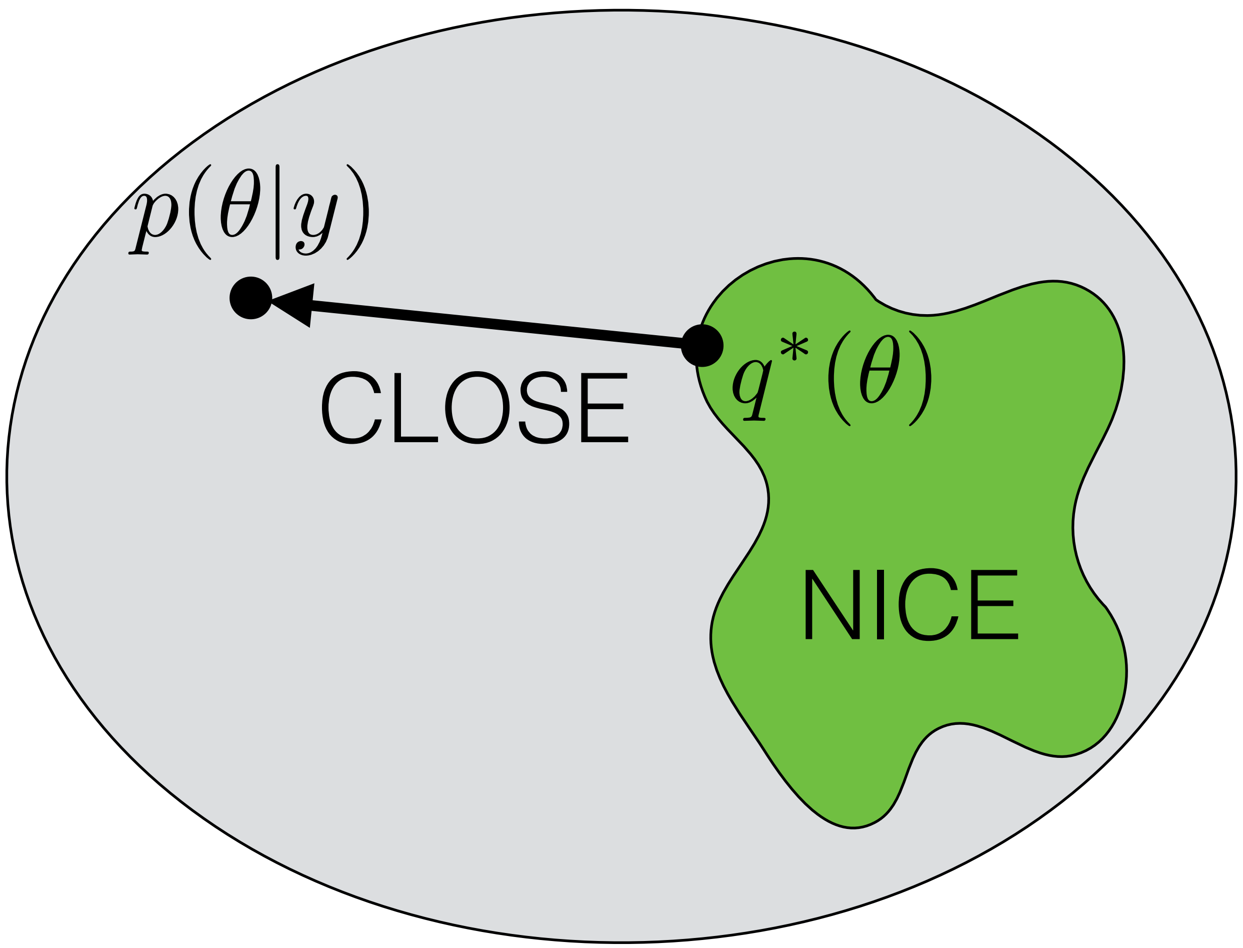
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success: point estimates and prediction, fast

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

- Variational Bayes (VB): *f* is Kullback-Leibler divergence

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

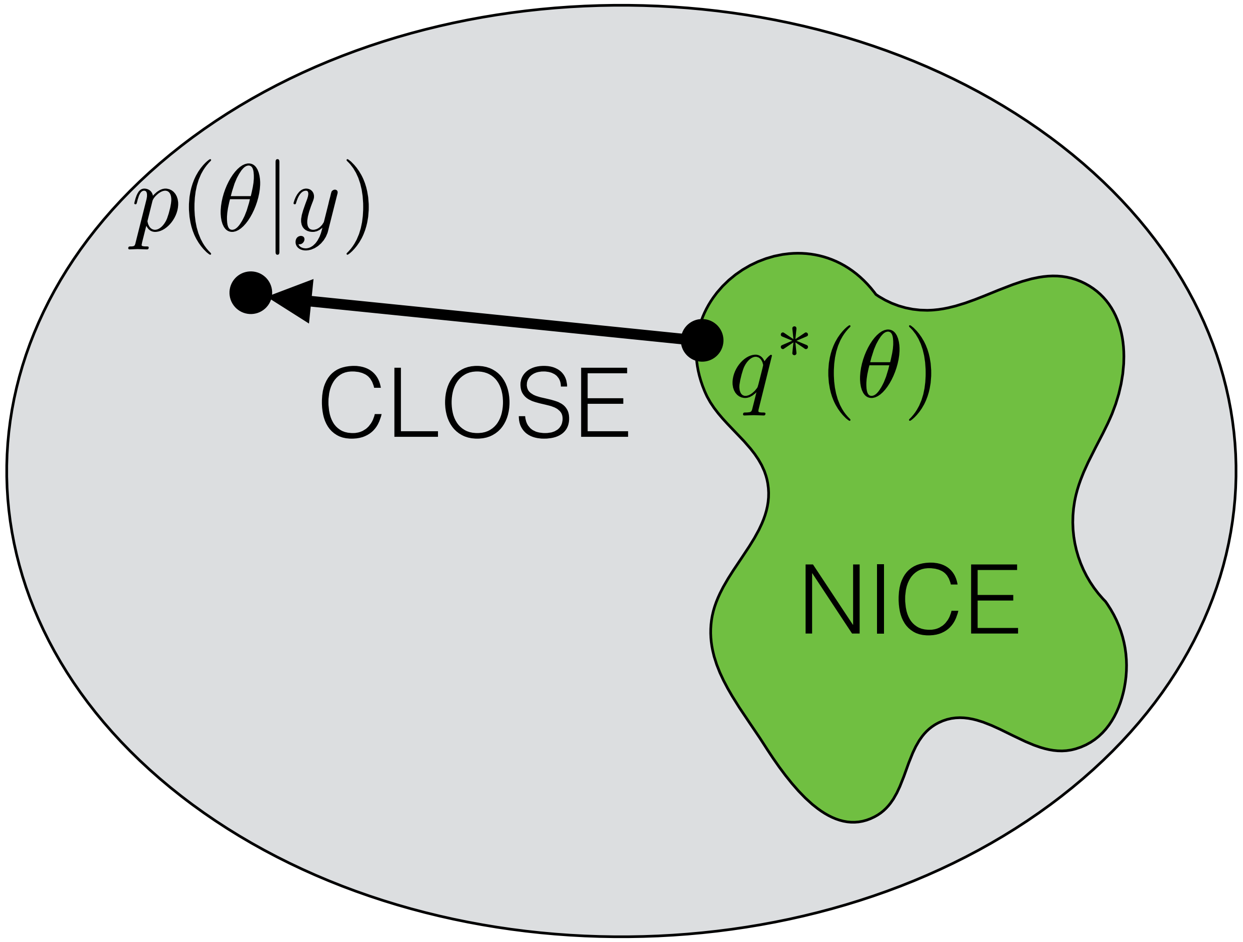
- VB practical success: point estimates and prediction, fast, streaming, distributed (3.6M Wikipedia, 350K Nature)

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

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

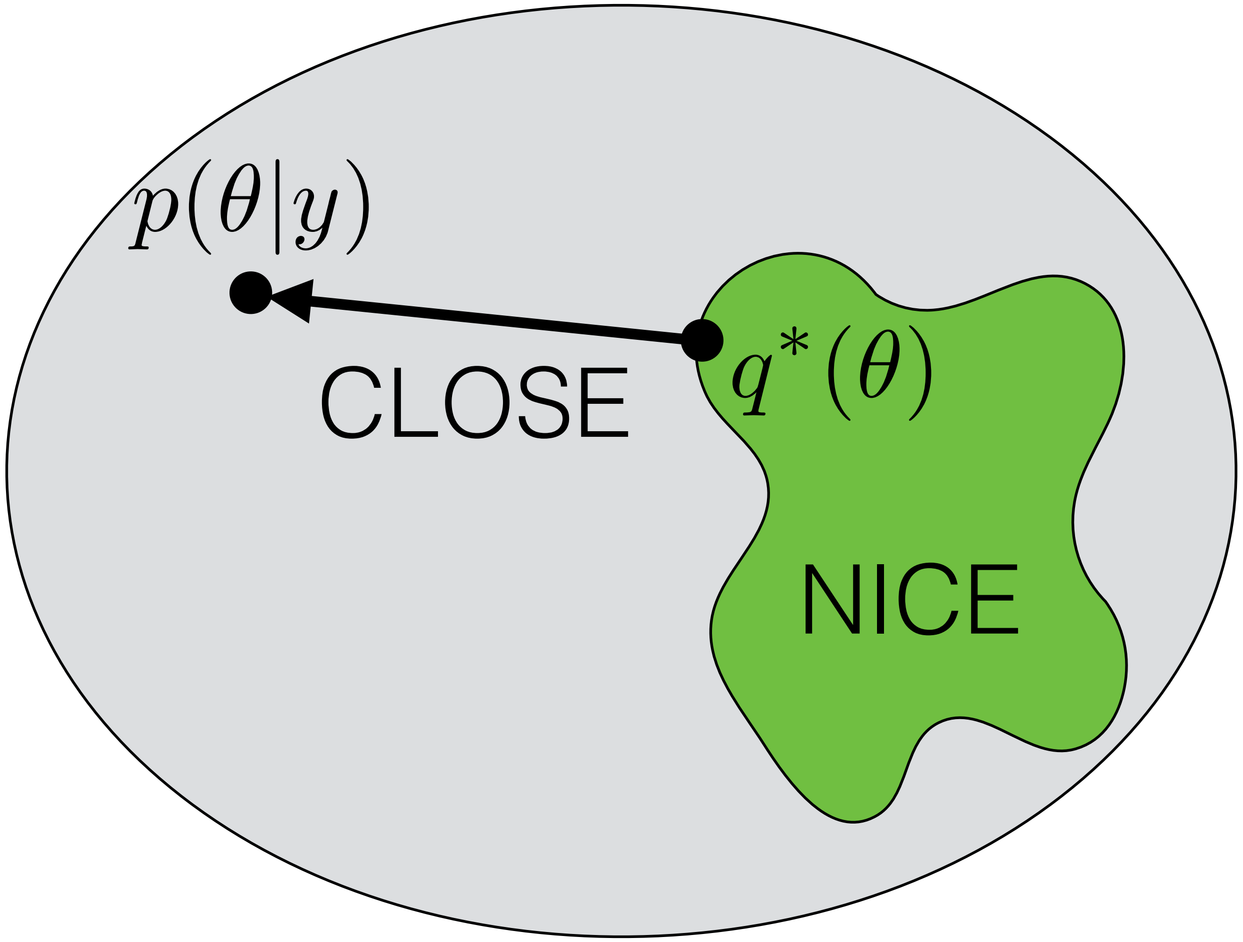


Why KL?

- Variational Bayes

$$q^* = \mathrm{argmin}_{q \in Q} \mathrm{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

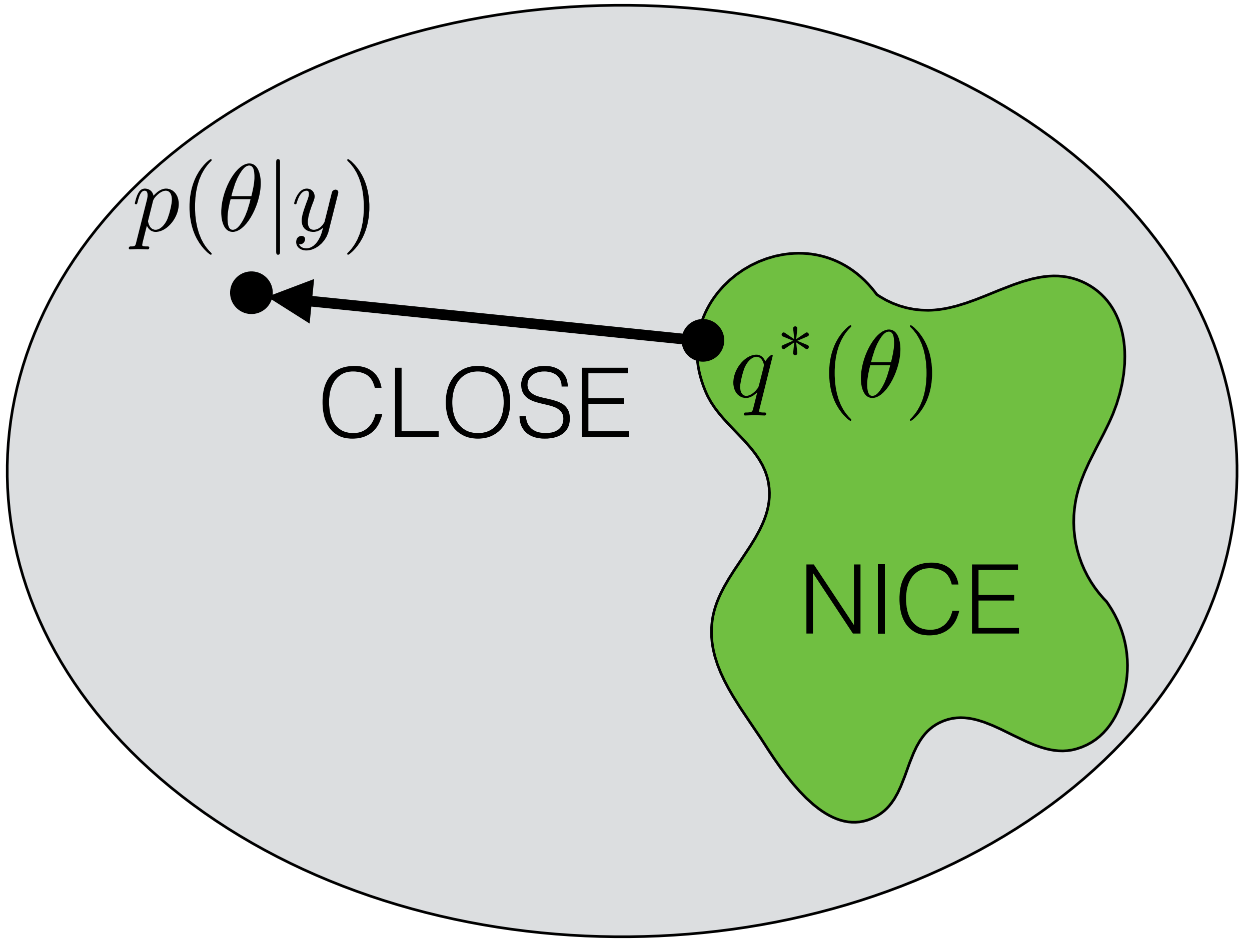
$$\mathrm{KL}\left(q(\cdot) || p(\cdot|y)\right)$$
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$



Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



$$\text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

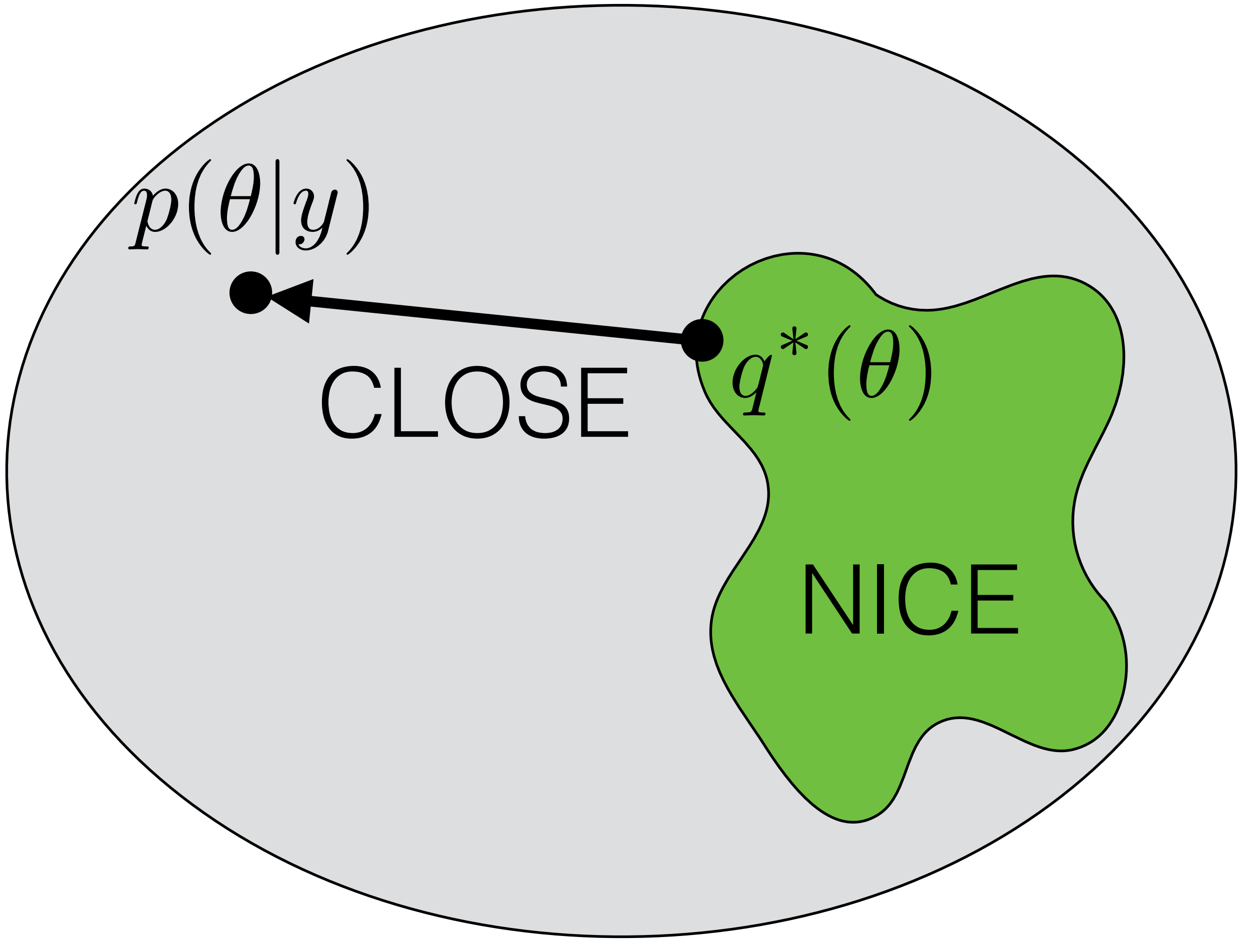
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta$$

Why KL?

- Variational Bayes

$$q^* = \mathrm{argmin}_{q \in Q} \mathrm{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



$$\mathrm{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

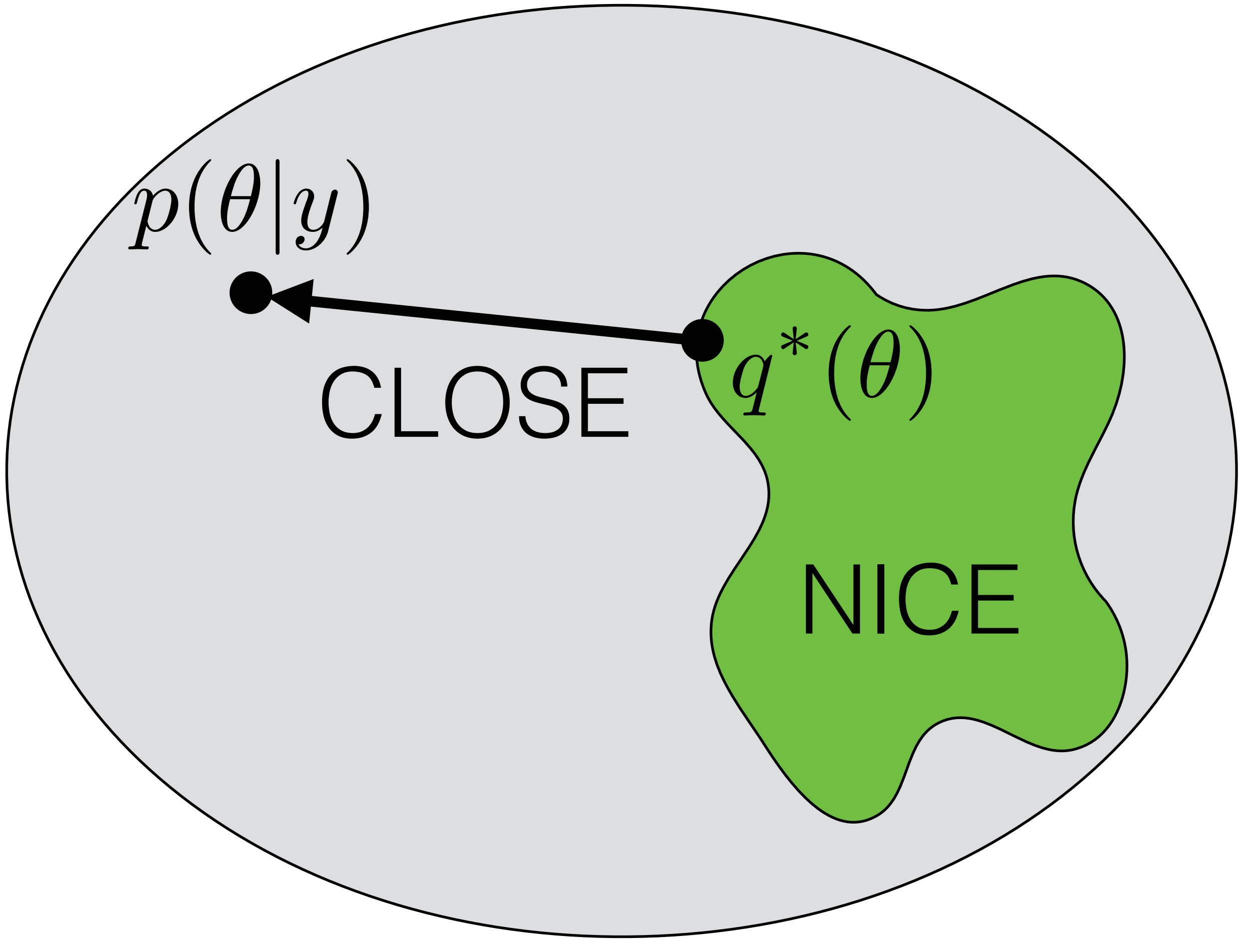
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot)||p(\cdot|y)\right)$$



$$\text{KL}\left(q(\cdot)||p(\cdot|y)\right)$$

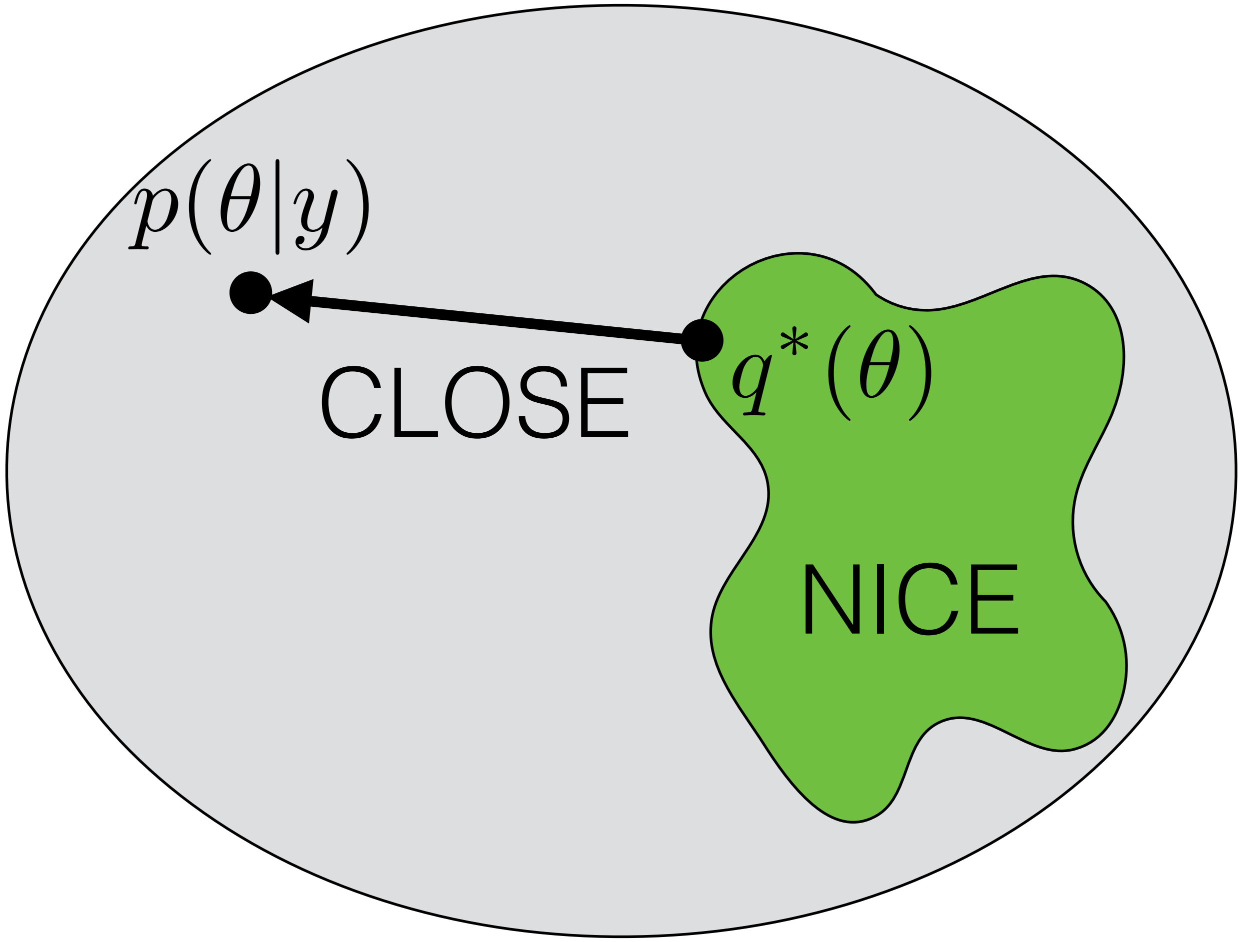
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



$$\text{KL}\left(q(\cdot)||p(\cdot|y)\right)$$

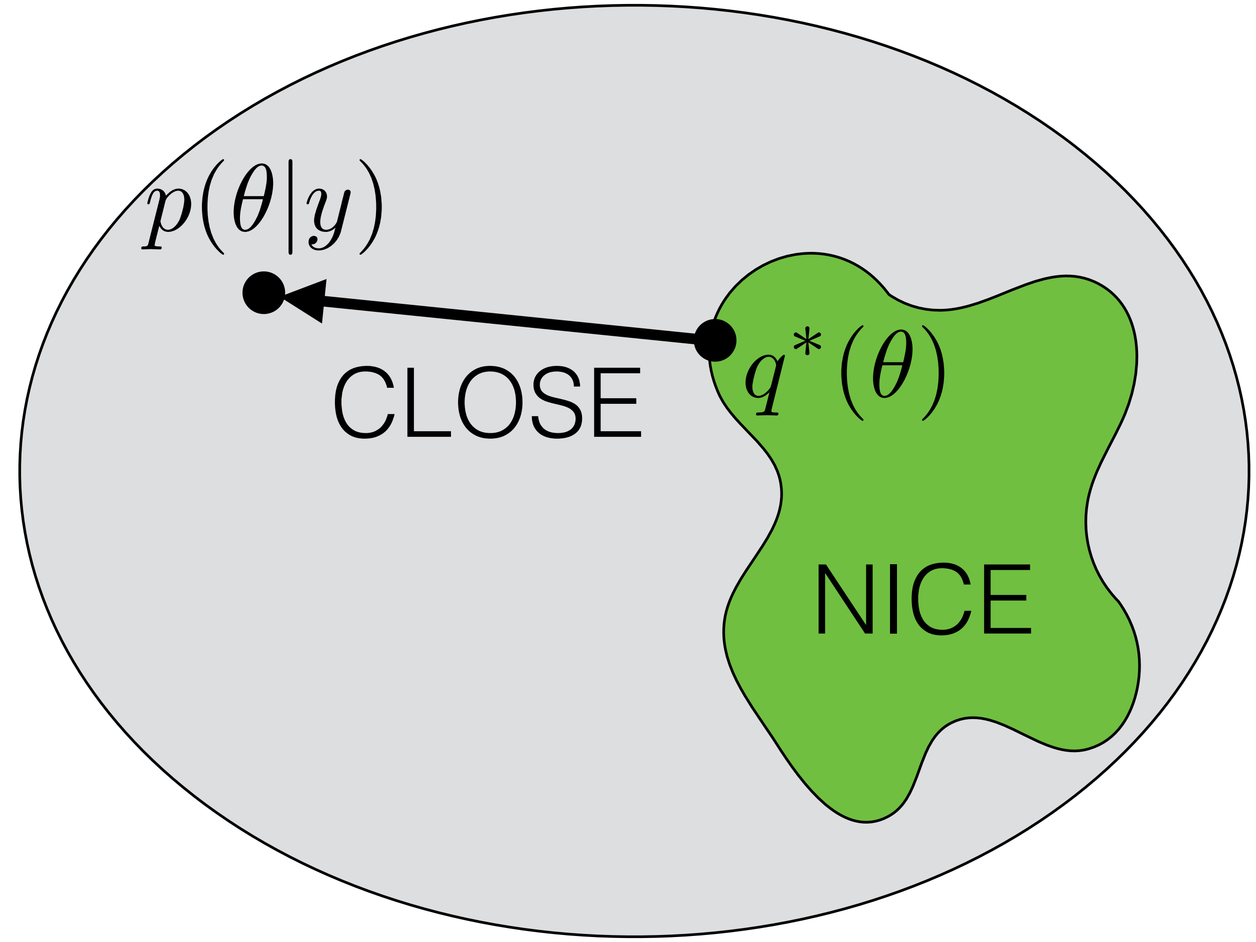
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



$$\text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

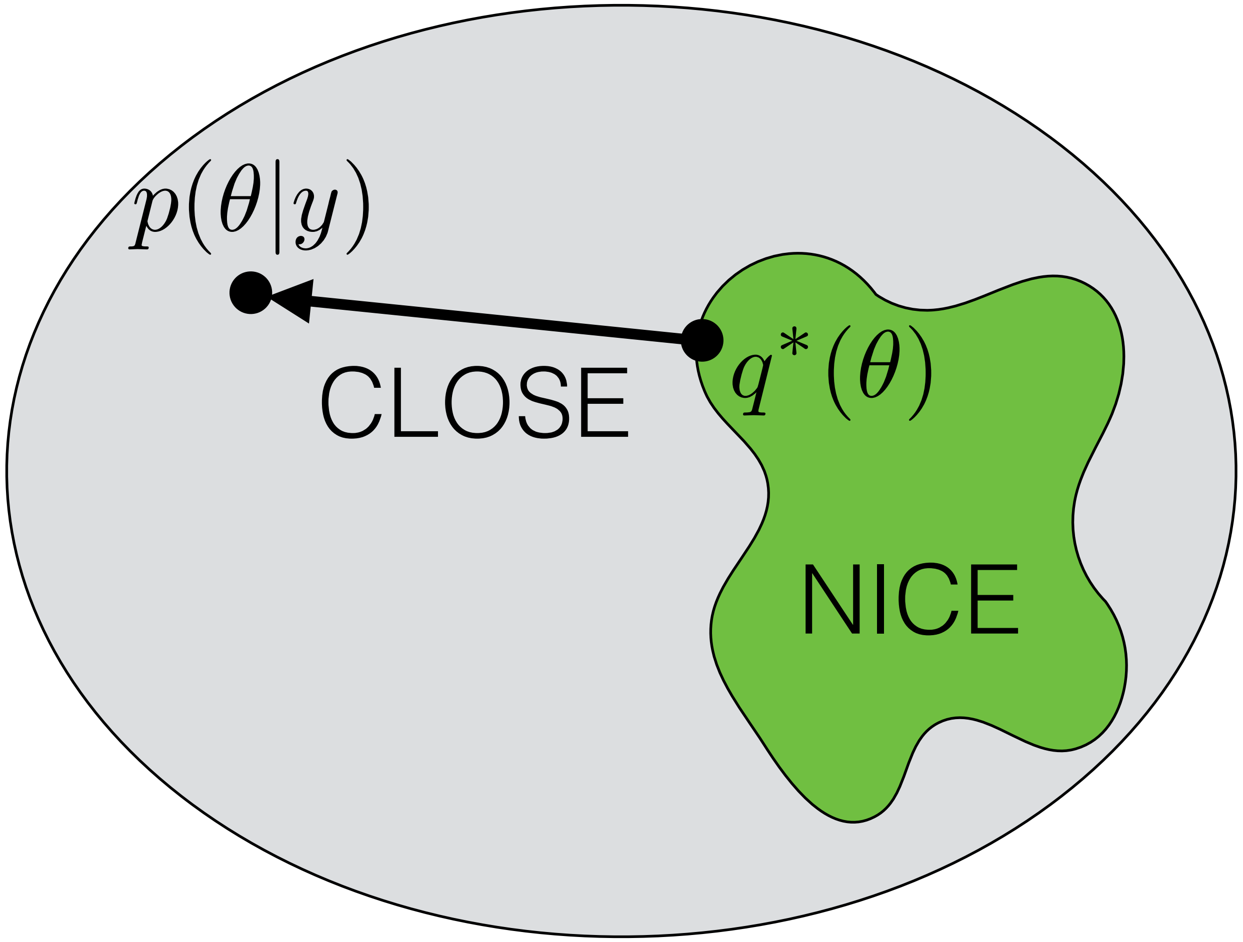
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

"Evidence lower bound" (ELBO)

