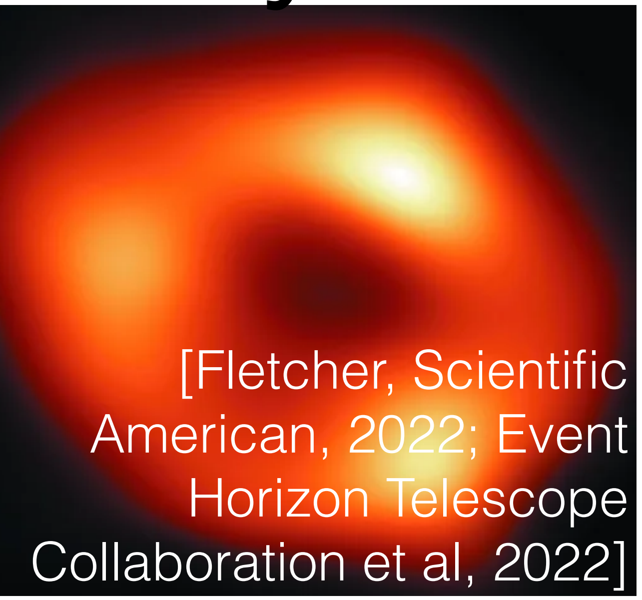
Variational Bayes and beyond: Foundations of scalable Bayesian inference

Tamara Broderick
Associate Professor
MIT

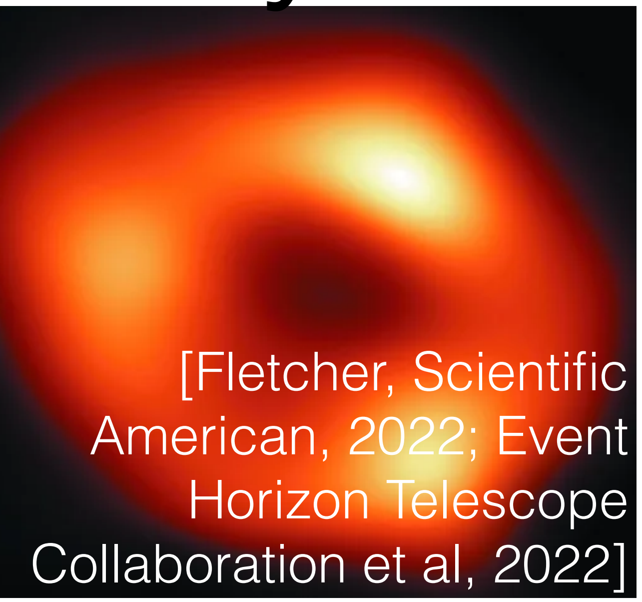
http://tamarabroderick.com/tutorial_2024_astroai.html

Bayesian inference

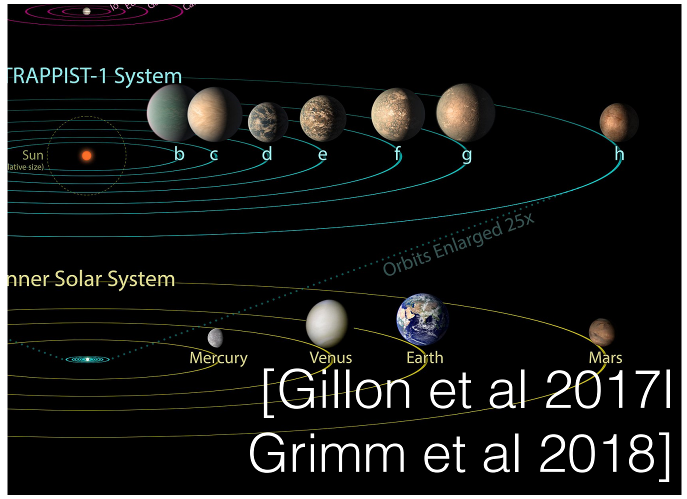
Bayesian inference



Bayesian inference

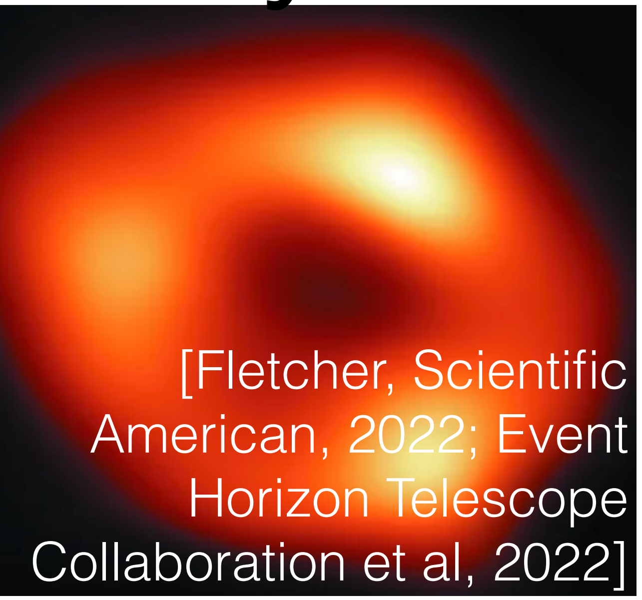


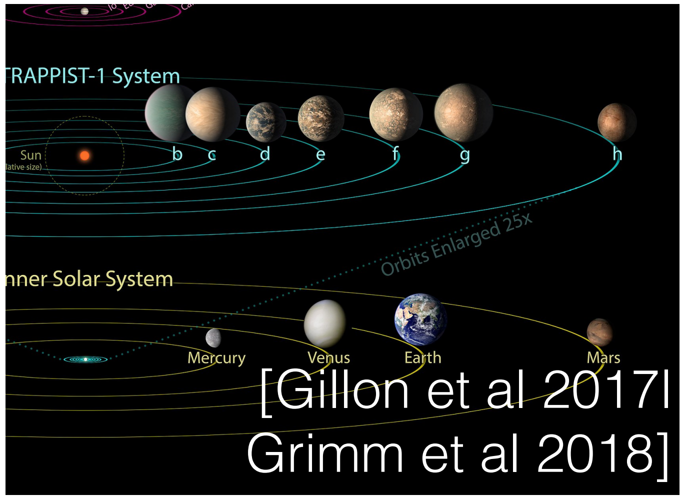
[Fletcher, Scientific American, 2022; Event Horizon Telescope Collaboration et al, 2022]



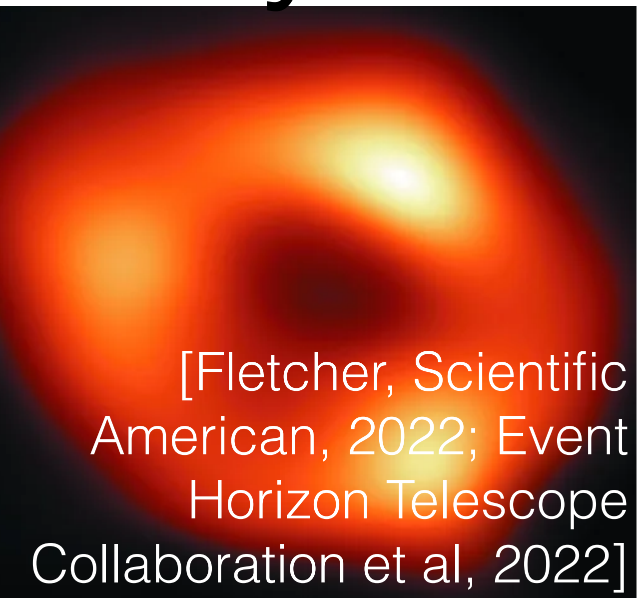
[Gillon et al 2017l Grimm et al 2018]

Bayesian inference

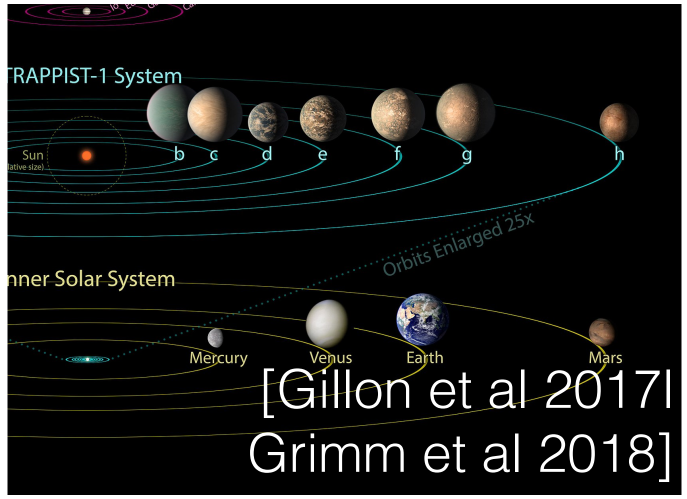




Bayesian inference

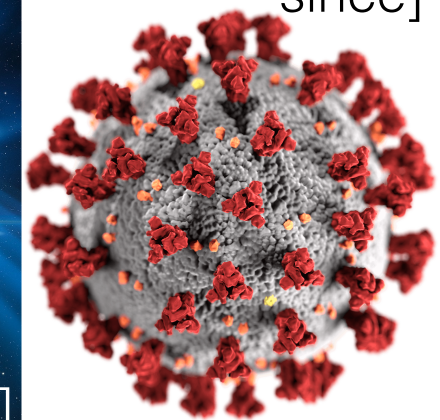


[Fletcher, Scientific American, 2022; Event Horizon Telescope Collaboration et al, 2022]



[Gillon et al 2017l Grimm et al 2018]

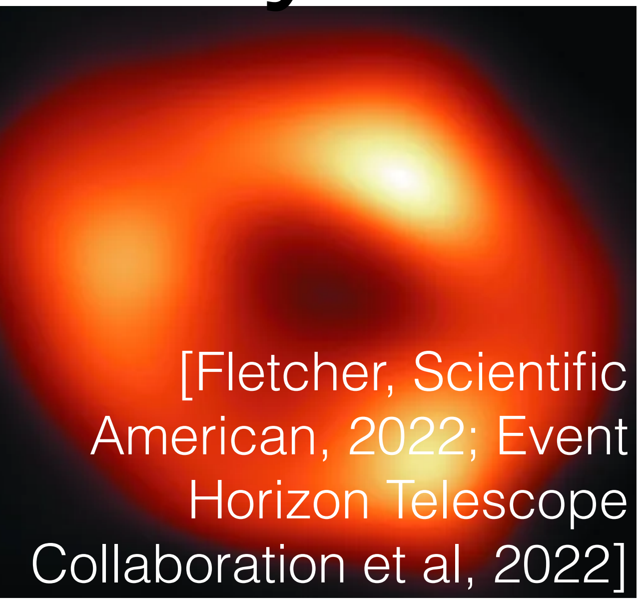
[ESO/ L. Calçada/ M. Kornmesser 2017] [Abbott et al 2016a,b]



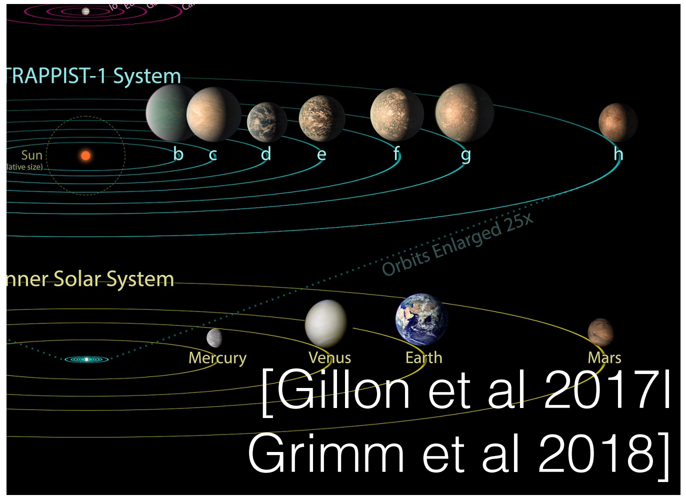
[2020 “Science Papers you should be Reading about the Coronavirus”; and many since]

Bayesian inference

[2020 "Science Papers you should be Reading about the Coronavirus"; and many since]

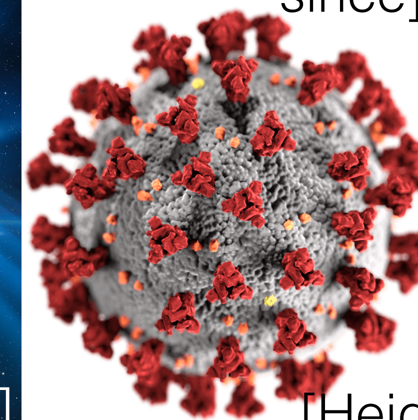


[Fletcher, Scientific American, 2022; Event Horizon Telescope Collaboration et al, 2022]



[Gillon et al 2017l Grimm et al 2018]

[ESO/ L. Calçada/ M. Kornmesser 2017]
[Abbott et al 2016a,b]

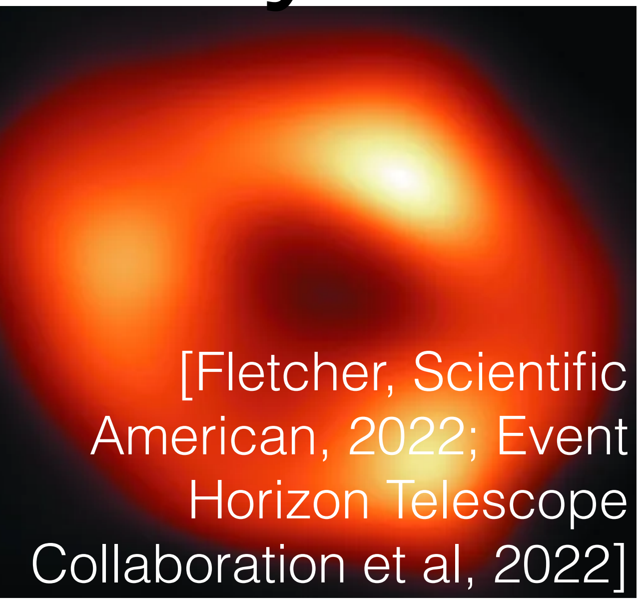


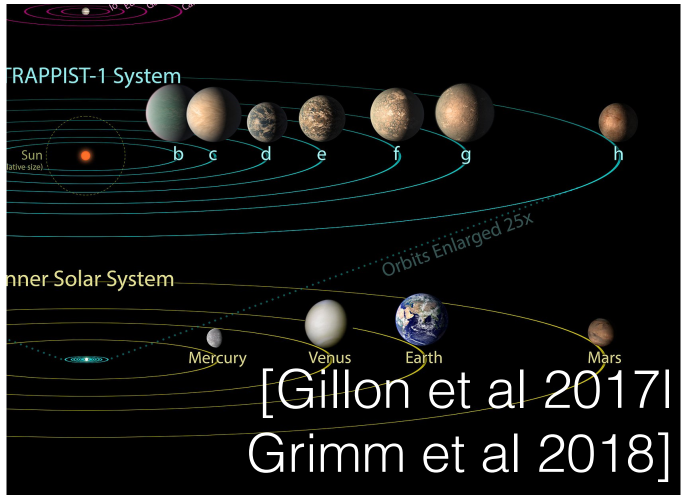
The Economist

[Heidemanns et al 2020

Bayesian inference

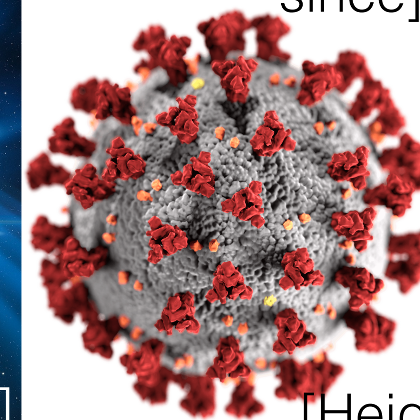
[2020 “Science Papers you should be Reading about the Coronavirus”; and many since]


[Fletcher, Scientific American, 2022; Event Horizon Telescope Collaboration et al, 2022]



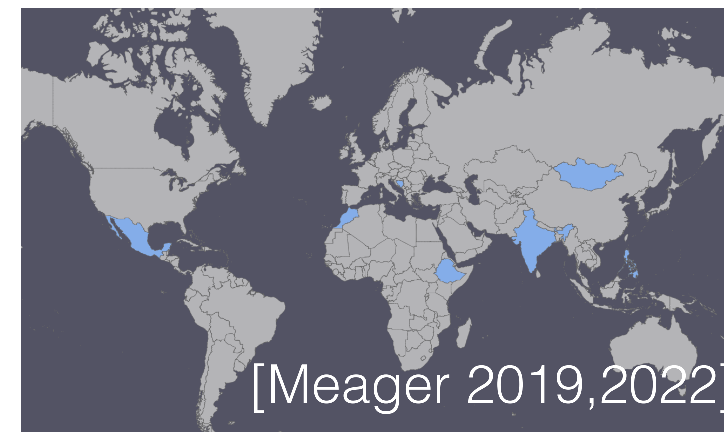
[Gillon et al 2017l Grimm et al 2018]


[ESO/ L. Calçada/ M. Kornmesser 2017]
[Abbott et al 2016a,b]



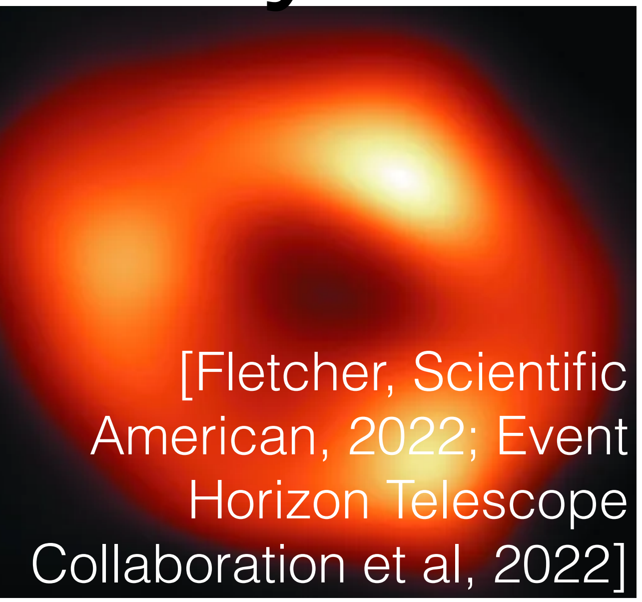
The Economist


[Heidemanns et al 2020

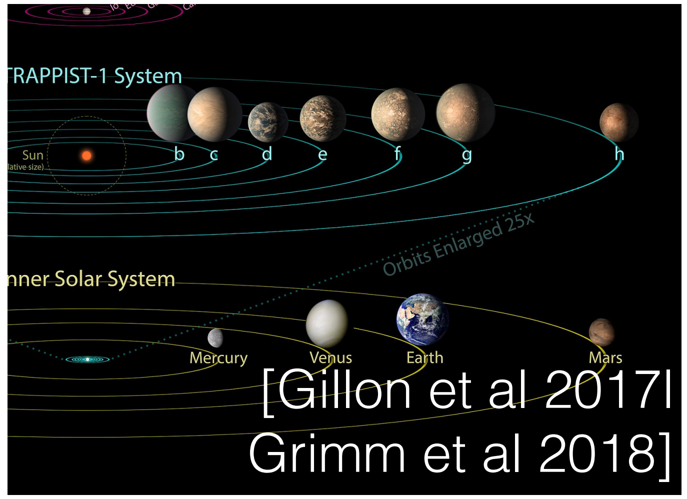

[Meager 2019,2022]

Bayesian inference

[2020 "Science Papers you should be Reading about the Coronavirus"; and many since]

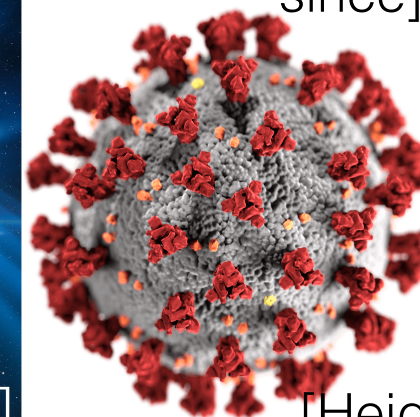


[Fletcher, Scientific American, 2022; Event Horizon Telescope Collaboration et al, 2022]



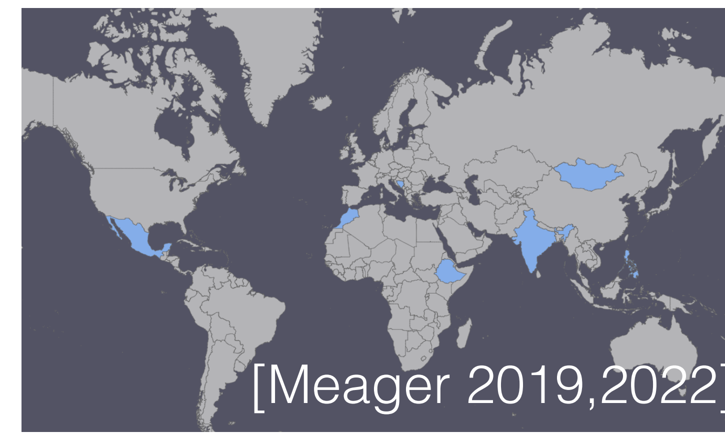
[Gillon et al 2017l Grimm et al 2018]

[ESO/ L. Calçada/ M. Kornmesser 2017] [Abbott et al 2016a,b]