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



$$\text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

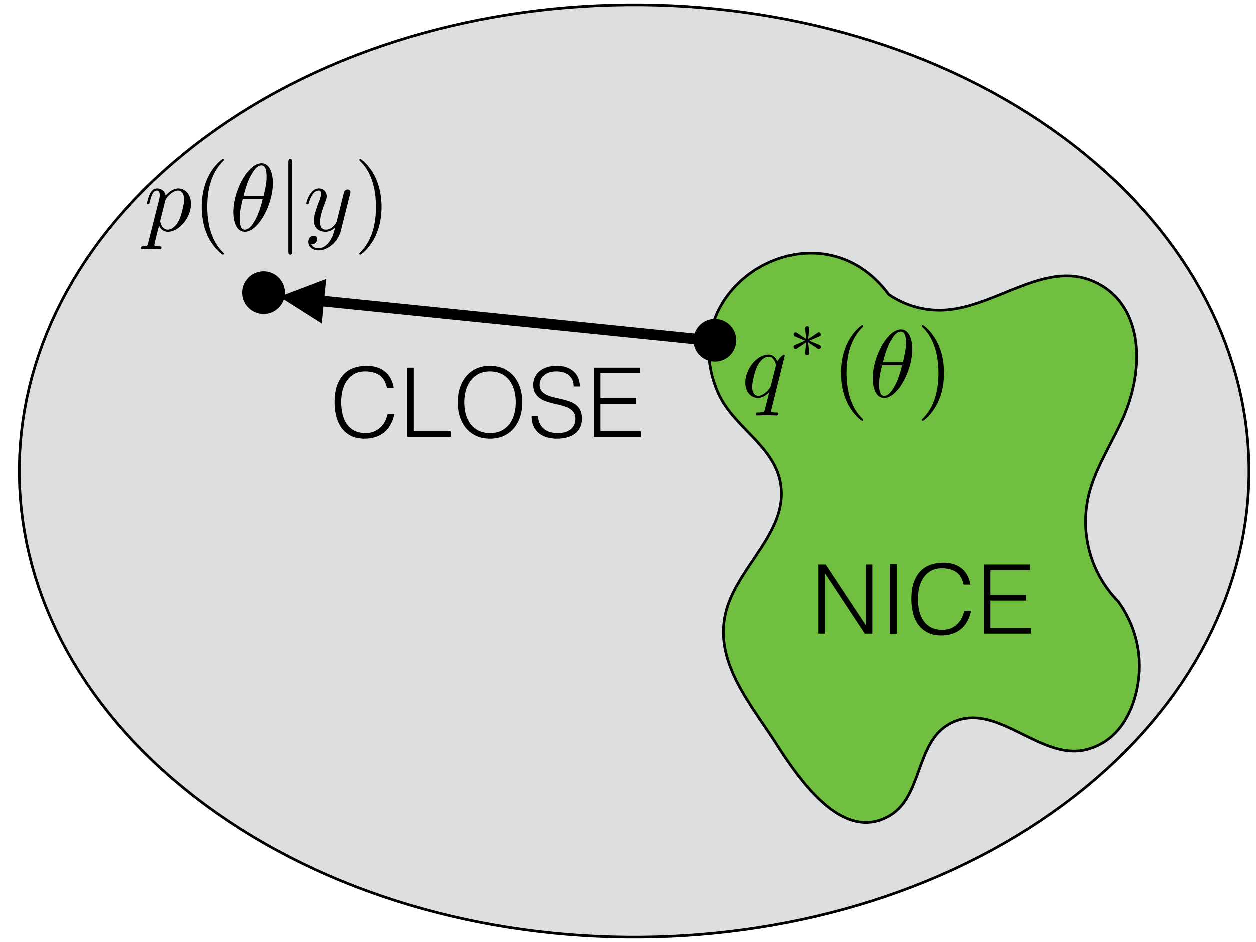
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

"Evidence lower bound" (ELBO)

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot)||p(\cdot|y)\right)$$



$$\text{KL}\left(q(\cdot)||p(\cdot|y)\right)$$

$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

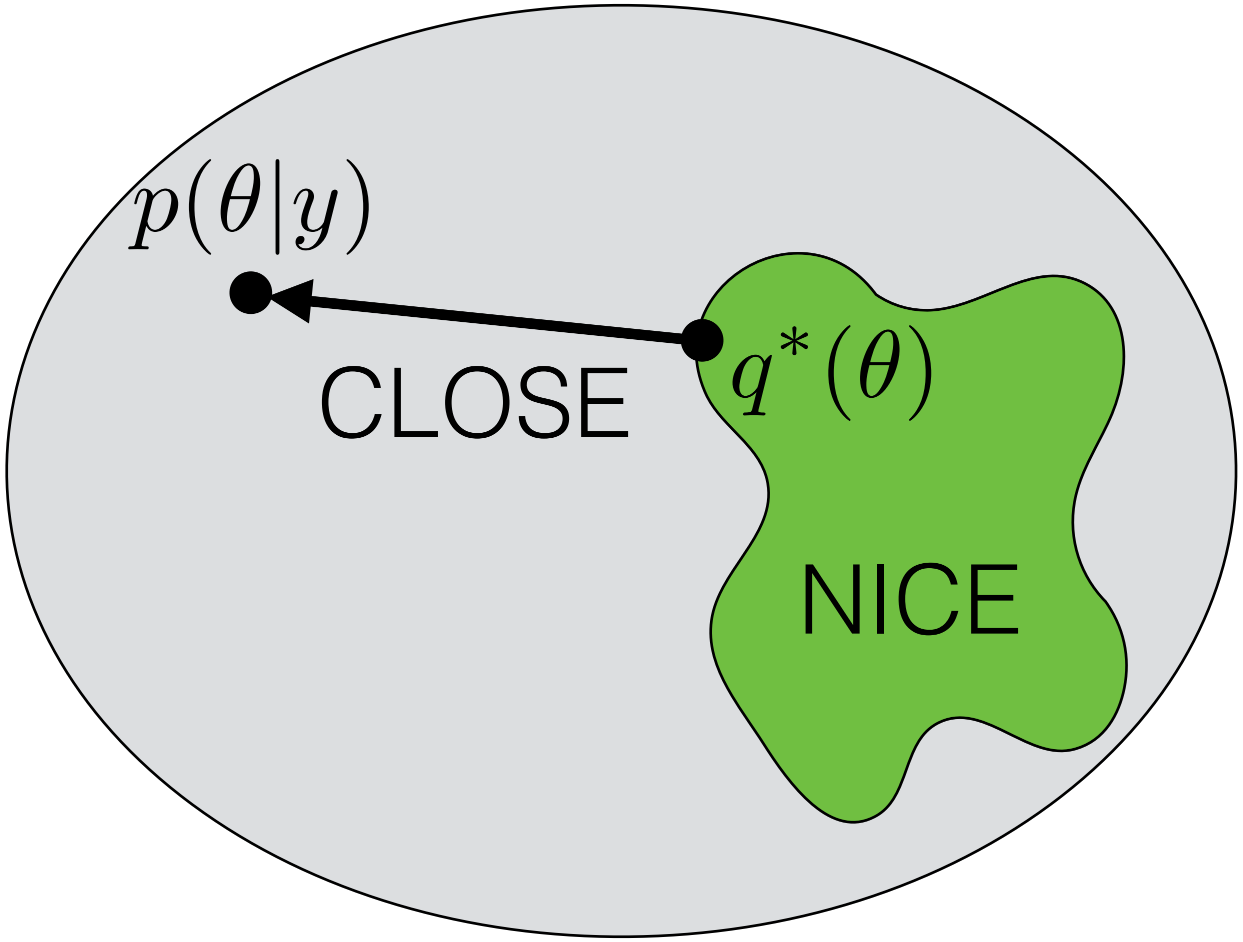
"Evidence lower bound" (ELBO)

- Exercise: Show $\text{KL} \geq 0$ [Bishop 2006, Sec 1.6.1]

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \mathrm{KL}\left(q(\cdot) || p(\cdot | y)\right)$$



$$\mathrm{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

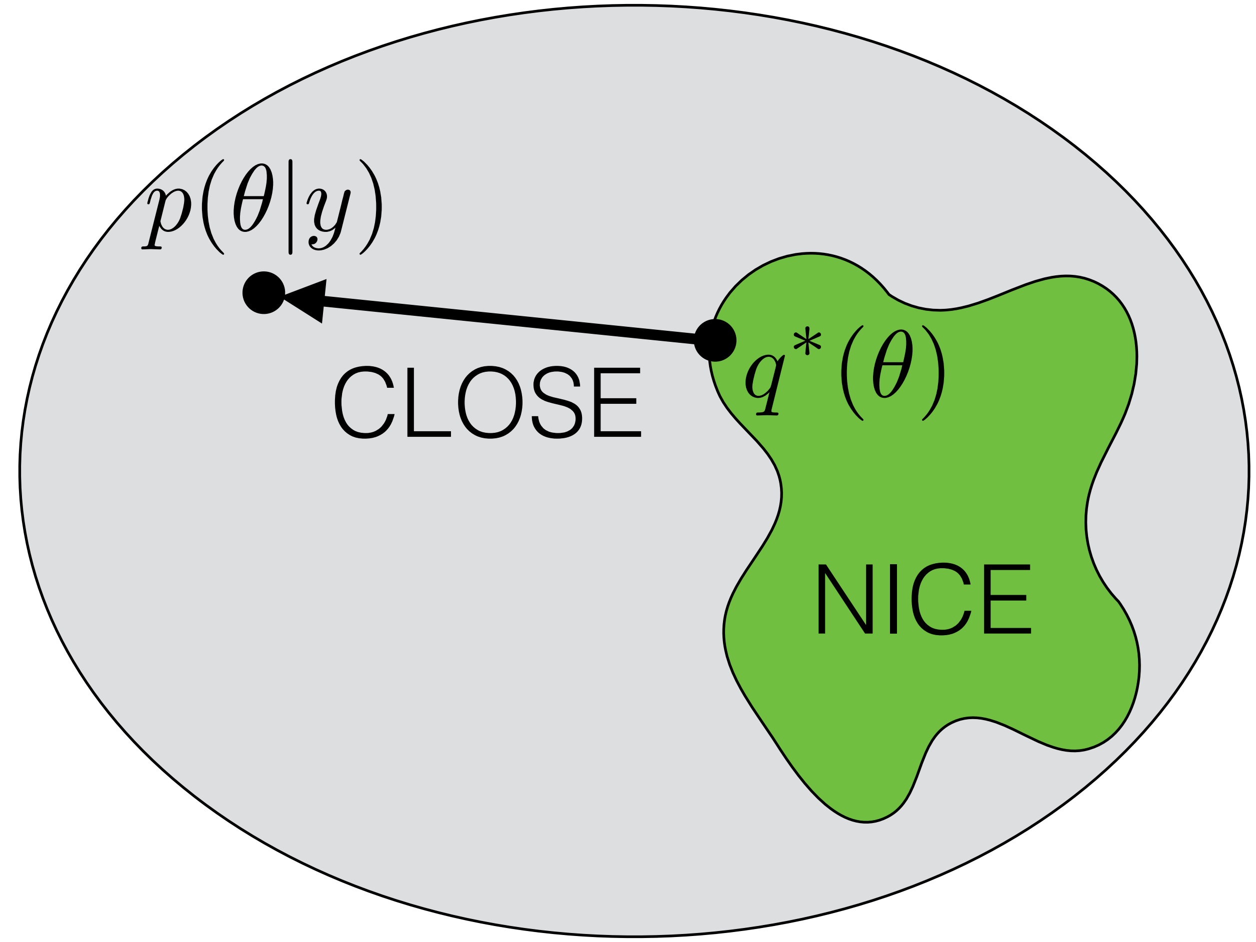
"Evidence lower bound" (ELBO)

- Exercise: Show $\mathrm{KL} \geq 0$ [Bishop 2006, Sec 1.6.1]
- $\mathrm{KL} \geq 0 \Rightarrow \log p(y) \geq \mathrm{ELBO}$

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$



$$\text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

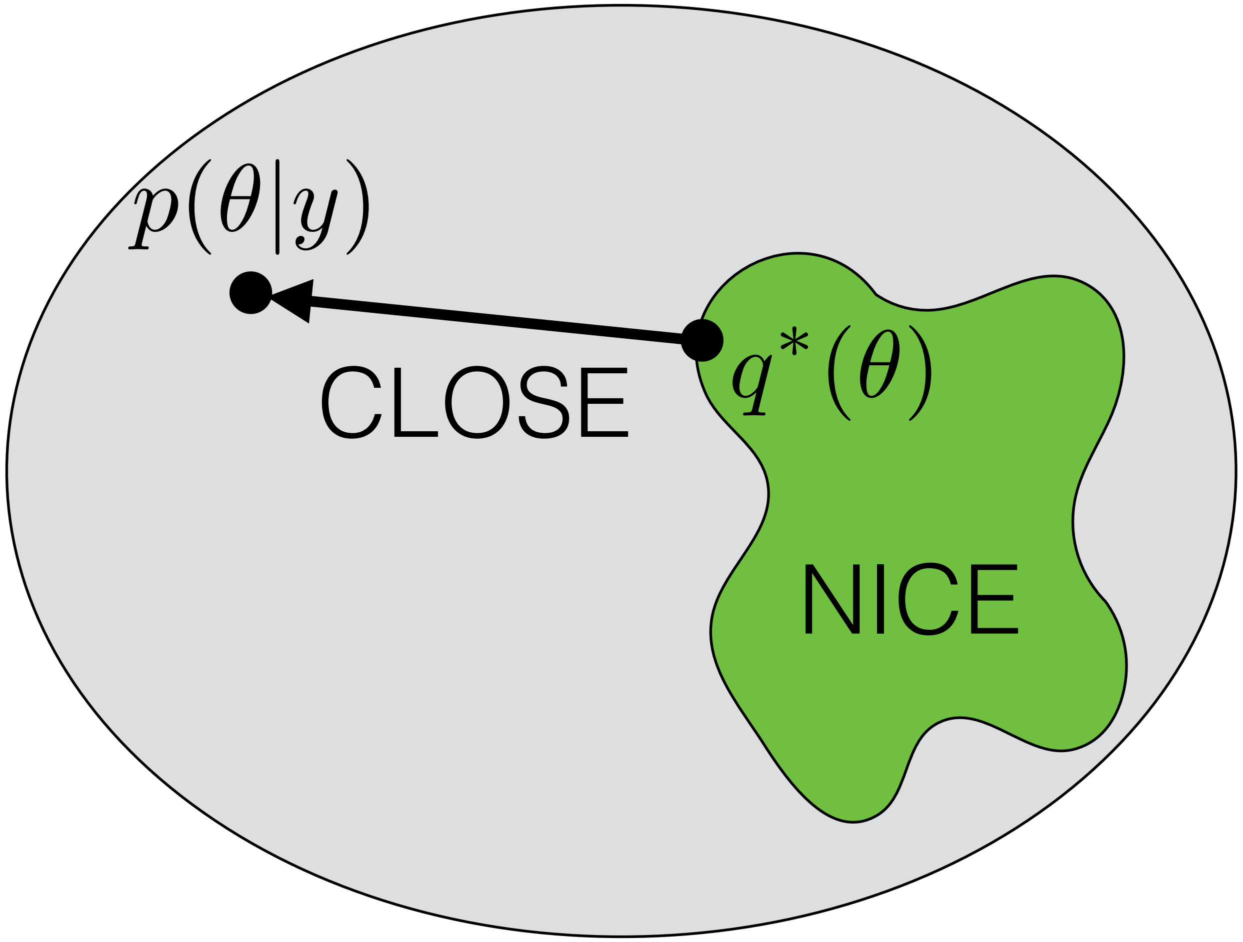
"Evidence lower bound" (ELBO)

- Exercise: Show $\text{KL} \geq 0$ [Bishop 2006, Sec 1.6.1]
- $\text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO}$
- $q^* = \text{argmax}_{q \in Q} \text{ELBO}(q)$

Why KL?

- Variational Bayes

$$q^* = \operatorname{argmin}_{q \in Q} \mathrm{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



$$\mathrm{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

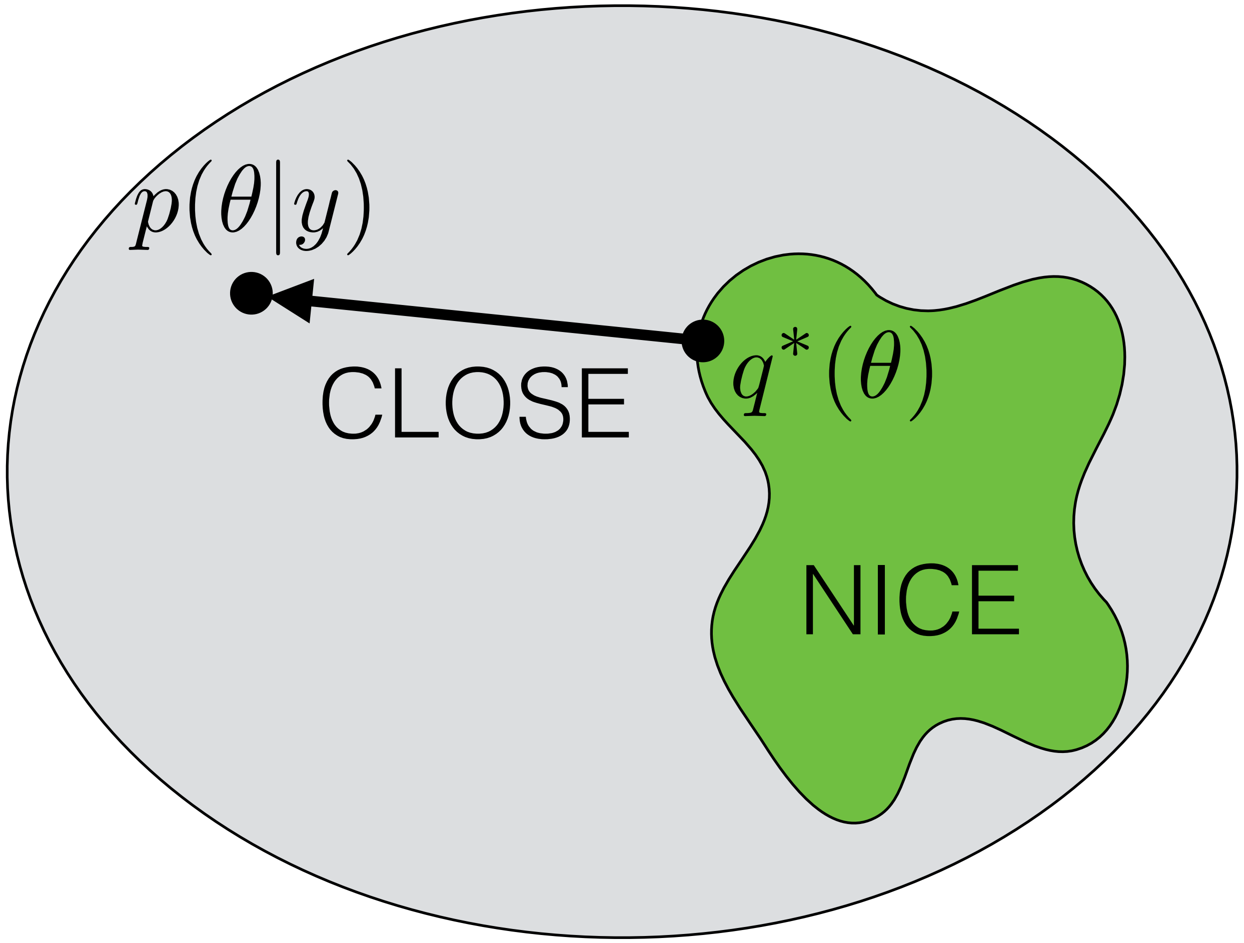
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

“Evidence lower bound” (ELBO)

- Exercise: Show $\mathrm{KL} \geq 0$ [Bishop 2006, Sec 1.6.1]
- $\mathrm{KL} \geq 0 \Rightarrow \log p(y) \geq \mathrm{ELBO}$
- $q^* = \operatorname{argmax}_{q \in Q} \mathrm{ELBO}(q)$
- Why KL (in this direction)?

Variational Bayes

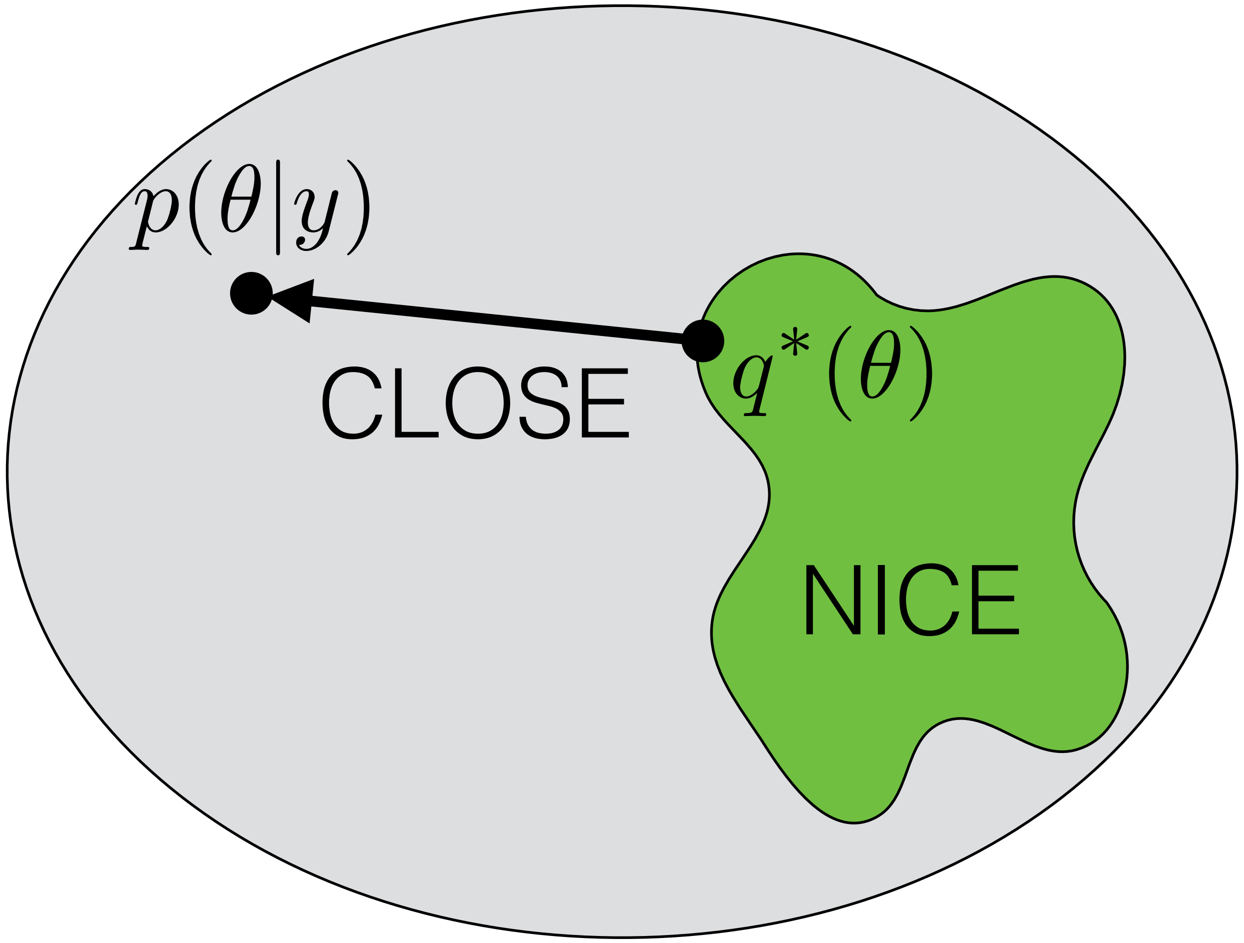
$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$



Variational Bayes

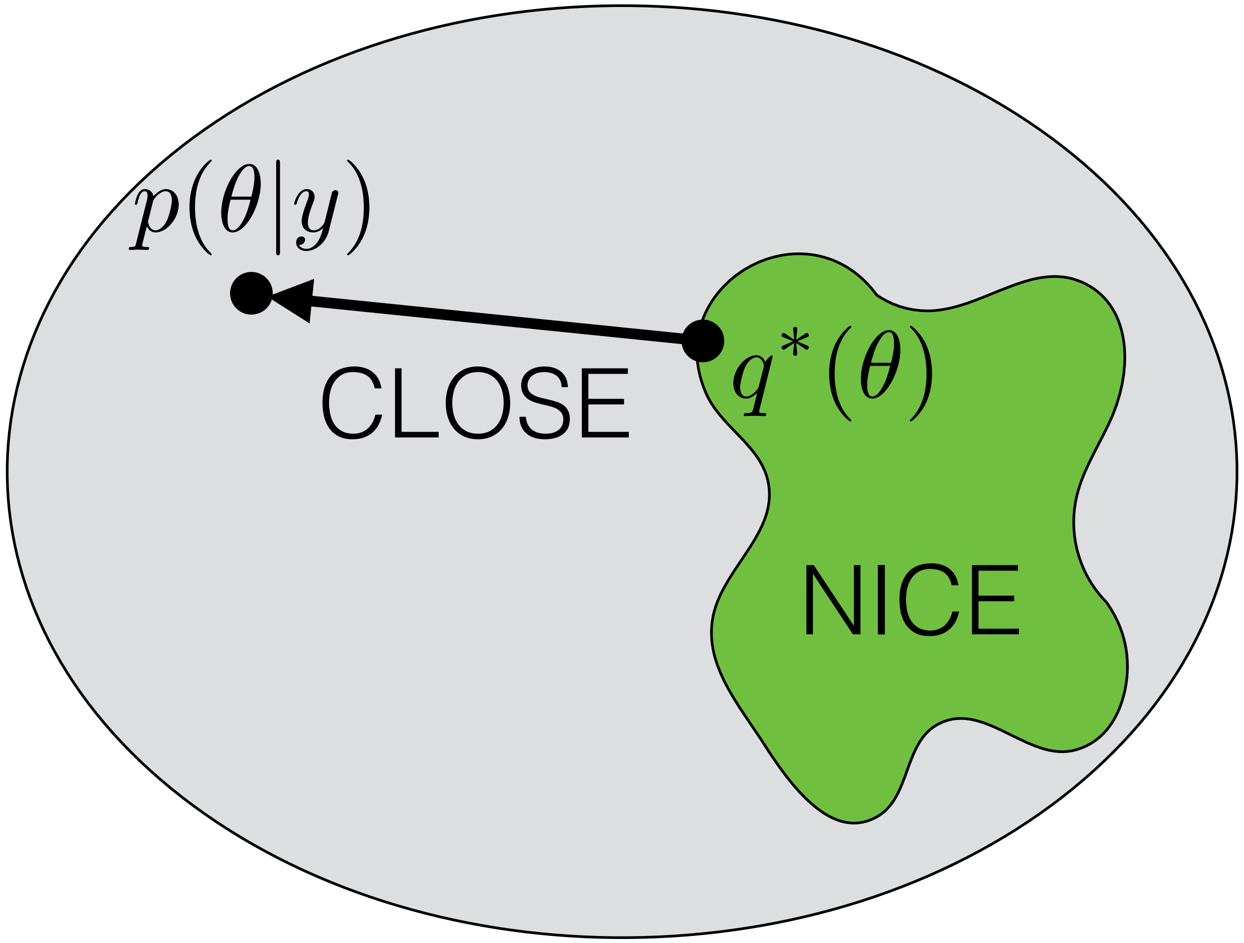
$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

Choose "NICE" distributions



Variational Bayes

$$q^* = \operatorname{argmin}_{q \in Q} \mathrm{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

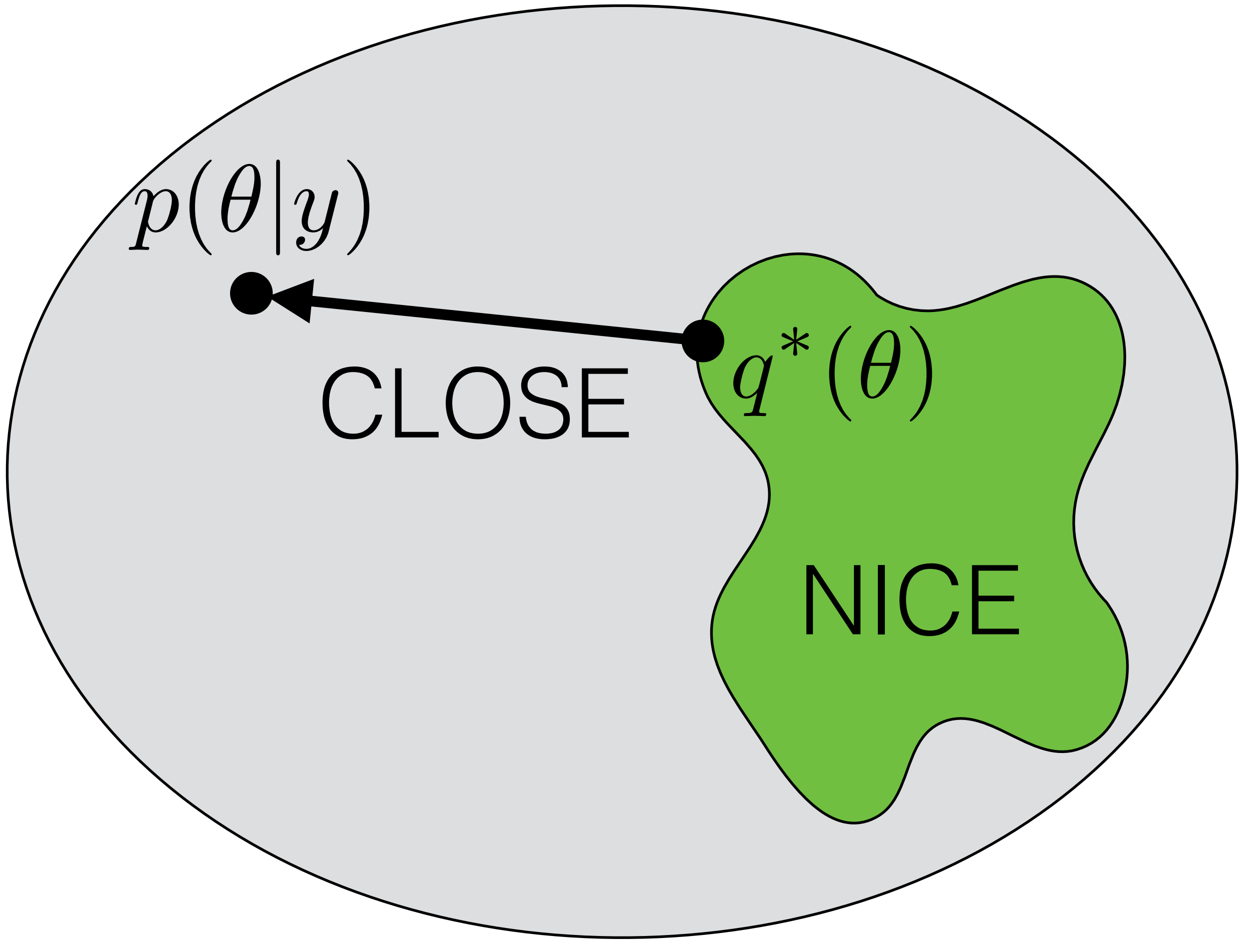


Choose "NICE" distributions

Variational Bayes

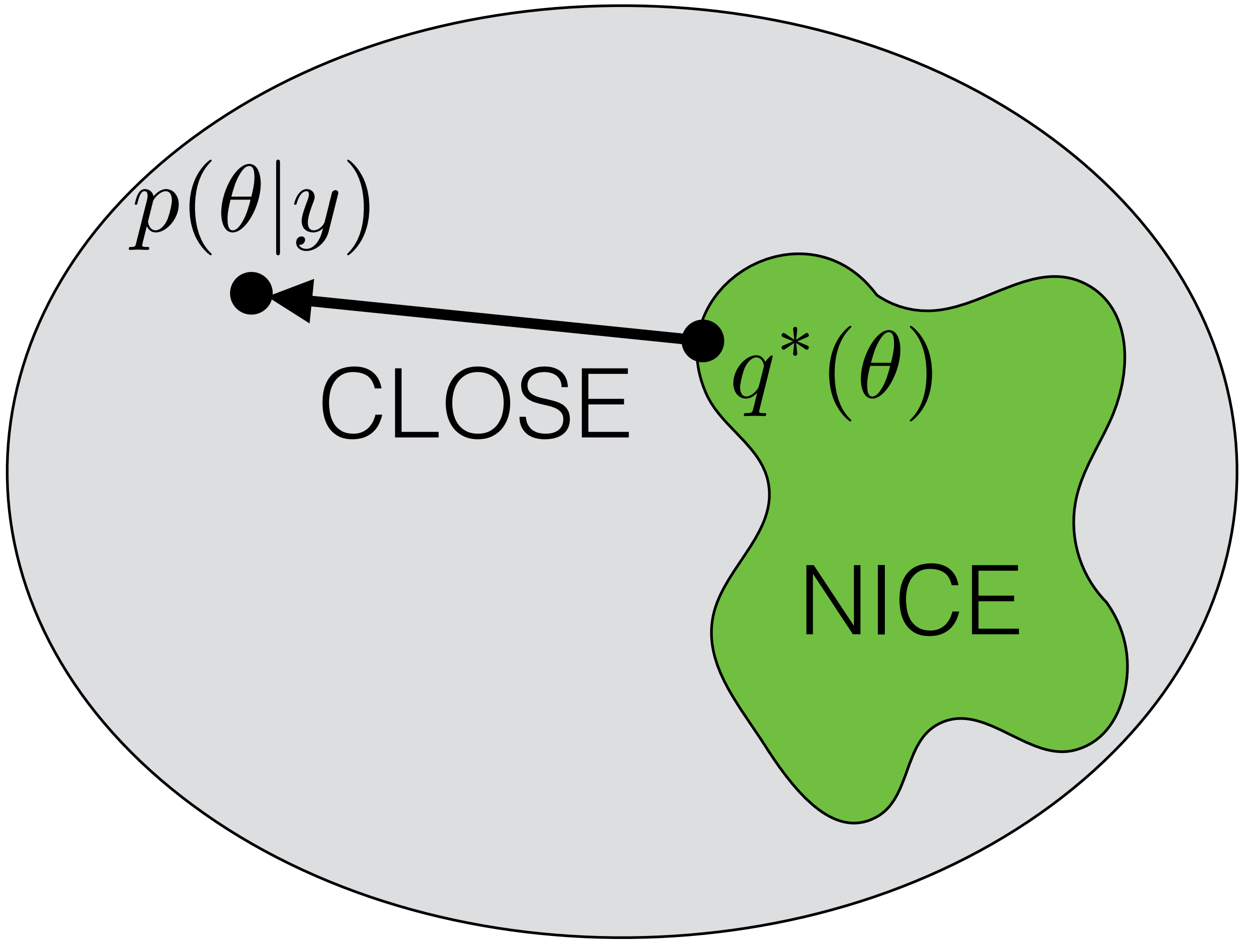
$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

Choose “NICE” distributions



Variational Bayes

$$q^* = \mathrm{argmin}_{q \in Q} \mathrm{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



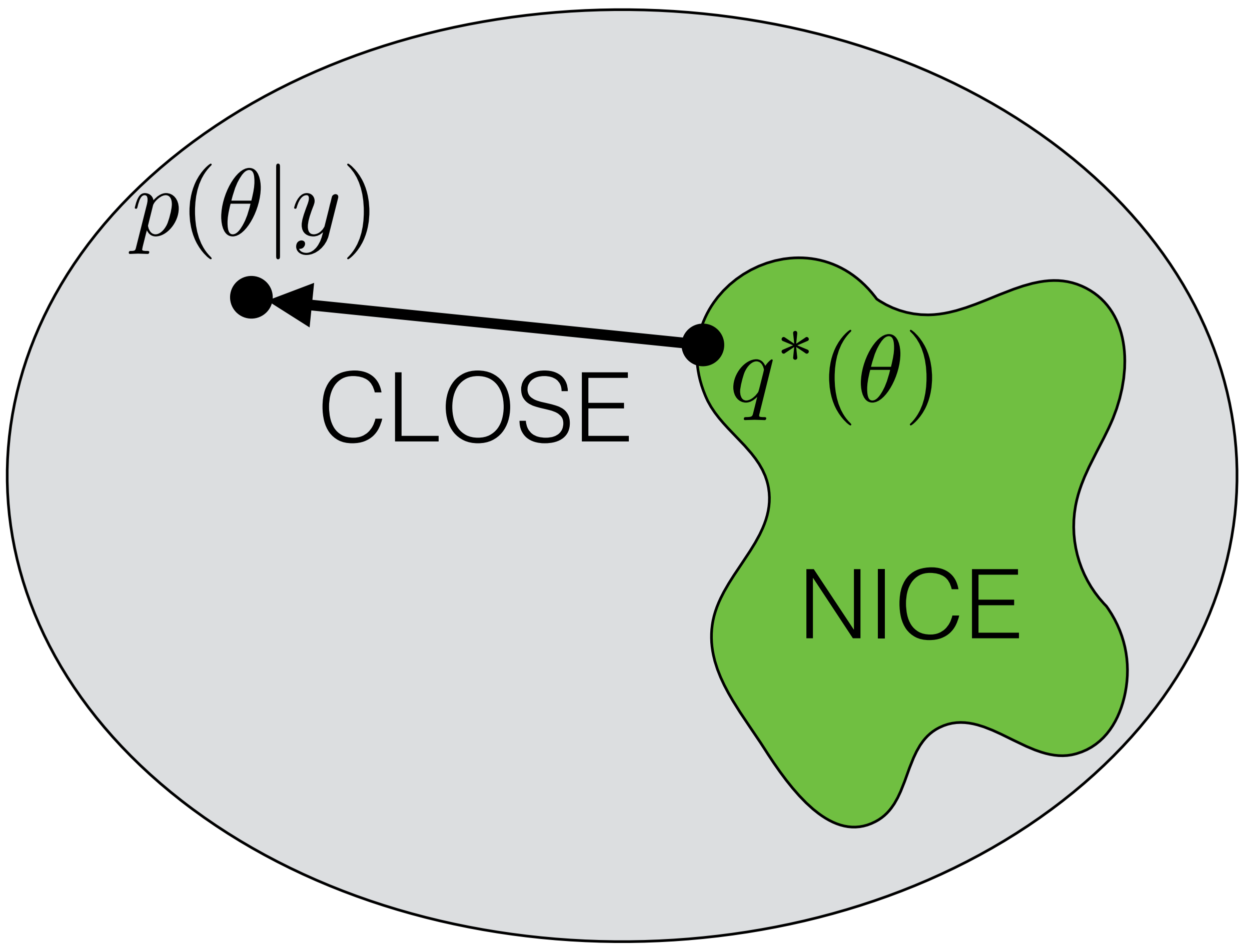
Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

Variational Bayes

$$q^* = \mathrm{argmin}_{q \in Q} \mathrm{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



Choose "NICE" distributions

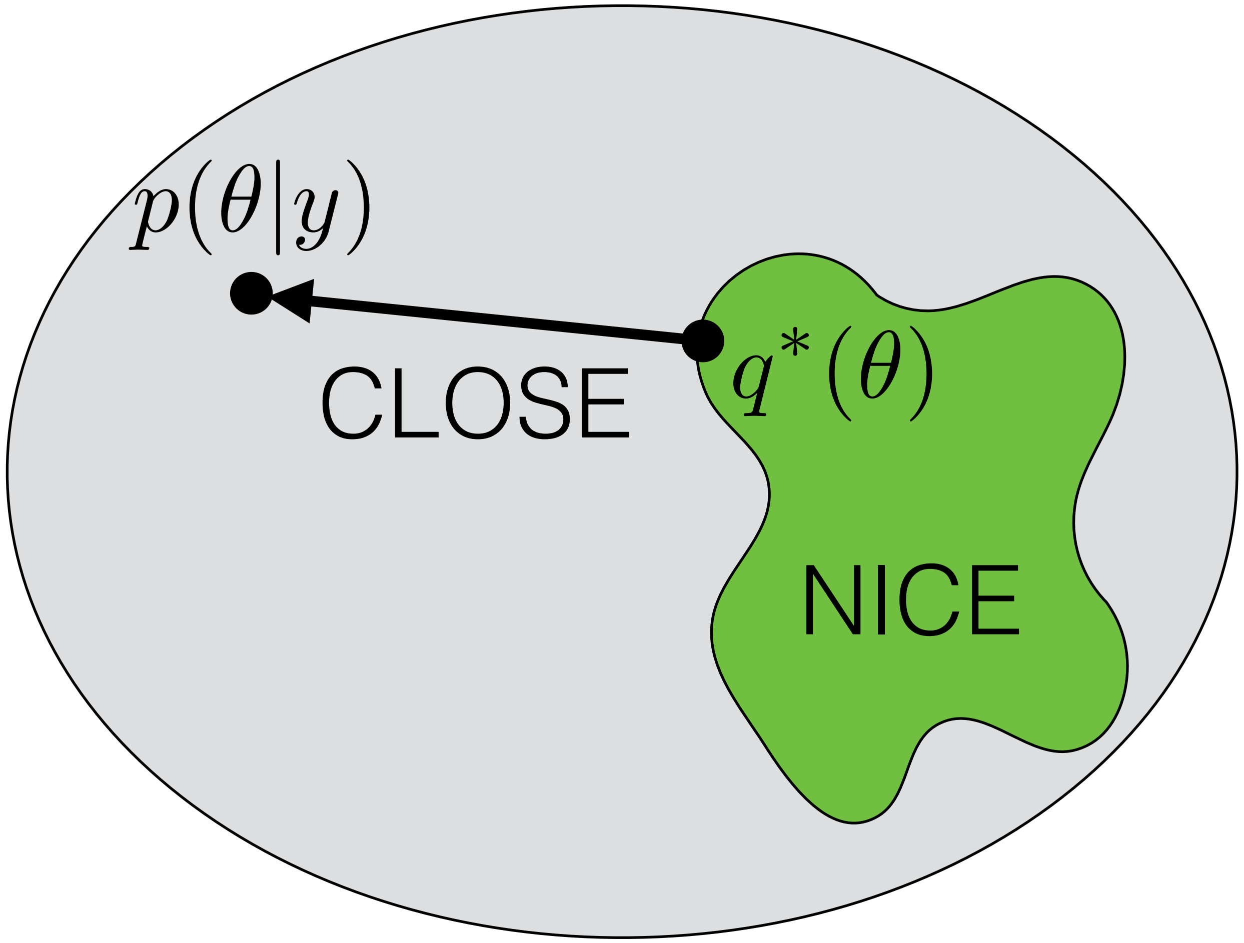
- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

- *Not* a modeling assumption

Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

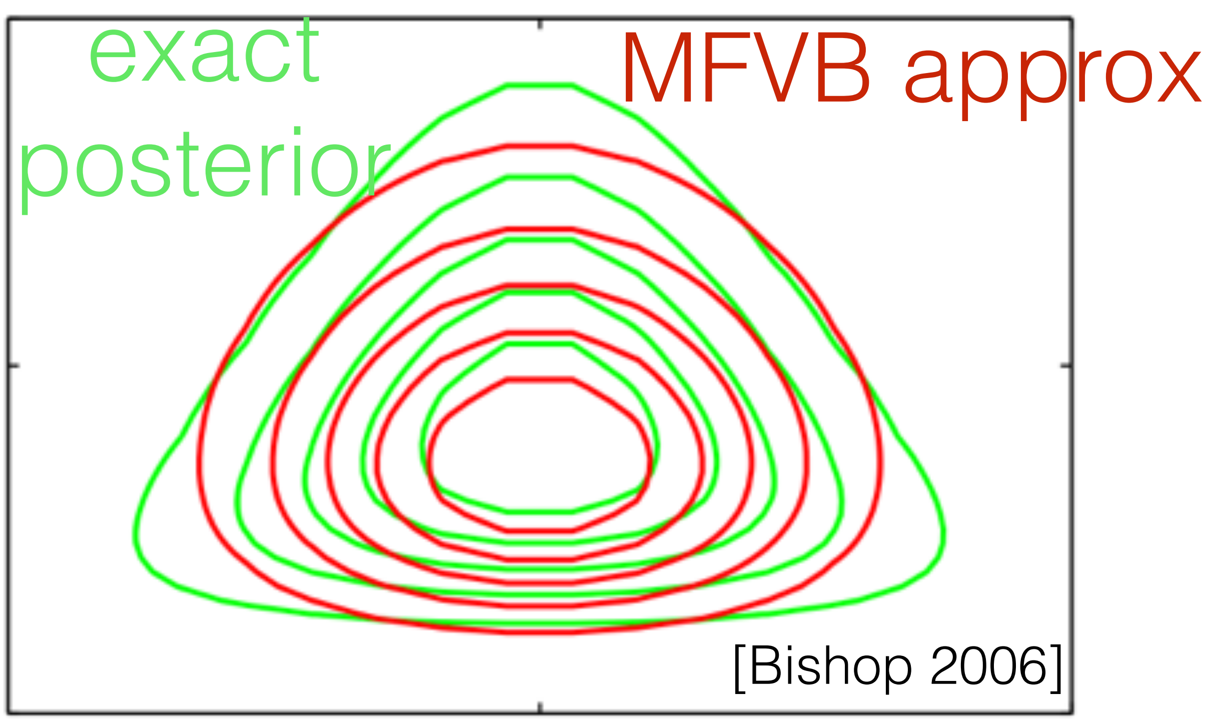


Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

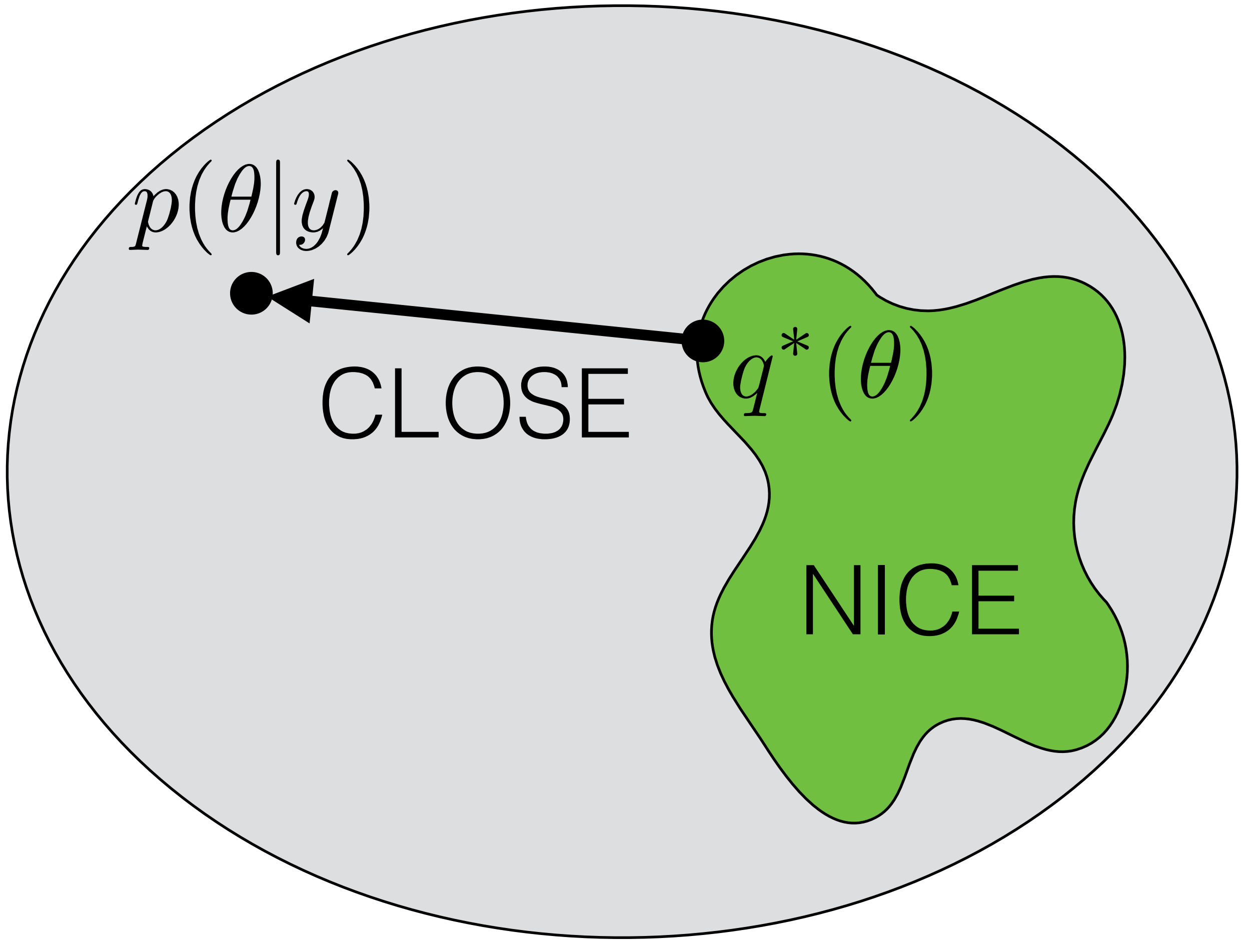
$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

- *Not* a modeling assumption



Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



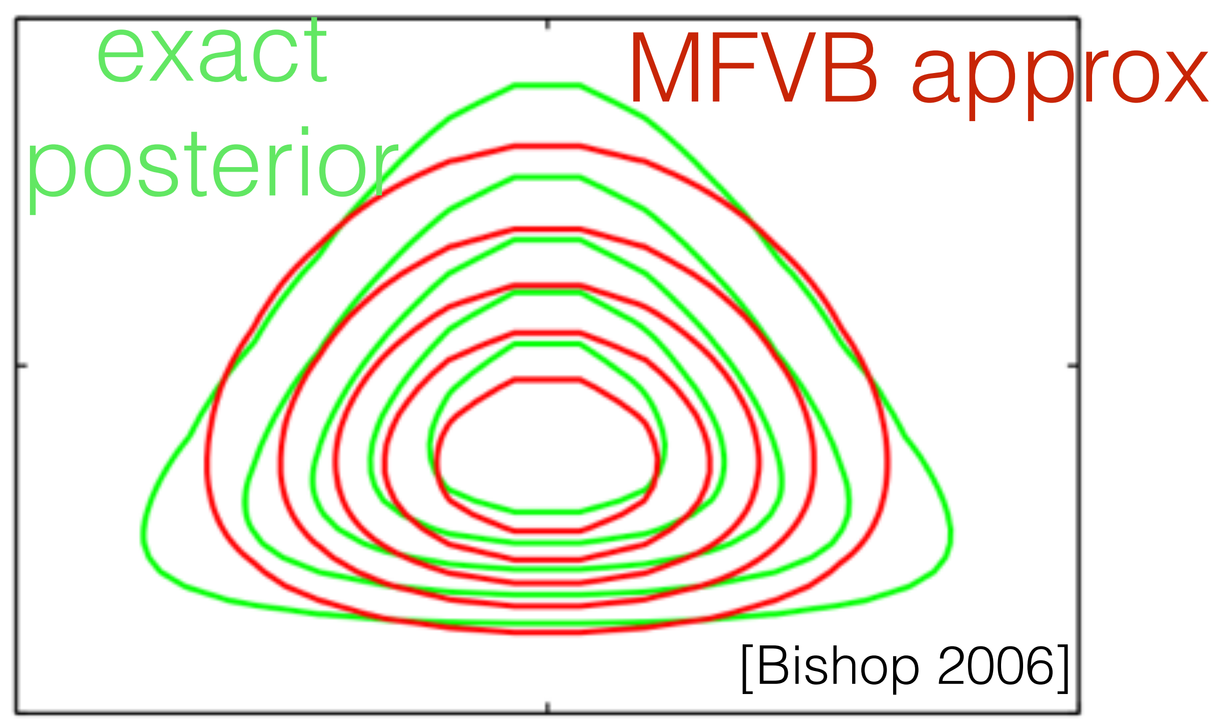
Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

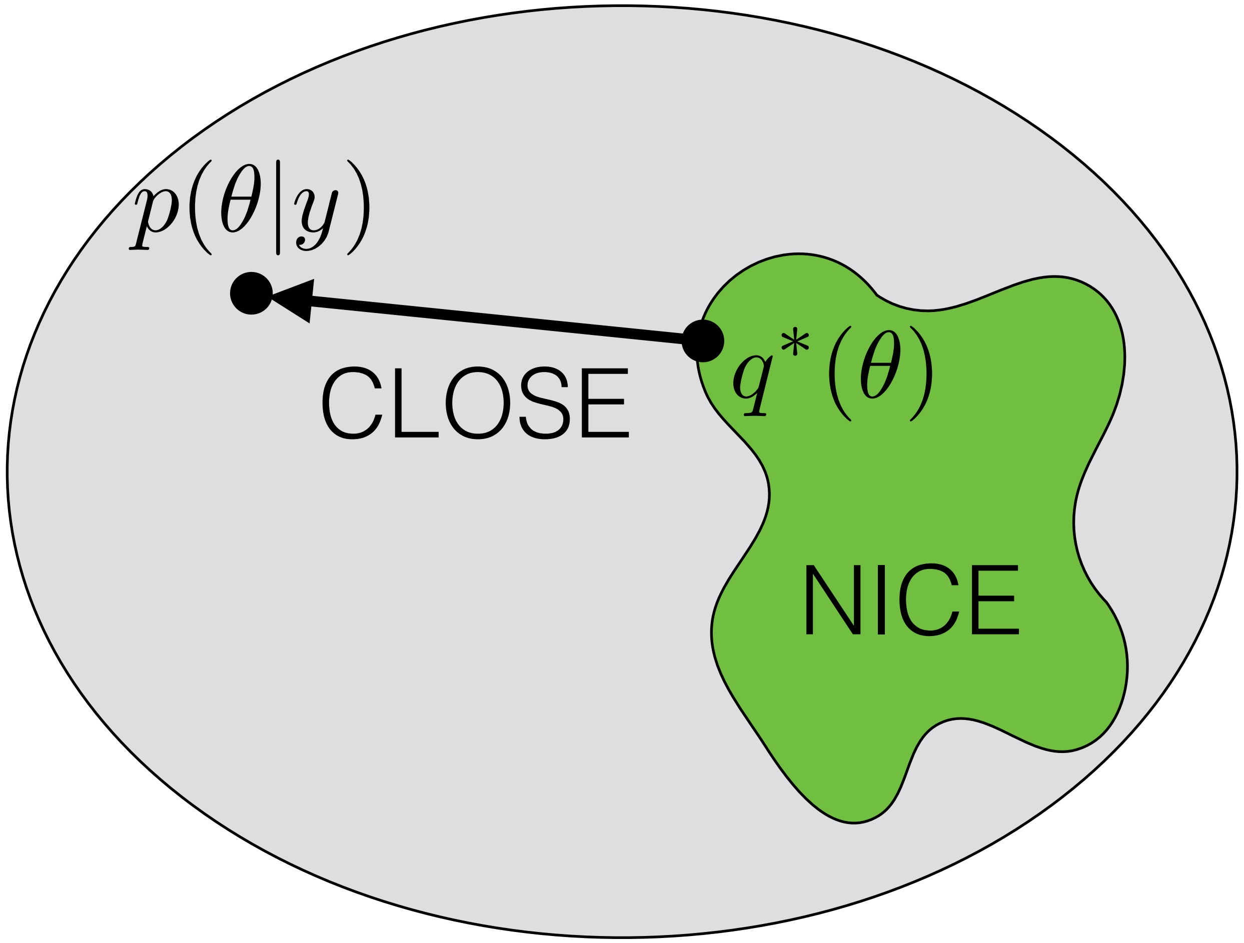
- *Not* a modeling assumption

Now we have an optimization problem; how to solve it?



Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

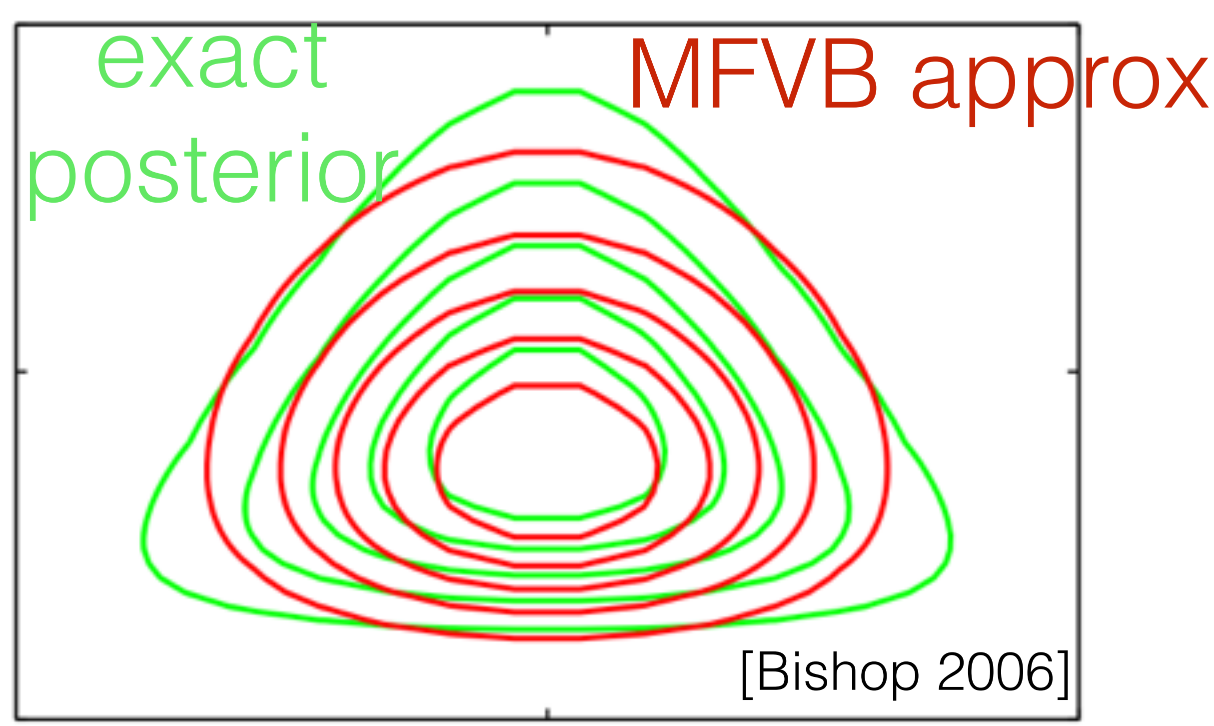


Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

- *Not* a modeling assumption



Now we have an optimization problem; how to solve it?

- One option: Coordinate descent in $q_1, \ldots, q_J$

Approximate Bayesian inference

Approximate Bayesian inference

Use q^* to approximate $p(\cdot|y)$

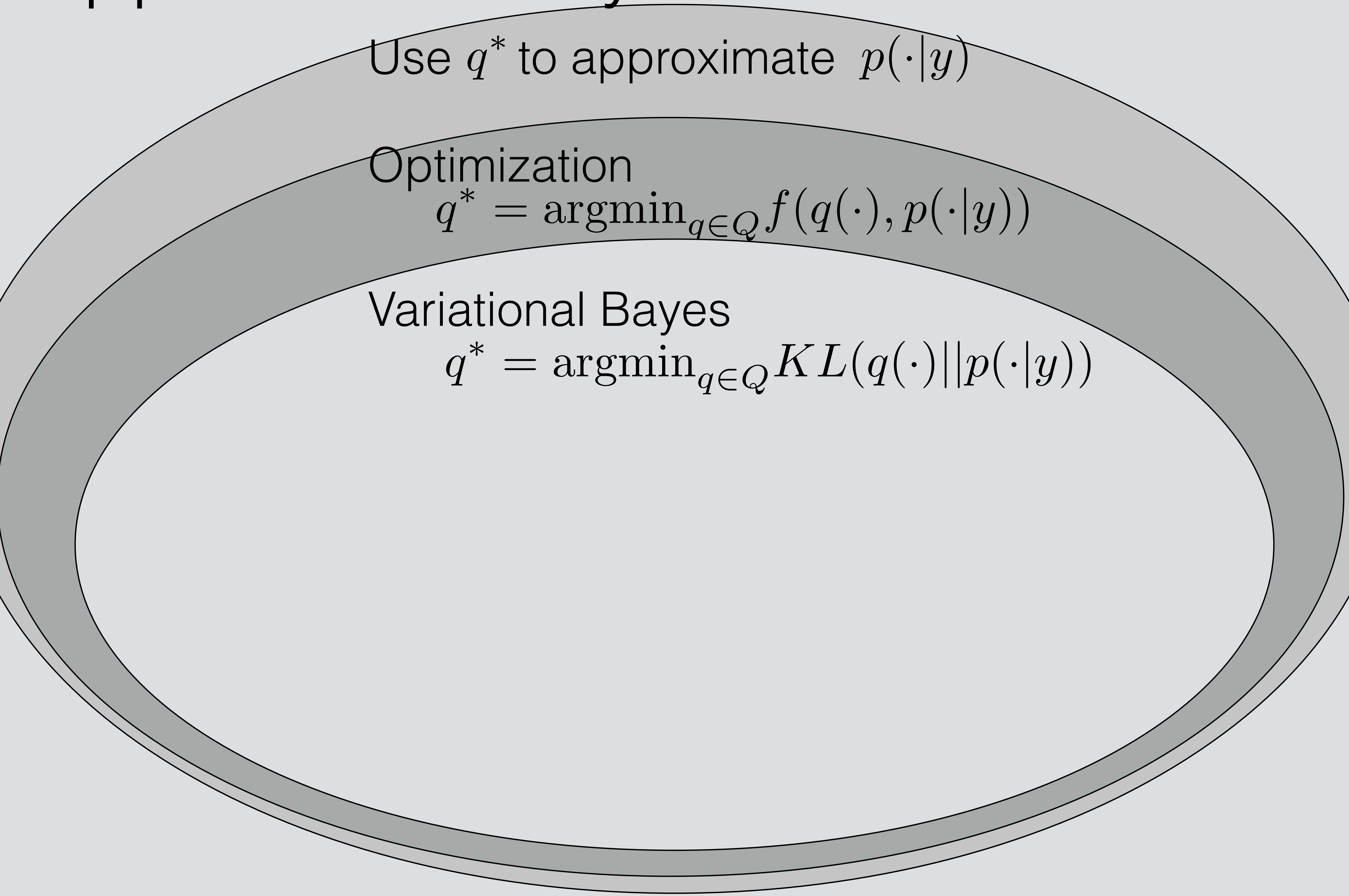
Approximate Bayesian inference

Use q^* to approximate $p(\cdot|y)$

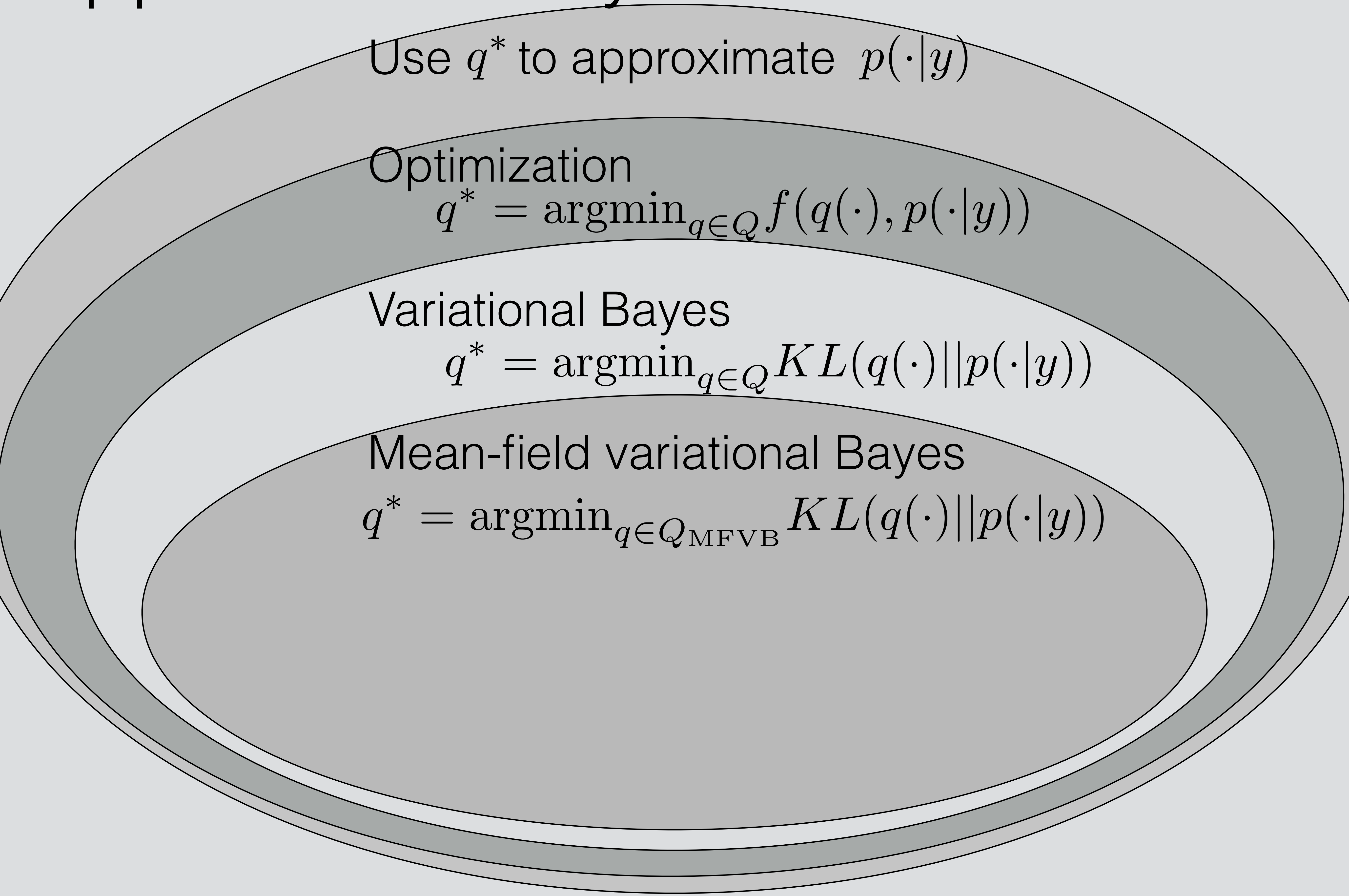
Optimization

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

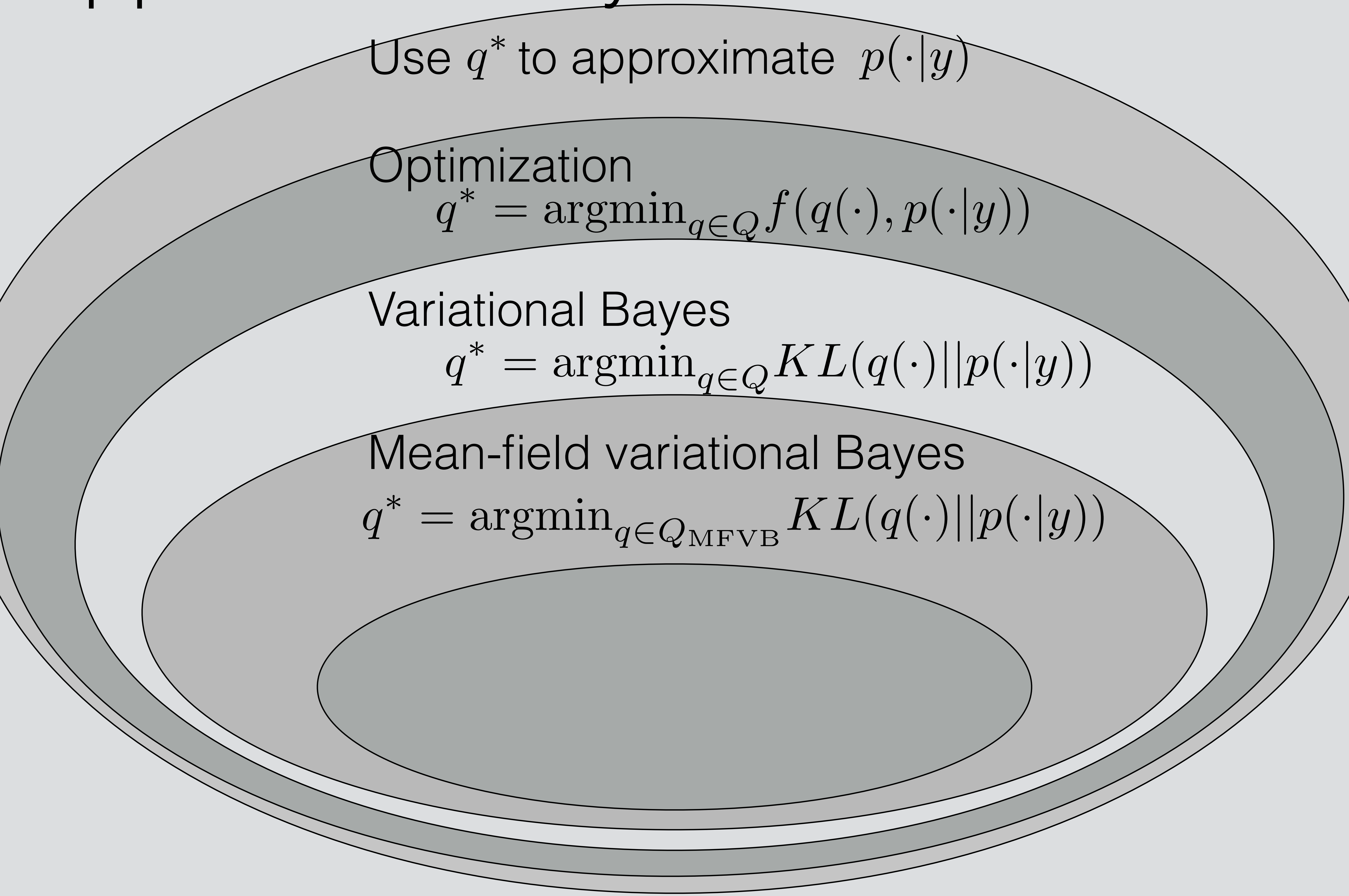
Approximate Bayesian inference



Approximate Bayesian inference



Approximate Bayesian inference



Approximate Bayesian inference

Use q^* to approximate $p(\cdot|y)$

Optimization

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Variational Bayes

$$q^* = \text{argmin}_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

Mean-field variational Bayes

$$q^* = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

- Coordinate descent

Approximate Bayesian inference

Use q^* to approximate $p(\cdot|y)$

Optimization

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Variational Bayes

$$q^* = \text{argmin}_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

Mean-field variational Bayes

$$q^* = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al 2013]