The Economist

[Heidemanns et al 2020

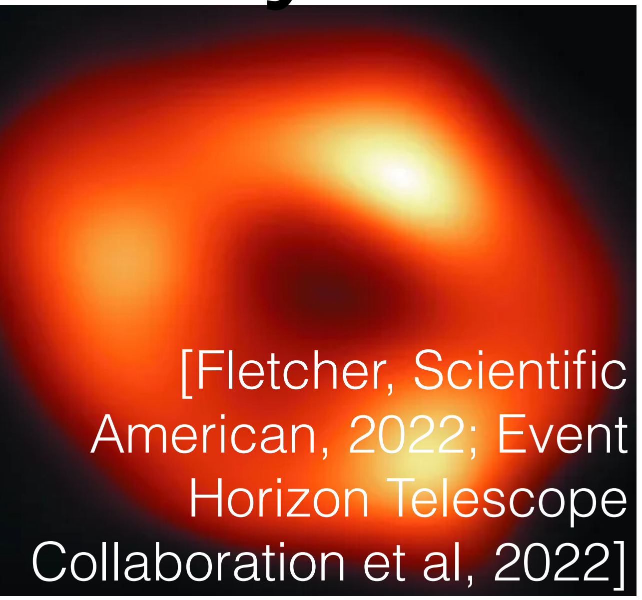


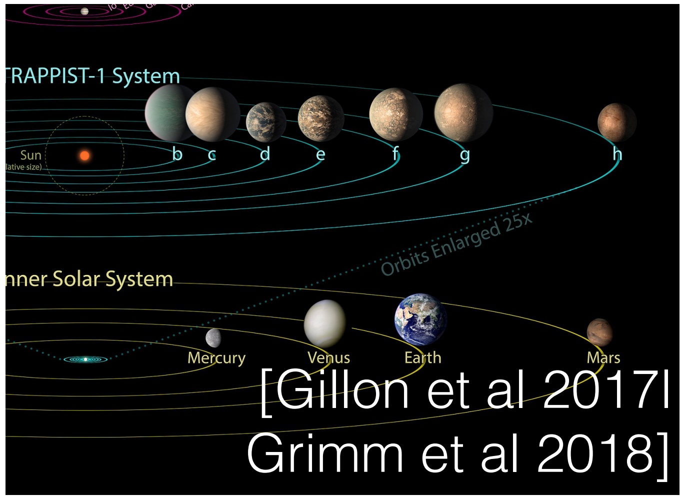
[Meager 2019,2022]

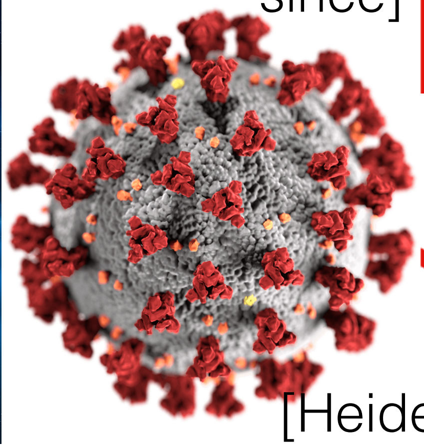
- Goal: good point estimates, uncertainty estimates
  - Also: share power, use expert info, different types of data

Bayesian inference

[2020 “Science Papers you should be Reading about the Coronavirus”; and many since]

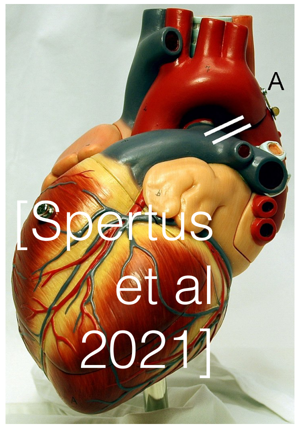


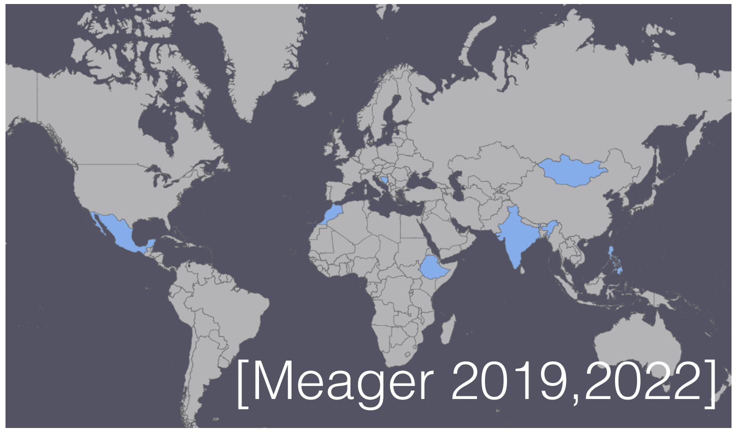




The Economist

[Heidemanns et al 2020

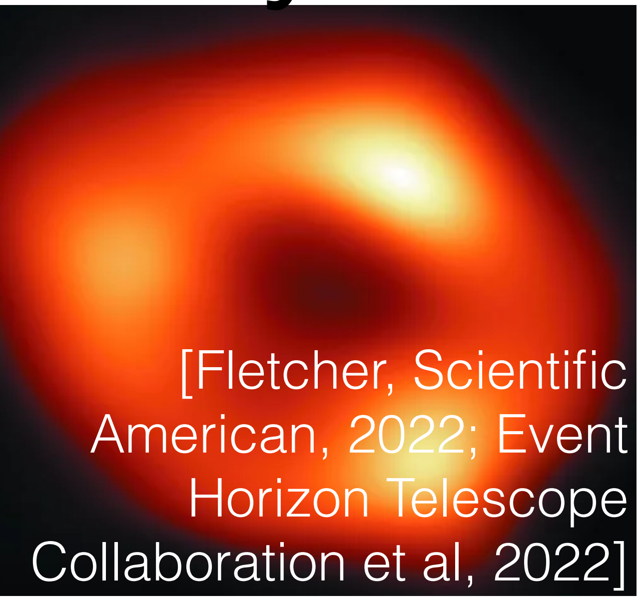




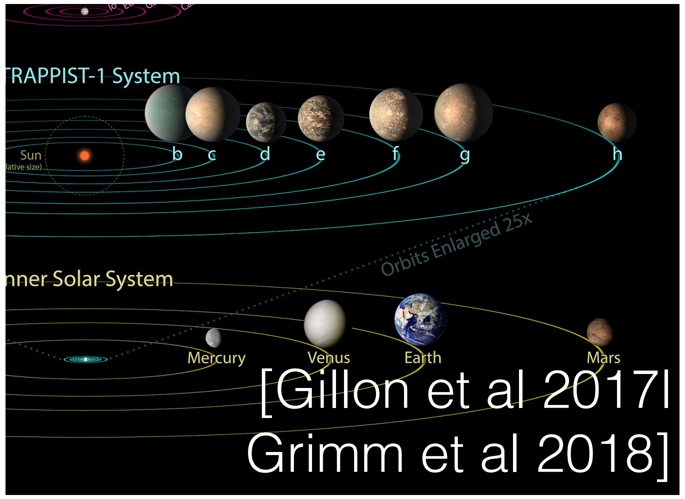
- Goal: good point estimates, uncertainty estimates
  - Also: share power, use expert info, different types of data

Bayesian inference

[2020 "Science Papers you should be Reading about the Coronavirus"; and many since]

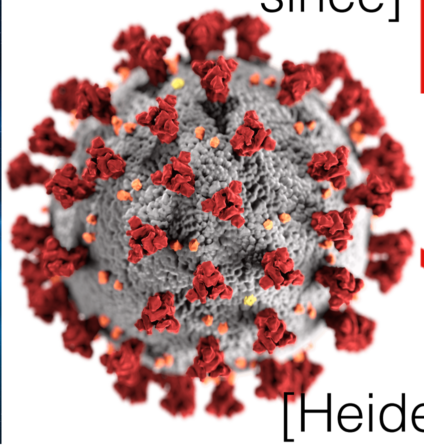


[Fletcher, Scientific American, 2022; Event Horizon Telescope Collaboration et al, 2022]



[Gillon et al 2017l Grimm et al 2018]

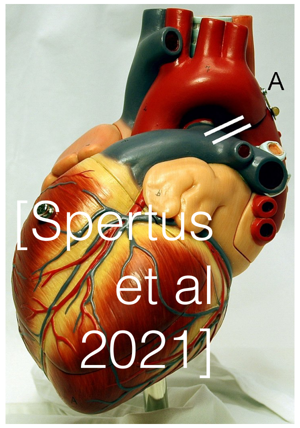
[ESO/ L. Calçada/ M. Kornmesser 2017] [Abbott et al 2016a,b]



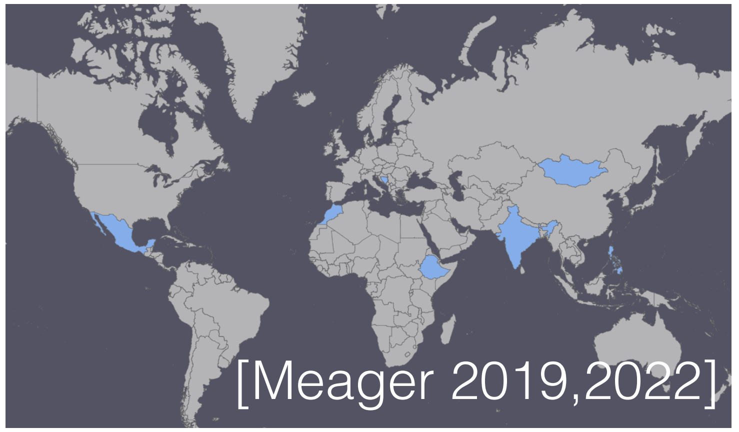
The Economist

[Heidemanns et al 2020

[Kuikka et al 2014] [Baltic Salmon Fund]



[Spertus et al 2021]



[Meager 2019,2022]

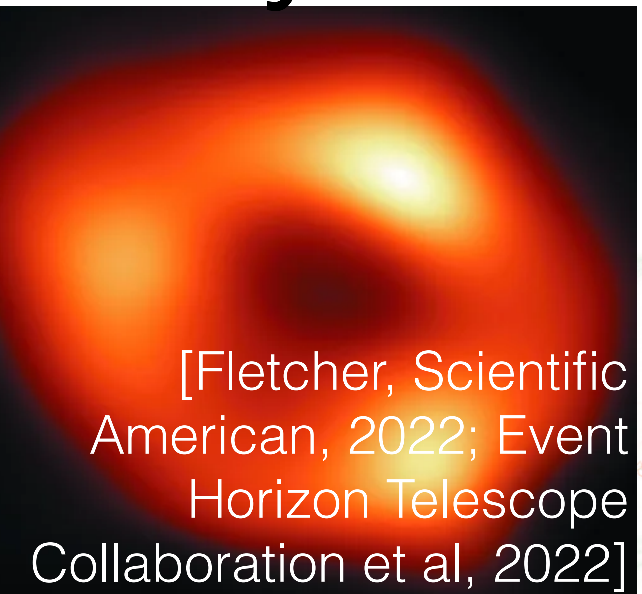
- Goal: good point estimates, uncertainty estimates
  - Also: share power, use expert info, different types of data

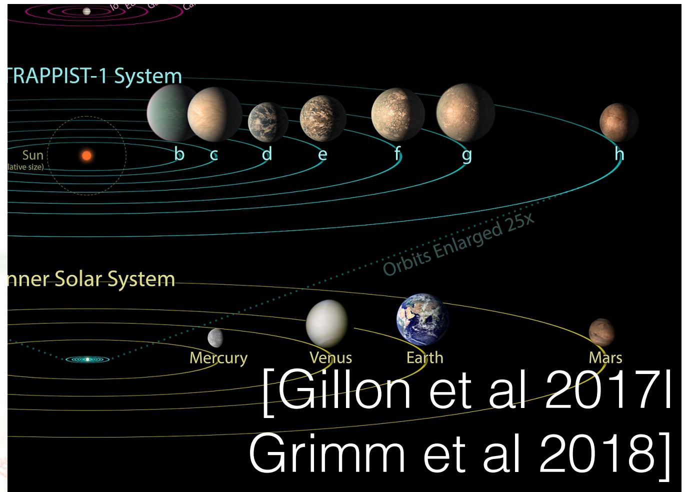
Bayesian inference

- Goal: good point estimates, uncertainty estimates
  - Also: share power, use expert info, different types of data

Bayesian inference

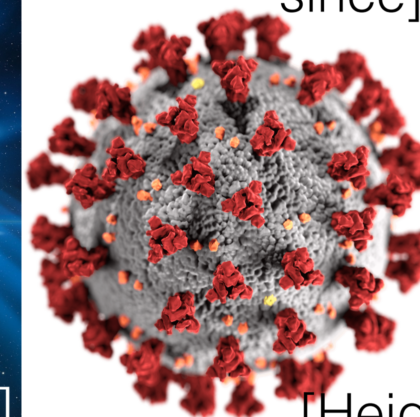
[2020 “Science Papers you should be Reading about the Coronavirus”; and many since]


[Fletcher, Scientific American, 2022; Event Horizon Telescope Collaboration et al, 2022]



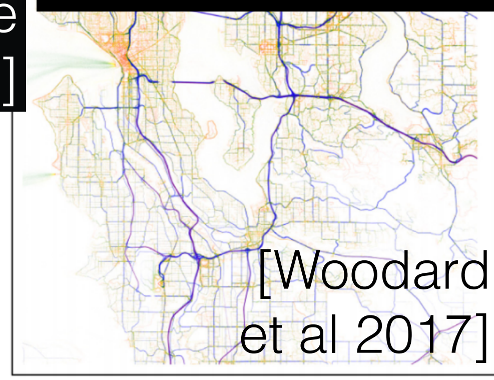
[Gillon et al 2017l Grimm et al 2018]


[ESO/ L. Calçada/ M. Kornmesser 2017] [Abbott et al 2016a,b]

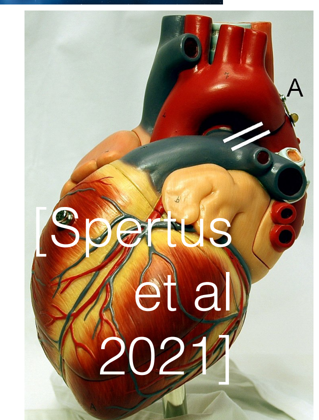


The Economist

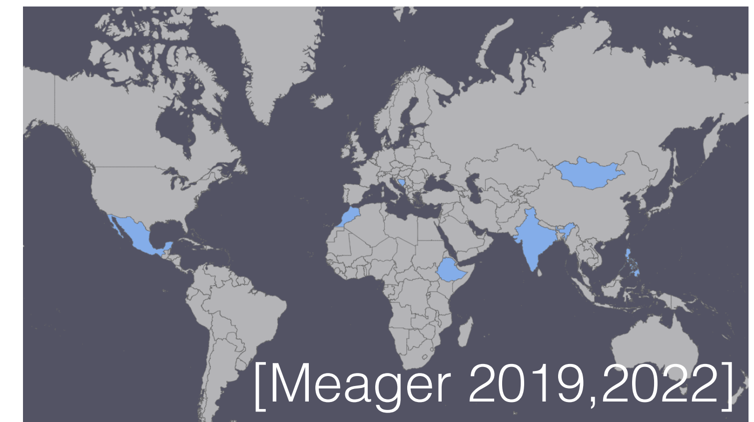

[Heidemanns et al 2020

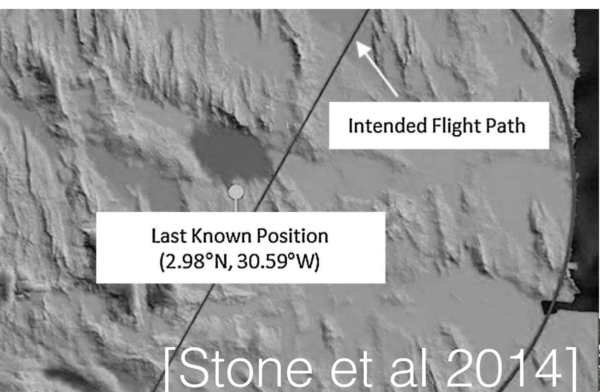

[Woodard et al 2017]


[Kuikka et al 2014] [Baltic Salmon Fund]



[Spertus et al 2021]


[Meager 2019,2022]



[Stone et al 2014]


[Poorter et al 2021]

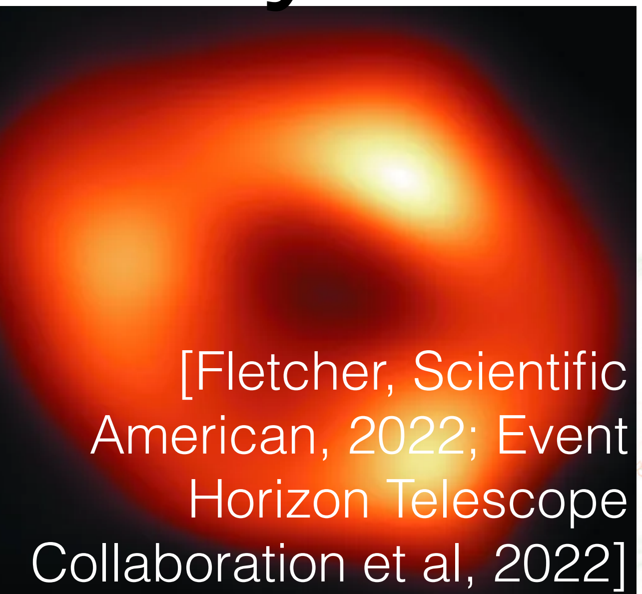

[McMahan, McFarland 2021]

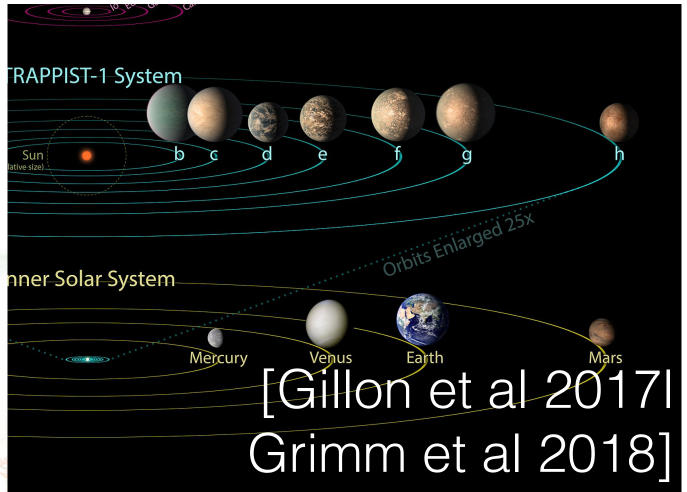

[Chati, Balakrishnan 2017] [Julian Hertzog 2016]

- Goal: good point estimates, uncertainty estimates
  - Also: share power, use expert info, different types of data

Bayesian inference

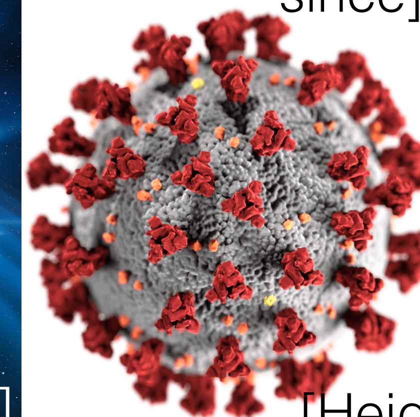
[2020 “Science Papers you should be Reading about the Coronavirus”; and many since]


[Fletcher, Scientific American, 2022; Event Horizon Telescope Collaboration et al, 2022]

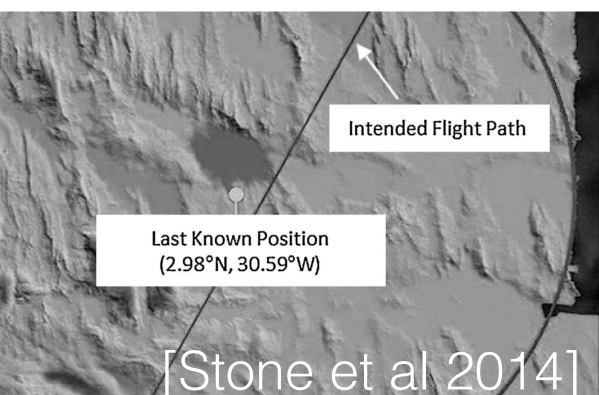


[Gillon et al 2017; Grimm et al 2018]

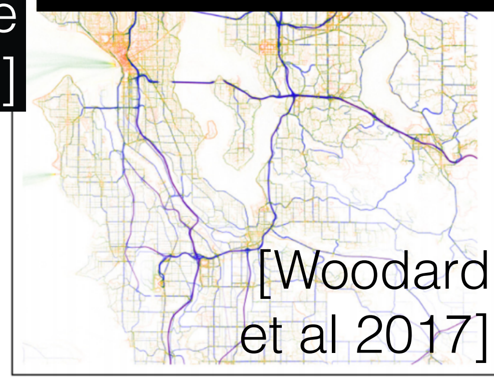

[ESO/ L. Calçada/ M. Kornmesser 2017]
[Abbott et al 2016a,b]



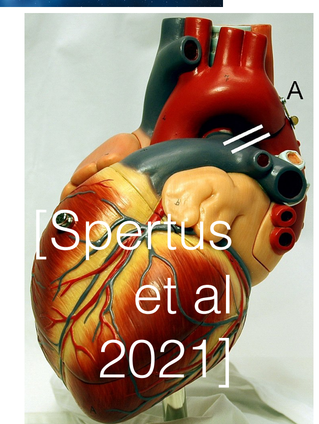
The Economist
[Heidemanns et al 2020

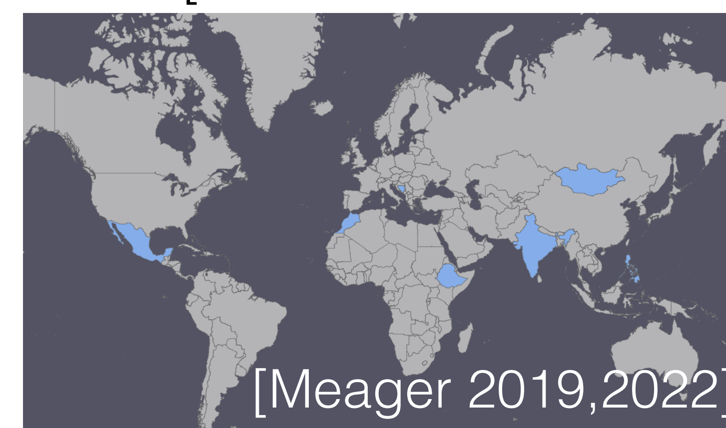


[Stone et al 2014]