Approximate Bayesian inference

Use q^* to approximate $p(\cdot|y)$

Optimization

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Variational Bayes

$$q^* = \text{argmin}_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

Mean-field variational Bayes

$$q^* = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al 2013]
- Automatic differentiation variational inference (ADVI) [Kucukelbir et al 2015, 2017]

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$

- Model:

$$p(y|\theta): \quad y_n \overset{iid}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N$$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta): \quad y_n \overset{iid}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N$$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta): \quad y_n \stackrel{iid}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta): \quad (\sigma^2)^{-1} \sim \mathrm{Gamma}(a_0, b_0)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)$$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta): \quad y_n \overset{iid}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta): \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)$$

- Exercise: check $p(\mu, \sigma^2|y) \neq f_1(\mu, y) f_2(\sigma^2, y)$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta): \quad y_n \overset{iid}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta): \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)$$

- Exercise: check $p(\mu, \sigma^2|y) \neq f_1(\mu, y) f_2(\sigma^2, y)$
- MFVB approximation:

$$q^*(\mu, \sigma) = q^*_\mu(\mu) q^*_{\sigma^2}(\sigma^2) = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta): \quad y_n \overset{iid}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta): \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)$$

- Exercise: check $p(\mu, \sigma^2|y) \neq f_1(\mu, y) f_2(\sigma^2, y)$
- MFVB approximation:

$$q^*(\mu, \sigma) = q^*_\mu(\mu) q^*_{\sigma^2}(\sigma^2) = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta): \quad y_n \stackrel{iid}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta): \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)$$

- Exercise: check $p(\mu, \sigma^2|y) \neq f_1(\mu, y) f_2(\sigma^2, y)$
- MFVB approximation:

$$q^*(\mu, \sigma) = q^*_\mu(\mu) q^*_{\sigma^2}(\sigma^2) = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]

$$q^*(\mu) = N(\mu|m_\mu, \rho^2_\mu) \qquad q^*((\sigma^2)^{-1}) = \text{Gamma}((\sigma^2)^{-1}|a_\sigma, b_\sigma)$$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta) : \quad y_n \overset{iid}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta) : \quad (\sigma^2)^{-1} \sim \mathrm{Gamma}(a_0, b_0)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)$$

- Exercise: check $p(\mu, \sigma^2|y) \neq f_1(\mu, y) f_2(\sigma^2, y)$
- MFVB approximation:

$$q^*(\mu, \sigma) = q^*_\mu(\mu) q^*_{\sigma^2}(\sigma^2) = \mathrm{argmin}_{q \in Q_{\mathrm{MFVB}}} KL(q(\cdot) || p(\cdot|y))$$

- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]

$$q^*(\mu) = N(\mu | m_\mu, \rho^2_\mu) \qquad q^*((\sigma^2)^{-1}) = \mathrm{Gamma}((\sigma^2)^{-1} | a_\sigma, b_\sigma)$$

"variational parameters"

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta): \quad y_n \stackrel{iid}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta): \quad (\sigma^2)^{-1} \sim \mathrm{Gamma}(a_0, b_0)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)$$

- Exercise: check $p(\mu, \sigma^2|y) \neq f_1(\mu, y) f_2(\sigma^2, y)$
- MFVB approximation:

$$q^*(\mu, \sigma) = q^*_\mu(\mu) q^*_{\sigma^2}(\sigma^2) = \mathrm{argmin}_{q \in Q_{\mathrm{MFVB}}} KL(q(\cdot) || p(\cdot|y))$$

- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]

$$q^*(\mu) = N(\mu | m_\mu, \rho_\mu^2) \qquad q^*((\sigma^2)^{-1}) = \mathrm{Gamma}((\sigma^2)^{-1} | a_\sigma, b_\sigma)$$

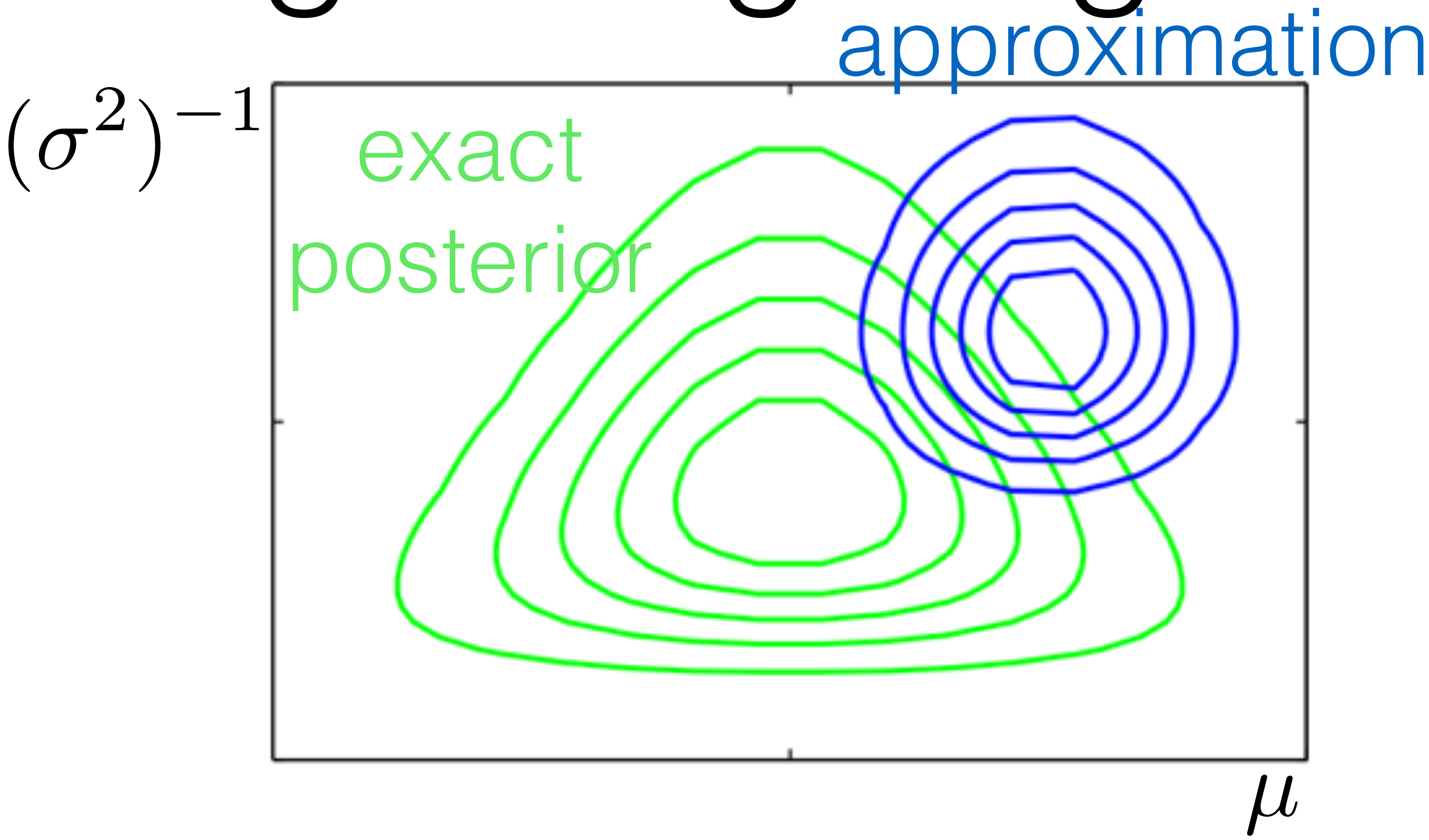
  - Iterate: $(m_\mu, \rho_\mu^2) = f(a_\sigma, b_\sigma)$

    $(a_\sigma, b_\sigma) = g(m_\mu, \rho_\mu^2)$

"variational parameters"

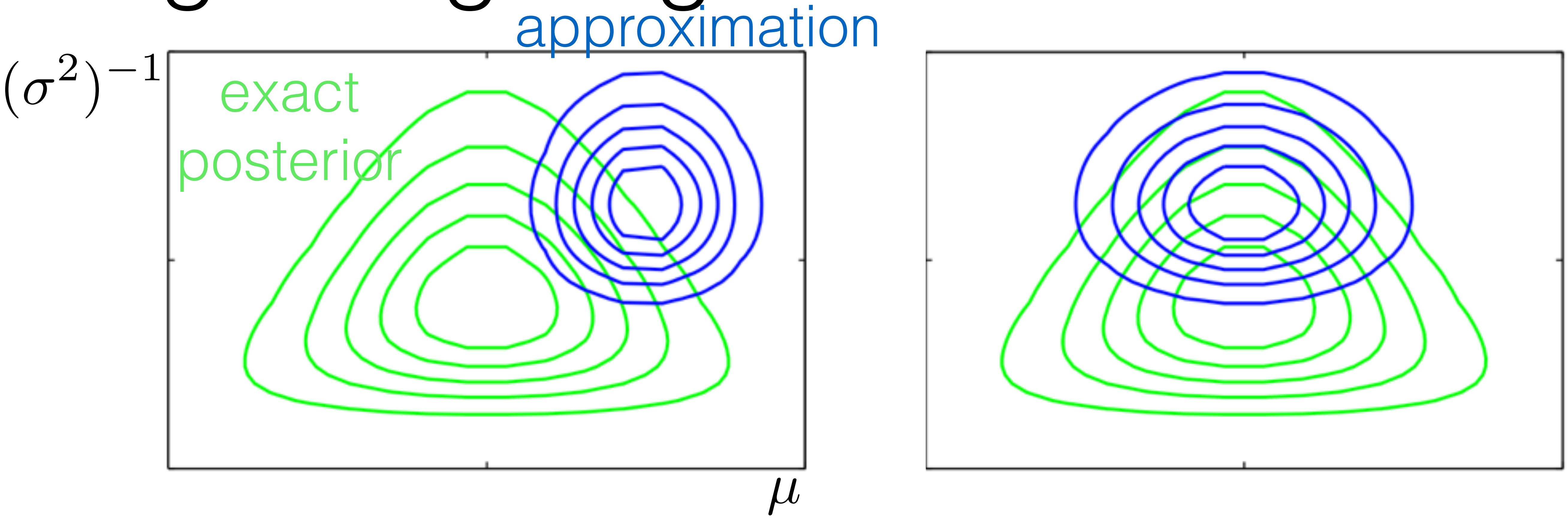
[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length



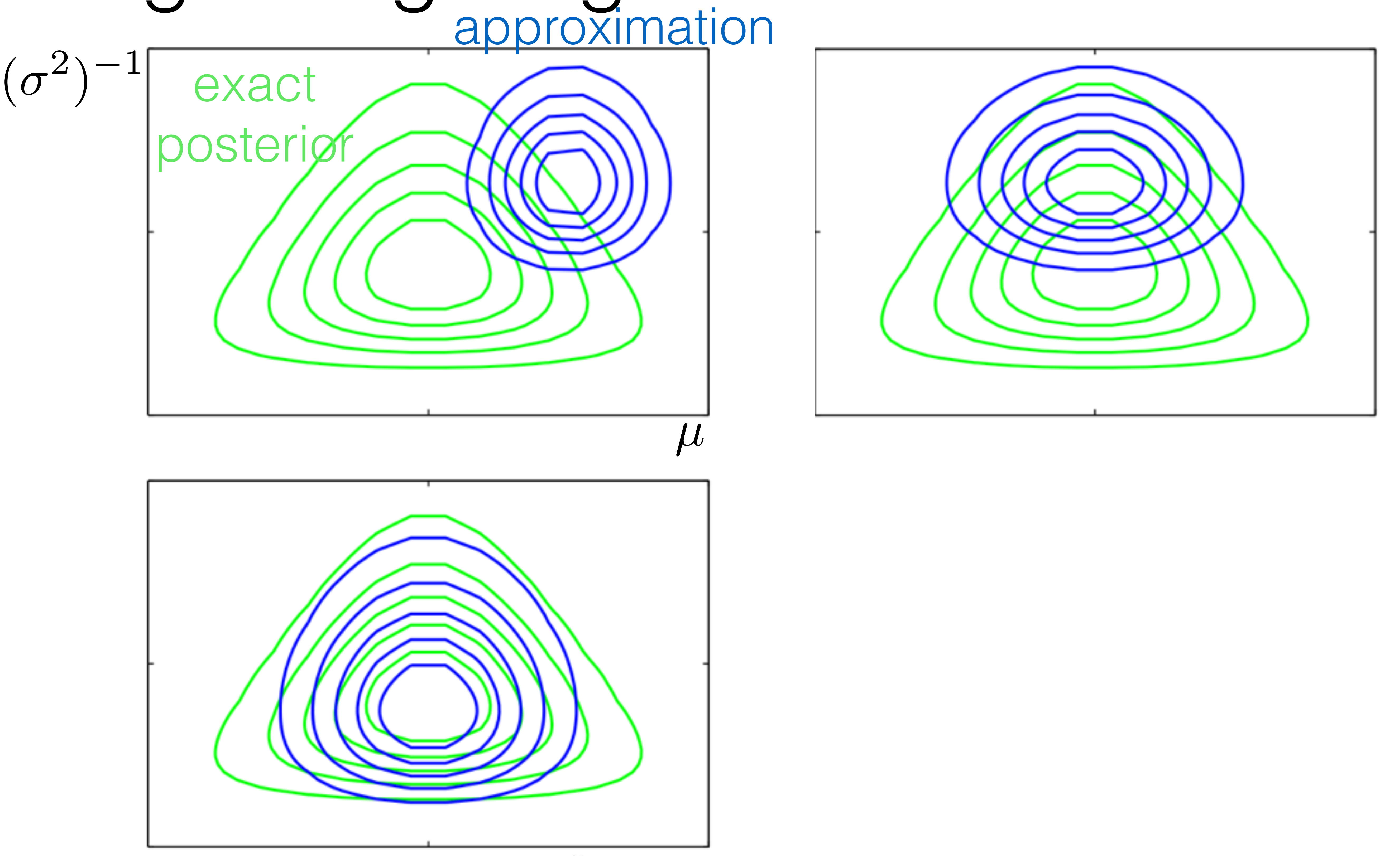
[Bishop 2006]

Midge wing length



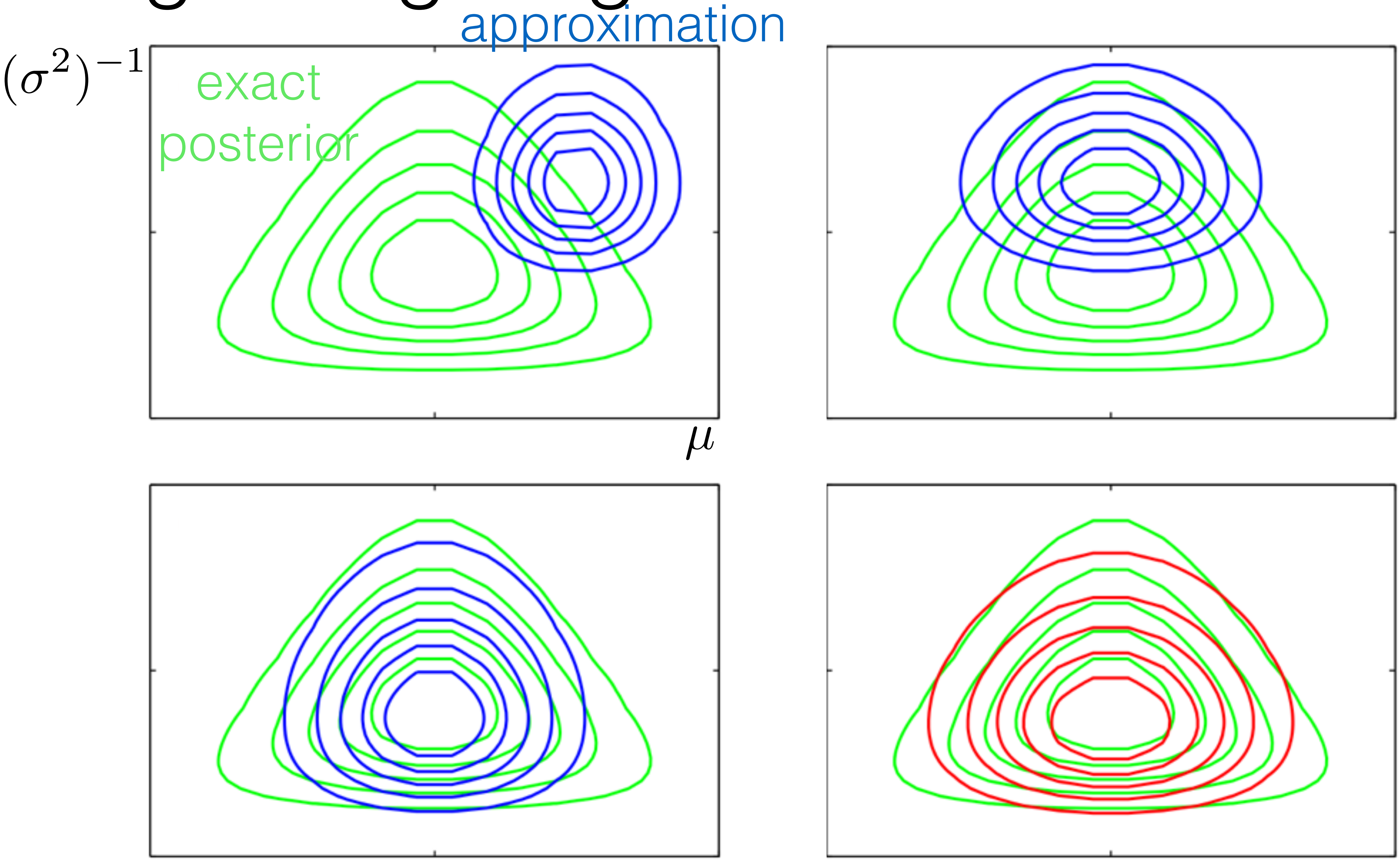
[Bishop 2006]

Midge wing length



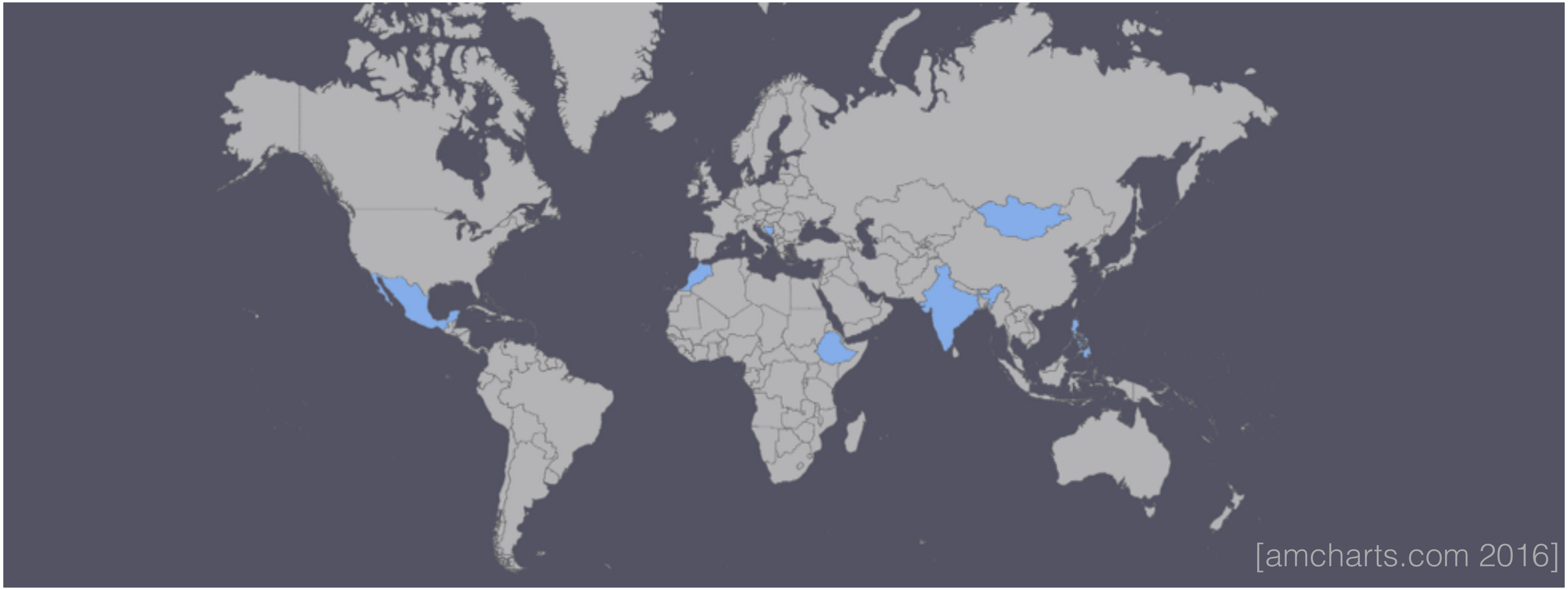
[Bishop 2006]

Midge wing length



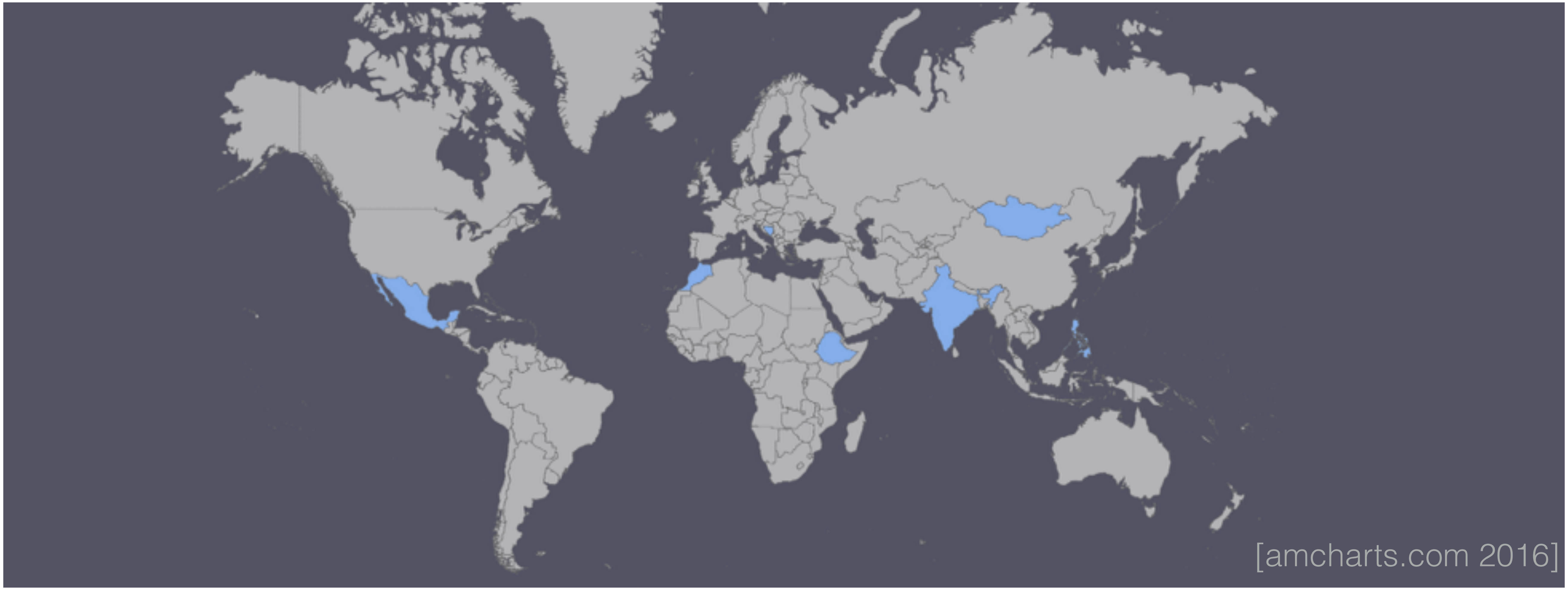
[Bishop 2006]

Microcredit Experiment



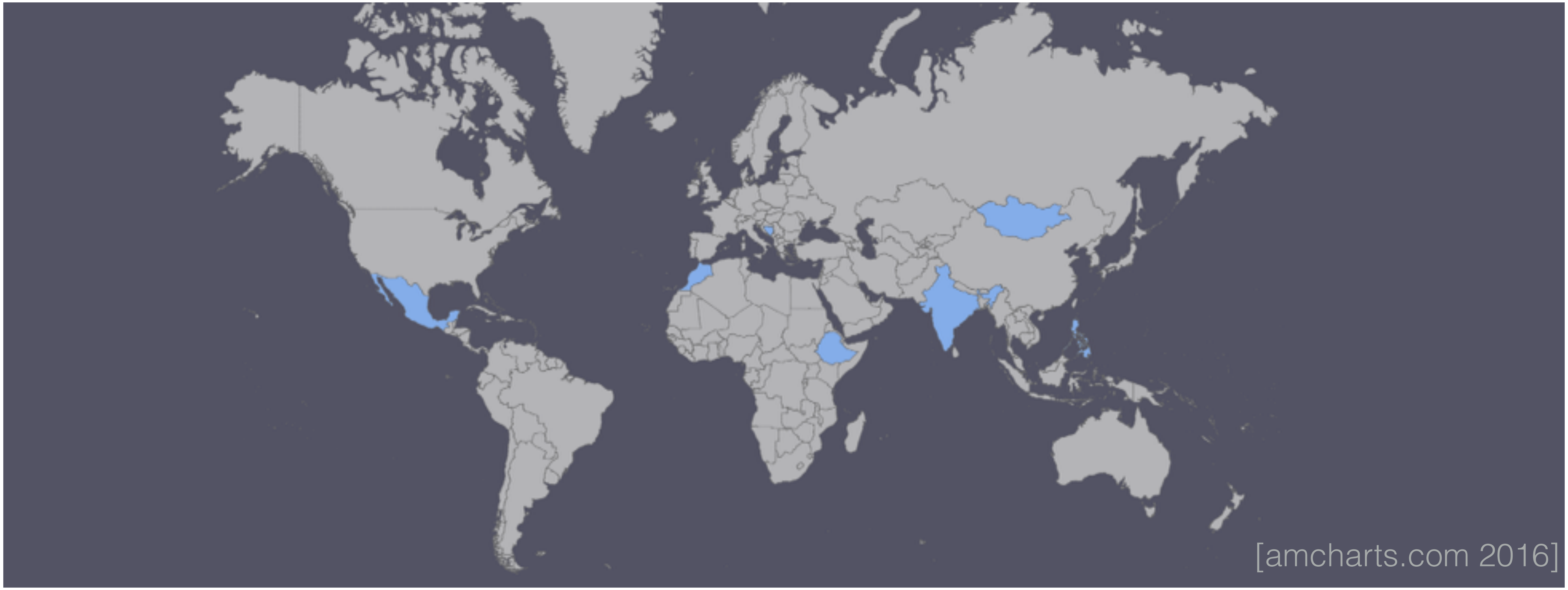
Microcredit Experiment

- Simplified from Meager (2018a)
- *K* microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)



Microcredit Experiment

- Simplified from Meager (2018a)
- K microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)



Microcredit Experiment

- Simplified from Meager (2018a)
- K microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)

Microcredit Experiment

- Simplified from Meager (2018a)
- 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:

Microcredit Experiment

- Simplified from Meager (2018a)
- 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}

Microcredit Experiment

- Simplified from Meager (2018a)
- *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

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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

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

Microcredit Experiment

- Simplified from Meager (2018a)
- *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

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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 → 1 if microcredit →

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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}; 1 if microcredit → T_{kn}

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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

1 if microcredit

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

- Priors and hyperpriors:

Microcredit Experiment

- Simplified from Meager (2018a)
- *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:

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

Microcredit Experiment

- Simplified from Meager (2018a)
- *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 (2018a)
- *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

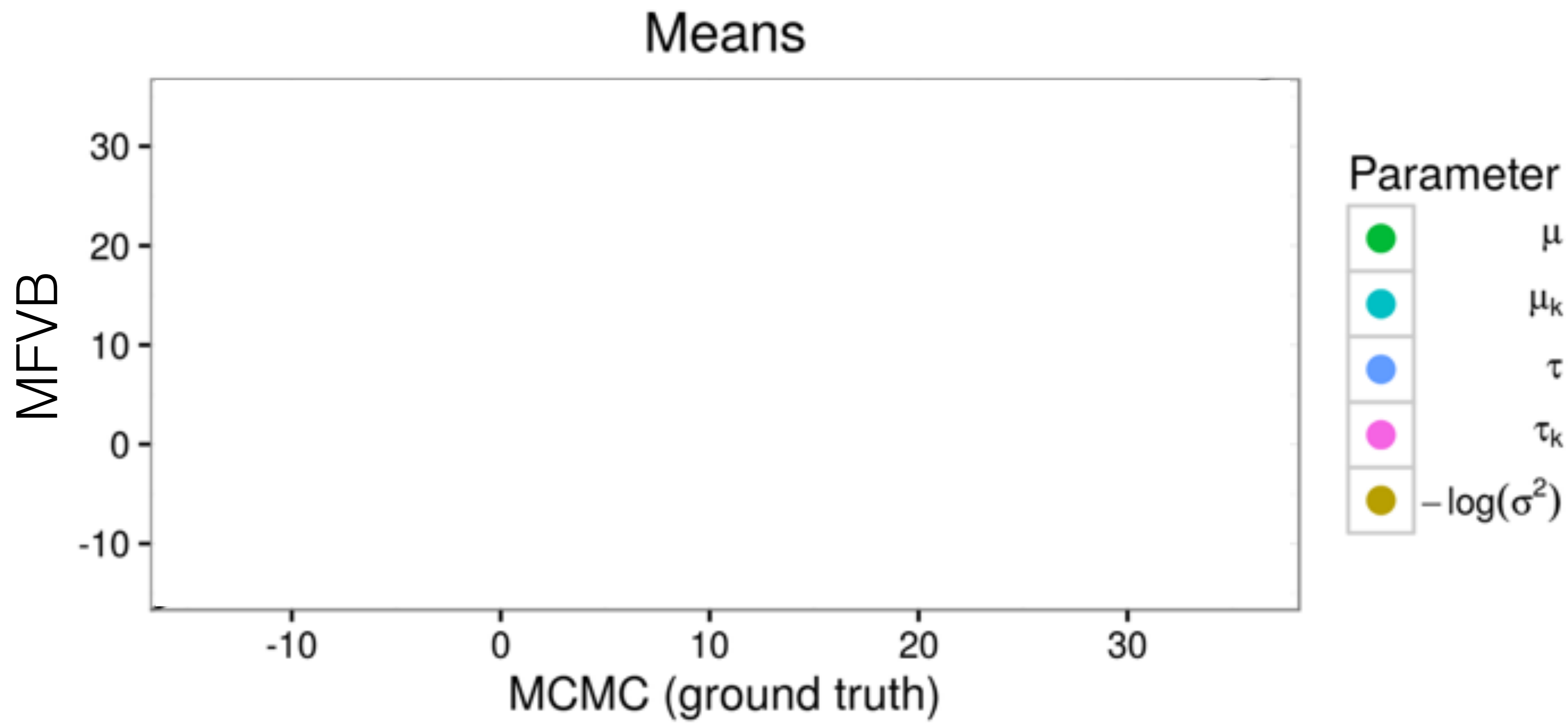
$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

- Priors and hyperpriors:

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

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

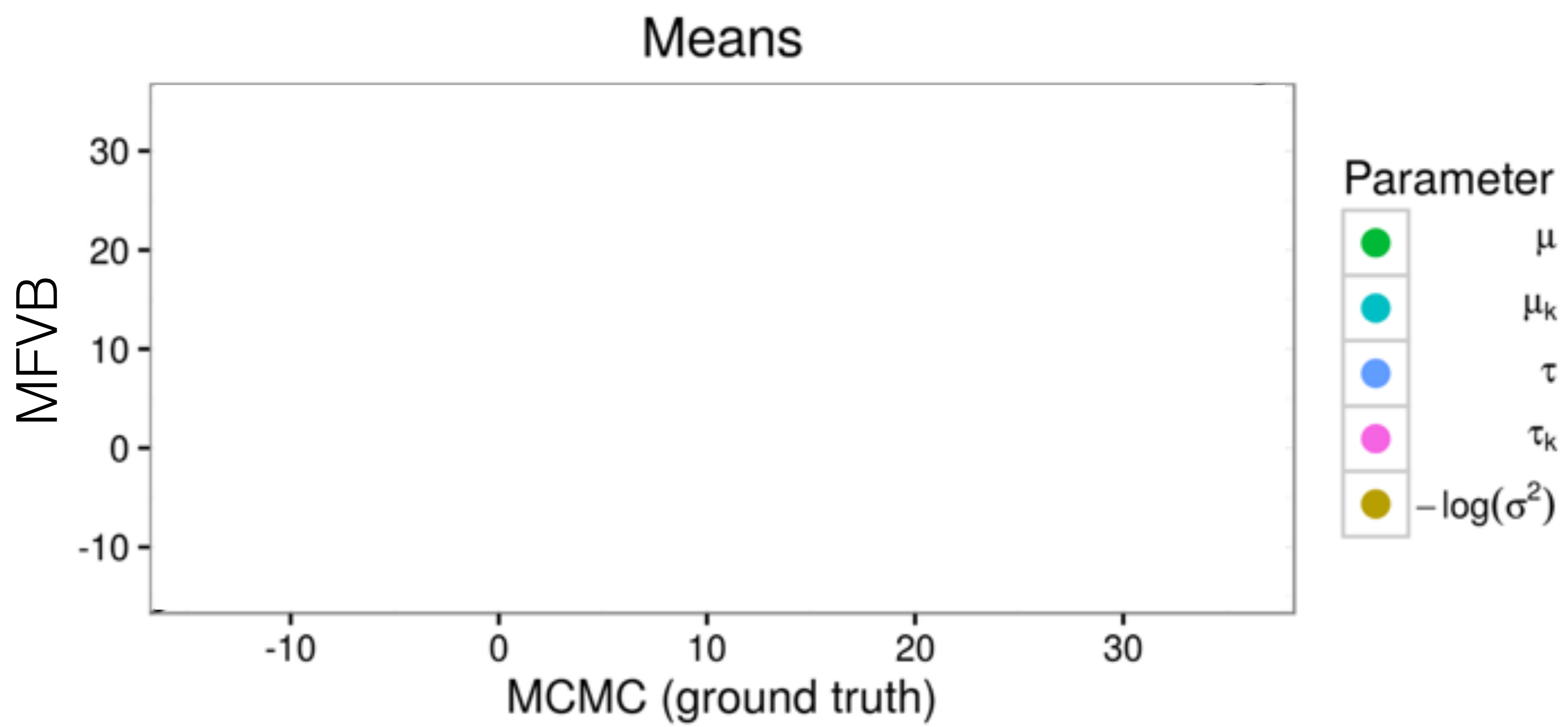
Microcredit



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

Microcredit

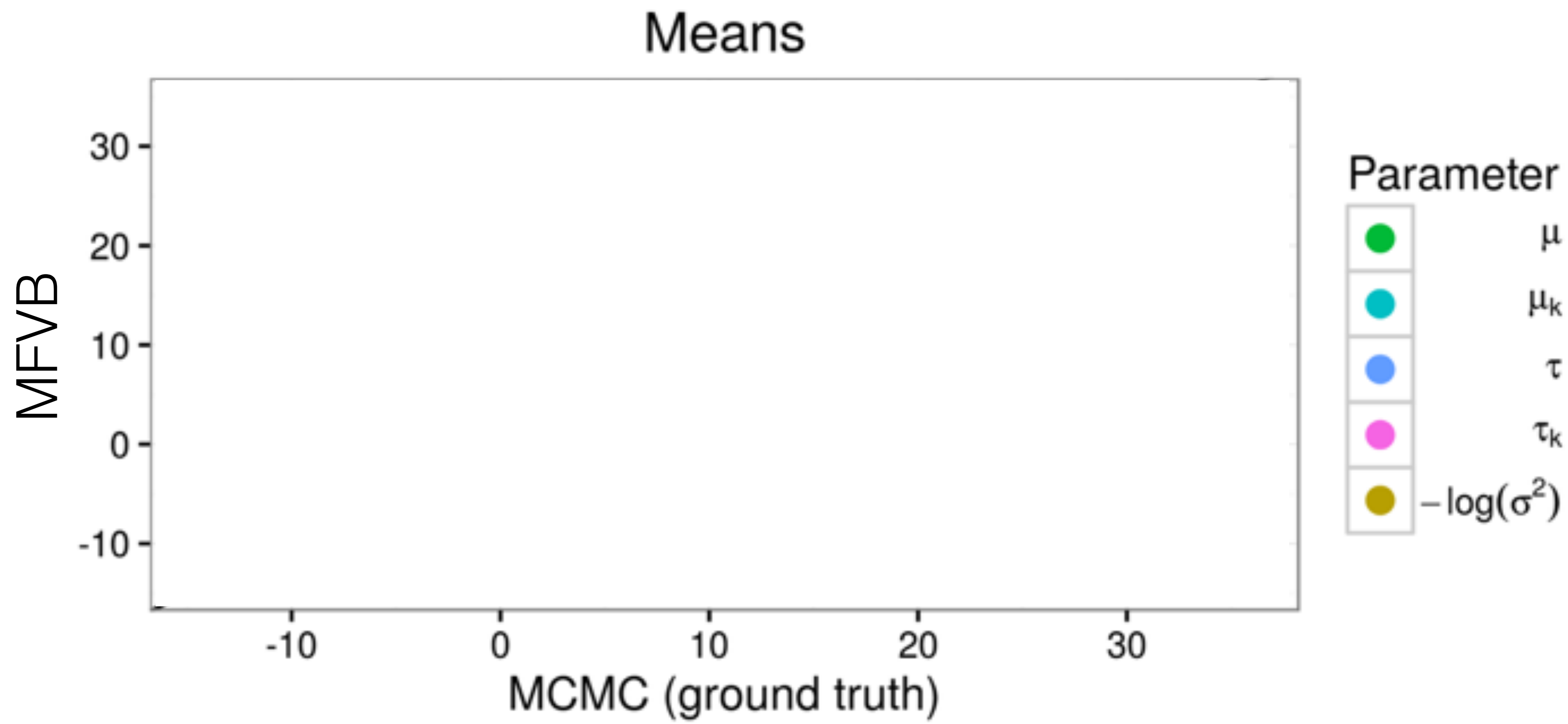
- *One set* of 2500 MCMC draws: **45 minutes**



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

Microcredit

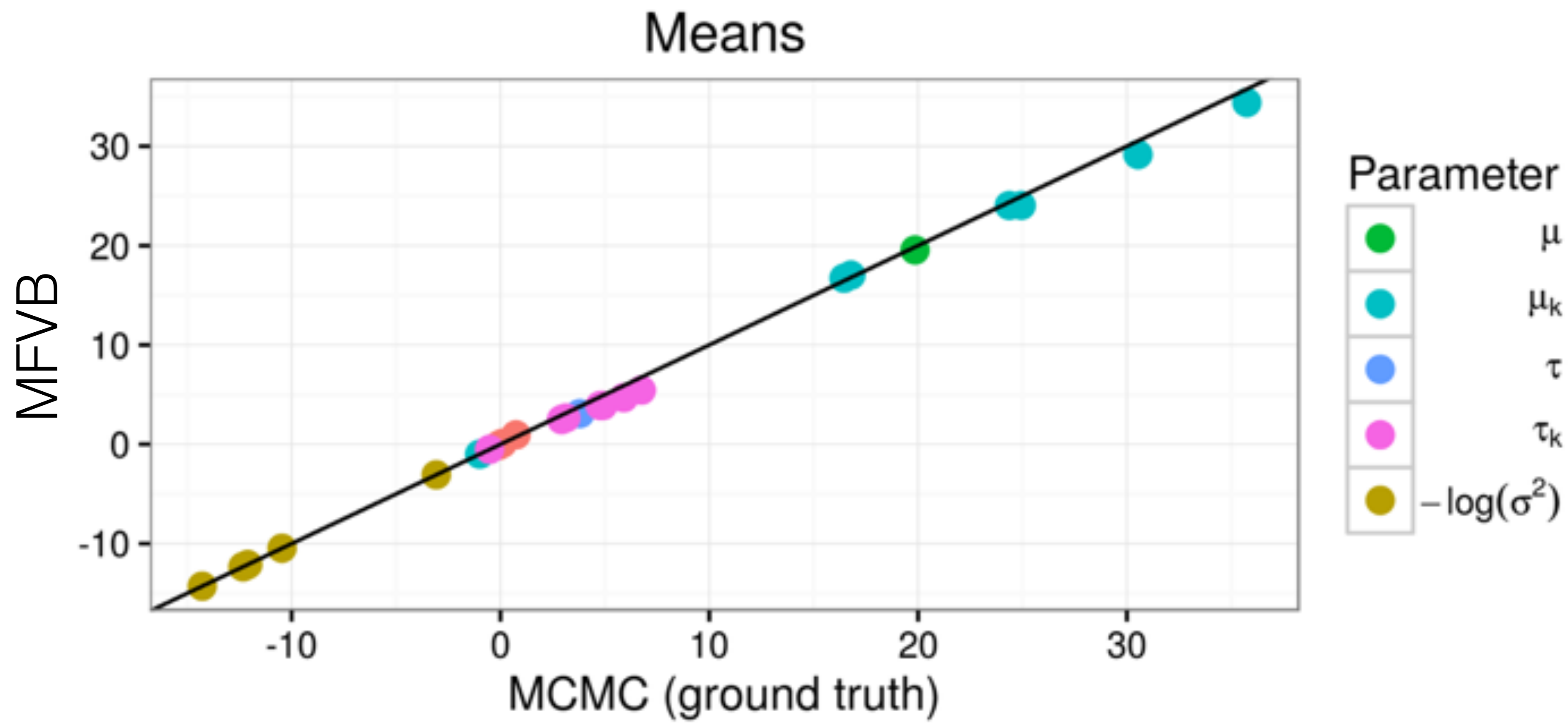
- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

Microcredit

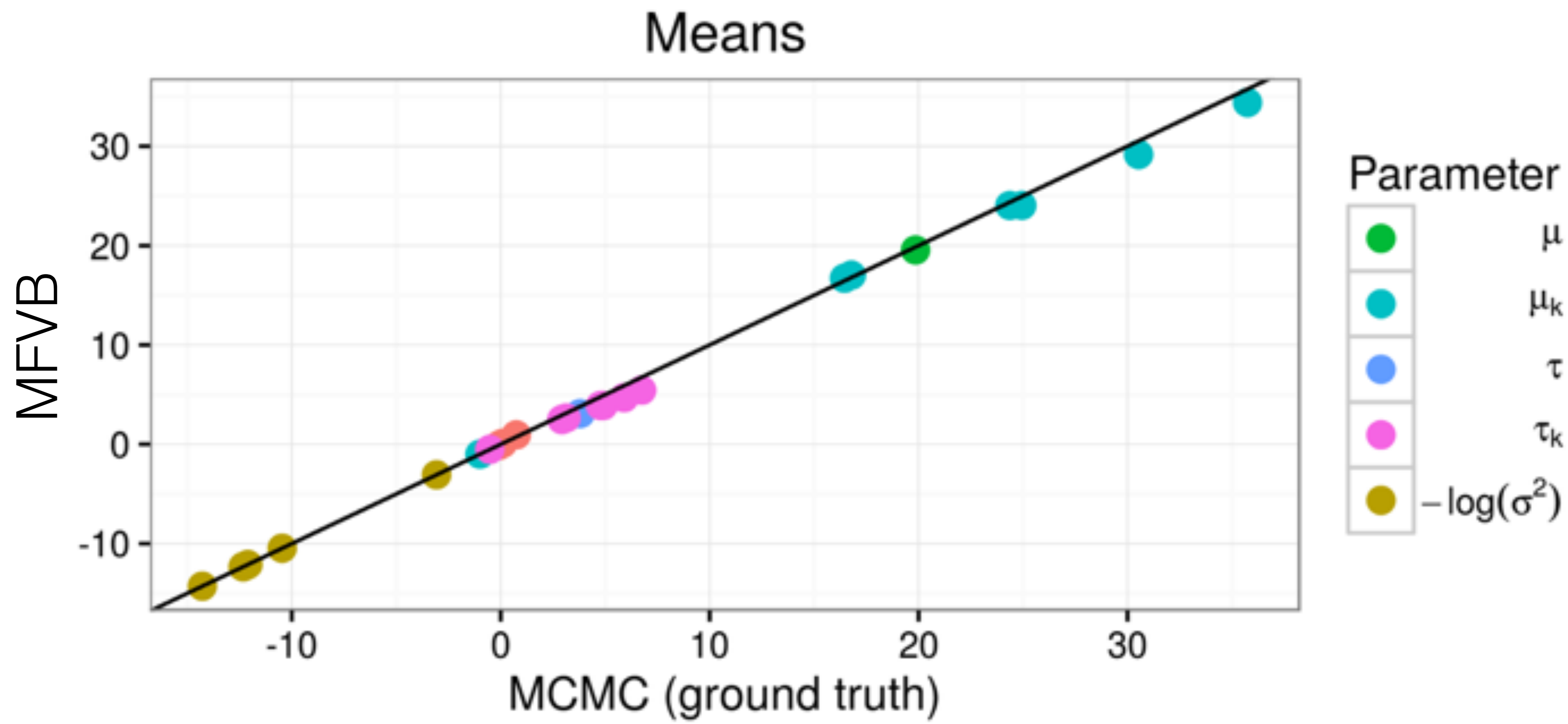
- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

Microcredit

- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



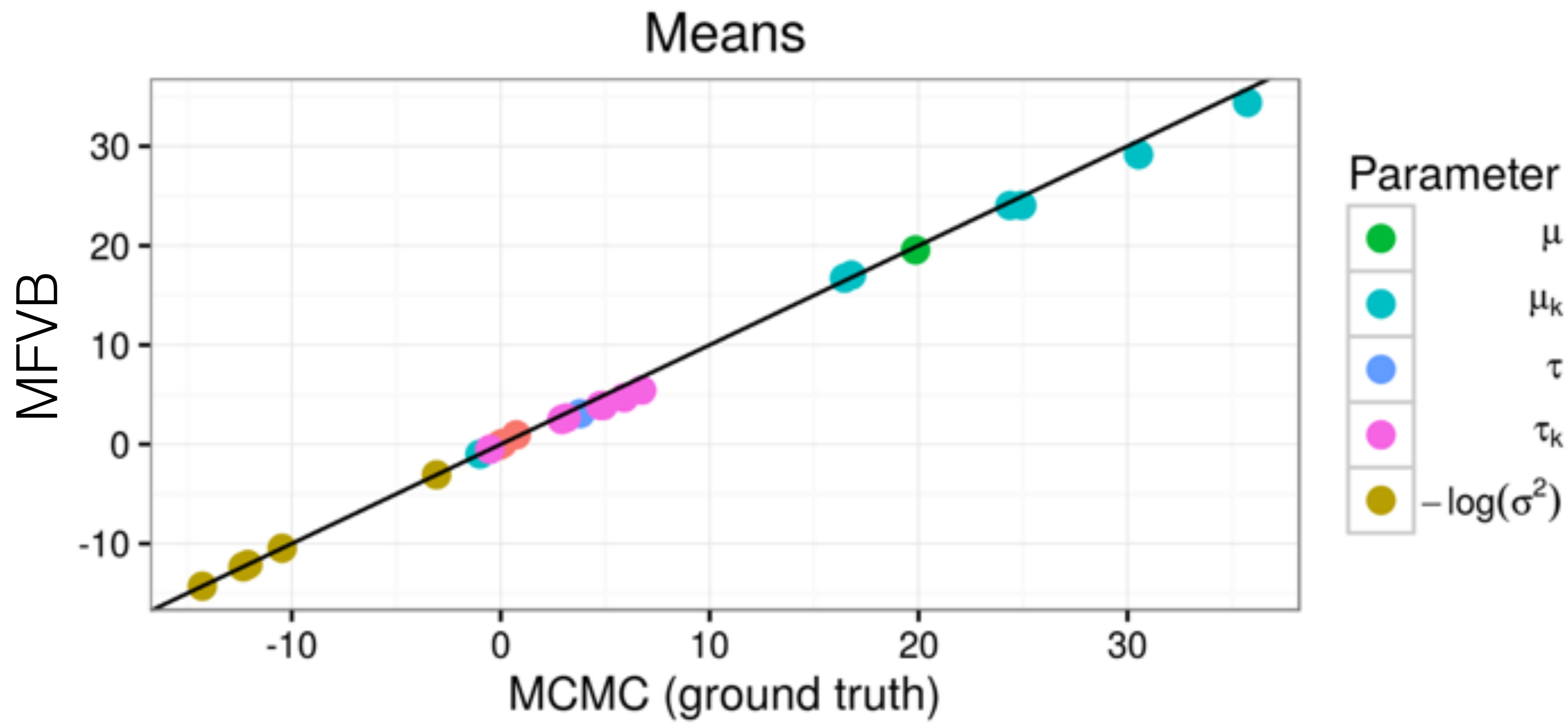
Criteo Online Ads Experiment

- Click-through conversion prediction
- Q: Will a customer (e.g.) buy a product after clicking?

[Giordano, Broderick, Meager, Huggins, Jordan 2016]

Microcredit

- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



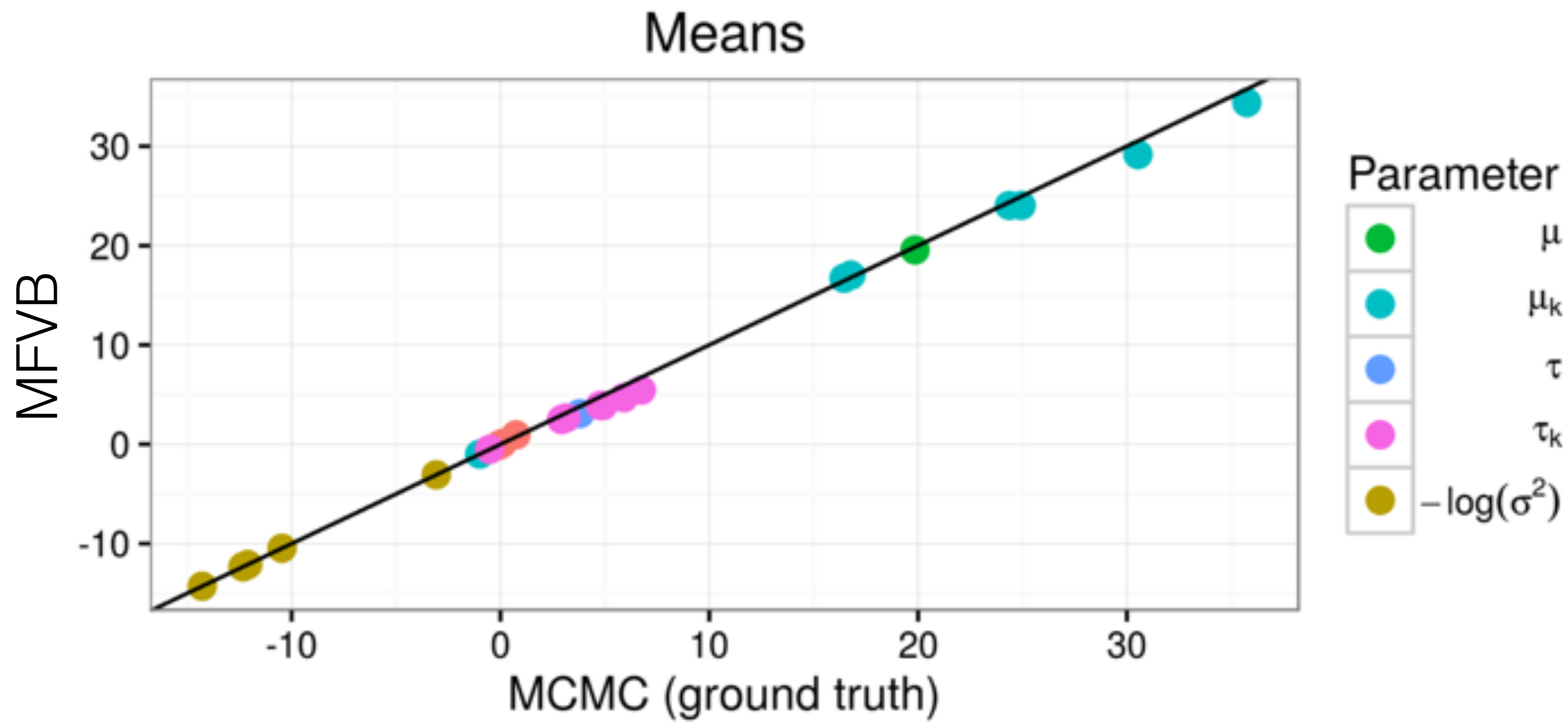
Criteo Online Ads Experiment

- Click-through conversion prediction
- Q: Will a customer (e.g.) buy a product after clicking?
- Q: How predictive of conversion are different features?

[Giordano, Broderick, Meager, Huggins, Jordan 2016]

Microcredit

- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



Criteo Online Ads Experiment

- Click-through conversion prediction
- Q: Will a customer (e.g.) buy a product after clicking?
- Q: How predictive of conversion are different features?
- Logistic GLMM; N = 61,895 subset to compare to MCMC

[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2017]

Criteo Online Ads Experiment

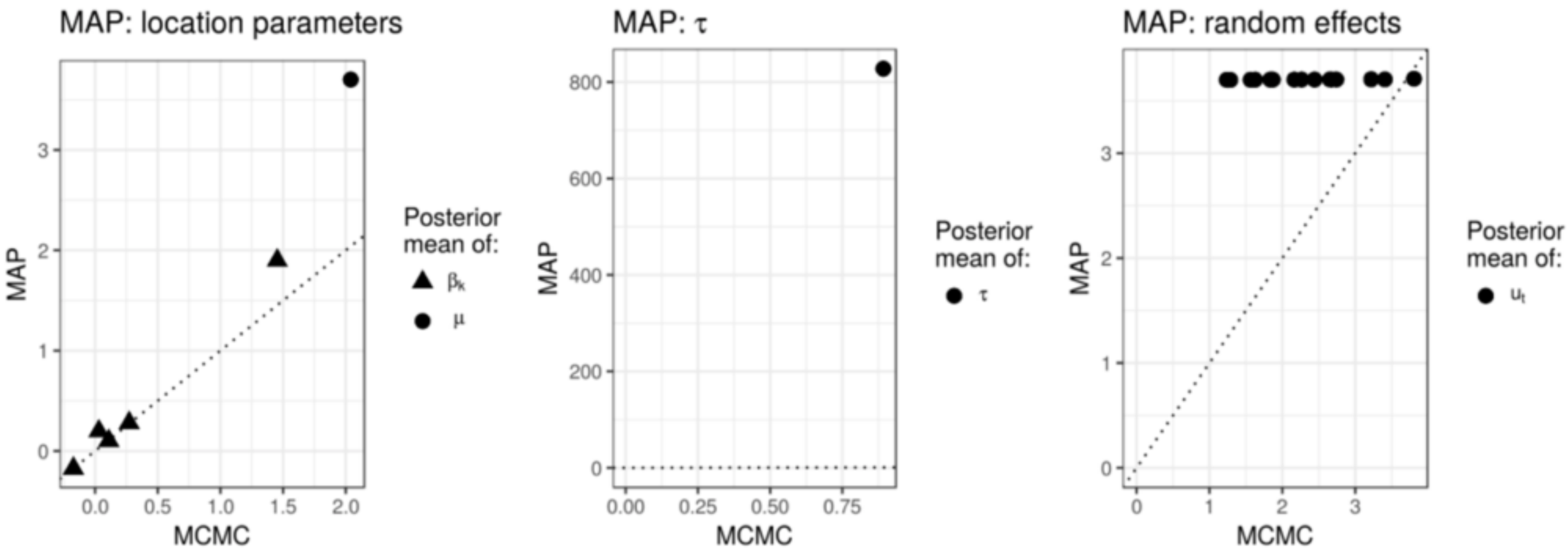
[Giordano, Broderick, Jordan 2017]

Criteo Online Ads Experiment

- MAP: **12 s**

[Giordano, Broderick, Jordan 2017]

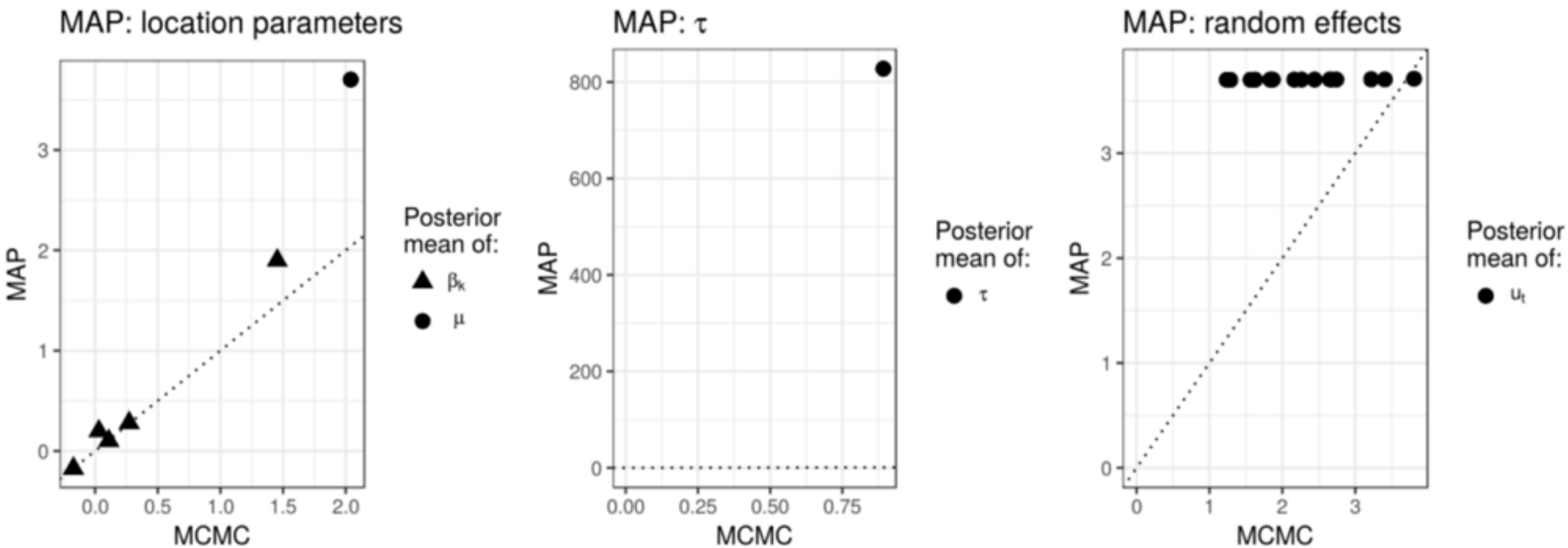
Criteo Online Ads Experiment



- MAP: **12 s**

[Giordano, Broderick, Jordan 2017]

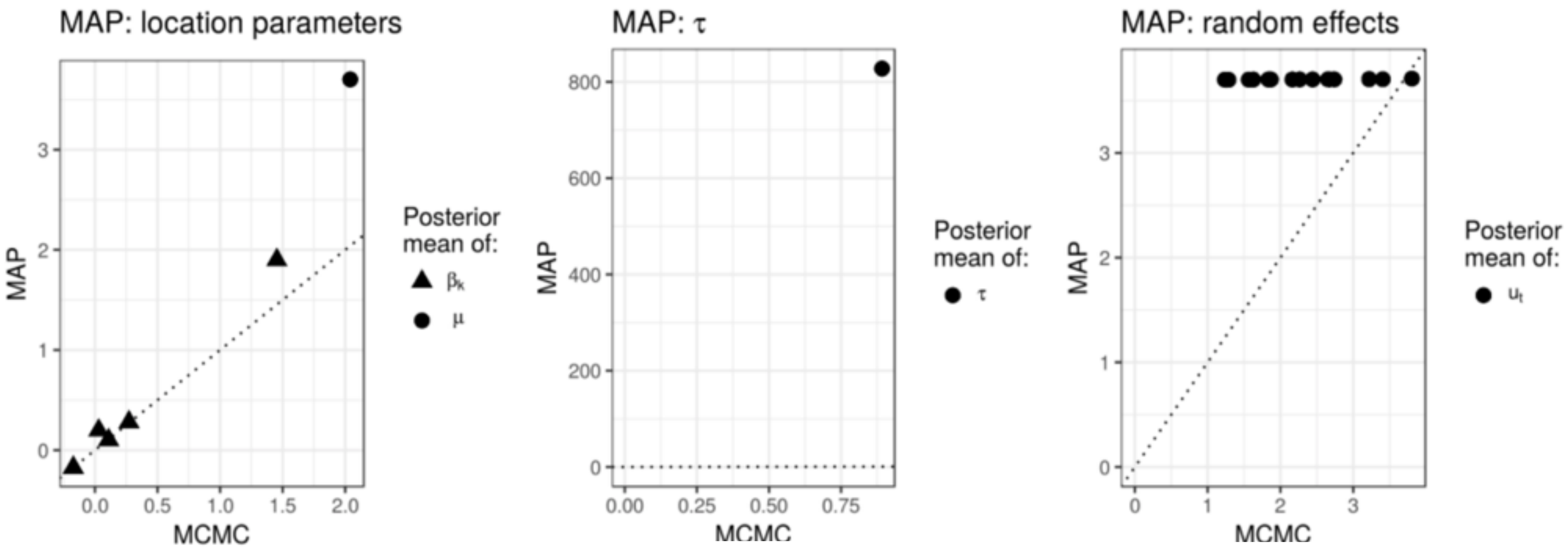
Criteo Online Ads Experiment



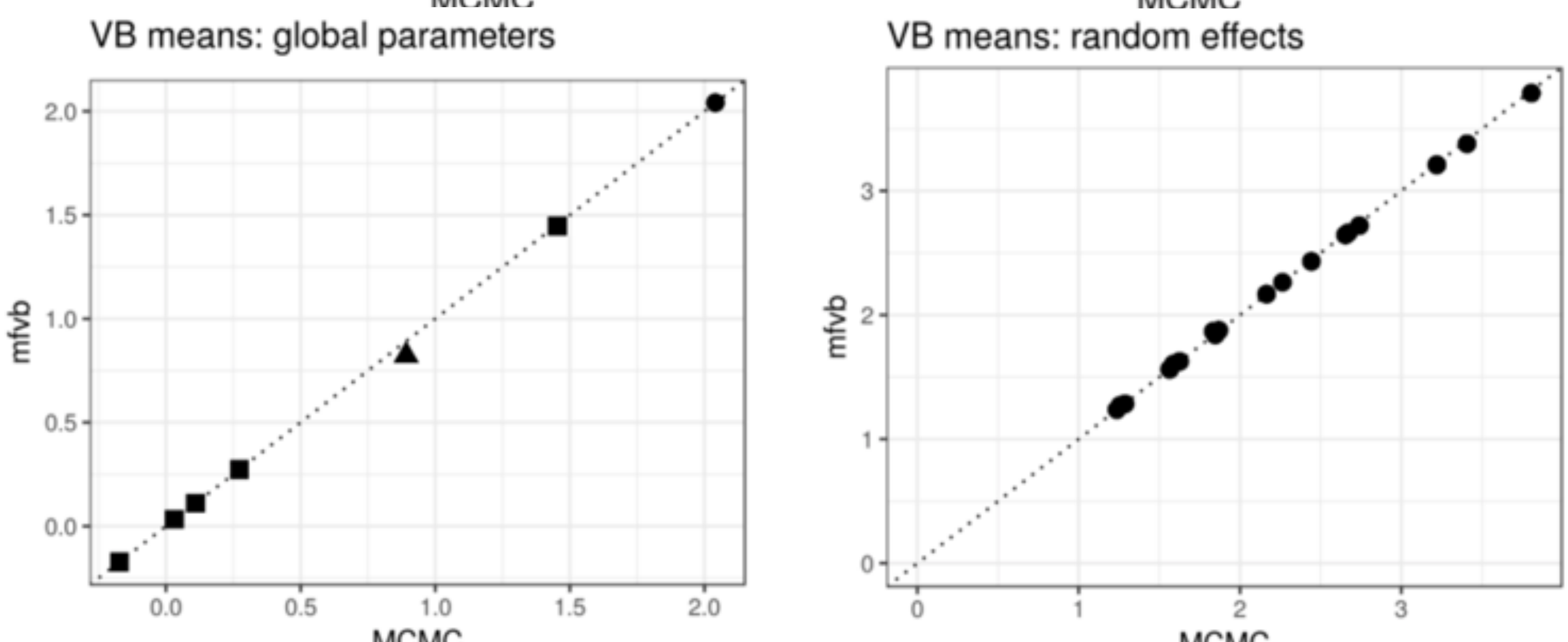
- MAP: **12 s**
- VB: **57 s**

[Giordano, Broderick, Jordan 2017]