[Woodard et al 2017]

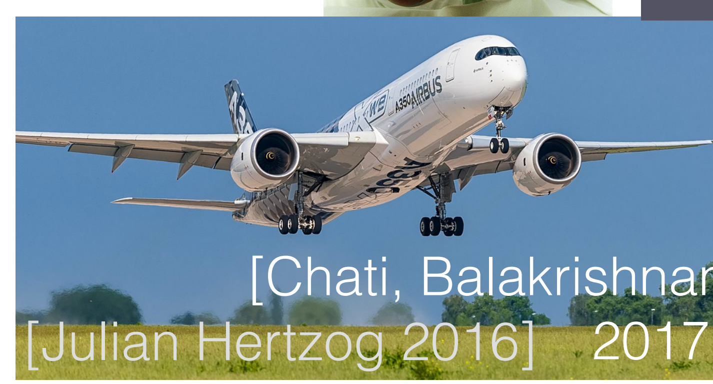

[Kuikka et al 2014]
[Baltic Salmon Fund]


[Spertus et al 2021]


[Meager 2019,2022]


[Poorter et al 2021]


[McMahan, McFarland 2021]


[Chati, Balakrishnan 2017]
[Julian Hertzog 2016]

[mc-stan.org]

PyMC

- Goal: good point estimates, uncertainty estimates
  - Also: share power, use expert info, different types of data

Bayesian inference

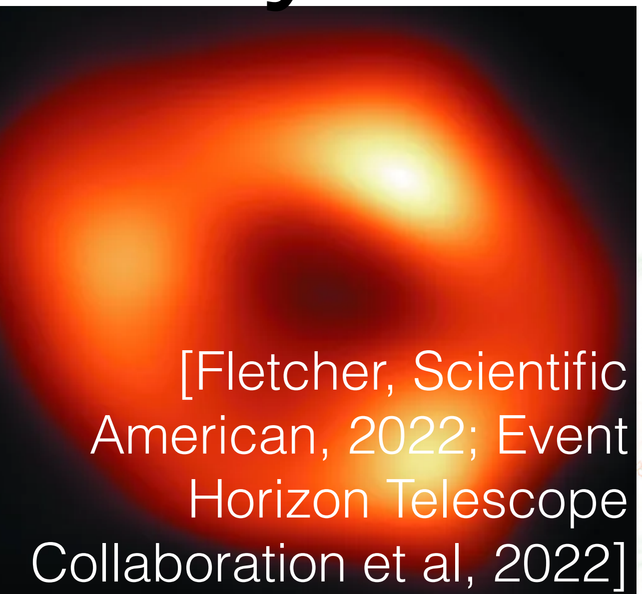
[2020 “Science Papers you should be Reading about the Coronavirus”; and many since]

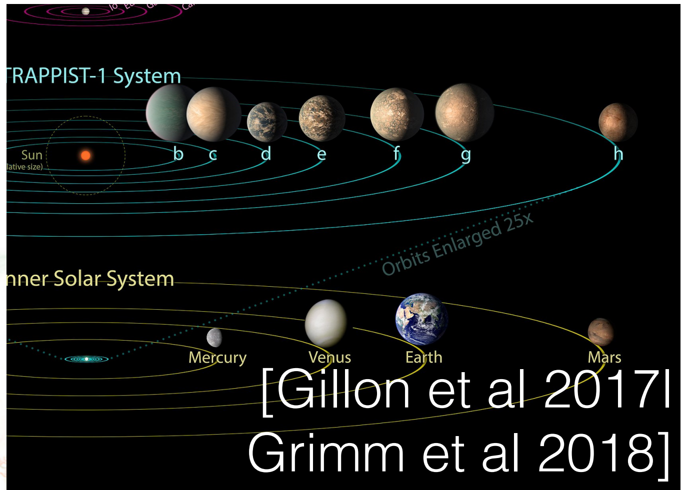
[mc-stan.org]

- Goal: good point estimates, uncertainty estimates
  - Also: share power, use expert info, different types of data
- Modern: often large data, dimensions (uncertainty remains)

Bayesian inference

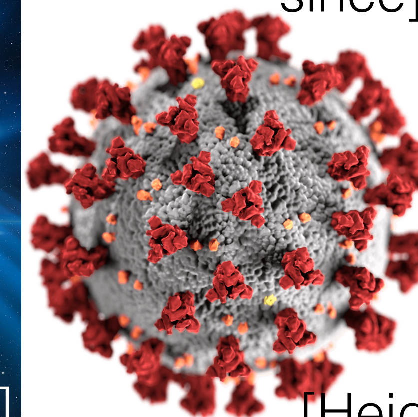
[2020 "Science Papers you should be Reading about the Coronavirus"; and many since]


[Fletcher, Scientific American, 2022; Event Horizon Telescope Collaboration et al, 2022]

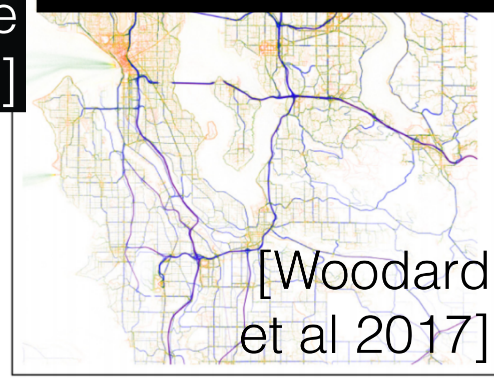


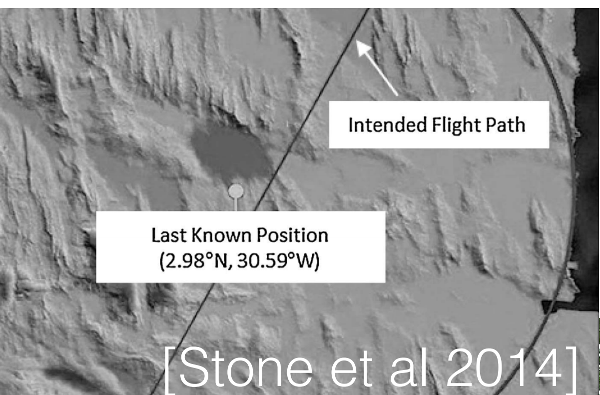
[Gillon et al 2017l Grimm et al 2018]


[ESO/ L. Calçada/ M. Kornmesser 2017]
[Abbott et al 2016a,b]



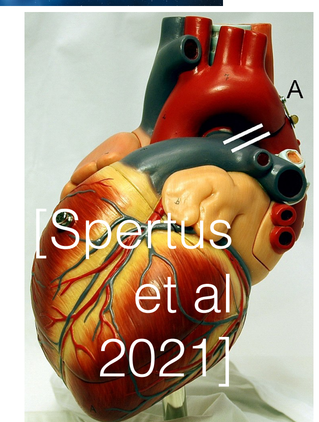
[Heidemanns et al 2020


[Woodard et al 2017]

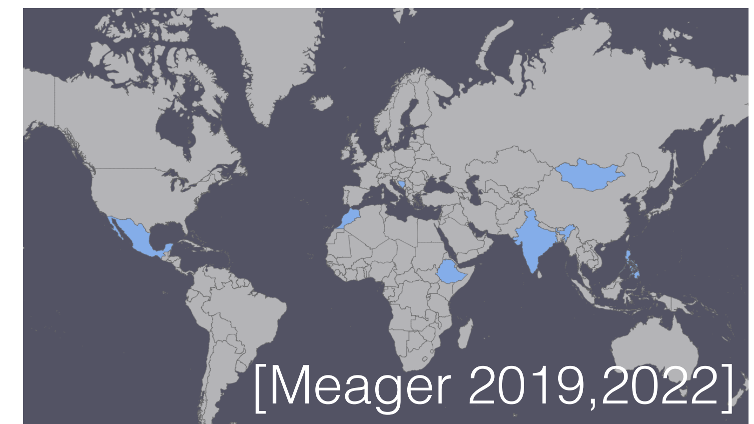


[Stone et al 2014]


[Kuikka et al 2014]
[Baltic Salmon Fund]

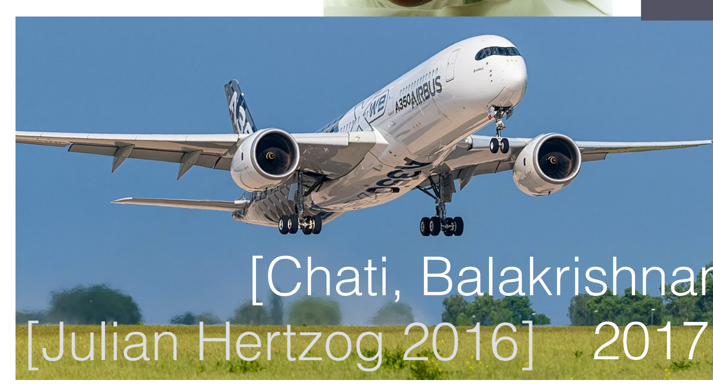


[Spertus et al 2021]


[Meager 2019,2022]


[Poorter et al 2021]


[McMahan, McFarland 2021]


[Chati, Balakrishnan 2017]
[Julian Hertzog 2016]

[mc-stan.org]

PyMC

- Goal: good point estimates, uncertainty estimates
  - Also: share power, use expert info, different types of data
- Modern: often large data, dimensions (uncertainty remains)
- Challenge: speed (compute, user), reliable inference

Bayesian inference

[2020 "Science Papers you should be Reading about the Coronavirus"; and many since]

- Goal: good point estimates, uncertainty estimates
  - Also: share power, use expert info, different types of data
- Modern: often large data, dimensions (uncertainty remains)
- Challenge: speed (compute, user), reliable inference
- Variational Bayes offers fast runtimes in modern regimes

Roadmap

Roadmap

- Bayes & Approximate Bayes setup

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB? Some VB successes (speed, accuracy)

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB? Some VB successes (speed, accuracy)
- Some VB failure modes, and partial solutions

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB? Some VB successes (speed, accuracy)
- Some VB failure modes, and partial solutions
- Ease of use / automation
  - Automatic differentiation variational inference (ADVI) and beyond

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB? Some VB successes (speed, accuracy)
- Some VB failure modes, and partial solutions
- Ease of use / automation
  - Automatic differentiation variational inference (ADVI) and beyond

Bayesian inference

Bayesian inference

θ

Bayesian inference

e.g. pollution level

θ

Bayesian inference

e.g. pollution level

parameters

θ

Bayesian inference

e.g. sensors e.g. pollution level

data parameters

θ

Bayesian inference

e.g. sensors — data → $y_{1:N}$

e.g. pollution level — parameters → θ

Bayesian inference

e.g. sensors — data → $y_{1:N}$

e.g. pollution level — parameters → $p(\theta)$

prior

Bayesian inference

e.g. sensors — data

e.g. pollution level — parameters

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

likelihood prior

Bayesian inference

e.g. sensors — data

e.g. pollution level — parameters

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

posterior — likelihood — prior

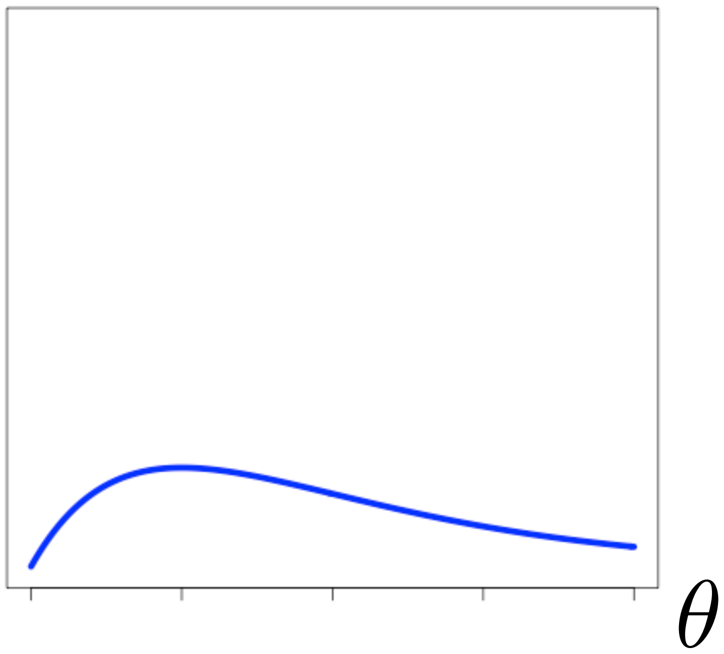
Bayesian inference

e.g. sensors — data

e.g. pollution level — parameters

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

posterior likelihood prior



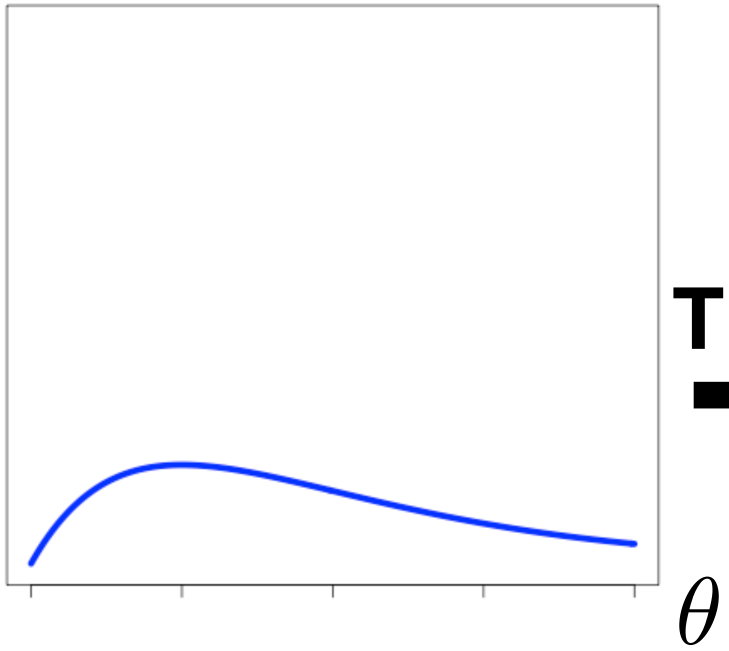
Bayesian inference

e.g. sensors — data

e.g. pollution level — parameters

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

posterior likelihood prior



Bayes Theorem

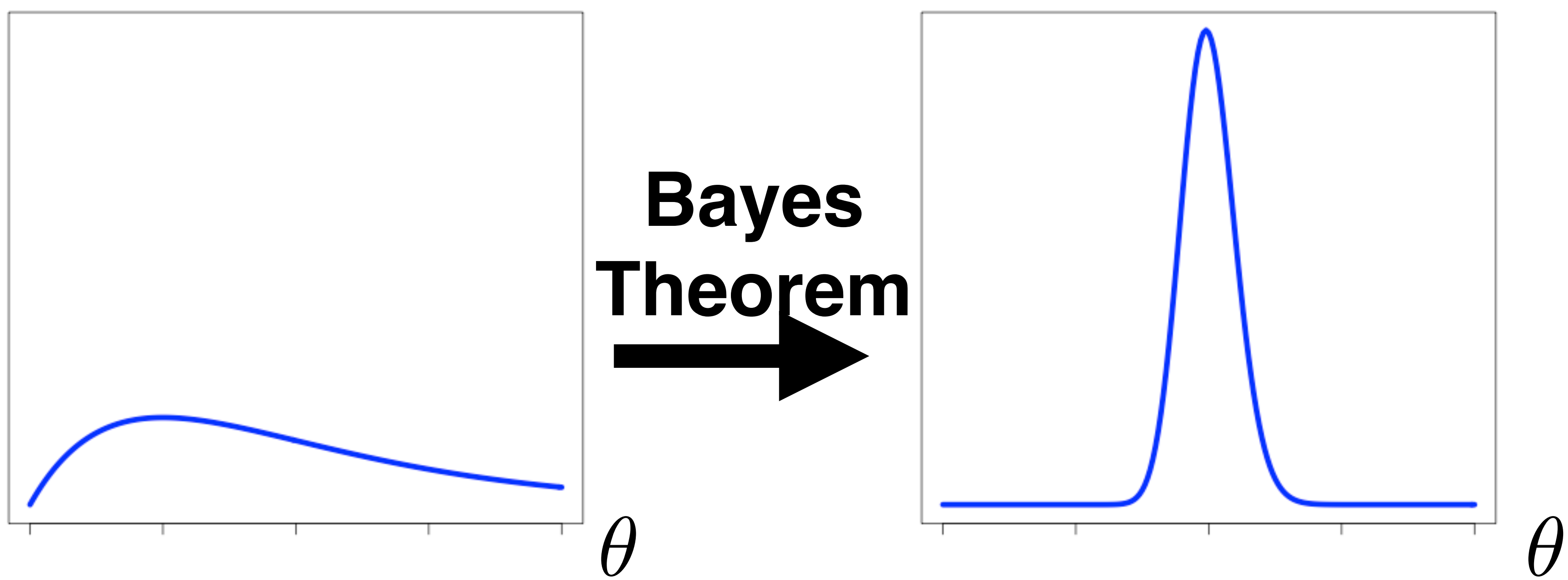
Bayesian inference

e.g. sensors — data

e.g. pollution level — parameters

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

posterior likelihood prior



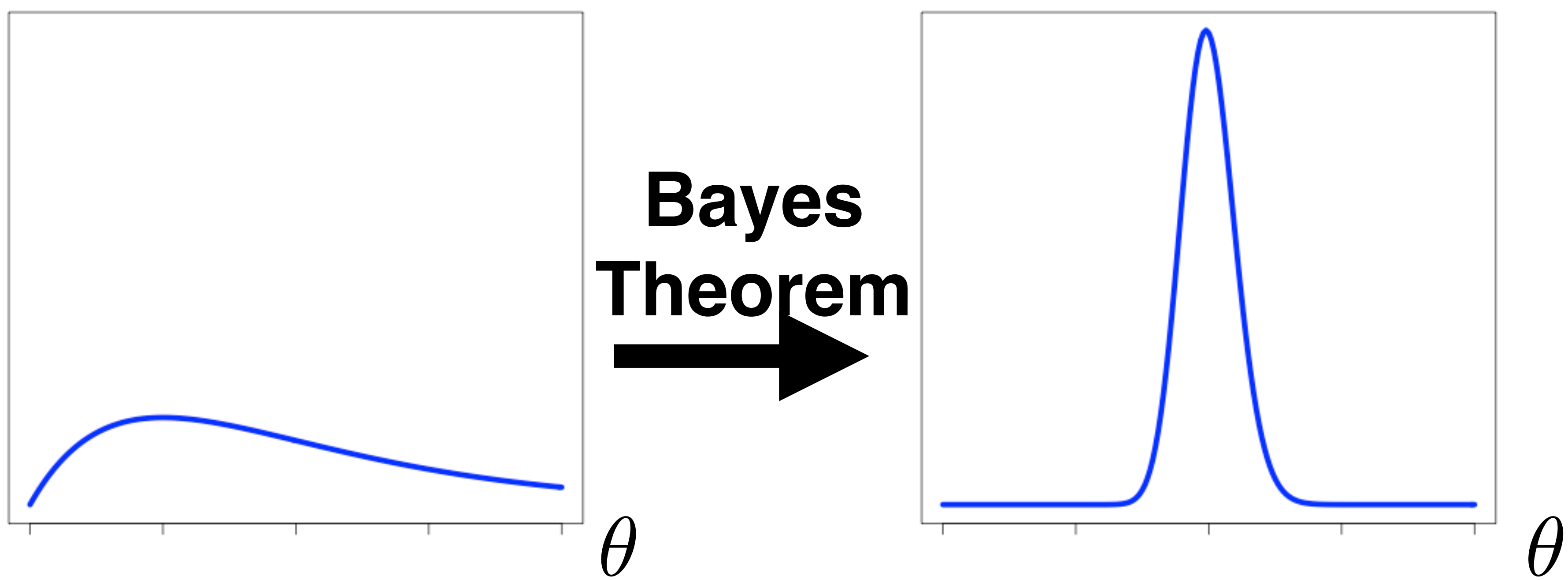
Bayesian inference

e.g. sensors — data

e.g. pollution level — parameters

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

posterior likelihood prior



0. Identify a data analysis goal e.g. estimate pollution level

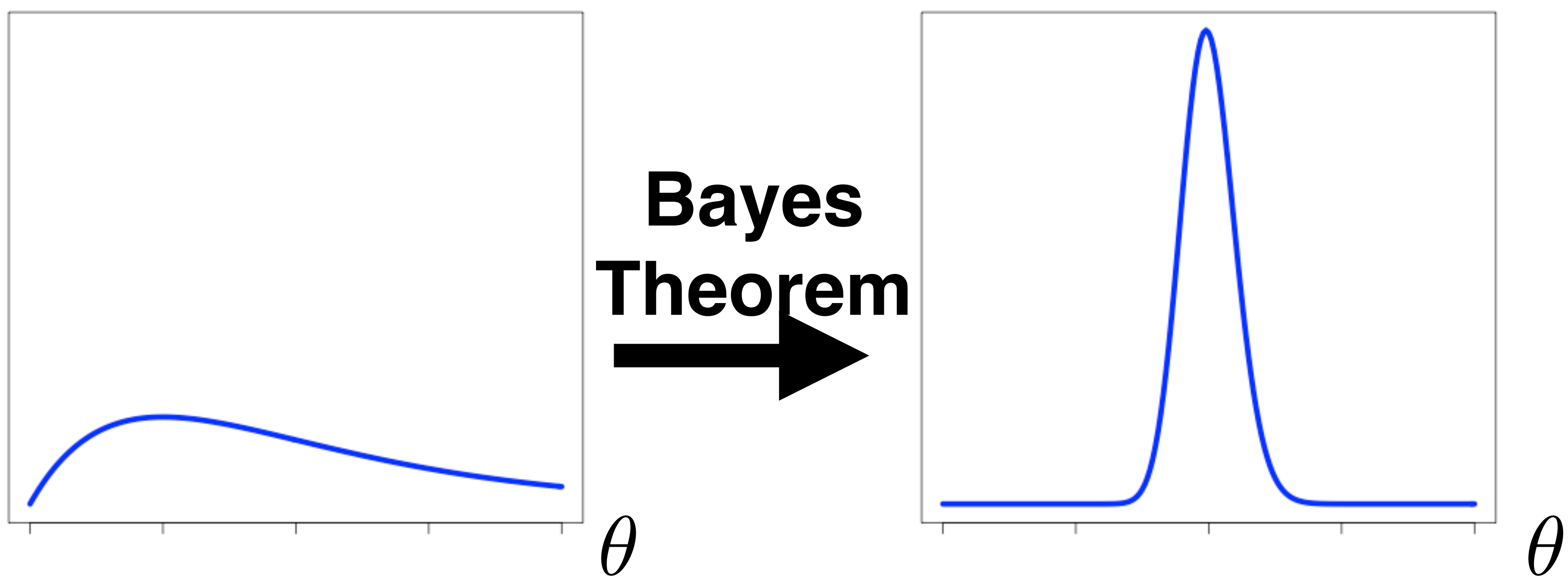
Bayesian inference

e.g. sensors — data

e.g. pollution level — parameters

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

posterior likelihood prior



0. Identify a data analysis goal e.g. estimate pollution level

1. Build a model: choose prior & choose likelihood

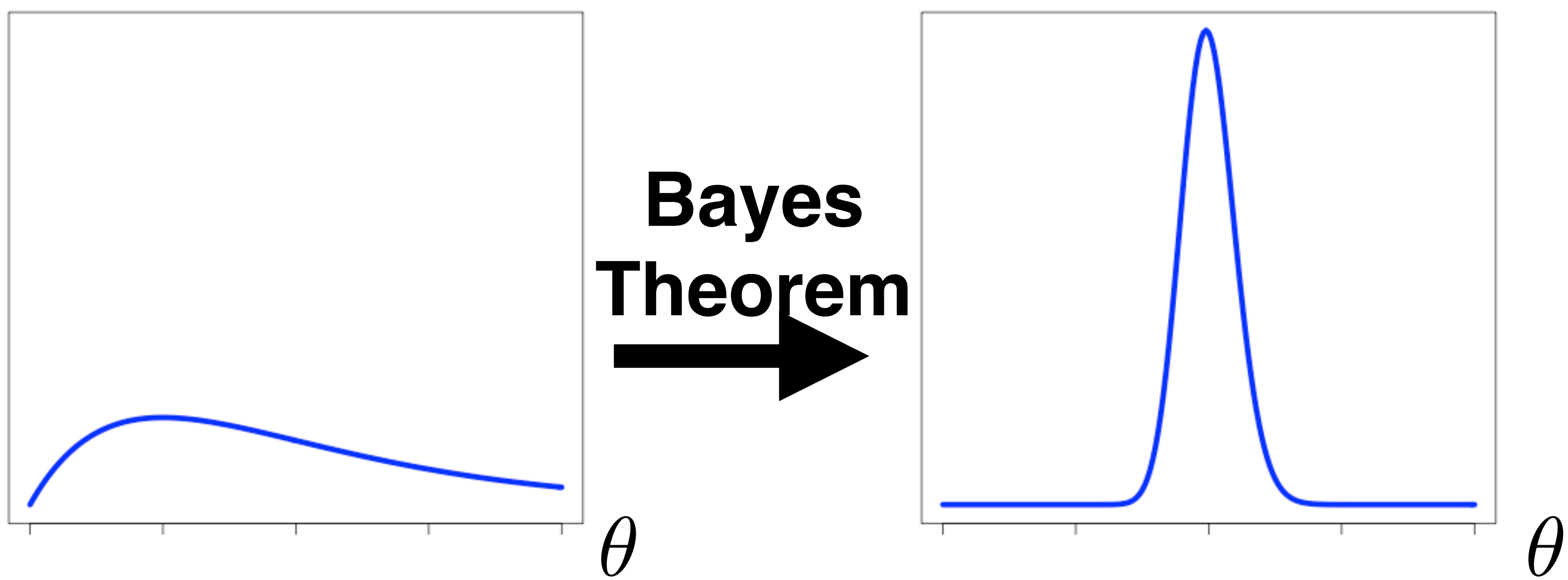
Bayesian inference

e.g. sensors e.g. pollution level

data parameters

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

posterior likelihood prior



0. Identify a data analysis goal e.g. estimate pollution level
1. Build a model: choose prior & choose likelihood
2. Report a posterior summary, e.g. means and (co)variances

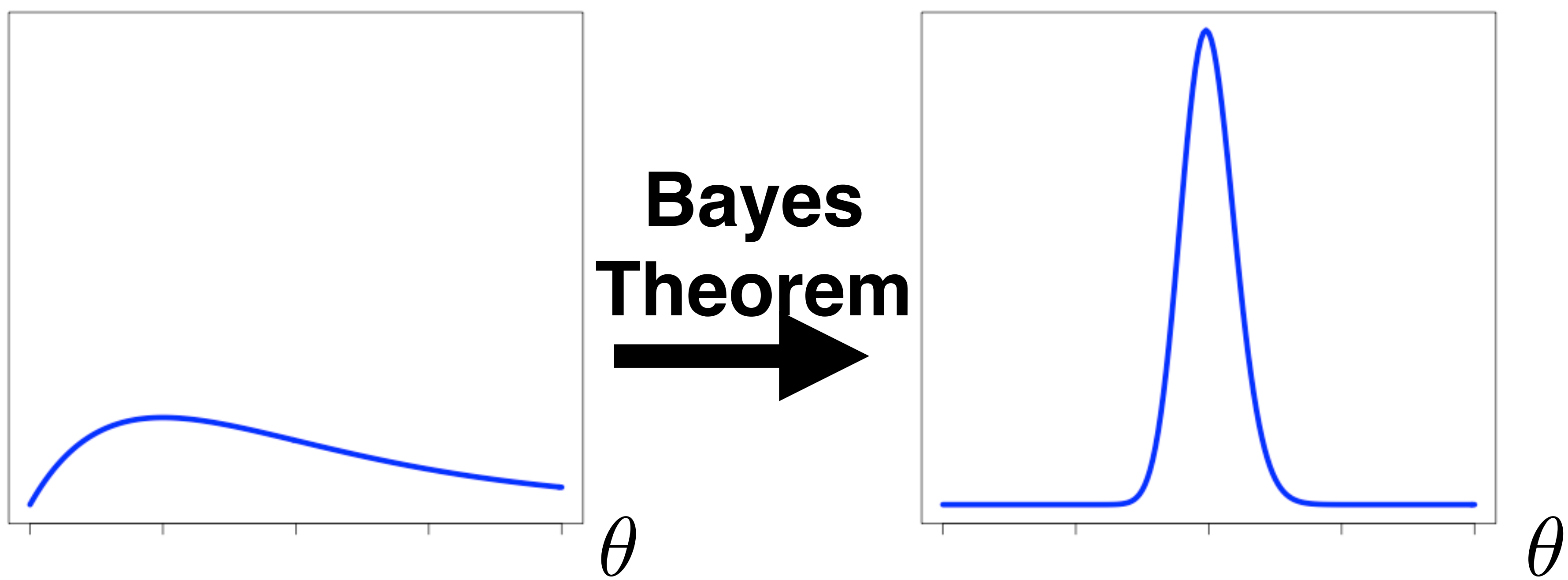
Bayesian inference

e.g. sensors e.g. pollution level

data parameters

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

posterior likelihood prior



0. Identify a data analysis goal e.g. estimate pollution level
1. Build a model: choose prior & choose likelihood
2. Report a posterior summary, e.g. means and (co)variances

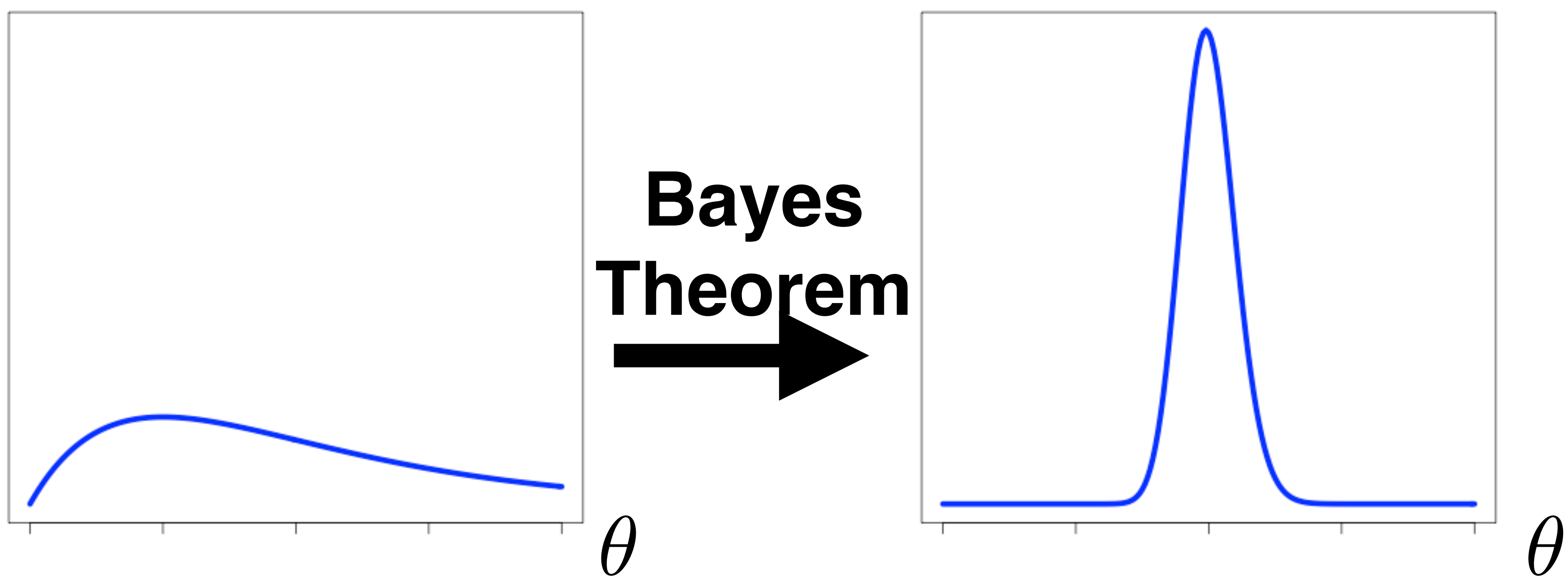
- Why is the final step hard?

Bayesian inference

e.g. sensors — data; e.g. pollution level — parameters

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

posterior — likelihood — prior



0. Identify a data analysis goal e.g. estimate pollution level
1. Build a model: choose prior & choose likelihood
2. Report a posterior summary, e.g. means and (co)variances

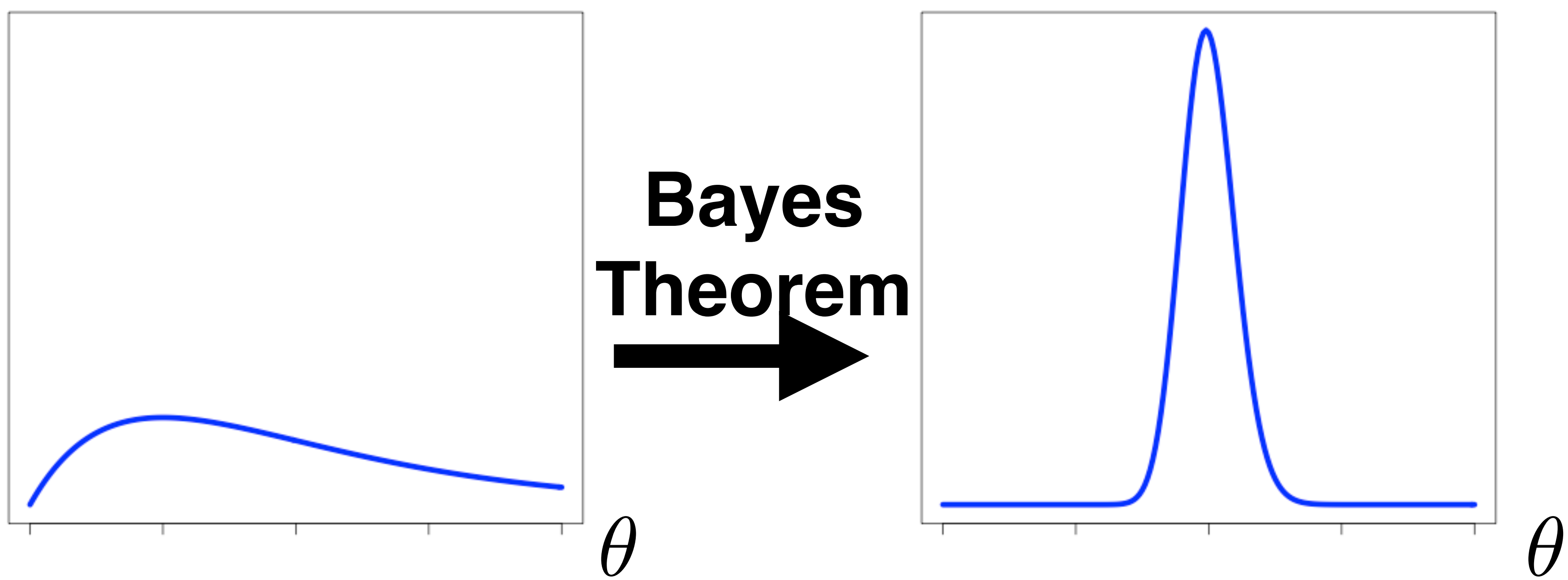
- Why is the final step hard?
  - Typically no closed form

Bayesian inference

e.g. sensors → data
e.g. pollution level → parameters

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

posterior likelihood prior



0. Identify a data analysis goal   e.g. estimate pollution level
1. Build a model: choose prior & choose likelihood
2. Report a posterior summary, e.g. means and (co)variances

- Why is the final step hard?
    - Typically no closed form, high-dimensional integration

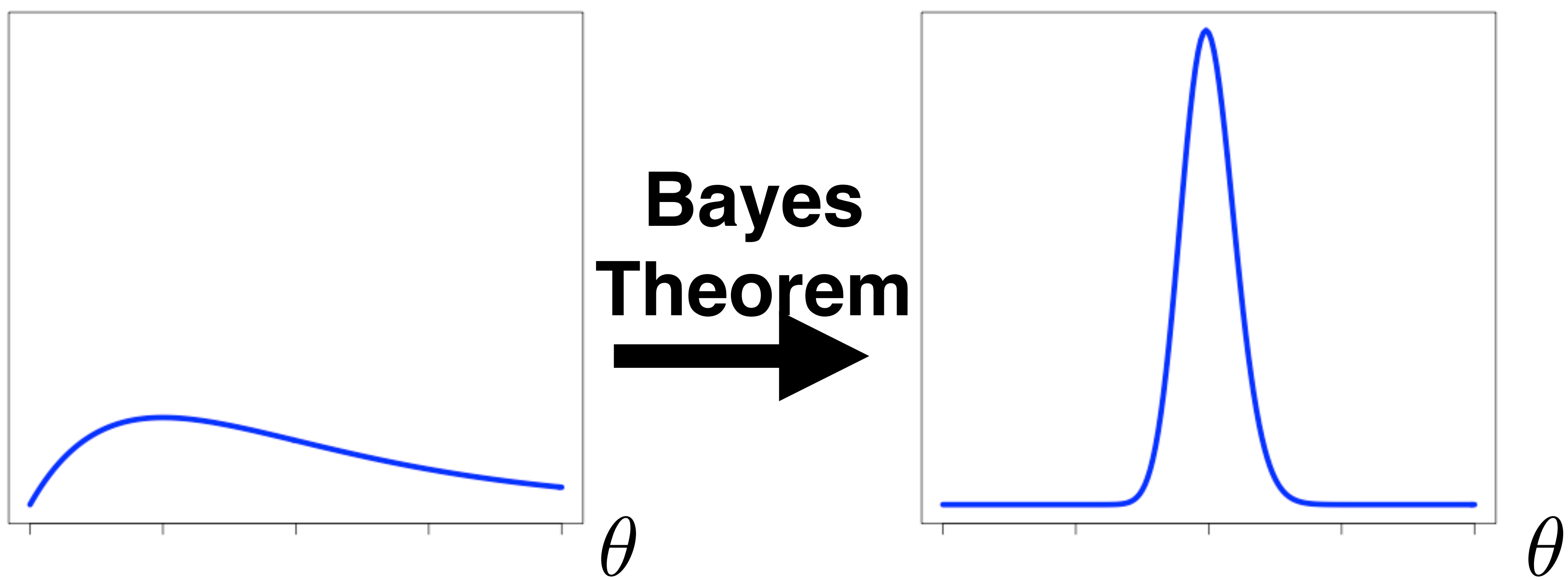
Bayesian inference

e.g. sensors e.g. pollution level

data parameters

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

posterior likelihood prior



0. Identify a data analysis goal e.g. estimate pollution level
1. Build a model: choose prior & choose likelihood
2. Report a posterior summary, e.g. means and (co)variances

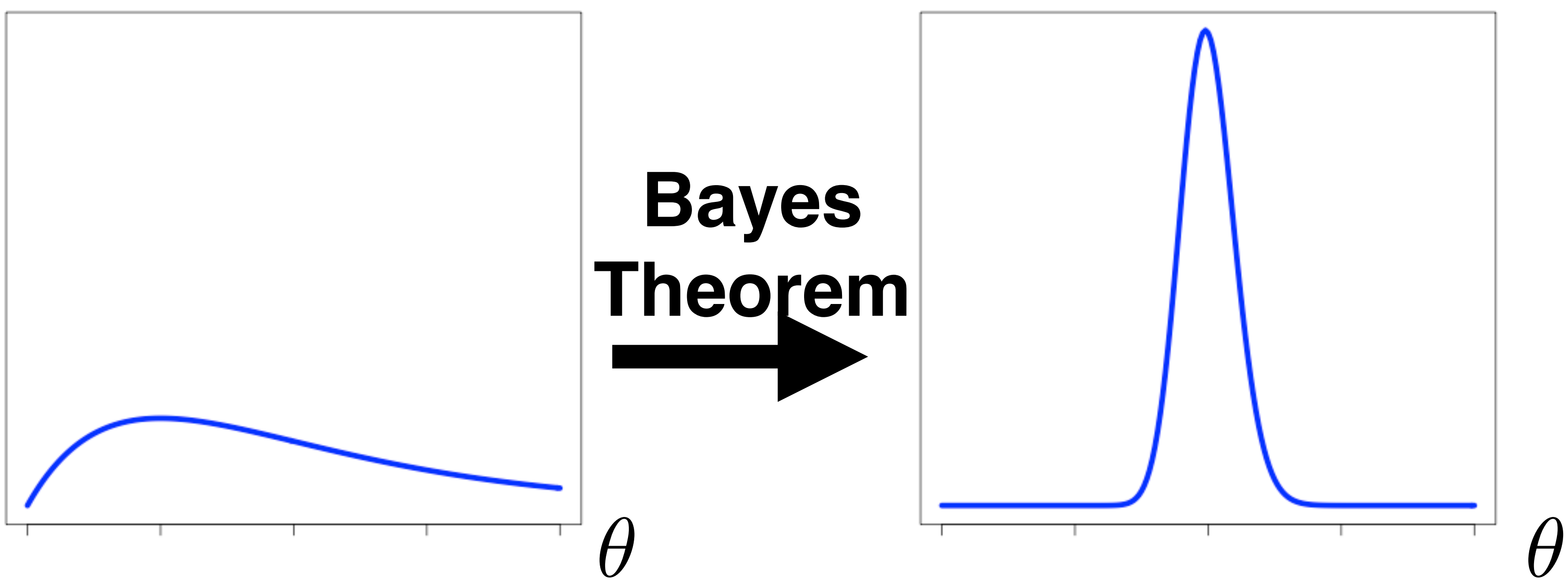
- Why is the final step hard?
  - Typically no closed form, high-dimensional integration

Bayesian inference

e.g. sensors — data; e.g. pollution level — parameters

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

posterior — likelihood — prior



0. Identify a data analysis goal — e.g. estimate pollution level
1. Build a model: choose prior & choose likelihood
2. Report a posterior summary, e.g. means and (co)variances

- Why is the final step hard?
  - Typically no closed form, high-dimensional integration