Criteo Online Ads Experiment

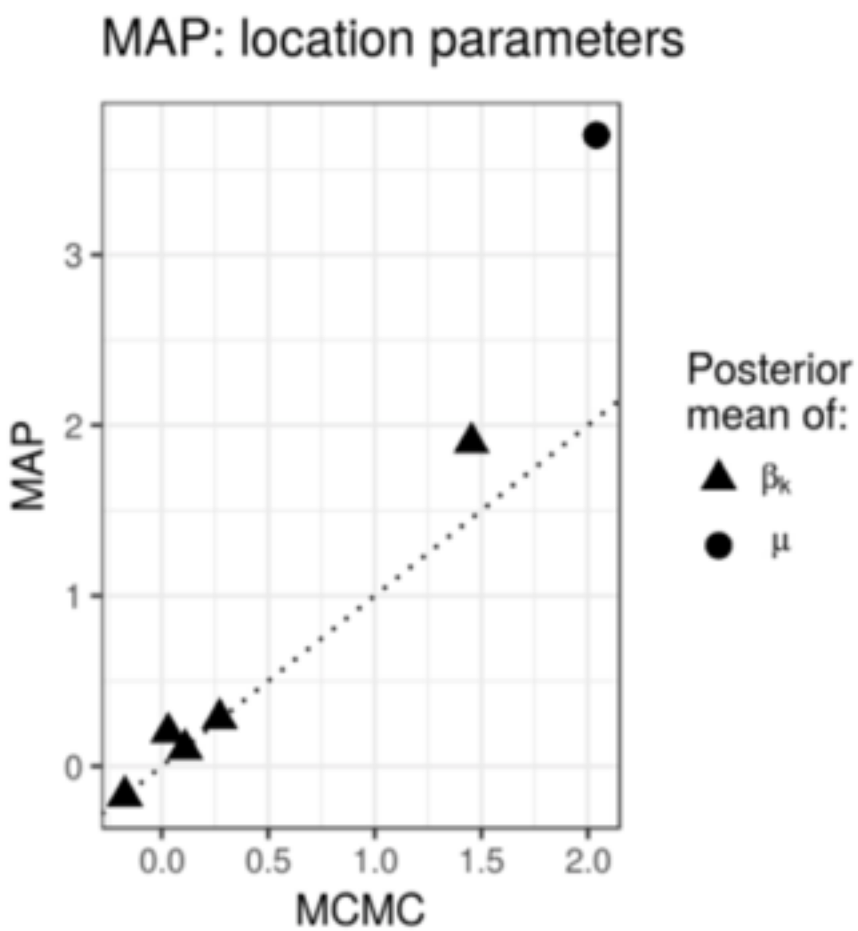


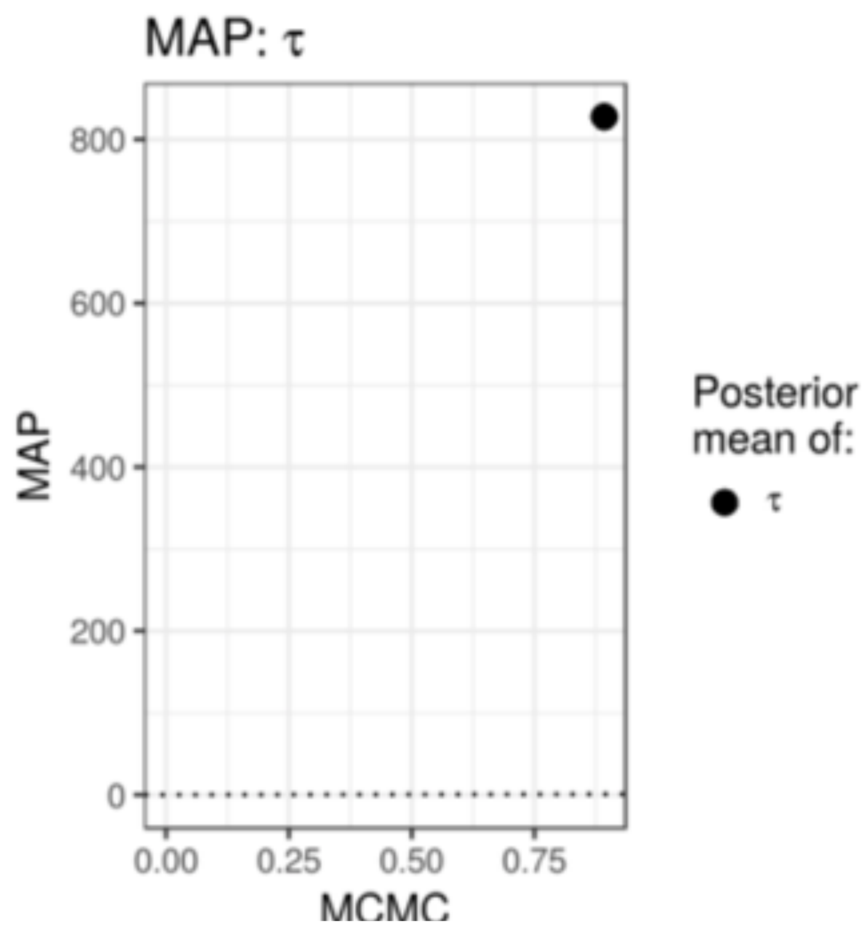
- MAP: **12 s**
- VB: **57 s**

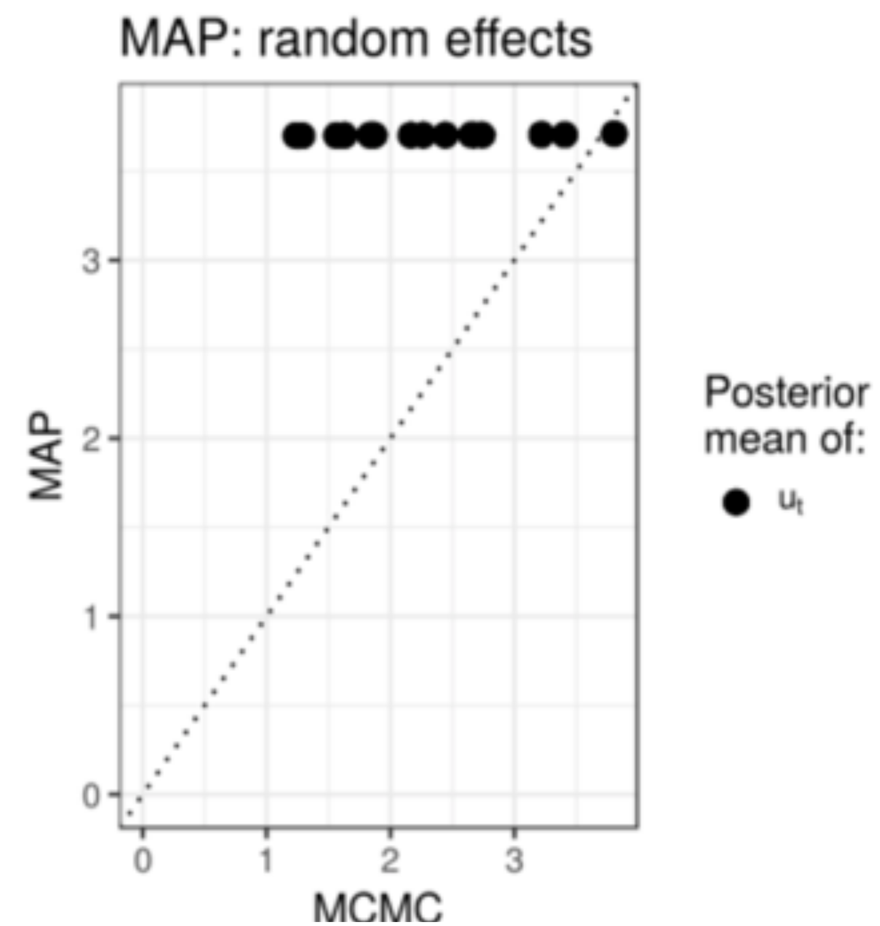


[Giordano, Broderick, Jordan 2017]

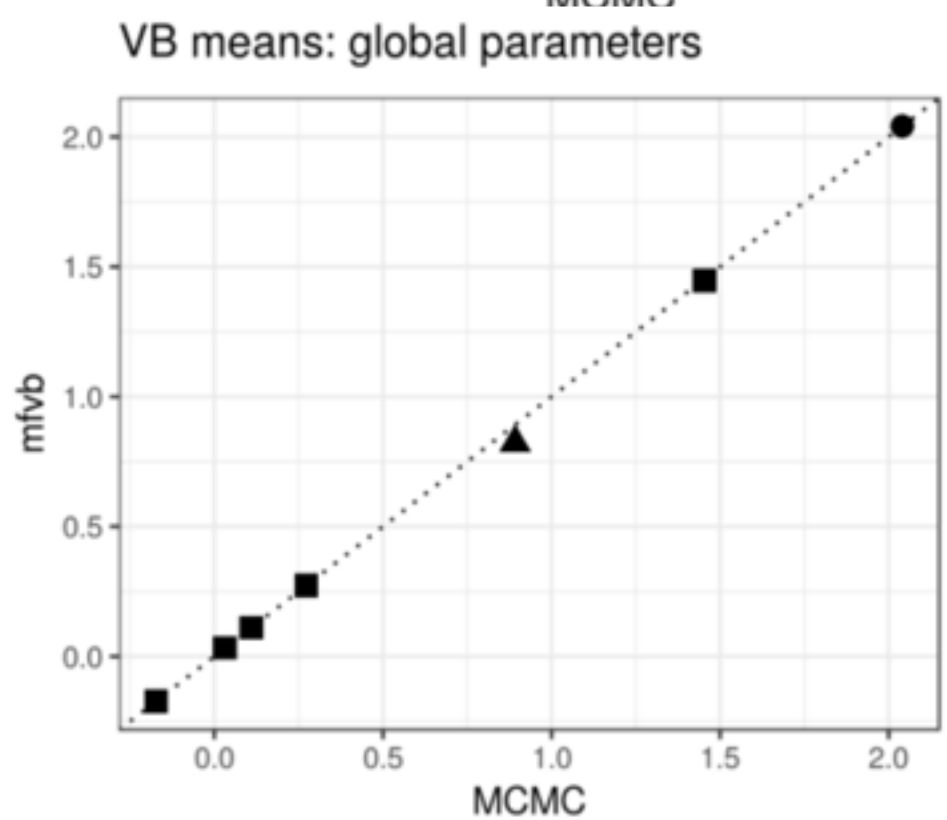
Criteo Online Ads Experiment

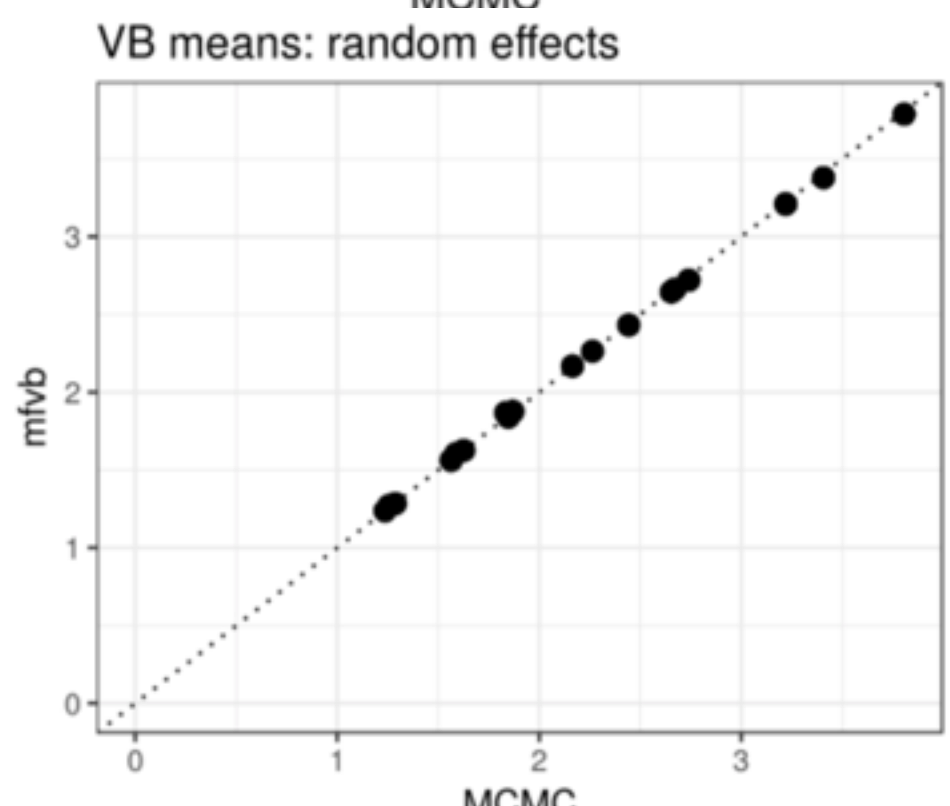






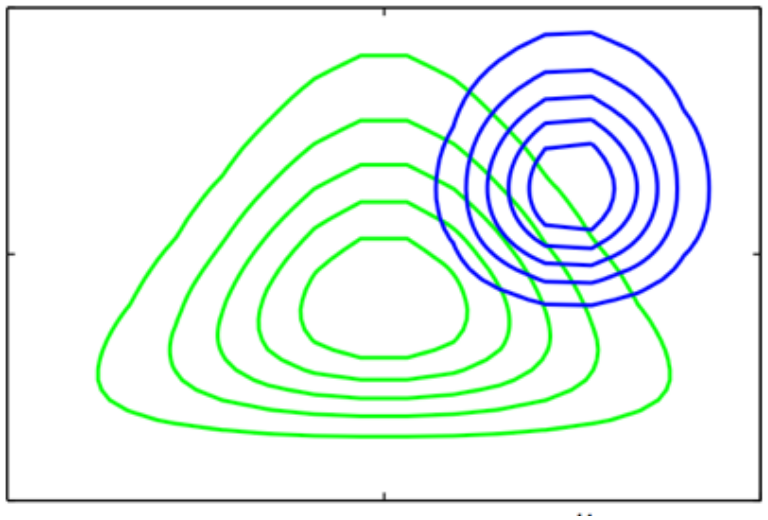
- MAP: **12 s**
- VB: **57 s**
- MCMC (5K samples): 21,066 s (**5.85 h**)



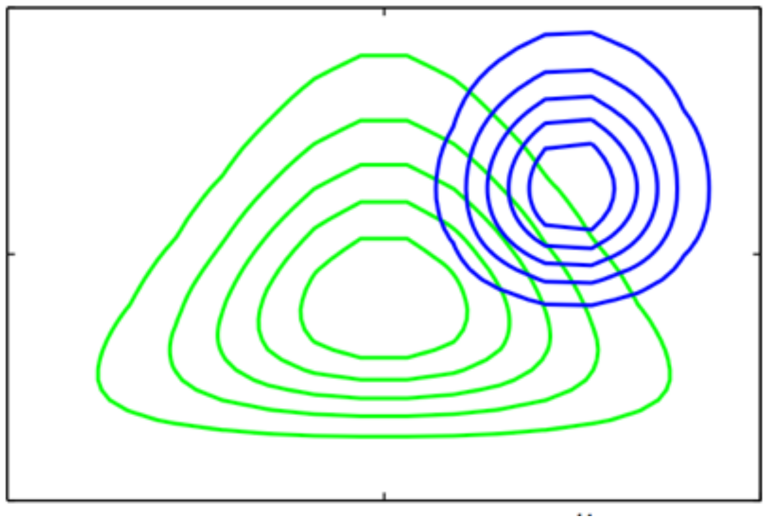


[Giordano, Broderick, Jordan 2017]

How to optimize: MFVB

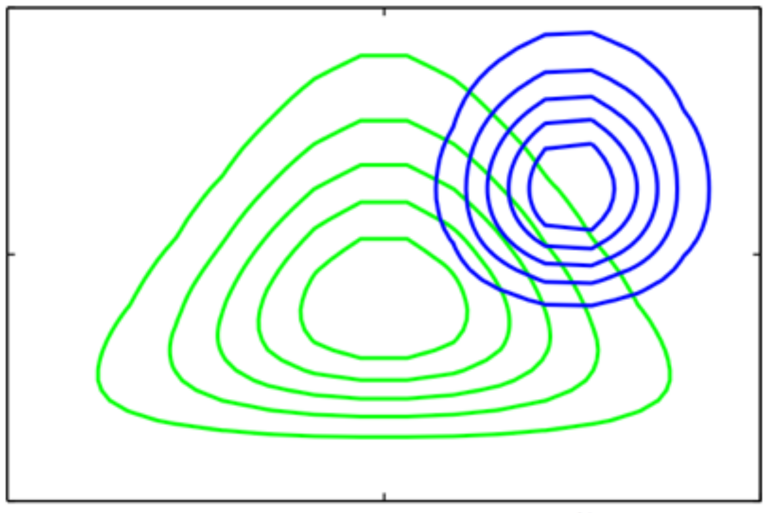


How to optimize: MFVB



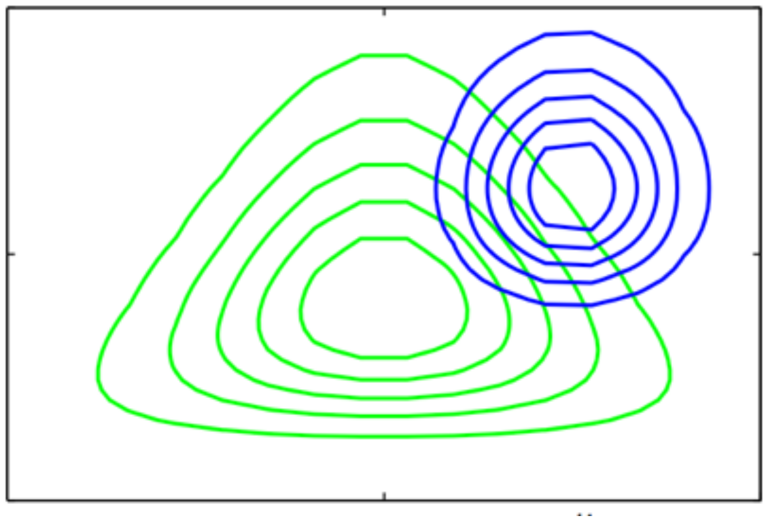
- Conditionally conjugate model
  - Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]

How to optimize: MFVB



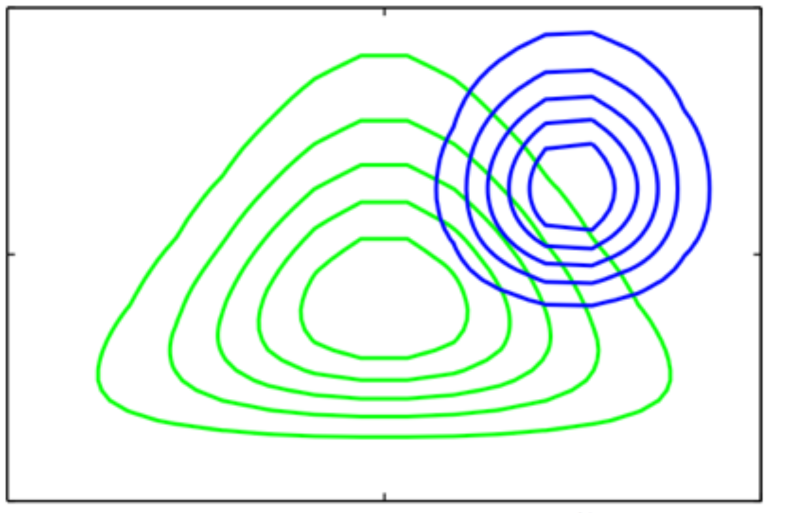
- Conditionally conjugate model
  - Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
- ELBO in closed form

How to optimize: MFVB



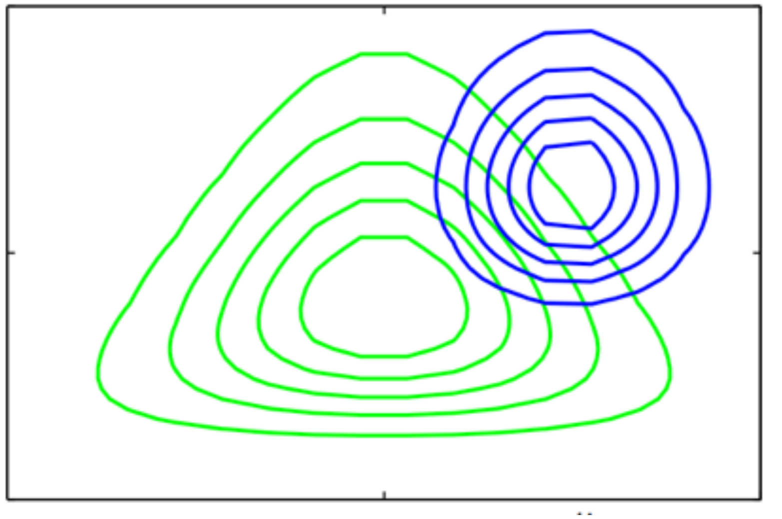
- Conditionally conjugate model
  - Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
- ELBO in closed form
  - E.g. $p(\cdot|y)$ nice enough (e.g. exponential family components) and familiar q (e.g. exp fam)

How to optimize: MFVB



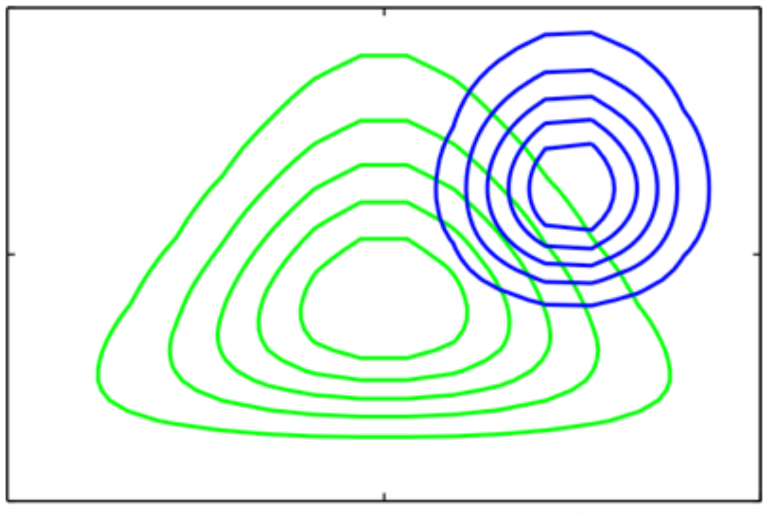
- Conditionally conjugate model
  - Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
- ELBO in closed form
  - E.g. $p(\cdot|y)$ nice enough (e.g. exponential family components) and familiar q (e.g. exp fam)
  - Optimize over variational params (method of choice)

How to optimize: MFVB



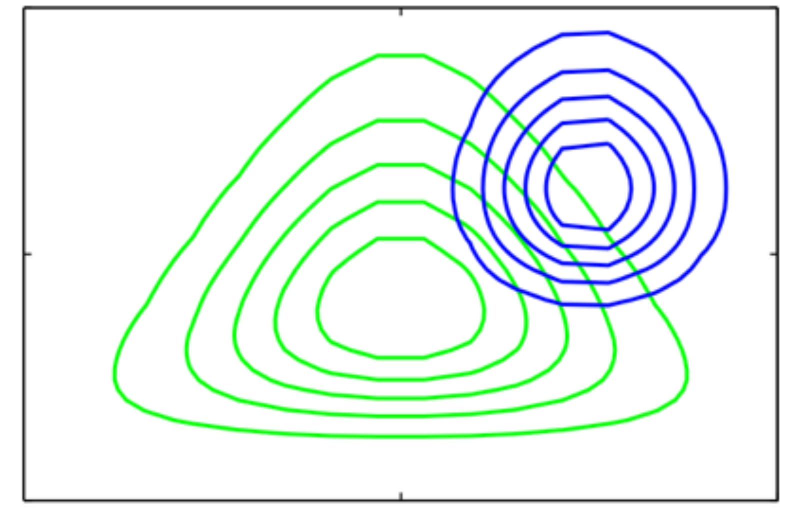
- Conditionally conjugate model
  - Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
- ELBO in closed form
  - E.g. $p(\cdot|y)$ nice enough (e.g. exponential family components) and familiar q (e.g. exp fam)
  - Optimize over variational params (method of choice)
- Continuous parameters

How to optimize: MFVB



- Conditionally conjugate model
  - Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
- ELBO in closed form
  - E.g. $p(\cdot|y)$ nice enough (e.g. exponential family components) and familiar *q* (e.g. exp fam)
  - Optimize over variational params (method of choice)
- Continuous parameters
  - Assume Gaussian *q* (possibly in transformed space)

How to optimize: MFVB

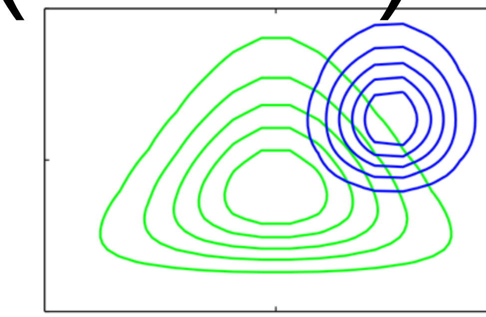


- Conditionally conjugate model
  - Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
- ELBO in closed form
  - E.g. $p(\cdot|y)$ nice enough (e.g. exponential family components) and familiar *q* (e.g. exp fam)
  - Optimize over variational params (method of choice)
- Continuous parameters
  - Assume Gaussian *q* (possibly in transformed space)
  - Automatic differentiation variational inference (ADVI)

[Kucukelbir et al 2015, 2017]

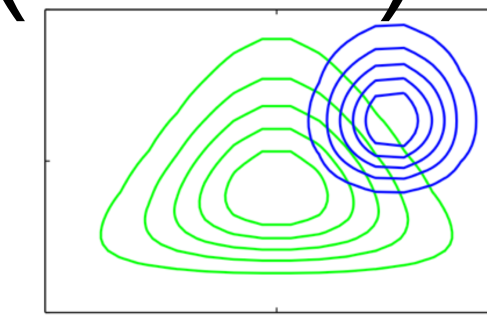
[Baydin et al 2018]

Stochastic gradient descent (SGD)



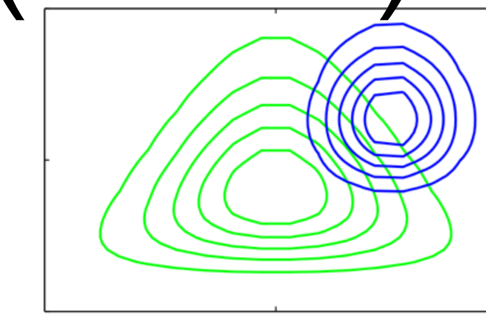
- MFVB: $\min_{\eta : q_\eta \in Q_{\mathrm{MFVB}}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta$

Stochastic gradient descent (SGD)



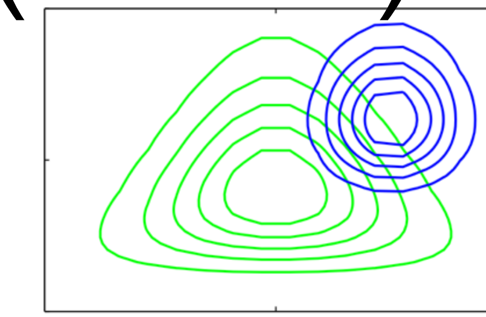
- MFVB: $\min_{\eta: q_\eta \in Q_{\text{MFVB}}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta$
- Recall: Stochastic gradient

Stochastic gradient descent (SGD)



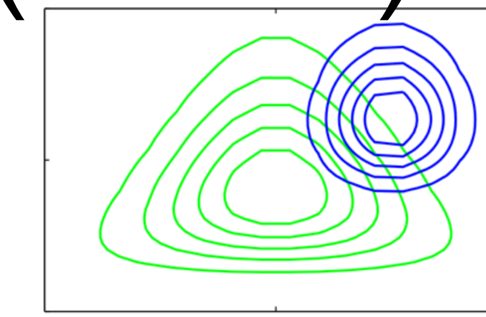
- MFVB: $\min_{\eta : q_\eta \in Q_{\mathrm{MFVB}}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta$
- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$

Stochastic gradient descent (SGD)



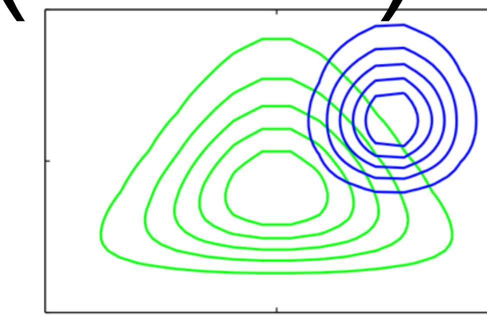
- MFVB: $\min_{\eta : q_\eta \in Q_{\text{MFVB}}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta$
- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \stackrel{iid}{\sim} Y$

Stochastic gradient descent (SGD)



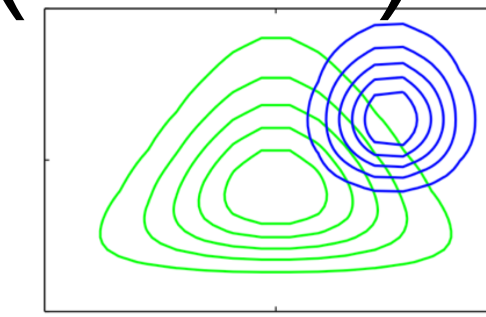
- MFVB: $\min_{\eta: q_\eta \in Q_{\mathrm{MFVB}}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta$
- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$

Stochastic gradient descent (SGD)



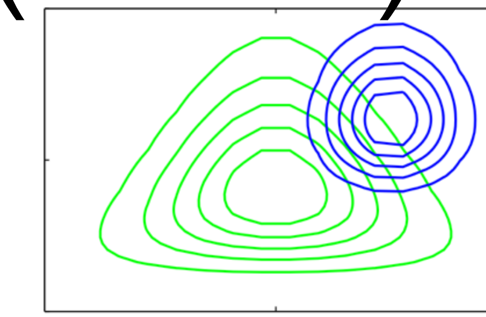
- MFVB: $\min_{\eta : q_\eta \in Q_{\mathrm{MFVB}}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta$
- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^N f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^N f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$

Stochastic gradient descent (SGD)



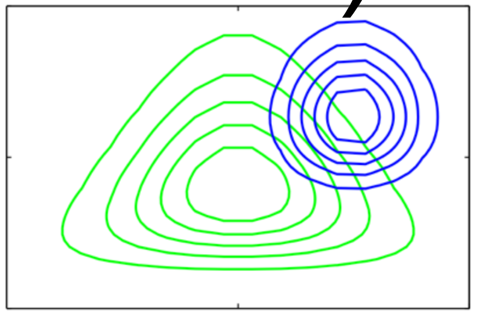
- MFVB: $\min_{\eta : q_\eta \in Q_{\text{MFVB}}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta$
- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^N f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^N f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}}$

Stochastic gradient descent (SGD)



- MFVB: $\min_{\eta: q_\eta \in Q_{\text{MFVB}}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta$
- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^N f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^N f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}}$
- Can we apply SGD to our problem? Assume indep data

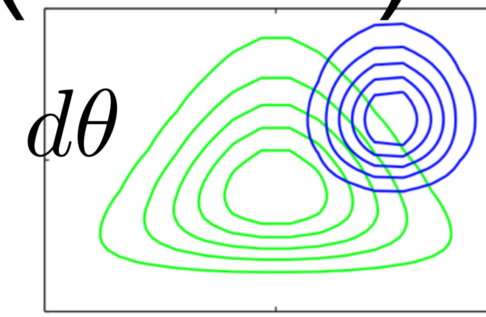
Stochastic gradient descent (SGD)



- MFVB: $\min_{\eta : q_\eta \in Q_{\mathrm{MFVB}}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta$
- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^N f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^N f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}}$
- Observe $y_n \overset{iid}{\sim} Y$
- Can we apply SGD to our problem? Assume indep data

Stochastic gradient descent (SGD)

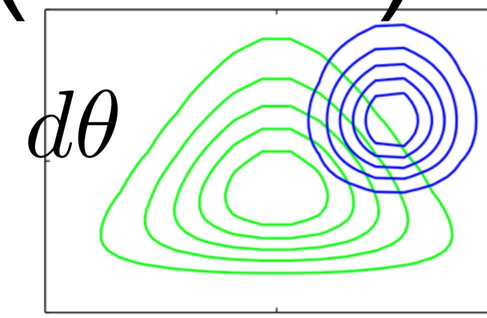
- MFVB: $\min_{\eta : q_\eta \in Q_{\text{MFVB}}} -\mathbb{E}_{q_\eta} \log \left[\prod_{n=1}^{N} p(y_n|\theta) \frac{p(\theta)}{q_\eta(\theta)} \right] d\theta$



- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^N f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^N f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}}$
- Can we apply SGD to our problem? Assume indep data

Stochastic gradient descent (SGD)

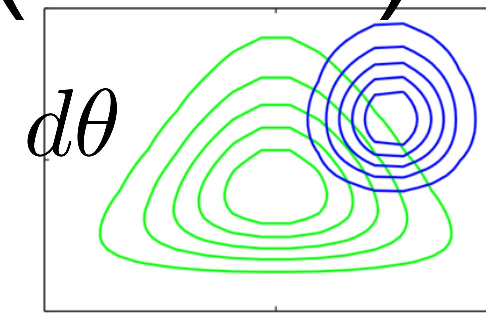
- MFVB: $\min_{\eta : q_\eta \in Q_{\mathrm{MFVB}}} -\mathbb{E}_{q_\eta} \log \left[\prod_{n=1}^{N} p(y_n | \theta) \frac{p(\theta)}{q_\eta(\theta)} \right] d\theta$



- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$
    - One option: $\min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta)$
    - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$
    - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}}$
- Can we apply SGD to our problem? Assume indep data

Stochastic gradient descent (SGD)

- MFVB: $\min_{\eta: q_\eta \in Q_{\text{MFVB}}} -\mathbb{E}_{q_\eta} \log \left[\prod_{n=1}^{N} p(y_n|\theta) \frac{p(\theta)}{q_\eta(\theta)} \right] d\theta$

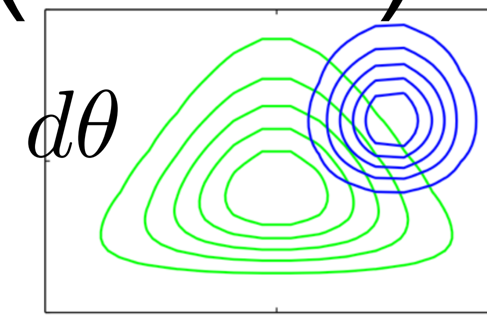


- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}}$
- Can we apply SGD to our problem? Assume indep data

$$\min_\eta N^{-1} \sum_{n=1}^{N} \left[-N \mathbb{E}_{q_\eta} \log p(y_n|\theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta)) \right]$$

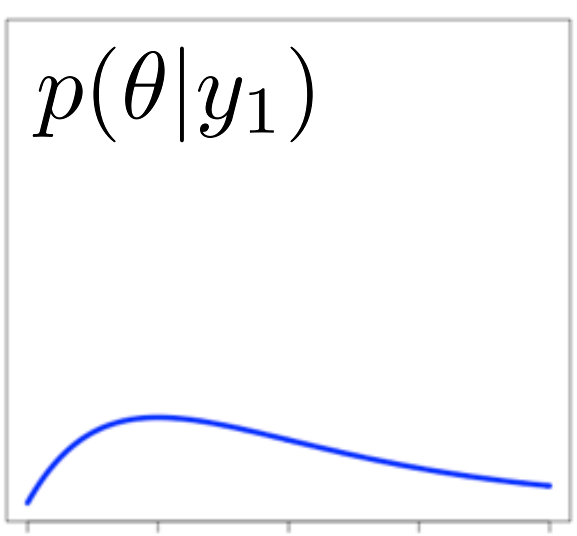
Stochastic gradient descent (SGD)

- MFVB: $\min_{\eta: q_\eta \in Q_{\mathrm{MFVB}}} -\mathbb{E}_{q_\eta} \log \left[\prod_{n=1}^{N} p(y_n|\theta) \frac{p(\theta)}{q_\eta(\theta)} \right] d\theta$



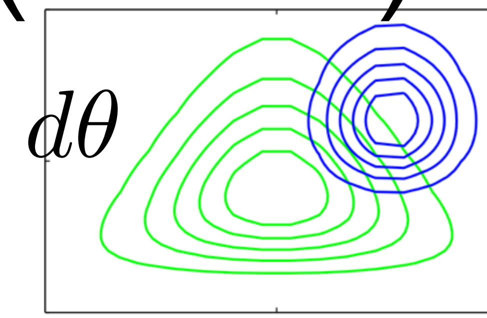
- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}}$
- Can we apply SGD to our problem? Assume indep data

$$\min_\eta N^{-1} \sum_{n=1}^{N} \left[-N \mathbb{E}_{q_\eta} \log p(y_n|\theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta)) \right]$$