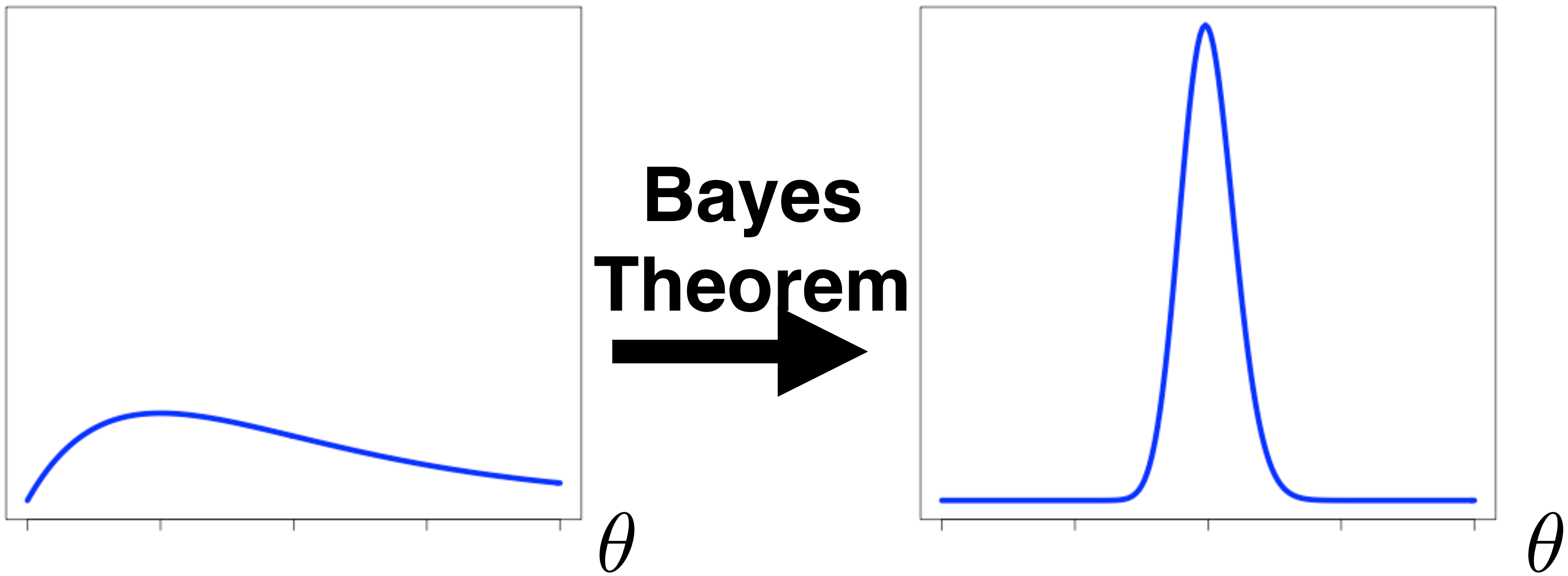
Bayesian inference

e.g. sensors e.g. pollution level

data parameters

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

posterior likelihood prior evidence



0. Identify a data analysis goal e.g. estimate pollution level
1. Build a model: choose prior & choose likelihood
2. Report a posterior summary, e.g. means and (co)variances

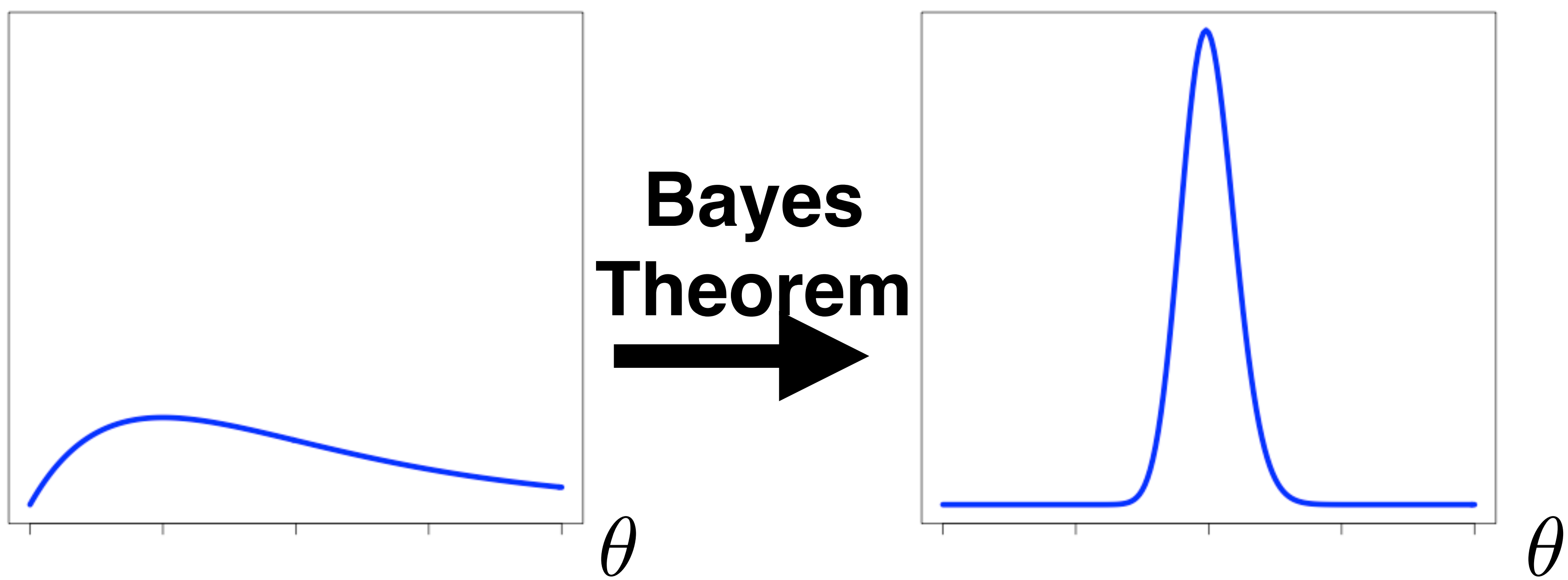
- Why is the final step hard?
  - Typically no closed form, high-dimensional integration

Bayesian inference

e.g. sensors — data; e.g. pollution level — parameters

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

posterior likelihood prior evidence



0. Identify a data analysis goal e.g. estimate pollution level
1. Build a model: choose prior & choose likelihood
2. Report a posterior summary, e.g. means and (co)variances

- Why is the final step hard?
  - Typically no closed form, high-dimensional integration

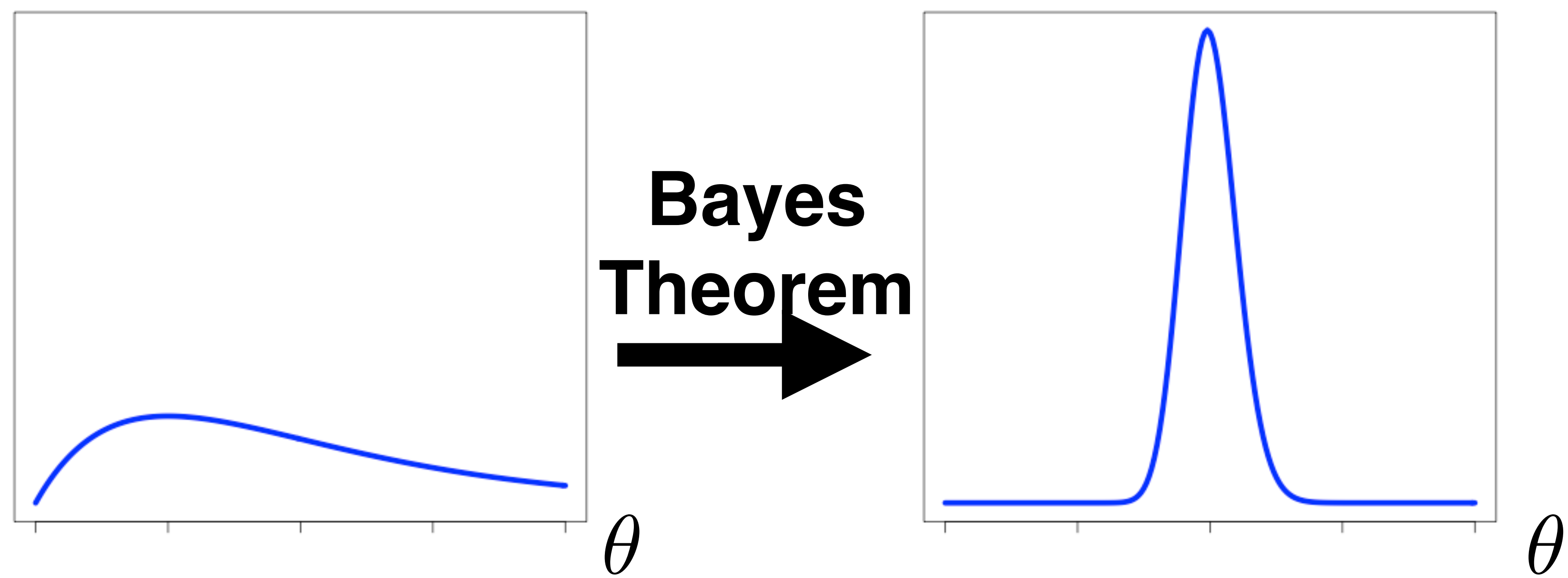
Bayesian inference

e.g. sensors — data

e.g. pollution level — parameters

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

posterior — likelihood — prior — evidence



0. Identify a data analysis goal — e.g. estimate pollution level
1. Build a model: choose prior & choose likelihood
2. Report a posterior summary, e.g. means and (co)variances

- Why is the final step hard?
  - Typically no closed form, high-dimensional integration

Approximate Bayesian Inference

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)

[Bardenet, Doucet, Holmes 2017]

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]

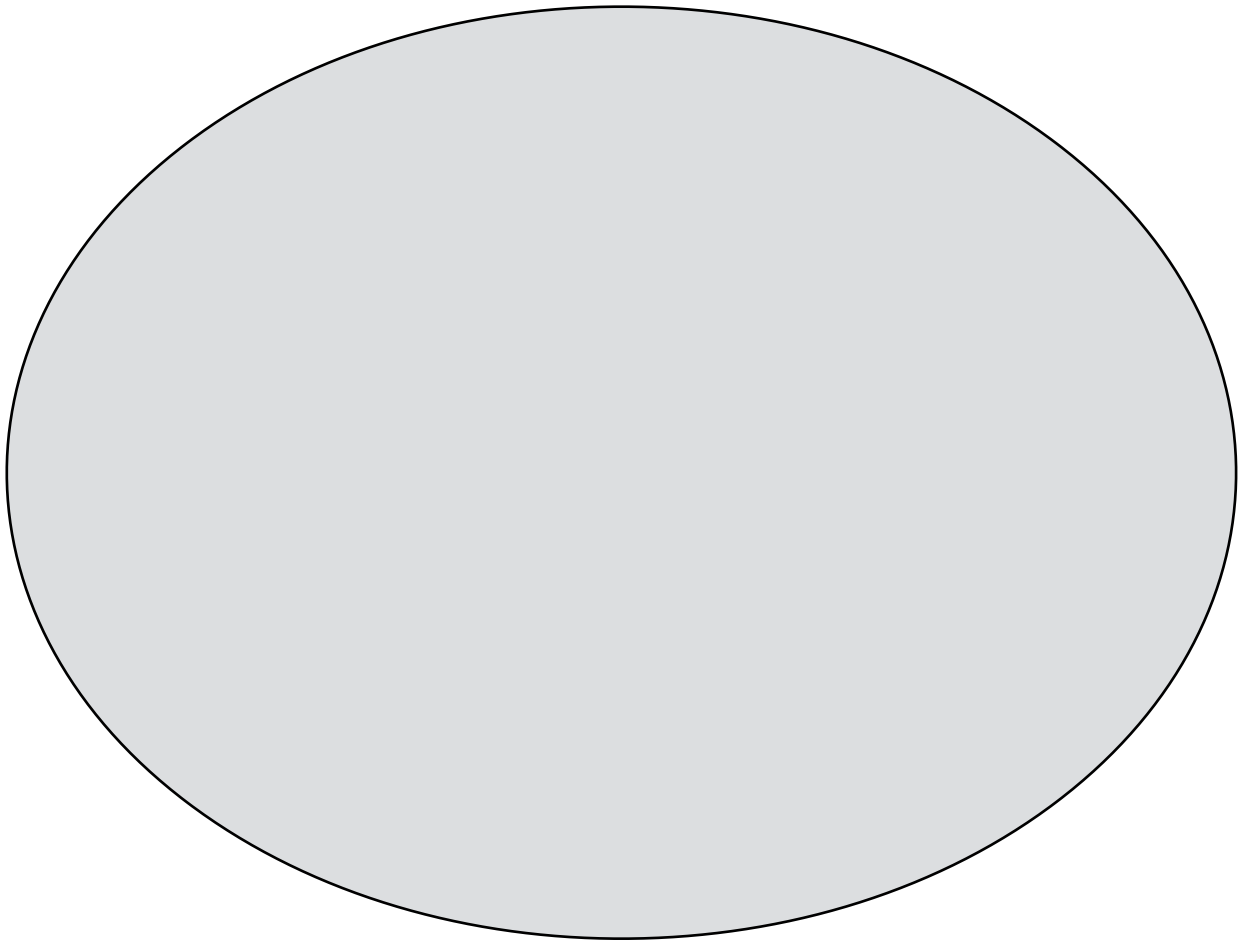
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



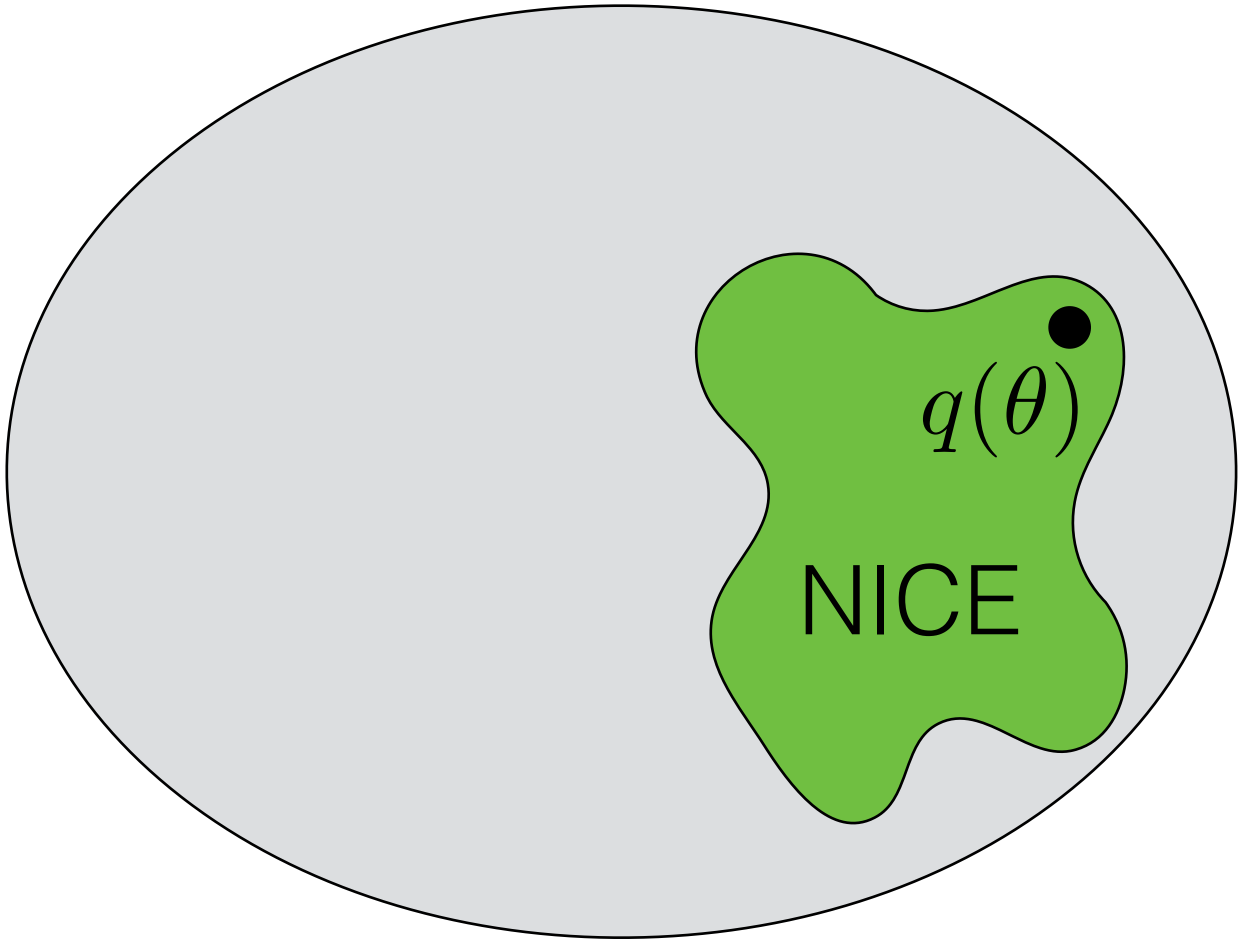
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



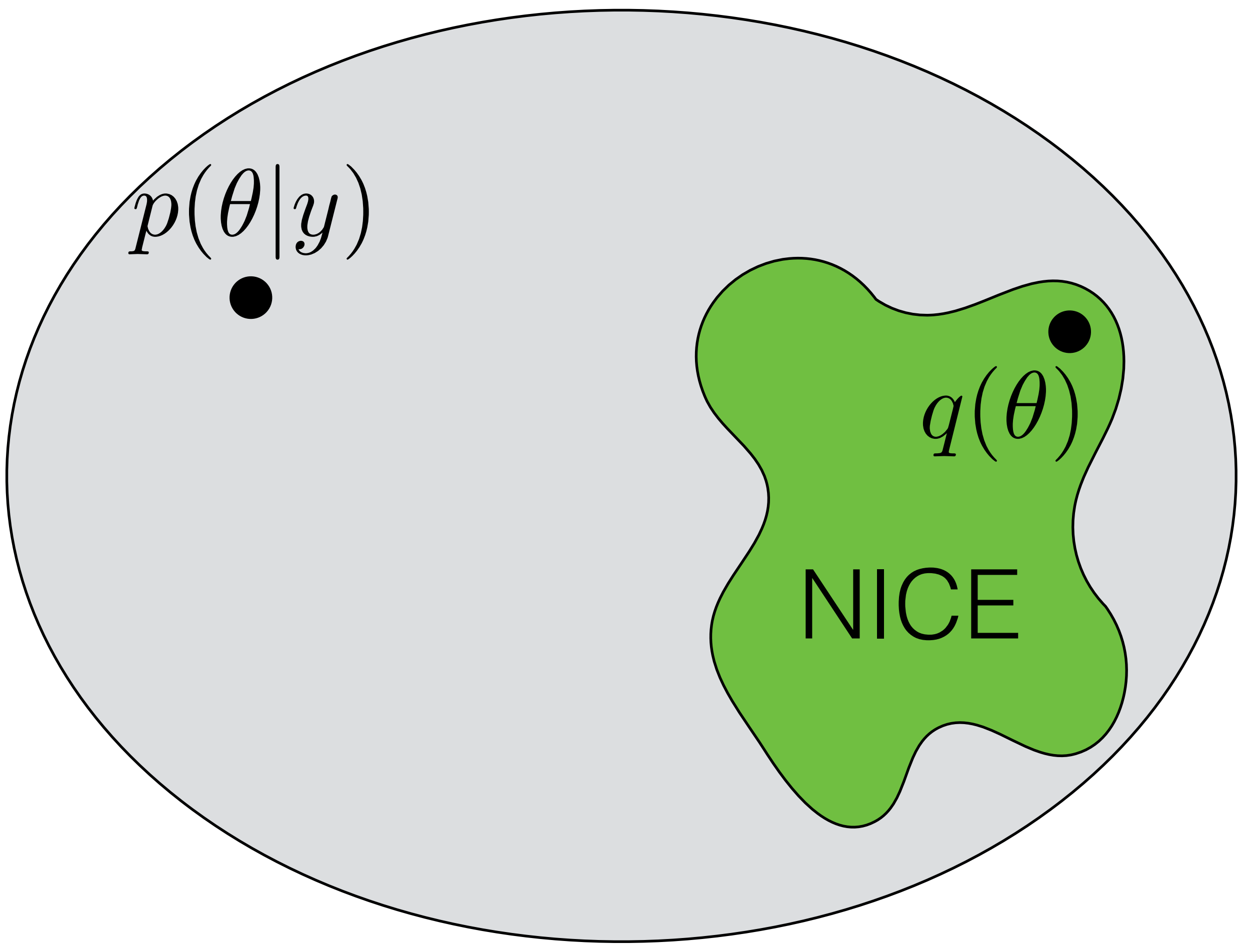
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



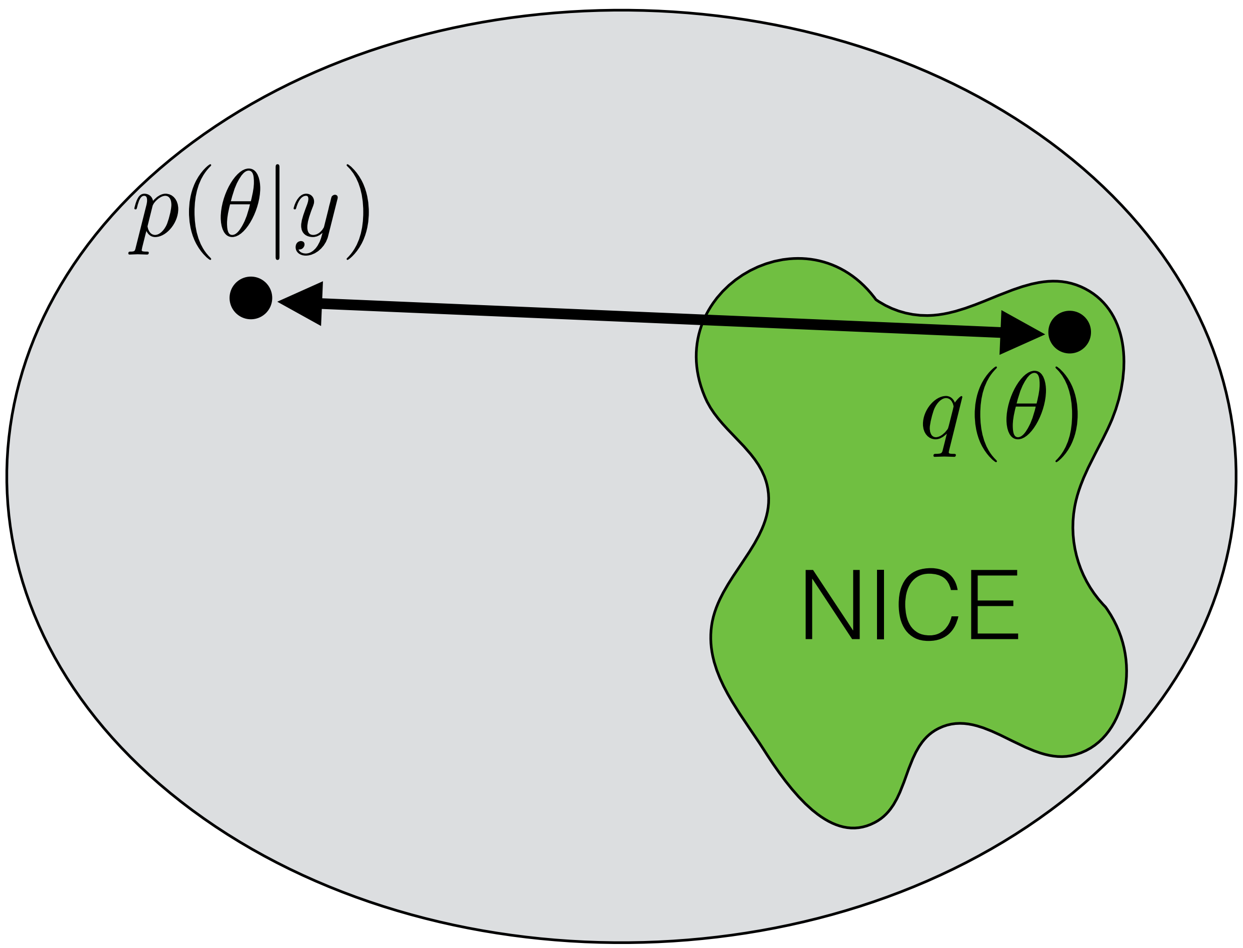
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