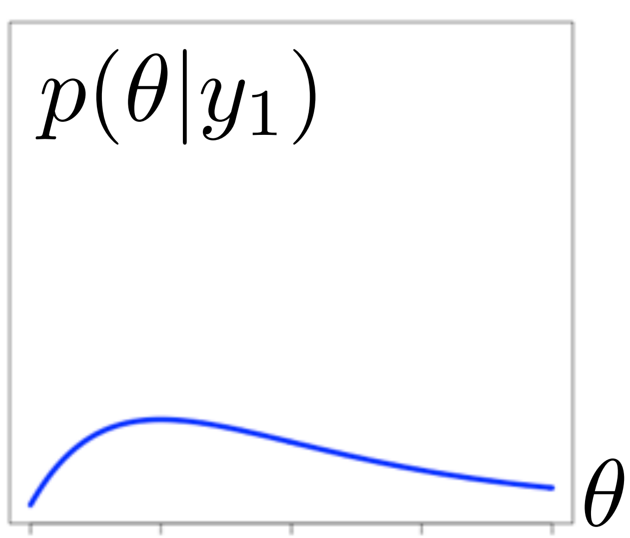
Stochastic gradient descent (SGD)

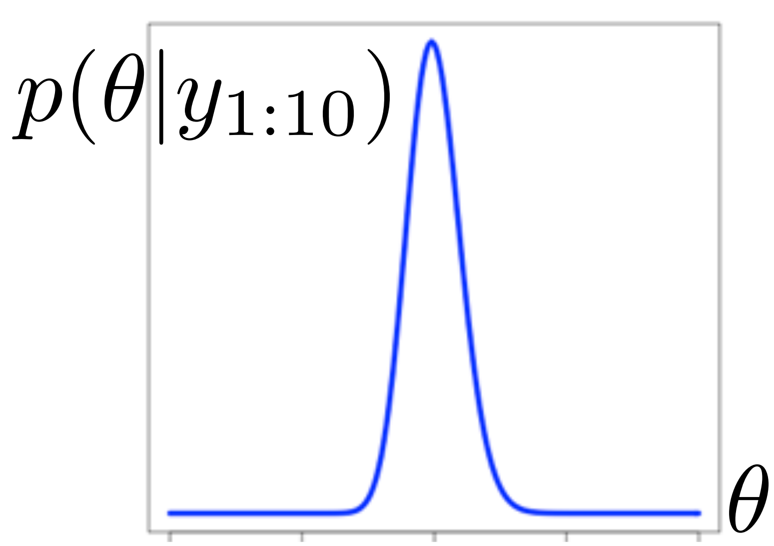
- MFVB: $\min_{\eta: q_\eta \in Q_{\mathrm{MFVB}}} -\mathbb{E}_{q_\eta} \log \left[\prod_{n=1}^{N} p(y_n|\theta) \frac{p(\theta)}{q_\eta(\theta)} \right] d\theta$



- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta=\eta^{(t)}}$
- Can we apply SGD to our problem? Assume indep data

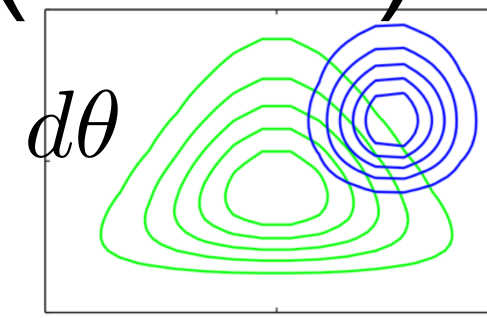
$$\min_\eta N^{-1} \sum_{n=1}^{N} \left[-N \mathbb{E}_{q_\eta} \log p(y_n|\theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta)) \right]$$





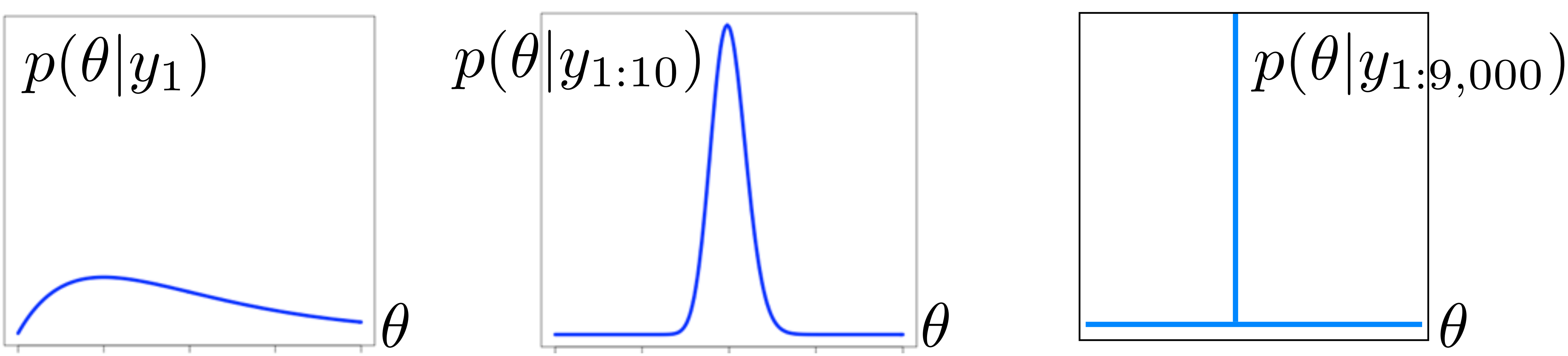
Stochastic gradient descent (SGD)

- MFVB: $\min_{\eta : q_\eta \in Q_{\text{MFVB}}} -\mathbb{E}_{q_\eta} \log \left[\prod_{n=1}^{N} p(y_n|\theta) \frac{p(\theta)}{q_\eta(\theta)} \right] d\theta$



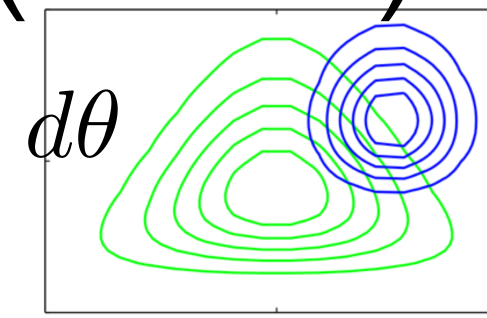
- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \stackrel{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta=\eta^{(t)}}$
- Can we apply SGD to our problem? Assume indep data

$$\min_\eta N^{-1} \sum_{n=1}^{N} \left[-N\mathbb{E}_{q_\eta} \log p(y_n|\theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta)) \right]$$



Stochastic gradient descent (SGD)

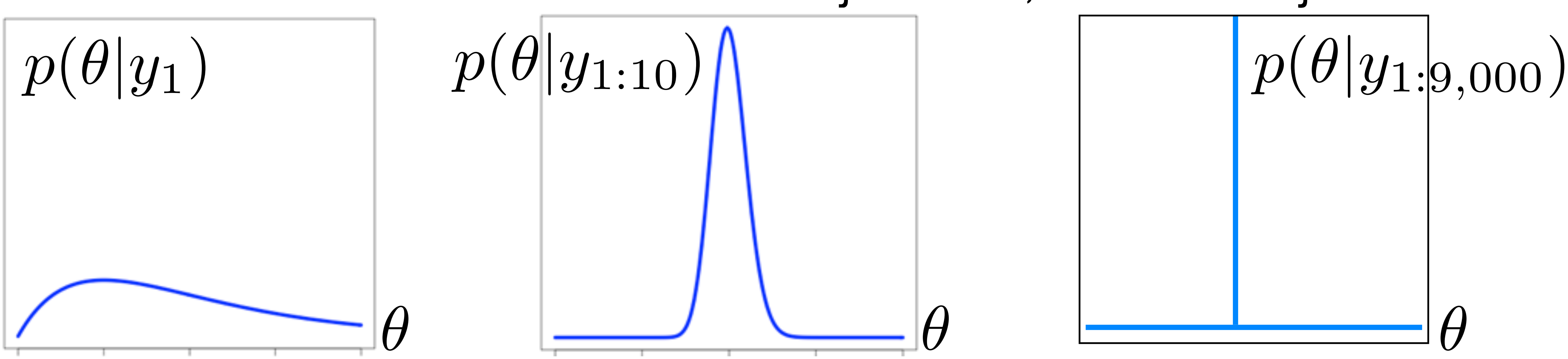
- MFVB: $\min_{\eta: q_\eta \in Q_{\text{MFVB}}} -\mathbb{E}_{q_\eta} \log \left[\prod_{n=1}^{N} p(y_n|\theta) \frac{p(\theta)}{q_\eta(\theta)} \right] d\theta$



- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}}$
- Can we apply SGD to our problem? Assume indep data

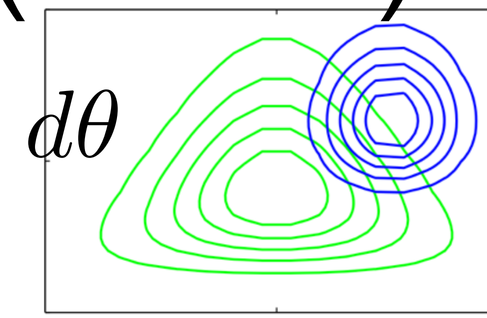
$$\min_\eta N^{-1} \sum_{n=1}^{N} \left[-N \mathbb{E}_{q_\eta} \log p(y_n|\theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta)) \right]$$

  - ELBO not an estimate of the objective; *is* the objective



Stochastic gradient descent (SGD)

- MFVB: $\min_{\eta: q_\eta \in Q_{\text{MFVB}}} -\mathbb{E}_{q_\eta} \log \left[\prod_{n=1}^{N} p(y_n|\theta) \frac{p(\theta)}{q_\eta(\theta)} \right] d\theta$

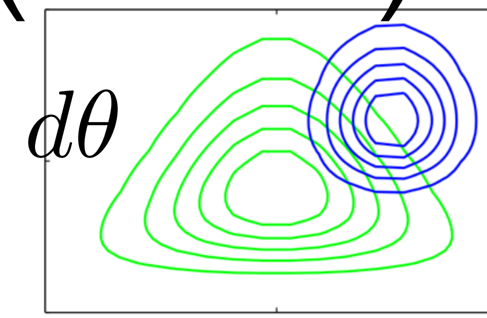


- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}}$
- Can we apply SGD to our problem? Assume indep data

  $\min_\eta N^{-1} \sum_{n=1}^{N} \left[-N \mathbb{E}_{q_\eta} \log p(y_n|\theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta)) \right]$
  - ELBO not an estimate of the objective; *is* the objective

Stochastic gradient descent (SGD)

- MFVB: $\min_{\eta : q_\eta \in Q_{\mathrm{MFVB}}} -\mathbb{E}_{q_\eta} \log \left[\prod_{n=1}^{N} p(y_n | \theta) \frac{p(\theta)}{q_\eta(\theta)} \right] d\theta$



- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}}$
- Can we apply SGD to our problem? Assume indep data

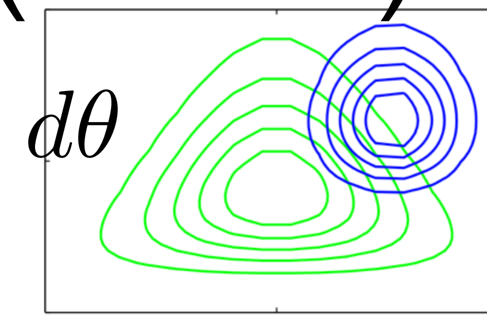
$$\min_\eta N^{-1} \sum_{n=1}^{N} \left[-N \mathbb{E}_{q_\eta} \log p(y_n | \theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta)) \right]$$

  - ELBO not an estimate of the objective; *is* the objective

$$f(\eta) = \mathbb{E}_Y f(Y, \eta) = -\mathbb{E}_{q_\eta} \log(p(\theta, y_{1:N})/q_\eta(\theta))$$

Stochastic gradient descent (SGD)

- MFVB: $\min_{\eta : q_\eta \in Q_{\text{MFVB}}} -\mathbb{E}_{q_\eta} \log \left[\prod_{n=1}^N p(y_n|\theta) \frac{p(\theta)}{q_\eta(\theta)} \right] d\theta$



- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \overset{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^N f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^N f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}}$
- Can we apply SGD to our problem? Assume indep data

$$\min_\eta N^{-1} \sum_{n=1}^N \left[-N \mathbb{E}_{q_\eta} \log p(y_n|\theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta)) \right]$$

  - ELBO not an estimate of the objective; *is* the objective

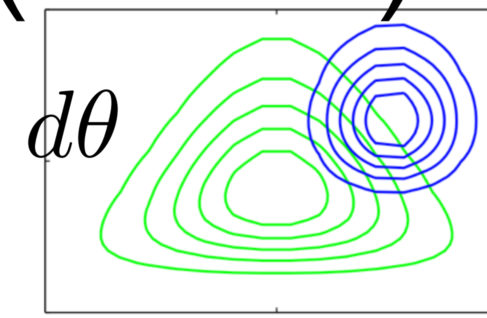
$$f(\eta) = \mathbb{E}_Y f(Y, \eta) = -\mathbb{E}_{q_\eta} \log(p(\theta, y_{1:N})/q_\eta(\theta))$$

$$f(\cdot, \eta) = -N \mathbb{E}_{q_\eta} \log p(\cdot|\theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta))$$

$$Y = y_n \text{ w.p. } 1/N, \quad n \in \{1, \ldots, N\}$$

Stochastic gradient descent (SGD)

- MFVB: $\min_{\eta: q_\eta \in Q_{\mathrm{MFVB}}} -\mathbb{E}_{q_\eta} \log \left[\prod_{n=1}^{N} p(y_n|\theta) \frac{p(\theta)}{q_\eta(\theta)} \right] d\theta$



- Recall: Stochastic gradient
- Goal: $\min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta)$
- Observe $y_n \stackrel{iid}{\sim} Y$
  - One option: $\min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta)$
  - GD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta = \eta^{(t)}}$
  - SGD: $\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[\nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}}$
- Can we apply SGD to our problem? Assume indep data

$$\min_\eta N^{-1} \sum_{n=1}^{N} \left[-N \mathbb{E}_{q_\eta} \log p(y_n|\theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta)) \right]$$

  - ELBO not an estimate of the objective; *is* the objective

$$f(\eta) = \mathbb{E}_Y f(Y, \eta) = -\mathbb{E}_{q_\eta} \log(p(\theta, y_{1:N})/q_\eta(\theta))$$

$$f(\cdot, \eta) = -N \mathbb{E}_{q_\eta} \log p(\cdot|\theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta))$$

$$Y = y_n \text{ w.p. } 1/N, \quad n \in \{1, \ldots, N\}$$

  - Stochastic variational inference [Hoffman et al 2013]

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

What about uncertainty?

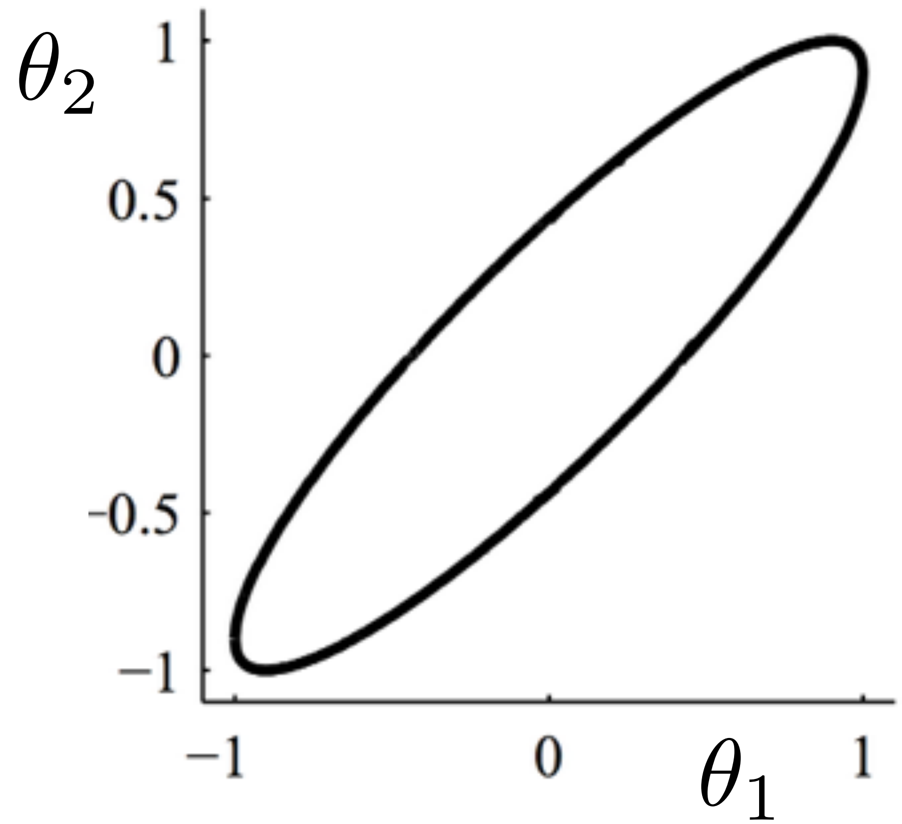
$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

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

What about uncertainty?

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

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

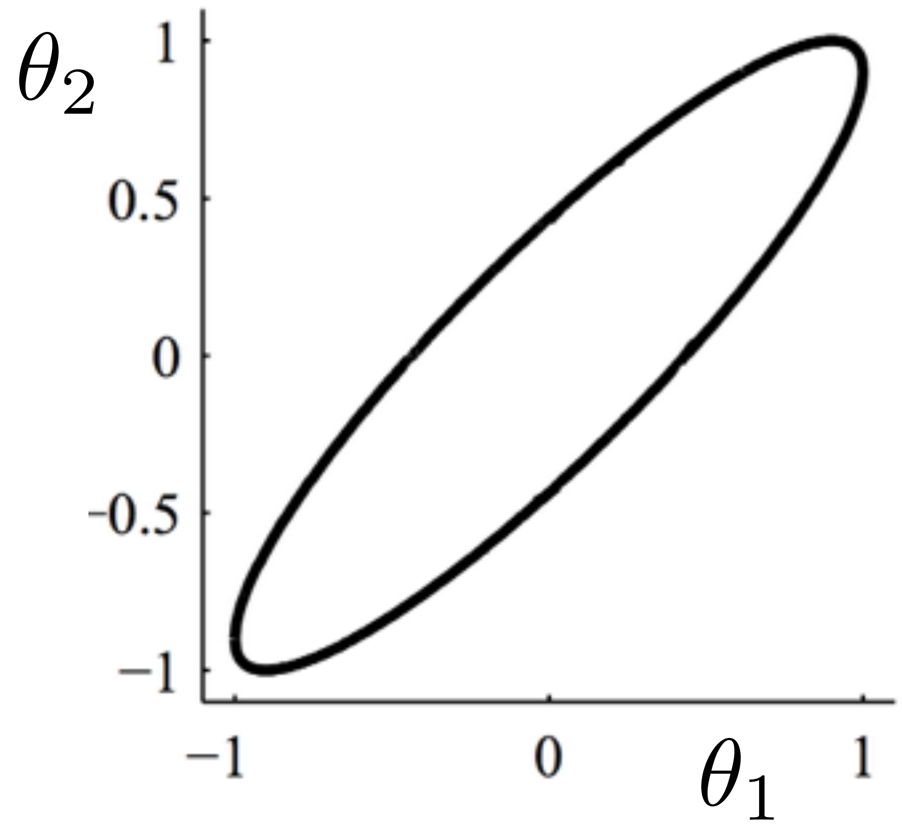


[Turner & Sahani 2011]

What about uncertainty?

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

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



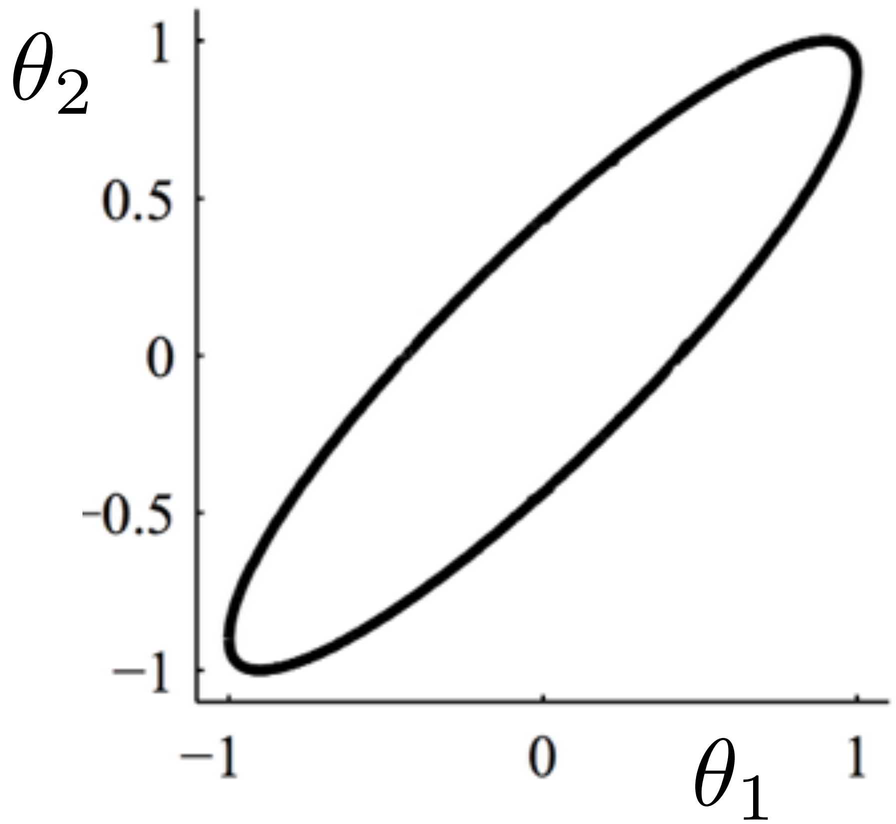
[Turner & Sahani 2011]

- Conjugate linear regression

What about uncertainty?

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

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

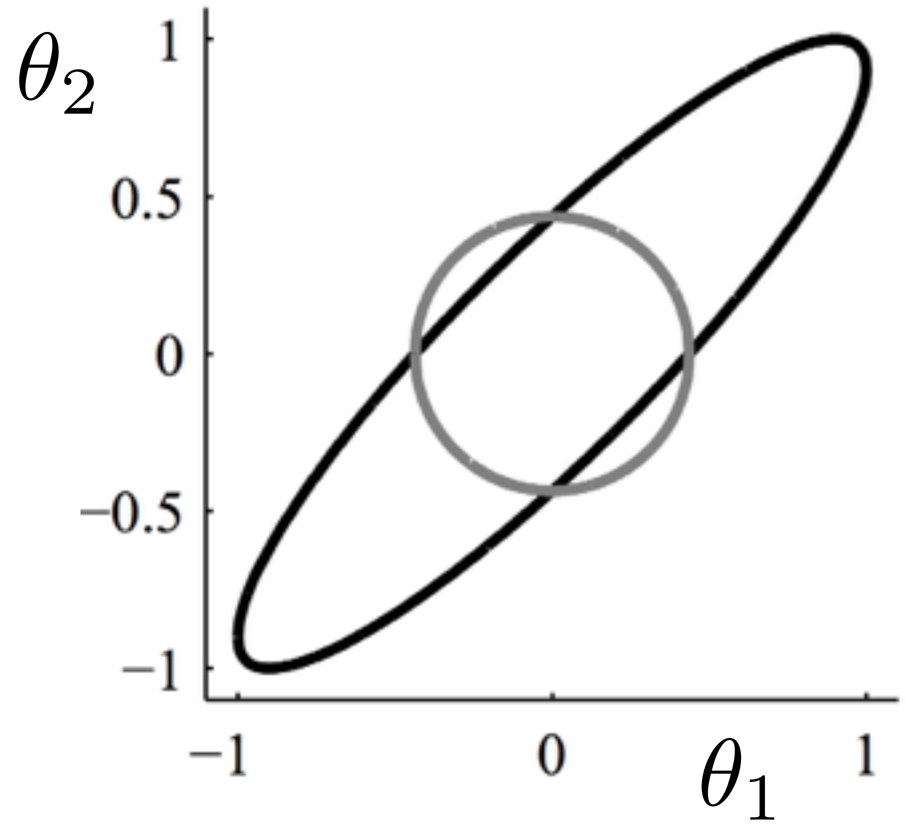


[Turner & Sahani 2011]

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta \qquad q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$



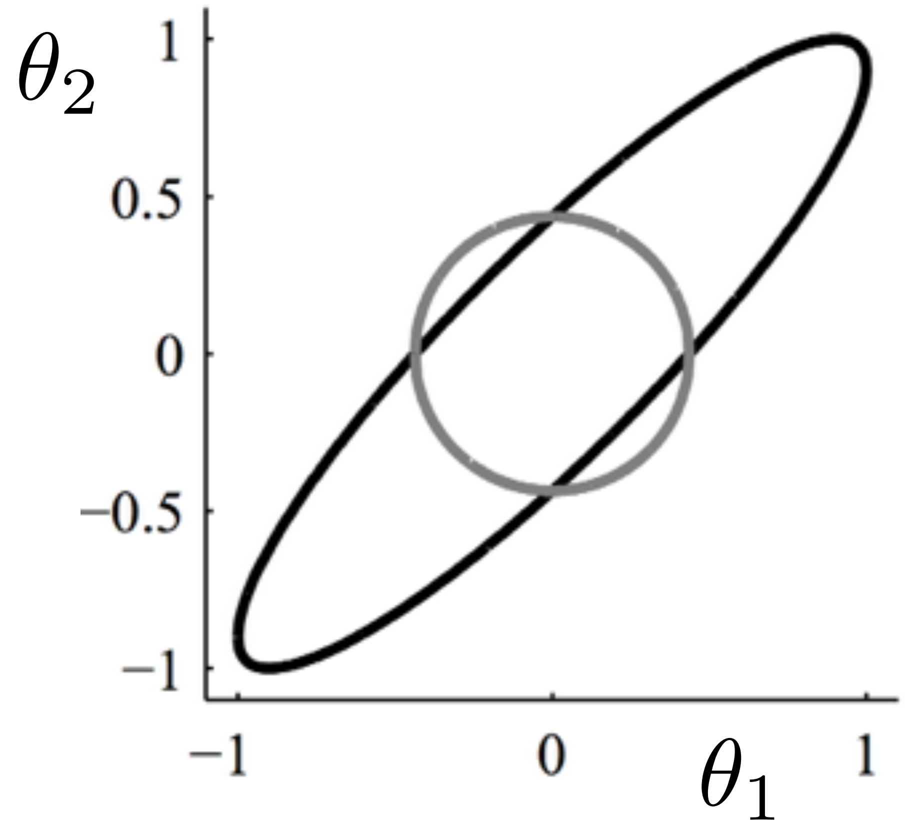
[Turner & Sahani 2011]

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

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

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



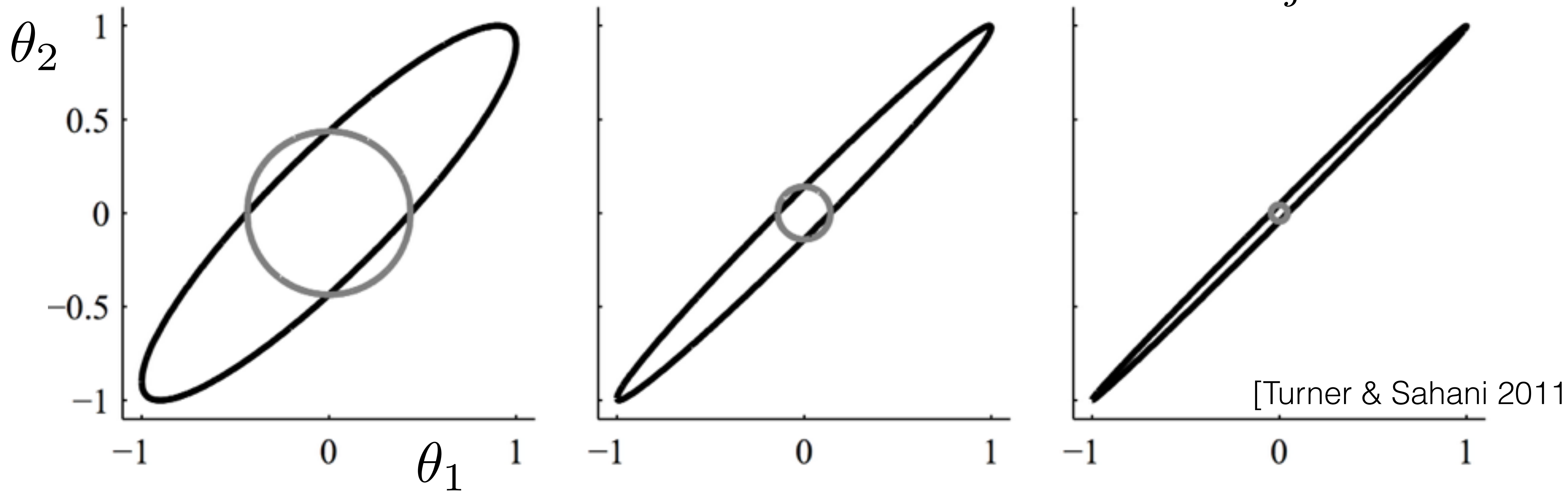
[Turner & Sahani 2011]

- Underestimates variance (sometimes severely)

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta \qquad q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$

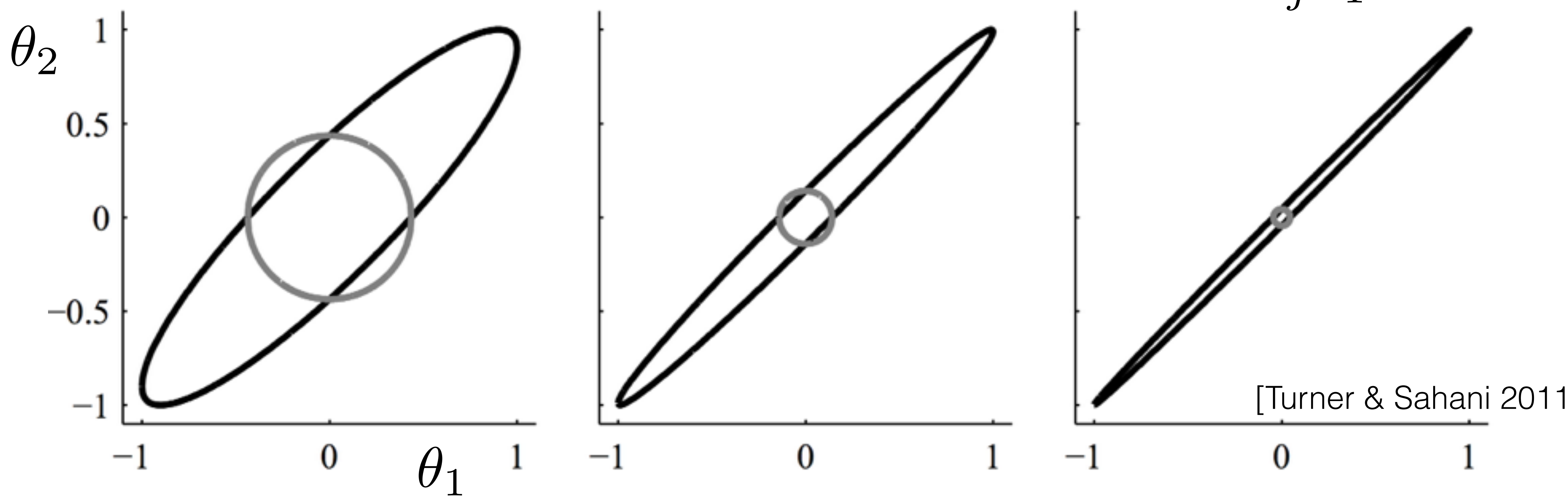


- Underestimates variance (sometimes severely)

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta \qquad q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$

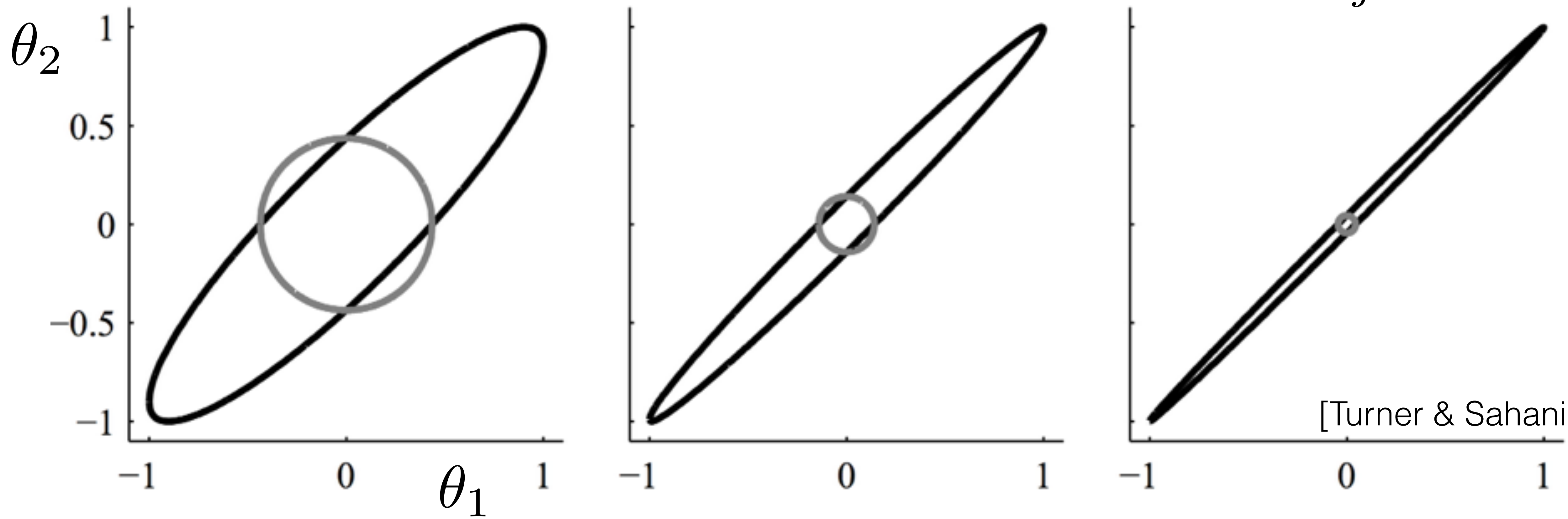


- Underestimates variance (sometimes severely)
- No covariance estimates

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta \qquad q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$



[Turner & Sahani 2011]

- Underestimates variance (sometimes severely)
- No covariance estimates

- Conjugate linear regression
- Bayesian central limit theorem
- Exercise: derive exact (closed) form of q^*

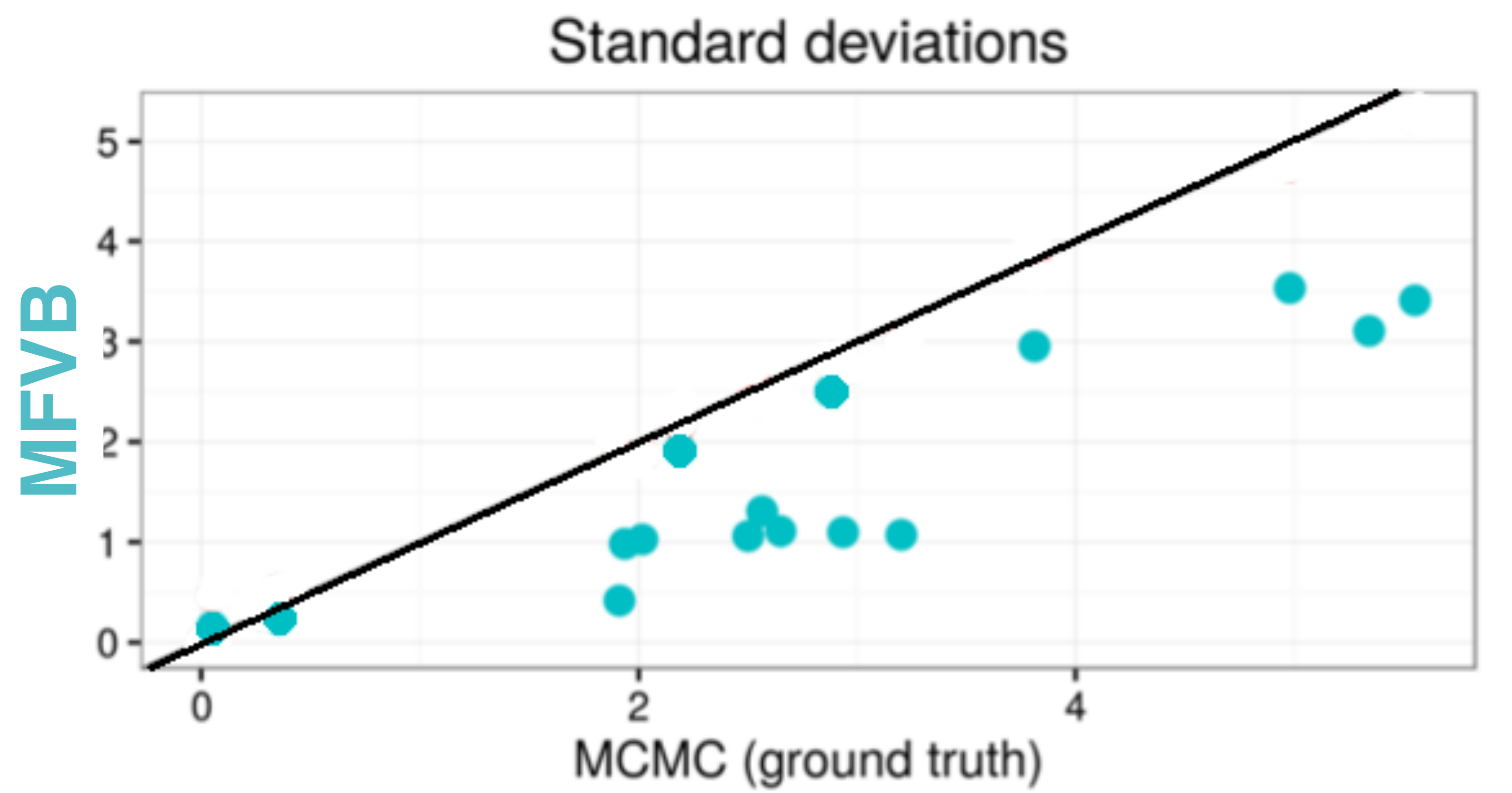
What about uncertainty?