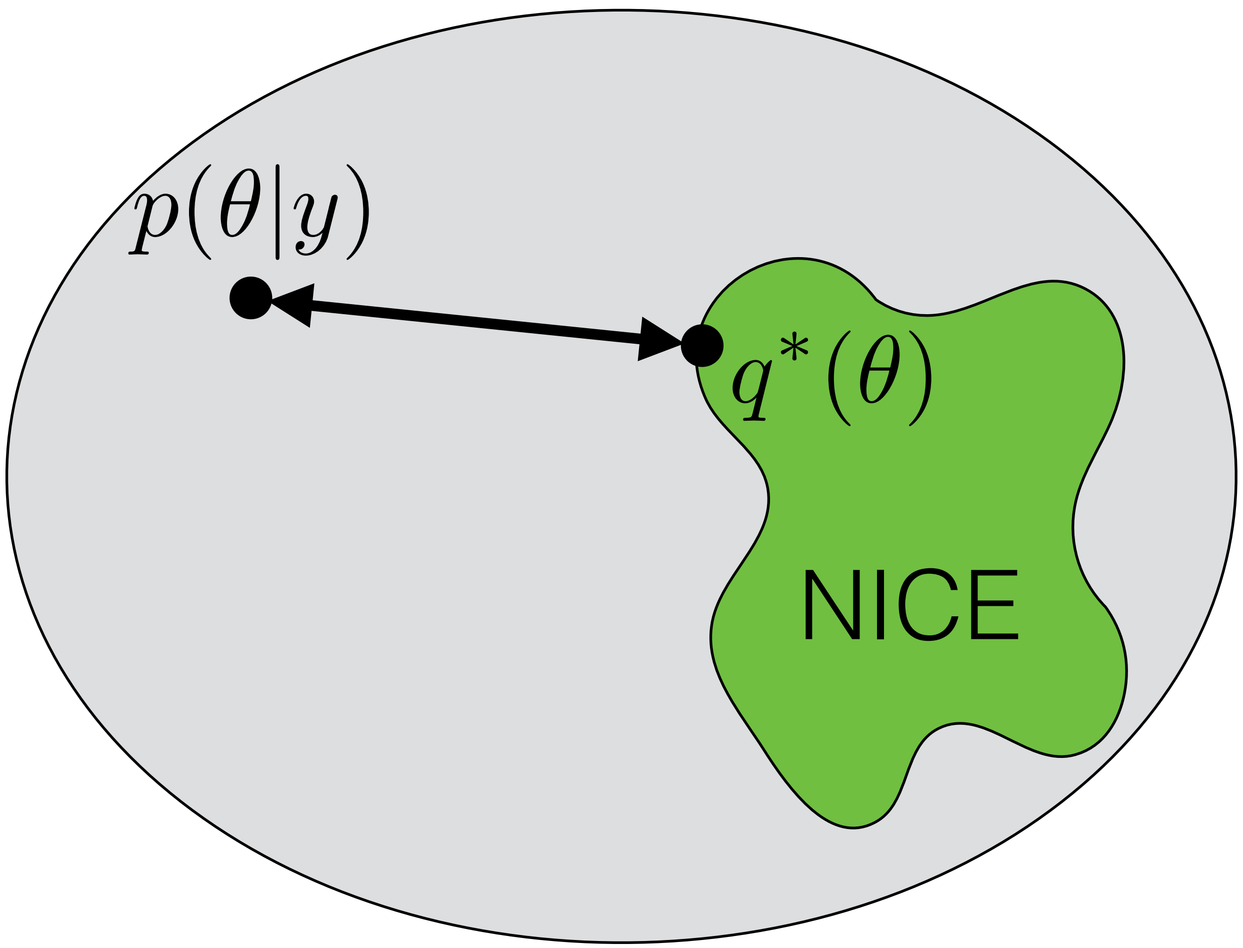
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



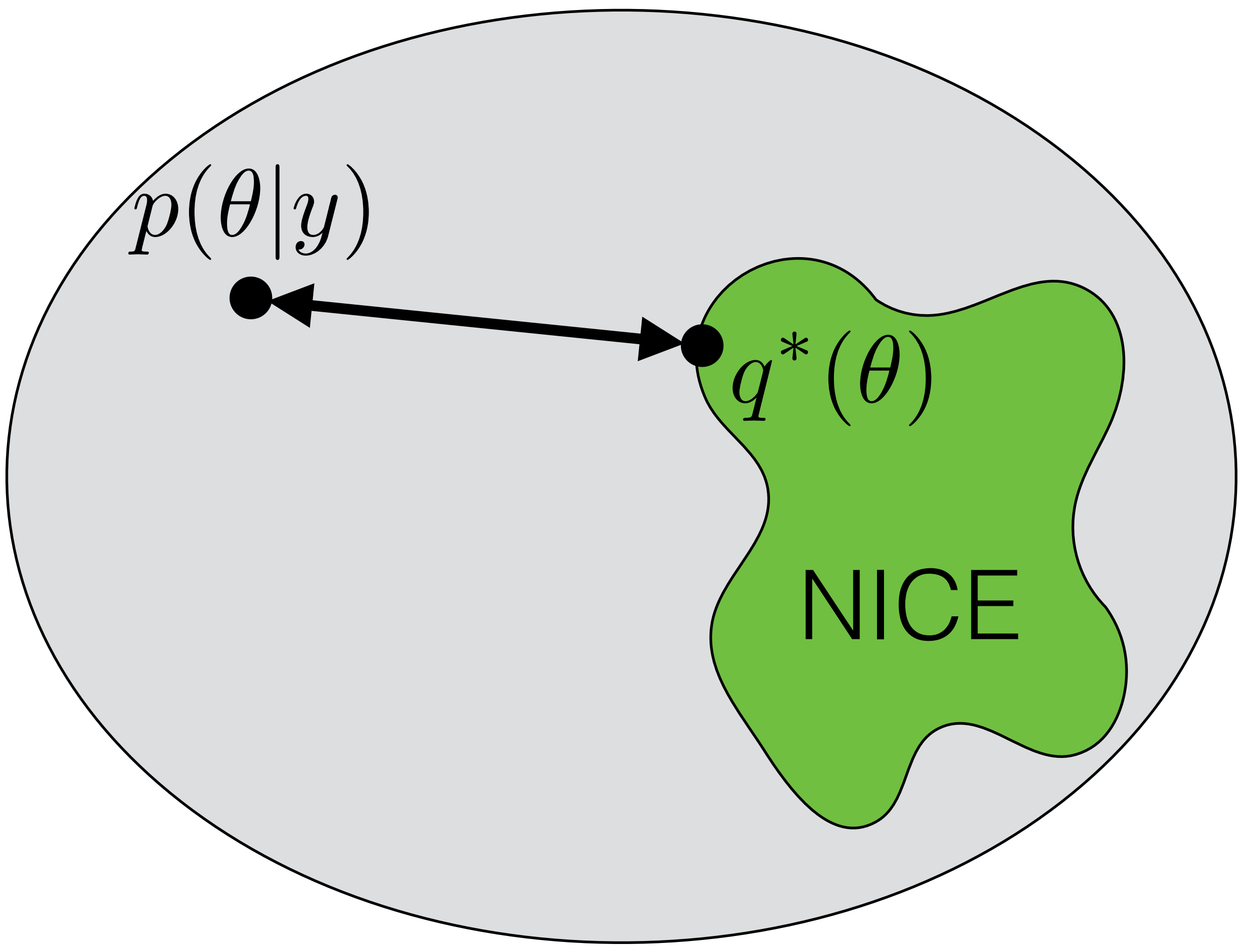
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