- Microcredit

[Giordano, Broderick, Meager, Huggins, Jordan 2016]

What about uncertainty?

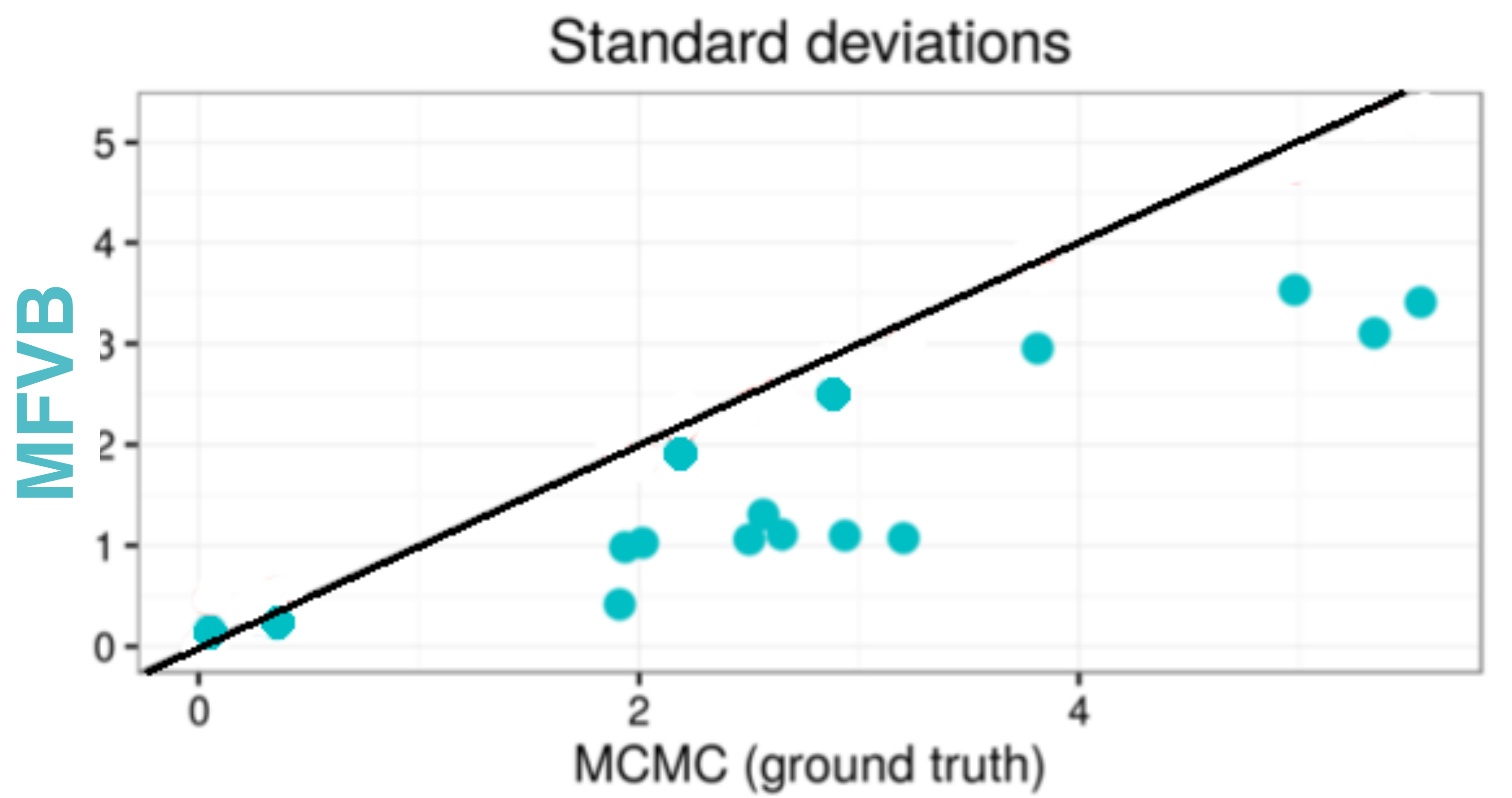
- Microcredit



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

What about uncertainty?

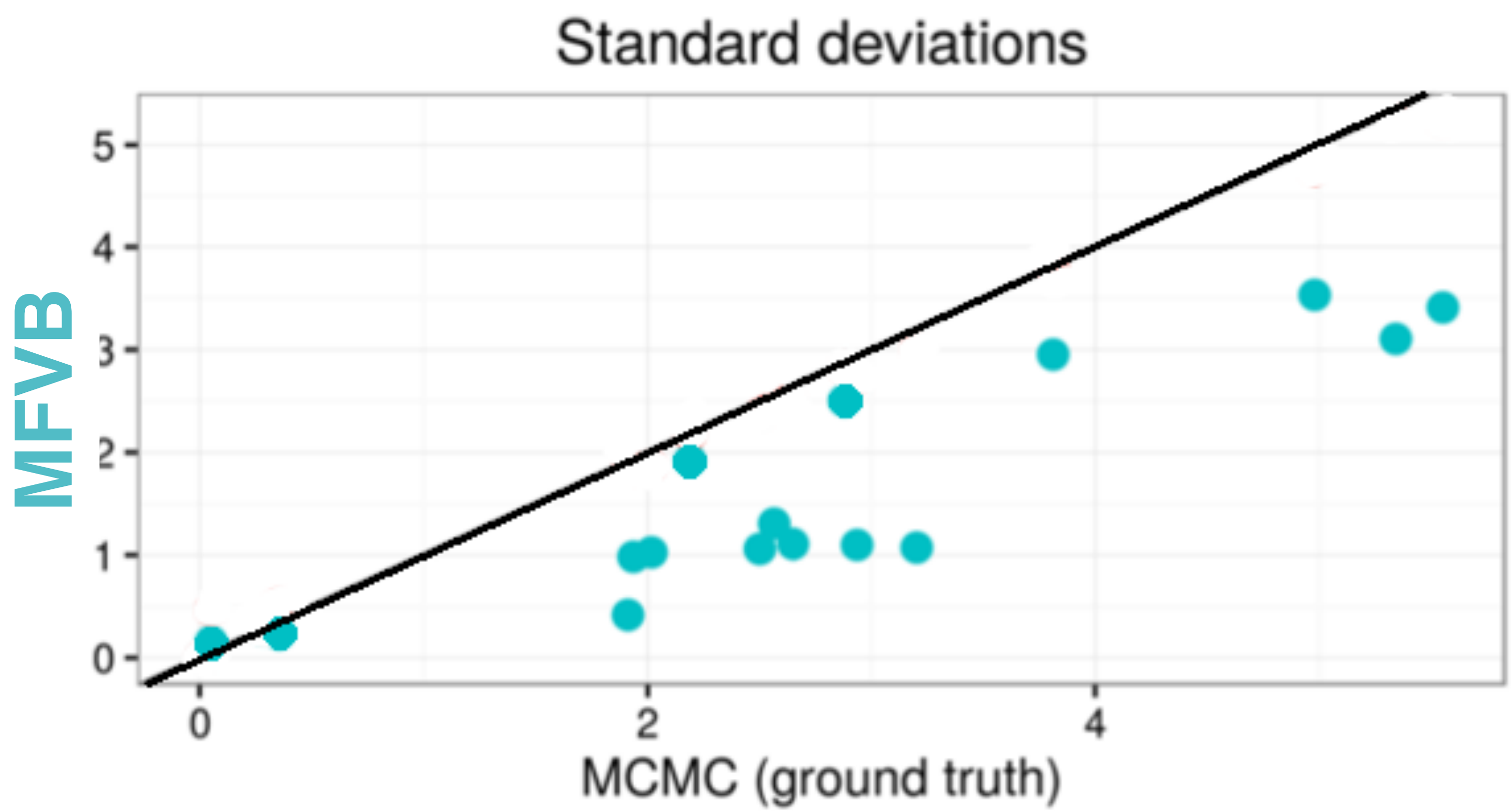
- Microcredit effect
- τ mean: 3.08 USD PPP



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

What about uncertainty?

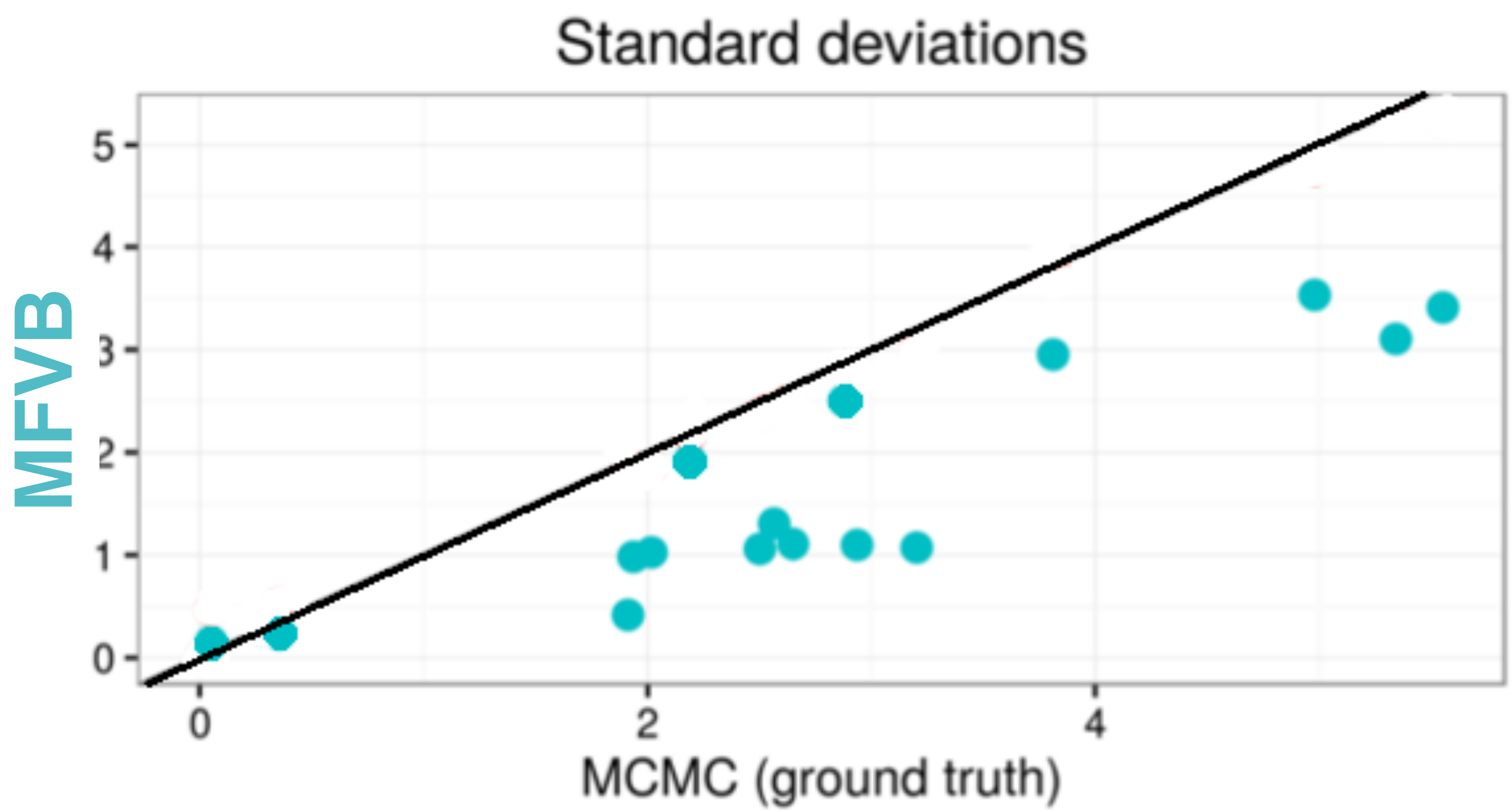
- Microcredit effect
- τ mean: 3.08 USD PPP
- τ std dev: 1.83 USD PPP



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

What about uncertainty?

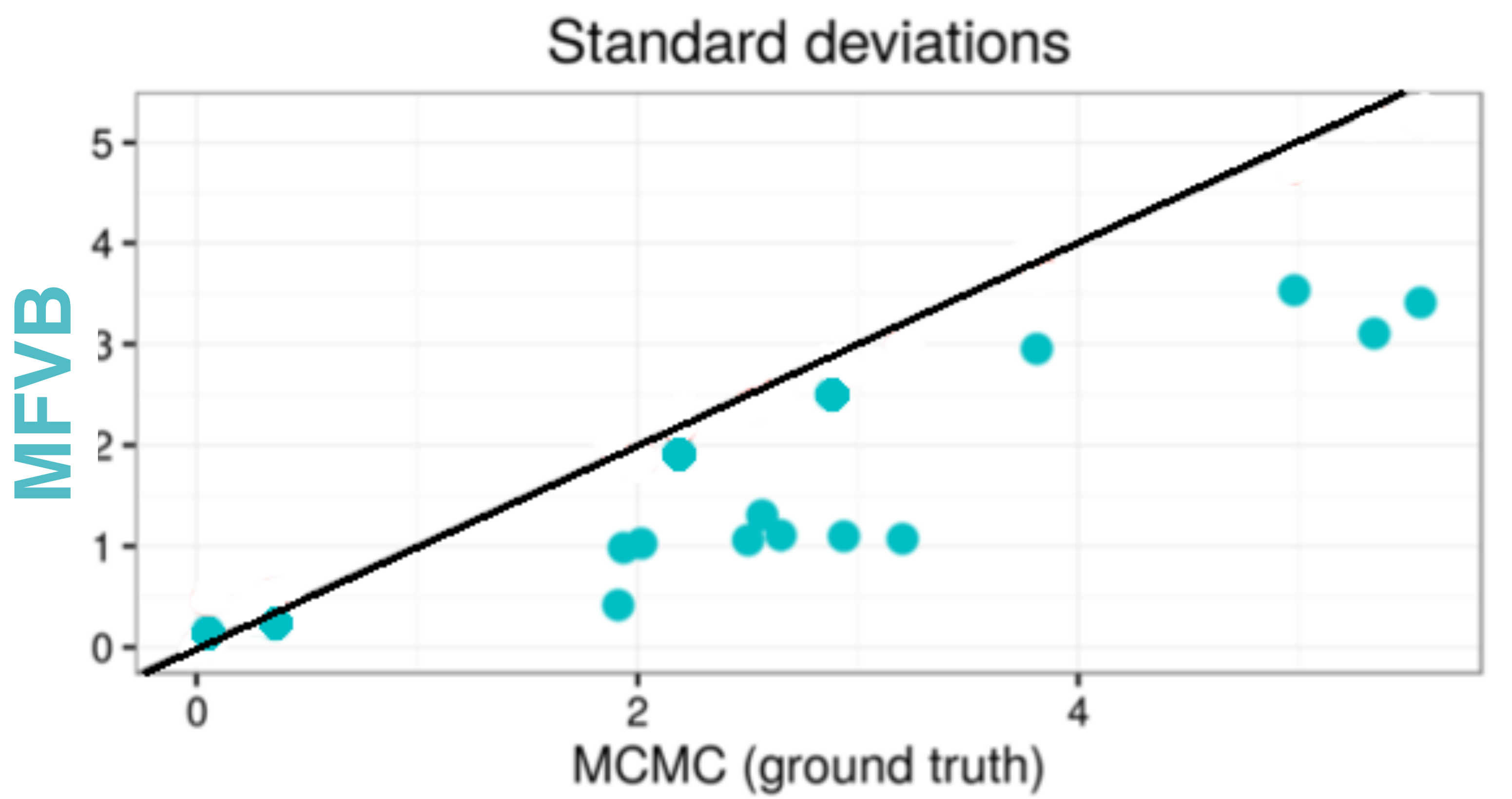
- Microcredit effect
- τ mean: 3.08 USD PPP
- τ std dev: 1.83 USD PPP
- Mean is 1.68 std dev from 0



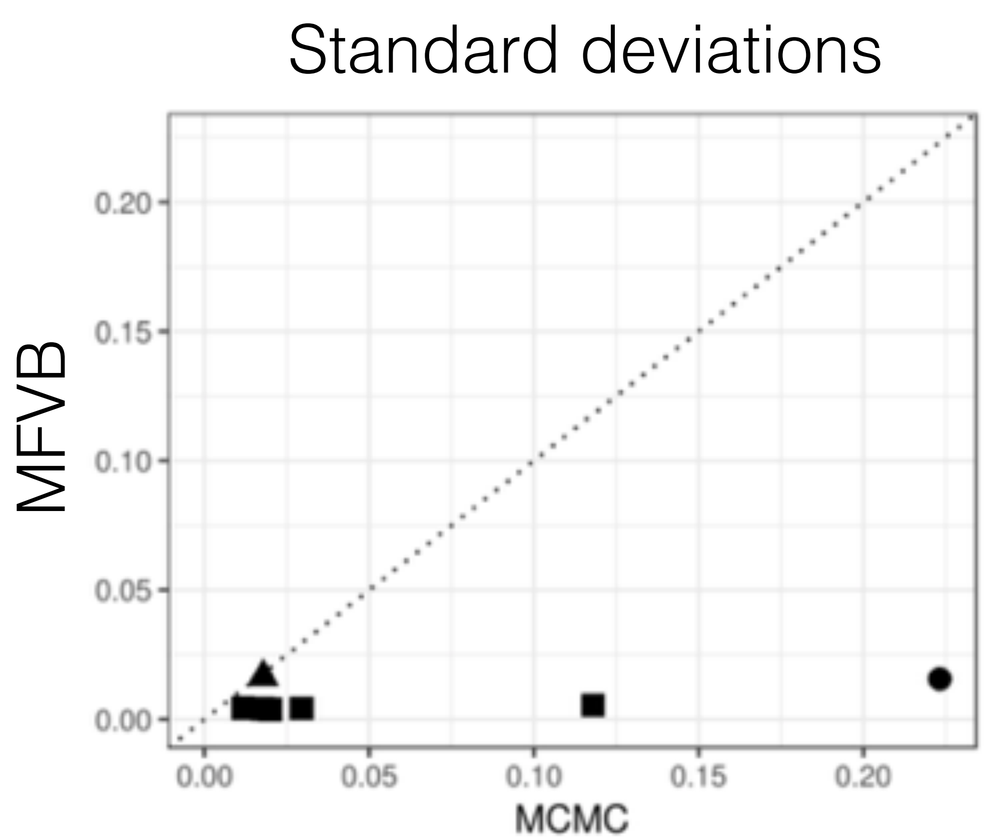
[Giordano, Broderick, Meager, Huggins, Jordan 2016]

What about uncertainty?

- Microcredit effect
- τ mean: 3.08 USD PPP
- τ std dev: 1.83 USD PPP
- Mean is 1.68 std dev from 0



- Criteo online ads experiment



[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2017]

Posterior means: revisited

- Want to predict college GPA y_n

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model: $y_n | \beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model: $y_n | \beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$

$$z_k | \rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2) \qquad (\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$$

$$\beta \sim \mathcal{N}(0, \Sigma) \qquad (\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$$

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model:

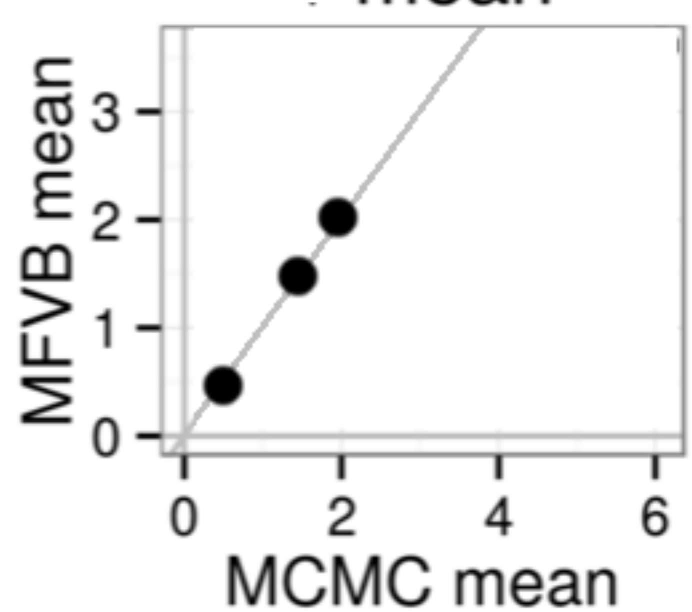
$$y_n | \beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$$

$$z_k | \rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2) \qquad (\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$$

$$\beta \sim \mathcal{N}(0, \Sigma) \qquad (\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$$

- Data simulated from model (3 data sets, 300 data points):

$(\rho^2)^{-1}$ mean



[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

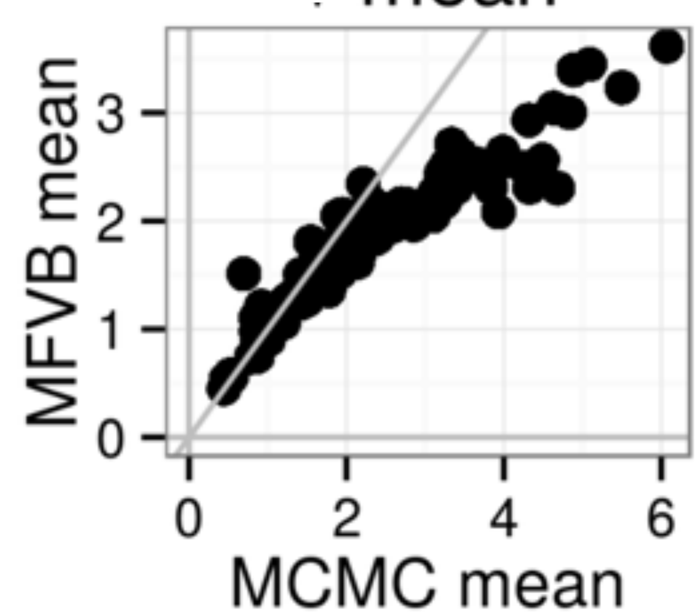
- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model: $y_n|\beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$

$$z_k|\rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2) \qquad (\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$$

$$\beta \sim \mathcal{N}(0, \Sigma) \qquad (\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$$

- Data simulated from model (100 data sets, 300 data points):

$(\rho^2)^{-1}$ mean



[Giordano, Broderick, Jordan 2015]

What can we do?

- Good evaluation methods

What can we do?

- Good evaluation methods

  [Gorham, Mackey 2015, 2017]

  [Talts et al 2018]

  [Yao et al 2018]

  etc.

What can we do?

- Good evaluation methods

    [Gorham, Mackey 2015, 2017]

    [Talts et al 2018]

    [Yao et al 2018]

    etc.

- A correction to MFVB (Part II)

    [Giordano, Broderick, Jordan 2015, 2017; Giordano, Broderick, Meager, Huggins, Jordan 2016]

What can we do?

- Good evaluation methods

  [Gorham, Mackey 2015, 2017]

  [Talts et al 2018]

  [Yao et al 2018]

  etc.

- A correction to MFVB (Part II)
  - Also VB & robustness quantification

  [Giordano, Broderick, Jordan 2015, 2017; Giordano, Broderick, Meager, Huggins, Jordan 2016]

What can we do?

- Good evaluation methods

  [Gorham, Mackey 2015, 2017]

  [Talts et al 2018]

  [Yao et al 2018]

  etc.

- A correction to MFVB (Part II)
  - Also VB & robustness quantification

  [Giordano, Broderick, Jordan 2015, 2017; Giordano, Broderick, Meager, Huggins, Jordan 2016]

- Data summarization for scalability (Part III)

  [Campbell, Broderick 2017, 2018]

  [Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick 2017]

References (1/6)

R Bardenet, A Doucet, and C Holmes. "On Markov chain Monte Carlo methods for tall data." *The Journal of Machine Learning Research* 18.1 (2017): 1515-1557.

AG Baydin, BA Pearlmutter, AA Radul, and JM Siskind. "Automatic differentiation in machine learning: a survey." ArXiv:1502.05767v4 (2018).

DM Blei, A Kucukelbir, and JD McAuliffe. "Variational inference: A review for statisticians." *Journal of the American Statistical Association* 112.518 (2017): 859-877.

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

CM Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

T Campbell and T Broderick. Automated scalable Bayesian inference via Hilbert coresets. Under review. ArXiv:1710.05053.

T Campbell and T Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. *ICML* 2018, to appear.

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

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

R Giordano, T Broderick, and MI Jordan. Covariances, robustness, and variational Bayes, 2017. Under review. ArXiv:1709.02536.

References (2/6)

J Gorham and L Mackey. "Measuring sample quality with Stein's method." *NIPS* 2015.

J Gorham, and L Mackey. "Measuring sample quality with kernels." ArXiv:1703.01717 (2017).

PD Hoff. *A first course in Bayesian statistical methods*. Springer Science & Business Media, 2009.

MD Hoffman, DM Blei, C Wang, and J Paisley. "Stochastic variational inference." *The Journal of Machine Learning Research* 14.1 (2013): 1303-1347.

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

JH Huggins, RP Adams, and T Broderick. PASS-GLM: Polynomial approximate sufficient statistics for scalable Bayesian GLM inference. *NIPS* 2017.

A Kucukelbir, R Ranganath, A Gelman, and D Blei. "Automatic variational inference in Stan." *NIPS* 2015.

A Kucukelbir, D Tran, R Ranganath, A Gelman, and DM Blei. "Automatic differentiation variational inference." *The Journal of Machine Learning Research* 18.1 (2017): 430-474.

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

Stan (open source software). http://mc-stan.org/ Accessed: 2018.

S Talts, M Betancourt, D Simpson, A Vehtari, and A Gelman. "Validating Bayesian Inference Algorithms with Simulation-Based Calibration." aArXiv:1804.06788 (2018).

References (3/6)

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.

Y Yao, A Vehtari, D Simpson, and A Gelman. "Yes, but Did It Work?: Evaluating Variational Inference." ArXiv:1802.02538 (2018).

Application References (4/6)

Abbott, Benjamin P., et al. "Observation of gravitational waves from a binary black hole merger." *Physical Review Letters* 116.6 (2016): 061102.

Abbott, Benjamin P., et al. "The rate of binary black hole mergers inferred from advanced LIGO observations surrounding GW150914." *The Astrophysical Journal Letters* 833.1 (2016): L1.

Airoldi, Edoardo M., David M. Blei, Stephen E. Fienberg, and Eric P. Xing. "Mixed membership stochastic blockmodels." *Journal of Machine Learning Research* 9.Sep (2008): 1981-2014.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation." *Journal of Machine Learning Research* 3.Jan (2003): 993-1022.

Chati, Yashovardhan Sushil, and Hamsa Balakrishnan. "A Gaussian process regression approach to model aircraft engine fuel flow rate." *Cyber-Physical Systems (ICCPS), 2017 ACM/IEEE 8th International Conference on*. IEEE, 2017.

Gershman, Samuel J., David M. Blei, Kenneth A. Norman, and Per B. Sederberg. "Decomposing spatiotemporal brain patterns into topographic latent sources." NeuroImage 98 (2014): 91-102.

Gillon, Michaël, et al. "Seven temperate terrestrial planets around the nearby ultracool dwarf star TRAPPIST-1." *Nature* 542.7642 (2017): 456.

Grimm, Simon L., et al. "The nature of the TRAPPIST-1 exoplanets." *Astronomy & Astrophysics* 613 (2018): A68.

Application References (5/6)

Grogan Jr, William L., and Willis W. Wirth. "A new American genus of predaceous midges related to Palpomyia and Bezzia (Diptera: Ceratopogonidae). Un nuevo género Americano de purrujas depredadoras relacionadas con Palpomyia y Bezzia (Diptera: Ceratopogonidae)." *Proceedings of the Biological Society of Washington*. 94.4 (1981): 1279-1305.

Kuikka, Sakari, Jarno Vanhatalo, Henni Pulkkinen, Samu Mäntyniemi, and Jukka Corander. "Experiences in Bayesian inference in Baltic salmon management." *Statistical Science* 29.1 (2014): 42-49.

Meager, Rachael. "Understanding the impact of microcredit expansions: A Bayesian hierarchical analysis of 7 randomized experiments." *AEJ: Applied*, to appear, 2018a.

Meager, Rachael. "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature." Working paper, 2018b.

Stegle, Oliver, Leopold Parts, Richard Durbin, and John Winn. "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies." PLoS computational biology 6.5 (2010): e1000770.

Stone, Lawrence D., Colleen M. Keller, Thomas M. Kratzke, and Johan P. Strumpfer. "Search for the wreckage of Air France Flight AF 447." *Statistical Science* (2014): 69-80.

Woodard, Dawn, Galina Nogin, Paul Koch, David Racz, Moises Goldszmidt, and Eric Horvitz. "Predicting travel time reliability using mobile phone GPS data." *Transportation Research Part C: Emerging Technologies* 75 (2017): 30-44.

Xing, Eric P., Wei Wu, Michael I. Jordan, and Richard M. Karp. "LOGOS: a modular Bayesian model for de novo motif detection." Journal of Bioinformatics and Computational Biology 2.01 (2004): 127-154.

Additional image references (6/6)

amCharts. Visited Countries Map. https://www.amcharts.com/visited_countries/ Accessed: 2016.

Baltic Salmon Fund. https://www.en.balticsalmonfund.org/about_us Accessed: 2018.

ESO/L. Calçada/M. Kornmesser. 16 October 2017, 16:00:00. Obtained from: https://commons.wikimedia.org/wiki/File:Artist%E2%80%99s_impression_of_merging_neutron_stars.jpg || Source: https://www.eso.org/public/images/eso1733a/ (Creative Commons Attribution 4.0 International License)

J. Herzog. 3 June 2016, 17:17:30. Obtained from: https://commons.wikimedia.org/wiki/File:Airbus_A350-941_F-WWCF_MSN002_ILA_Berlin_2016_17.jpg (Creative Commons Attribution 4.0 International License)

E. Xing. 2003. Slides "LOGOS: a modular Bayesian model for de novo motif detection." Obtained from: https://www.cs.cmu.edu/~epxing/papers/Old_papers/slide_CSB03/CSB1.pdf Accessed: 2018.