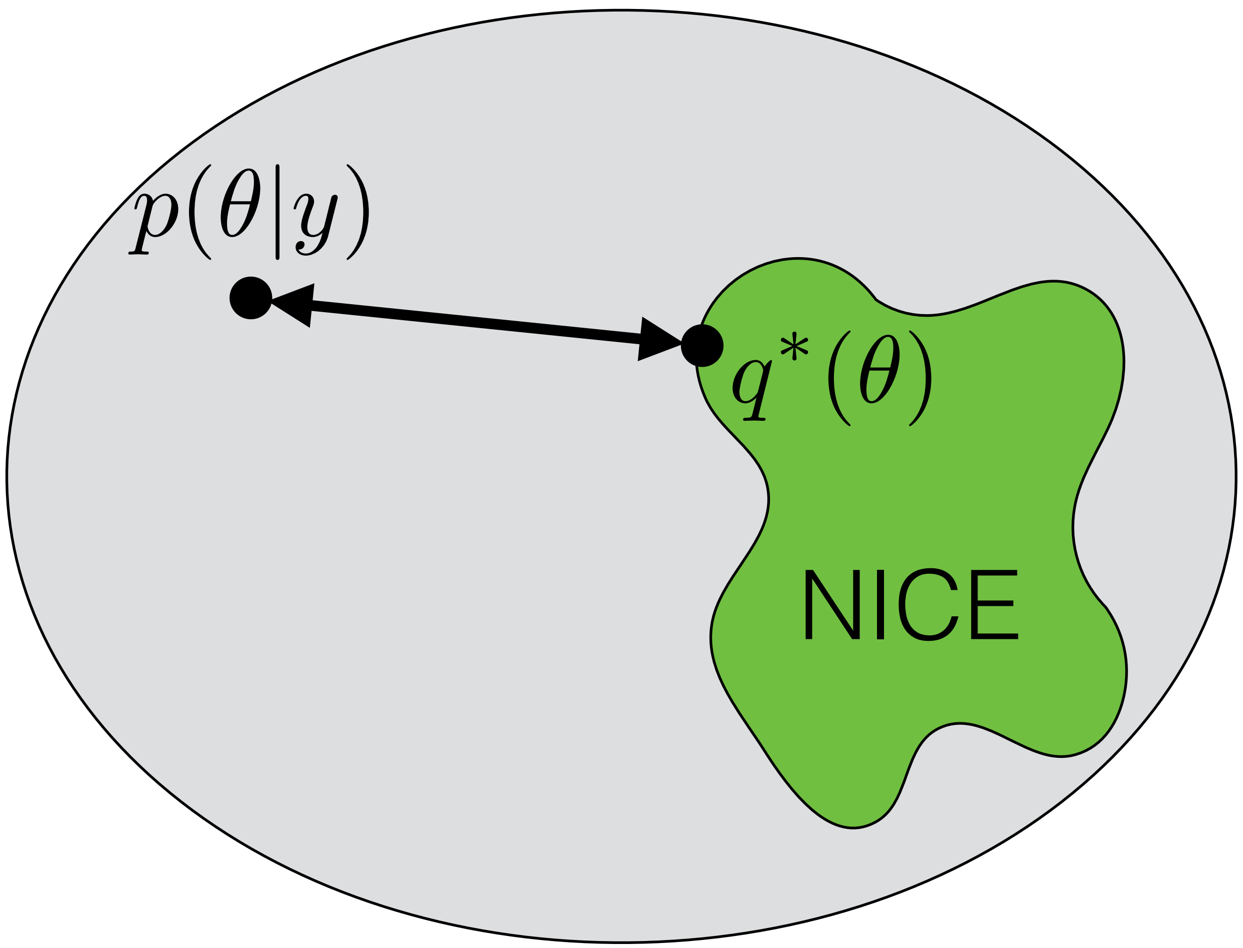
- Approximate posterior with q^*

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

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



Instead: an optimization approach

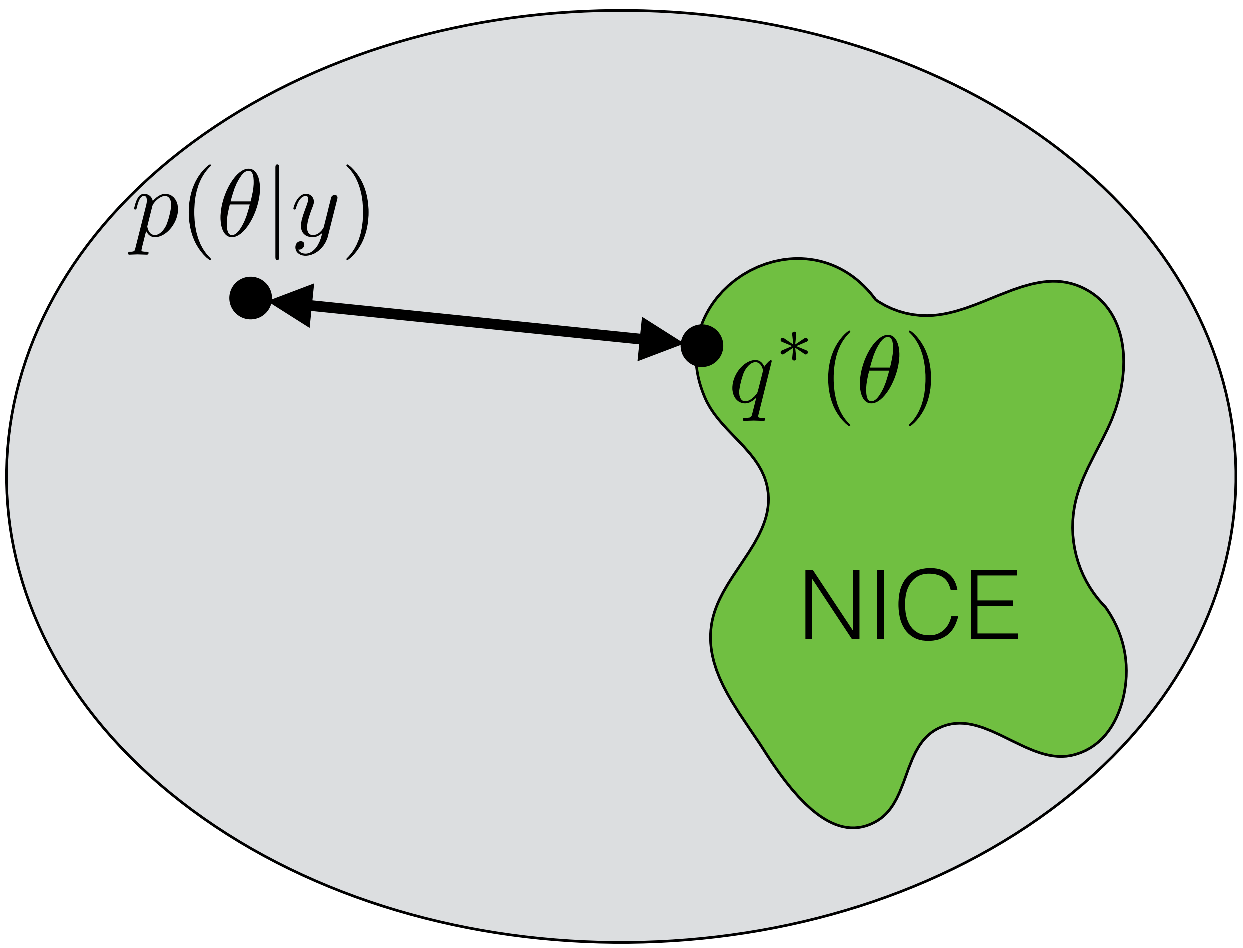
- Approximate posterior with q^*

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

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

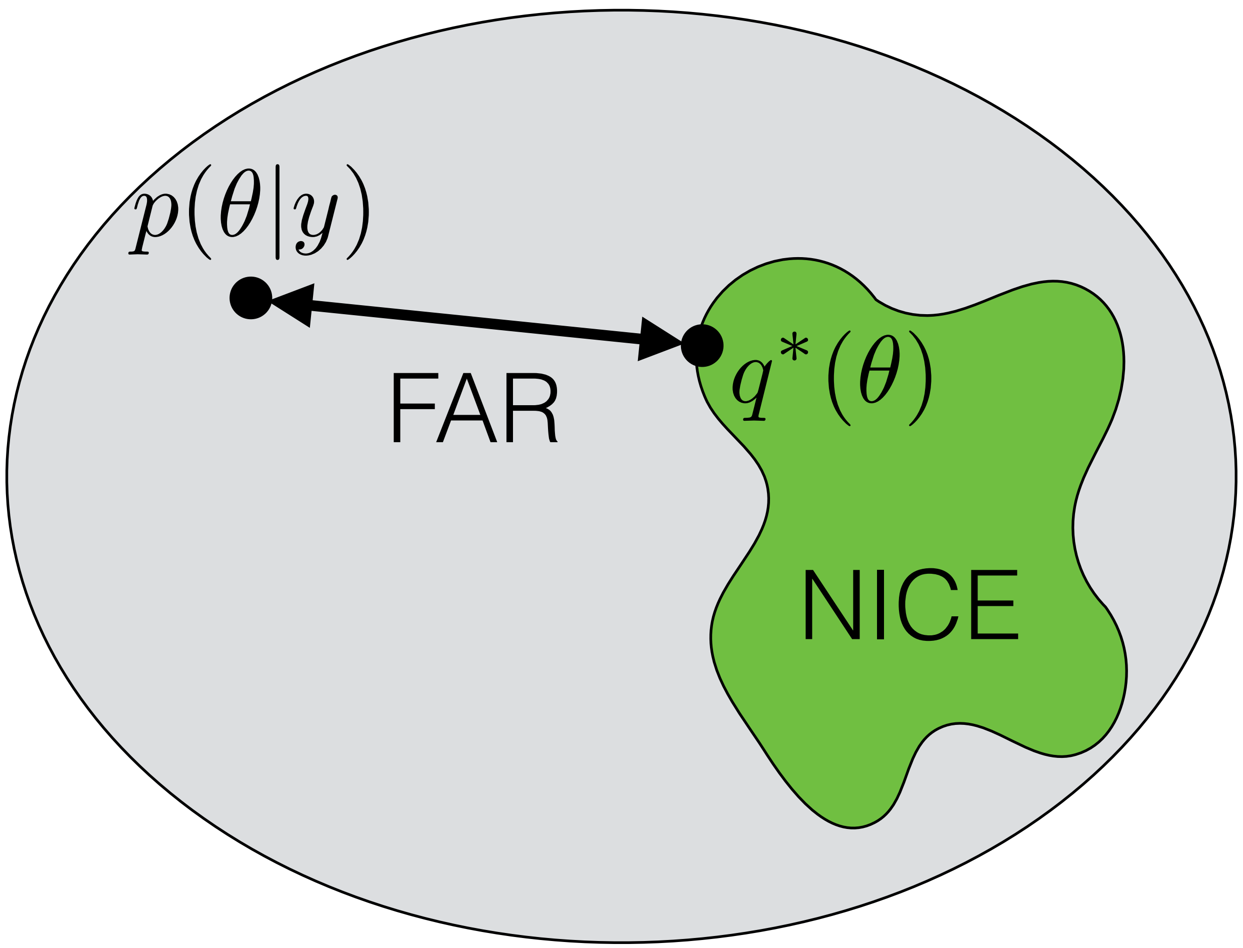
- Approximate posterior with q^*

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

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

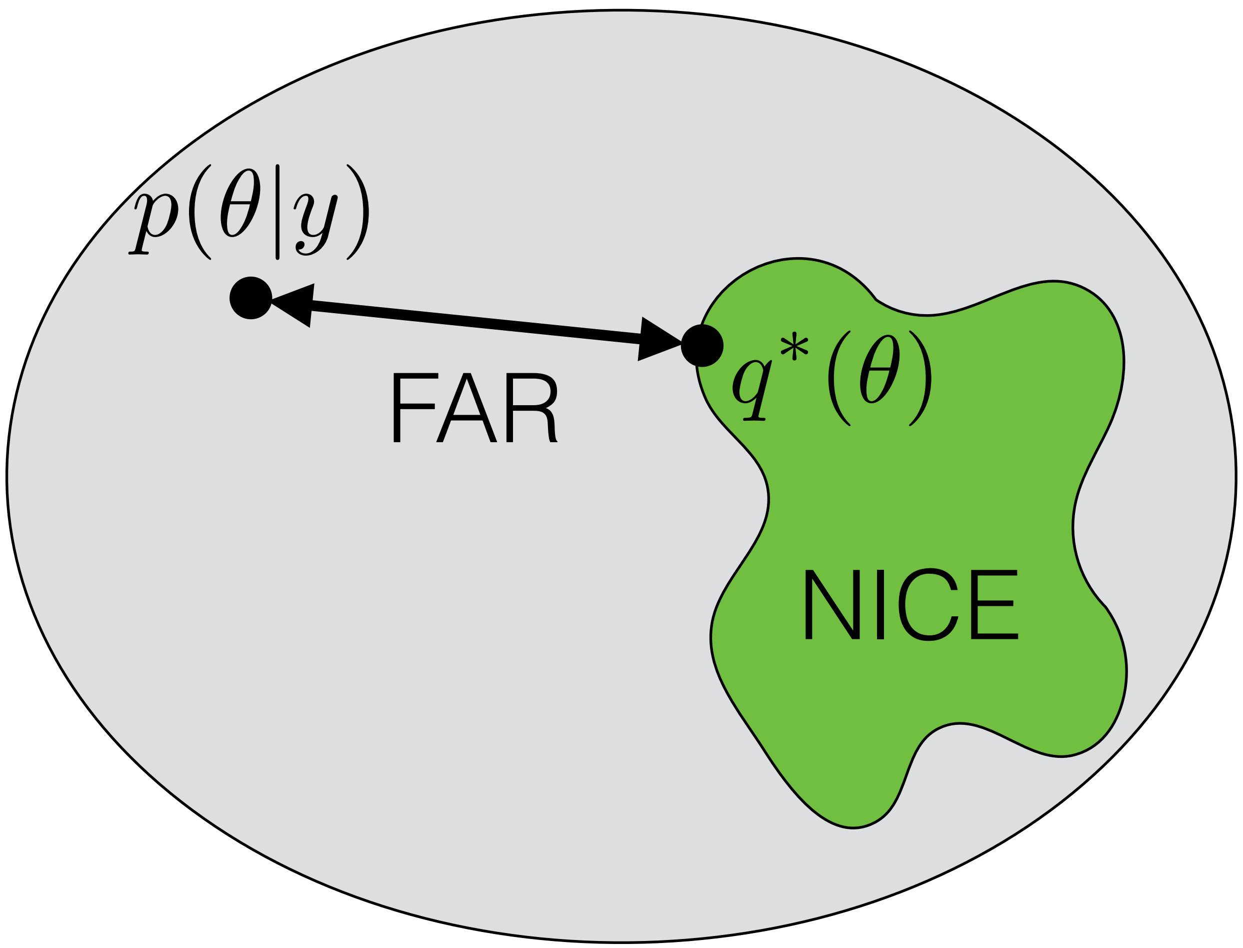
- Approximate posterior with q^*

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

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



Instead: an optimization approach

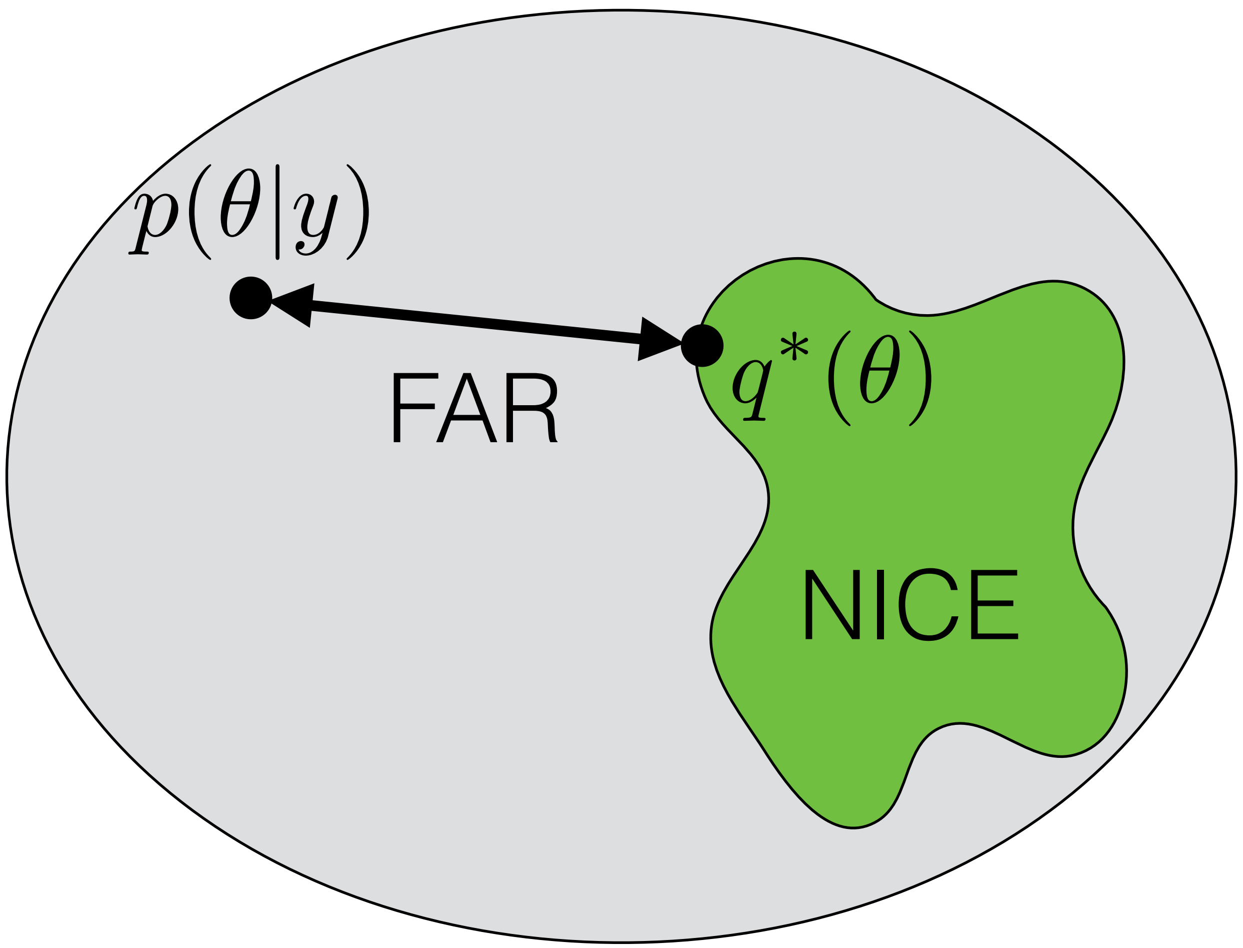
- Approximate posterior with q^*

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

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^*

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

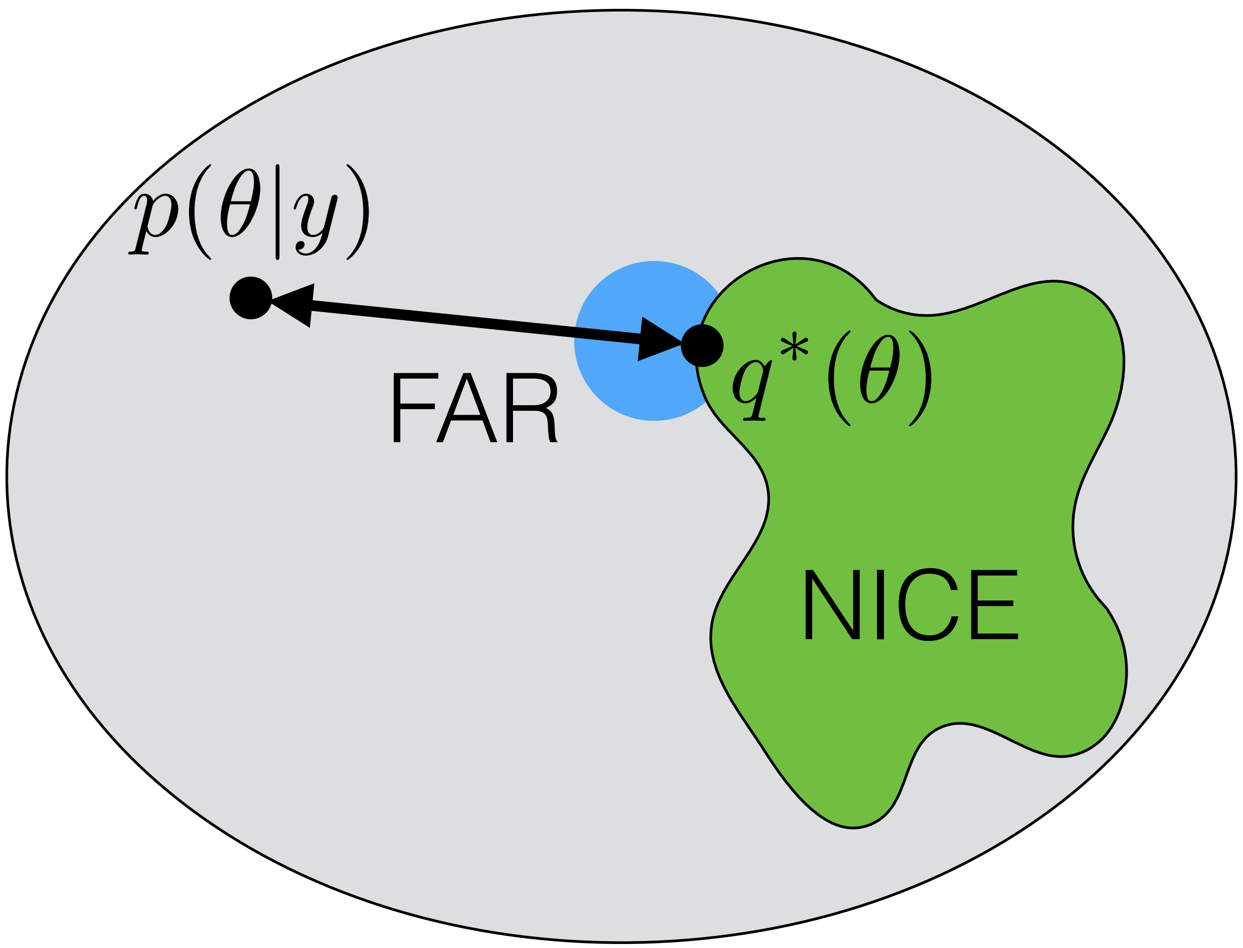
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

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

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^*

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

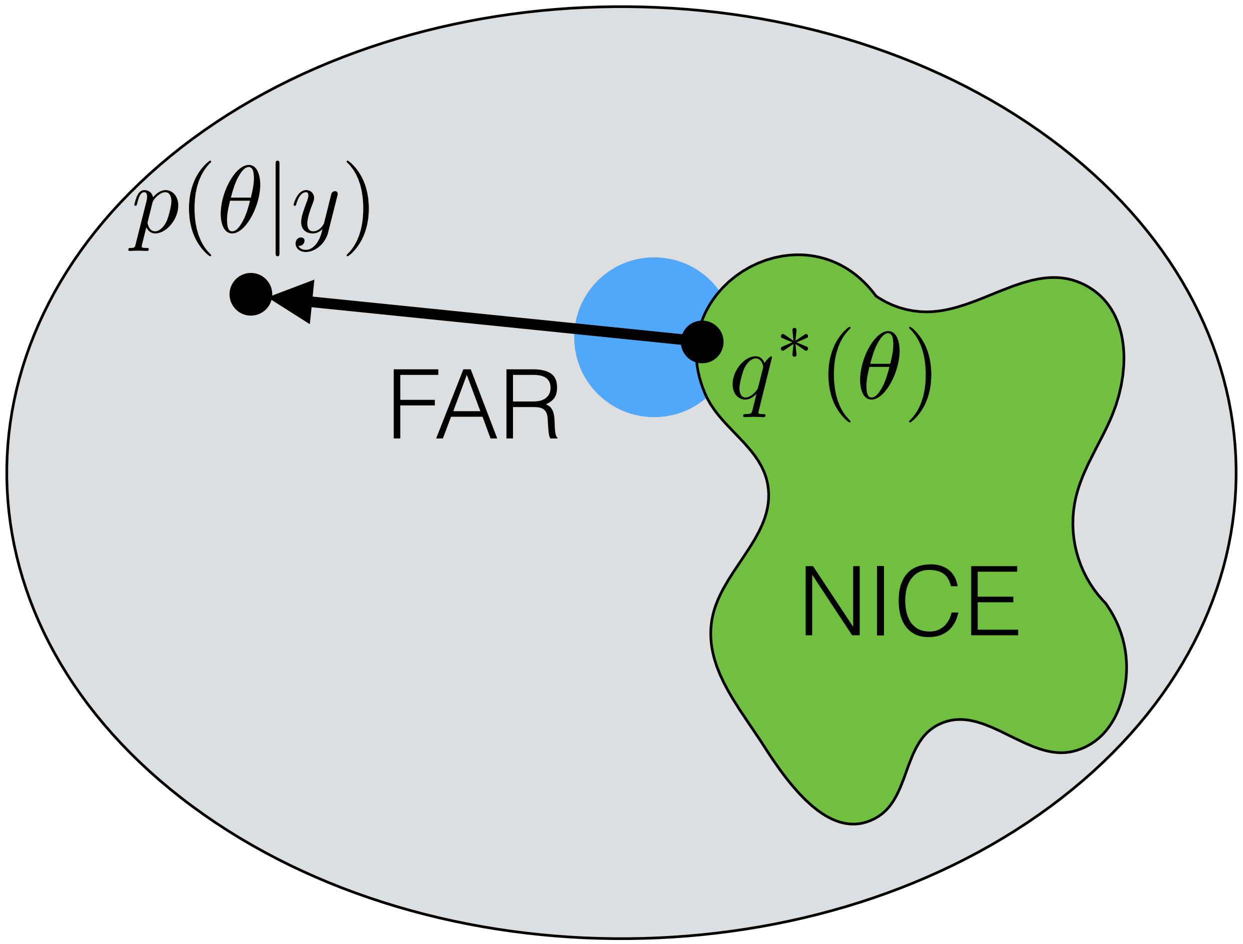
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

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

Approximate Bayesian Inference

- 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^*

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

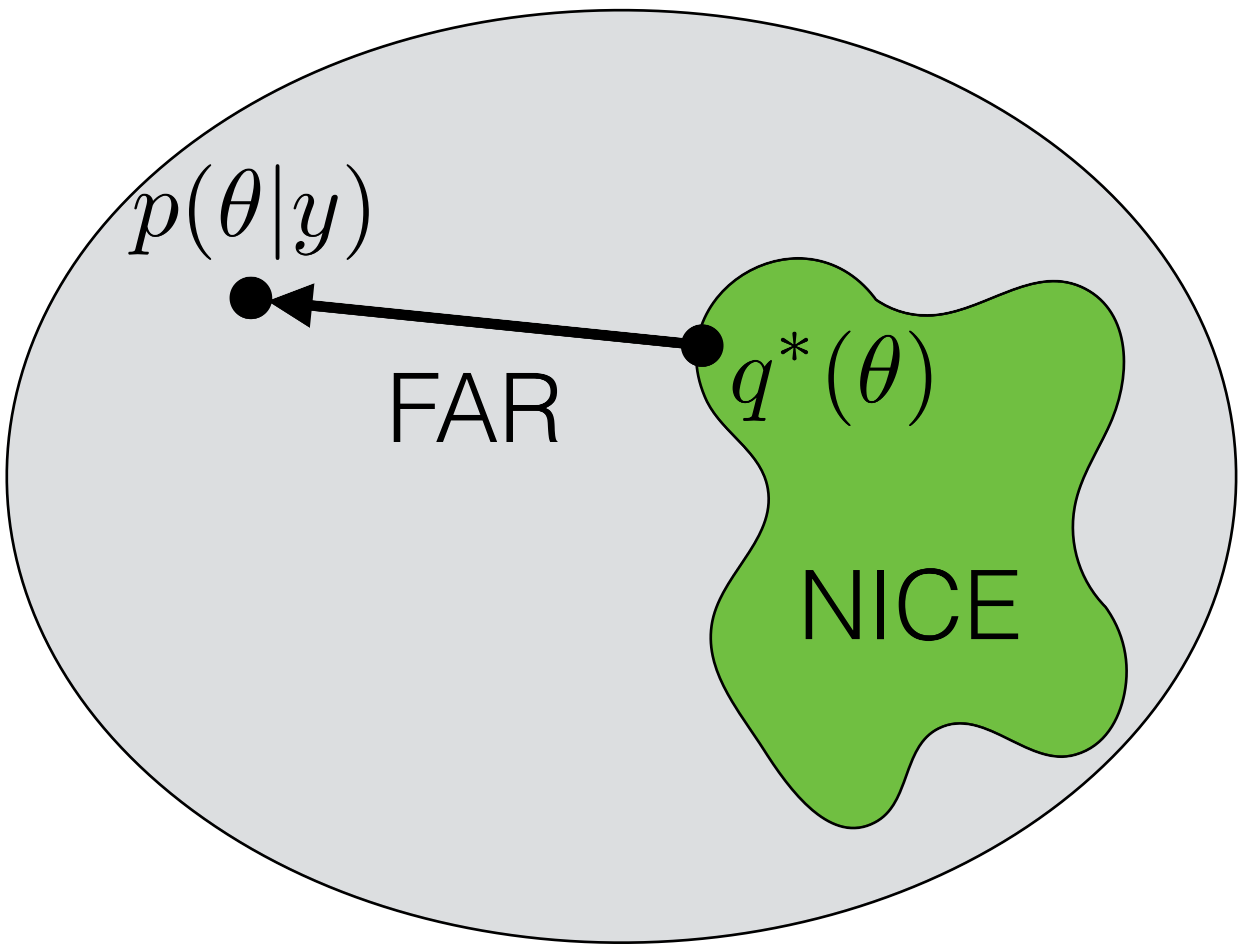
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

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

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^*

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

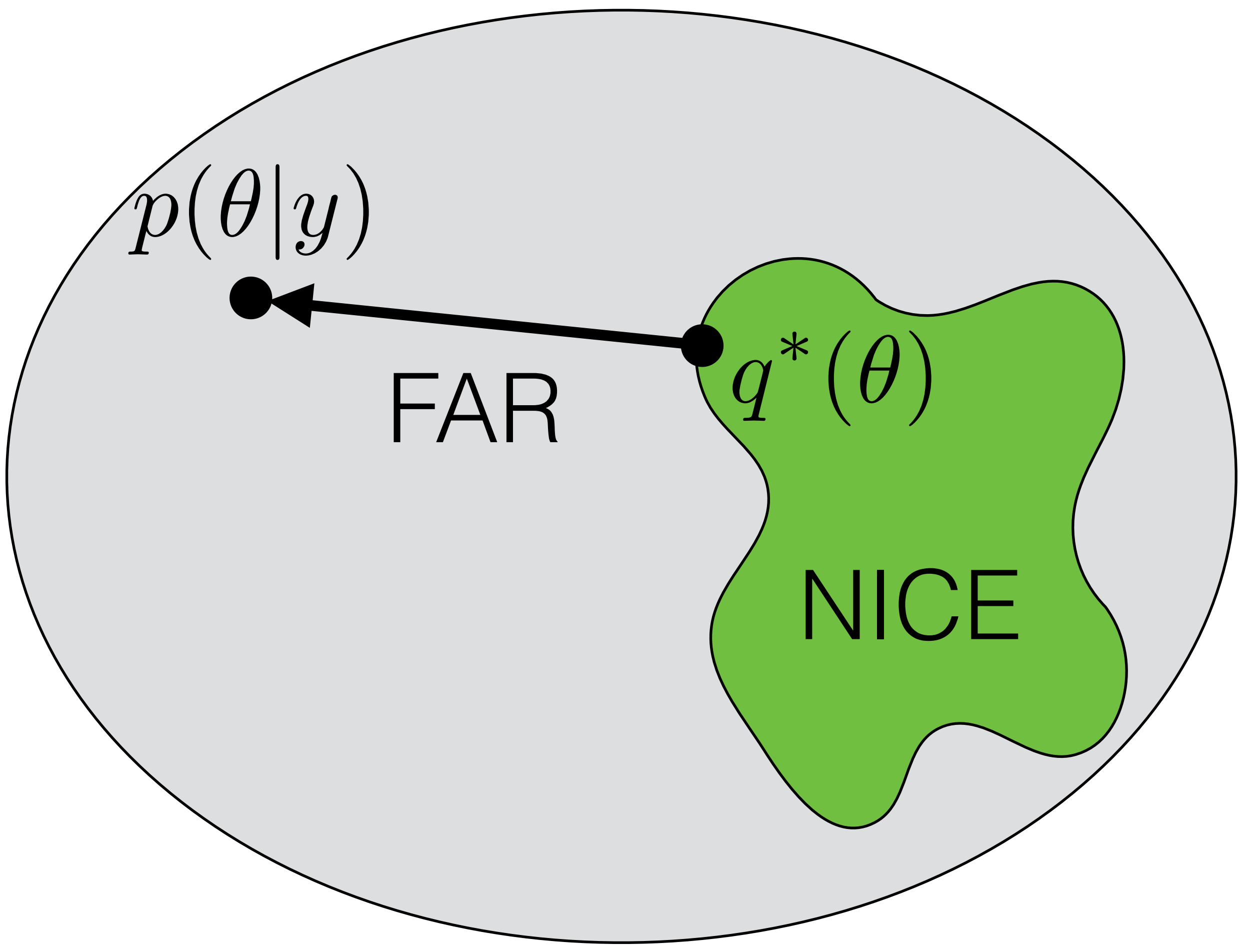
- Variational Bayes (VB): f is Kullback-Leibler divergence

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

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^*

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

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

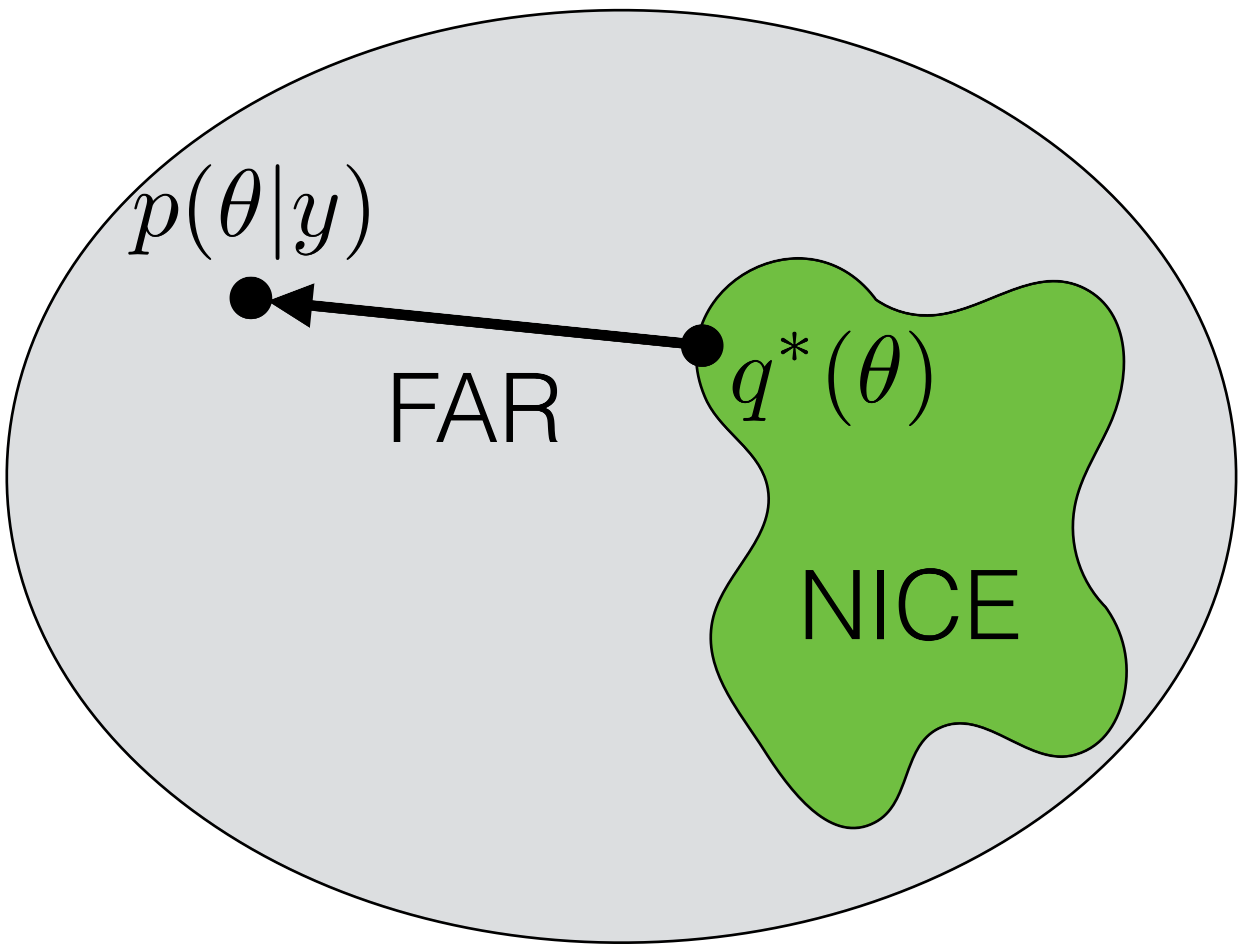
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success

Approximate Bayesian Inference

[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

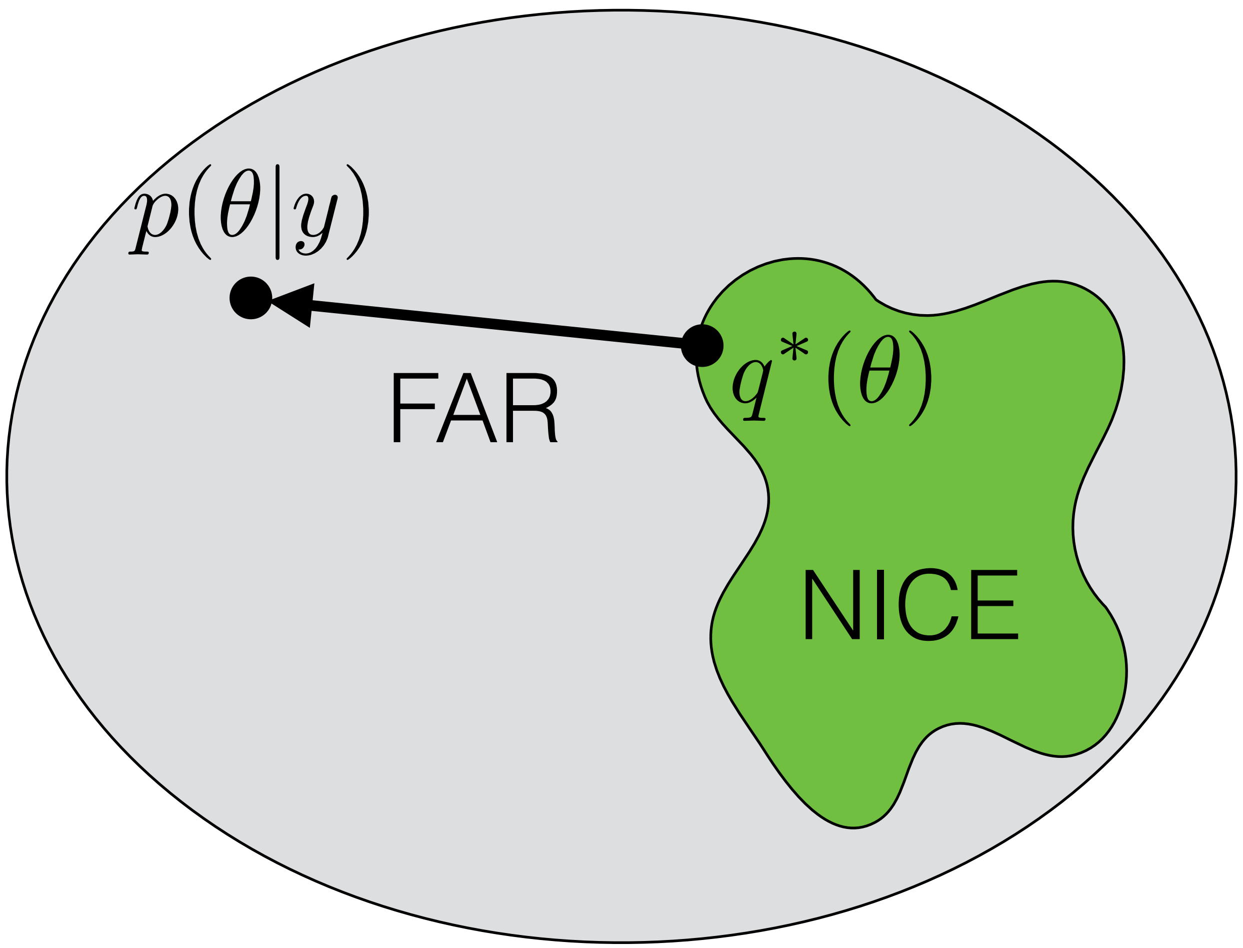
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success: point estimates and prediction

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

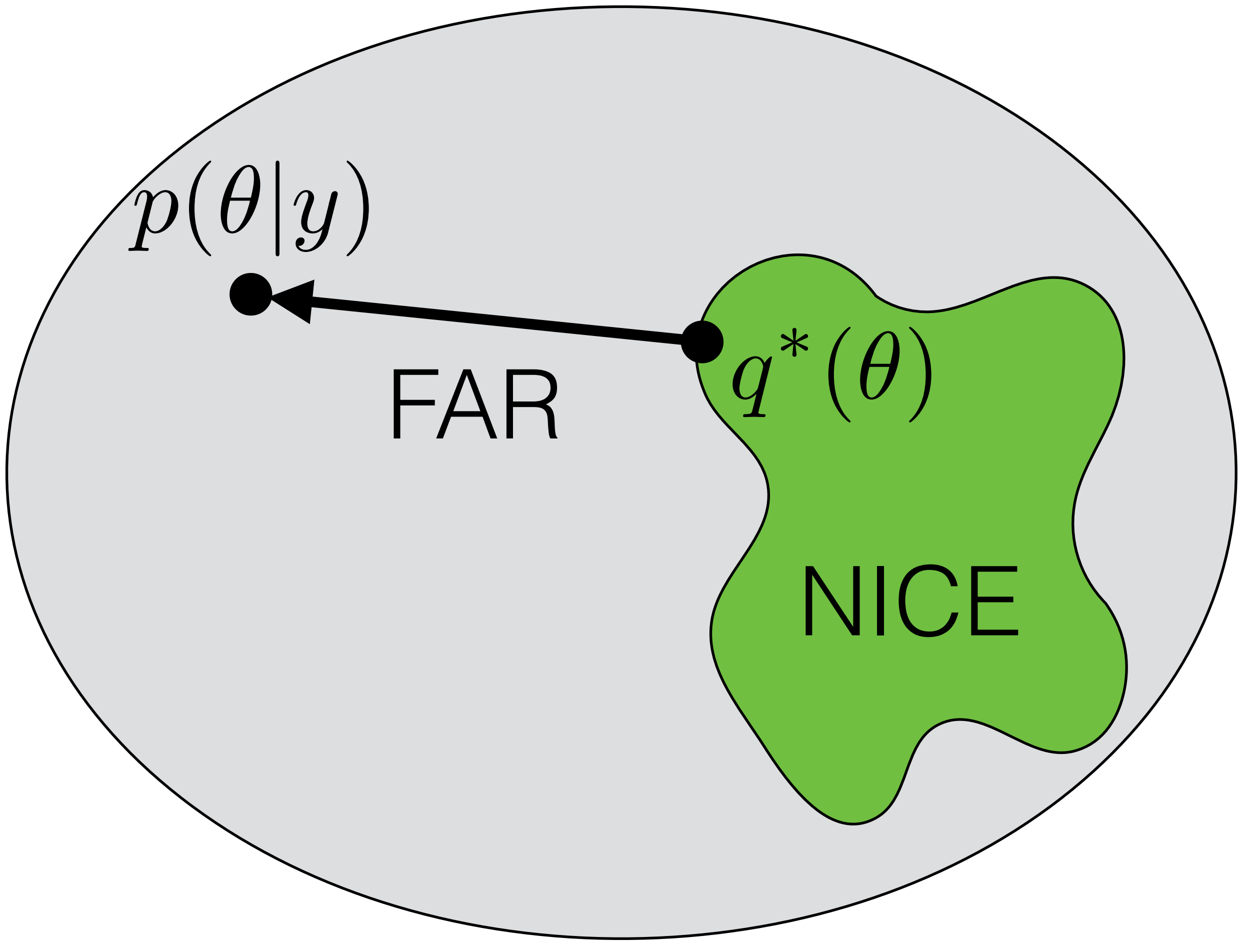
$$KL(q(\cdot)||p(\cdot|y))$$

- 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

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

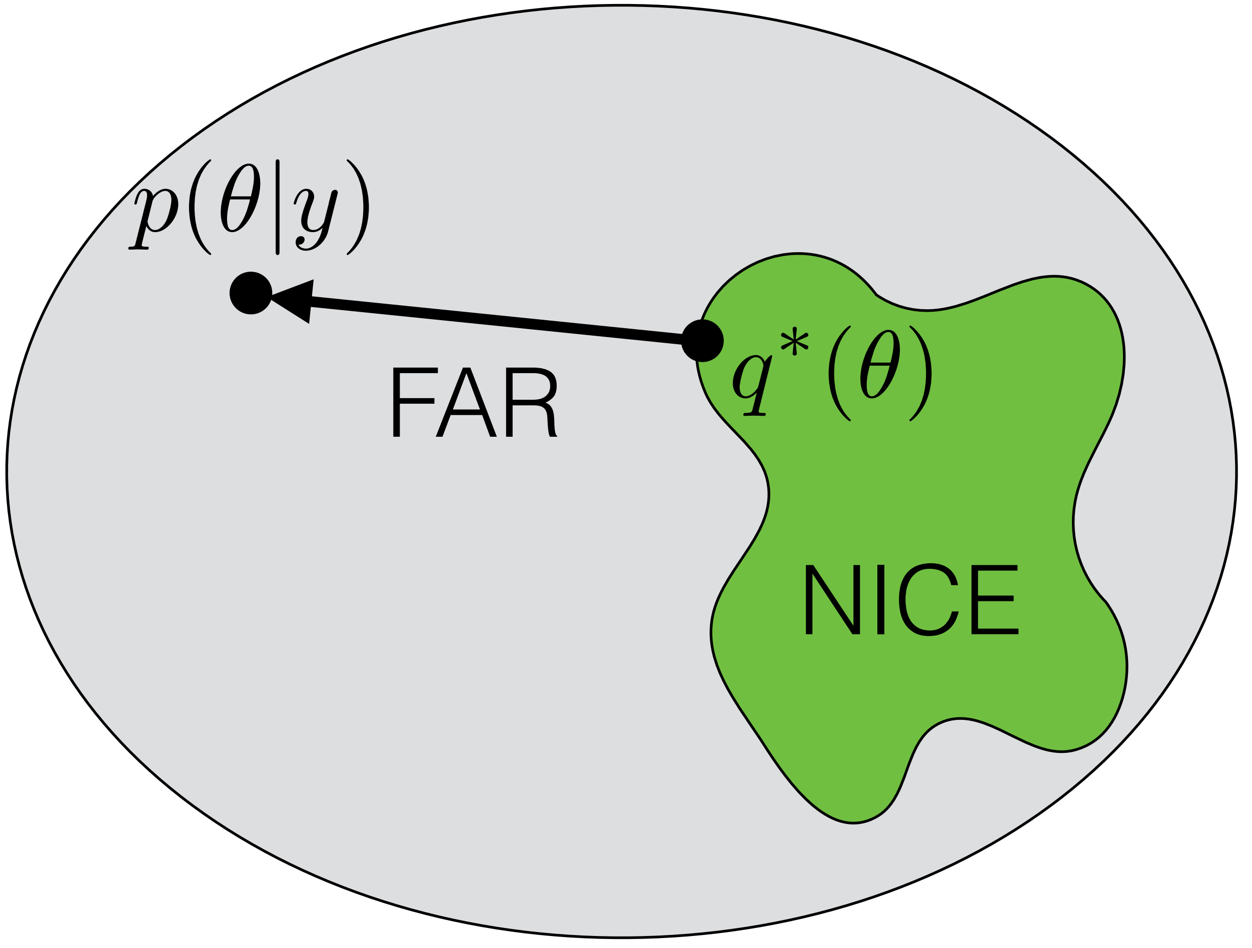
- VB practical success: point estimates and prediction, fast, streaming, distributed (3.6M Wikipedia, 350K Nature)

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

Why KL?

- Variational Bayes

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

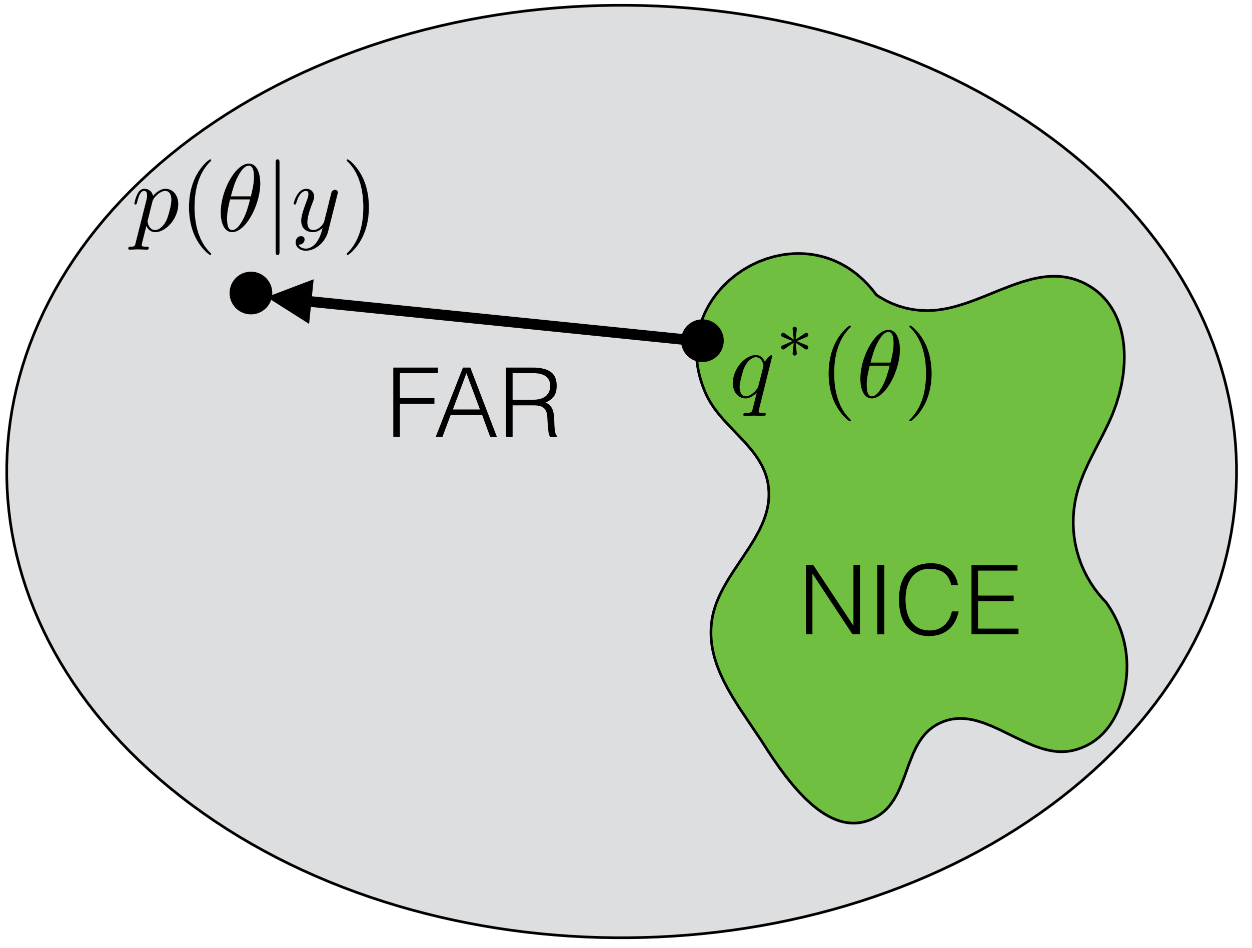


Why KL?

- Variational Bayes

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

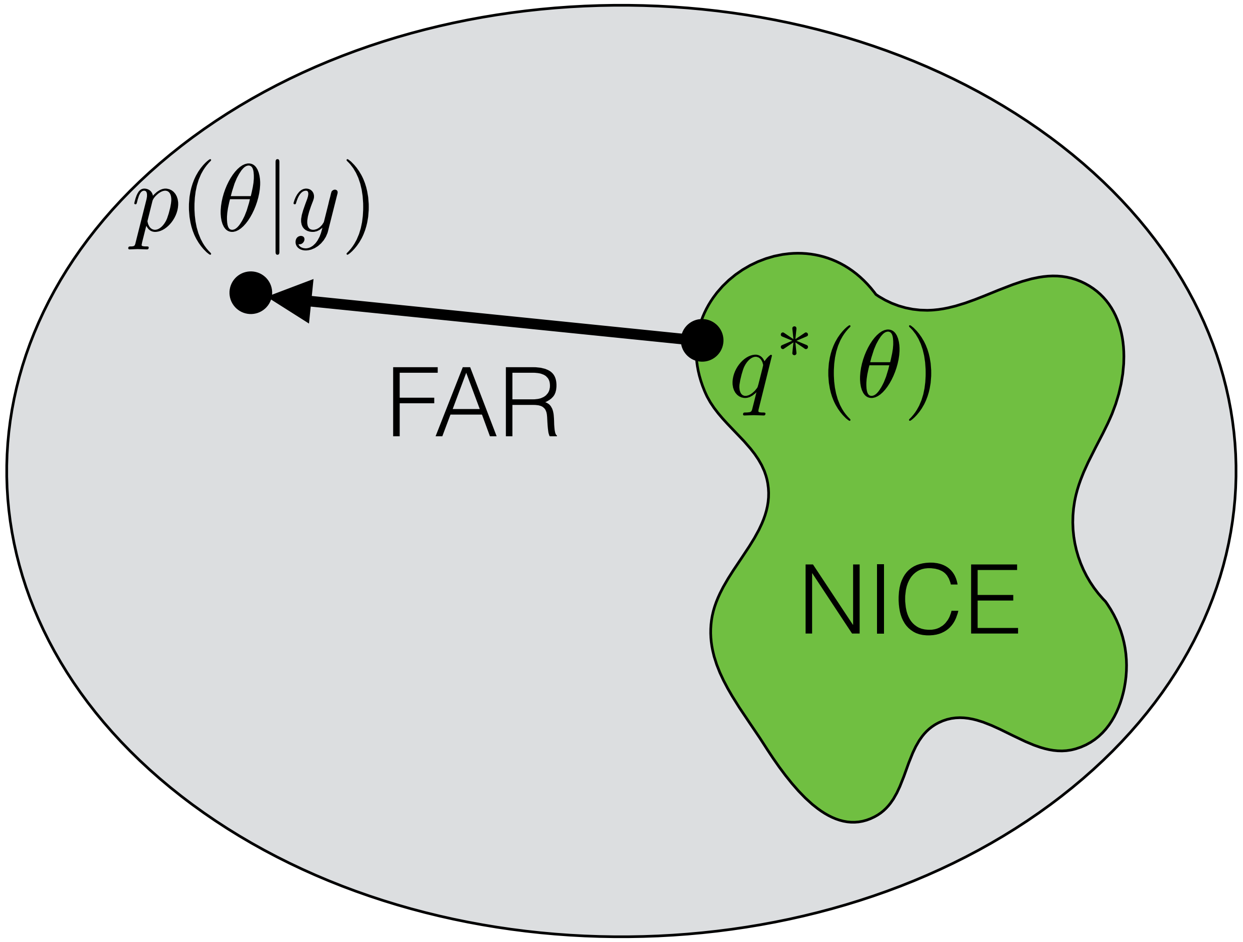
$$\text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$



Why KL?

- Variational Bayes

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



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

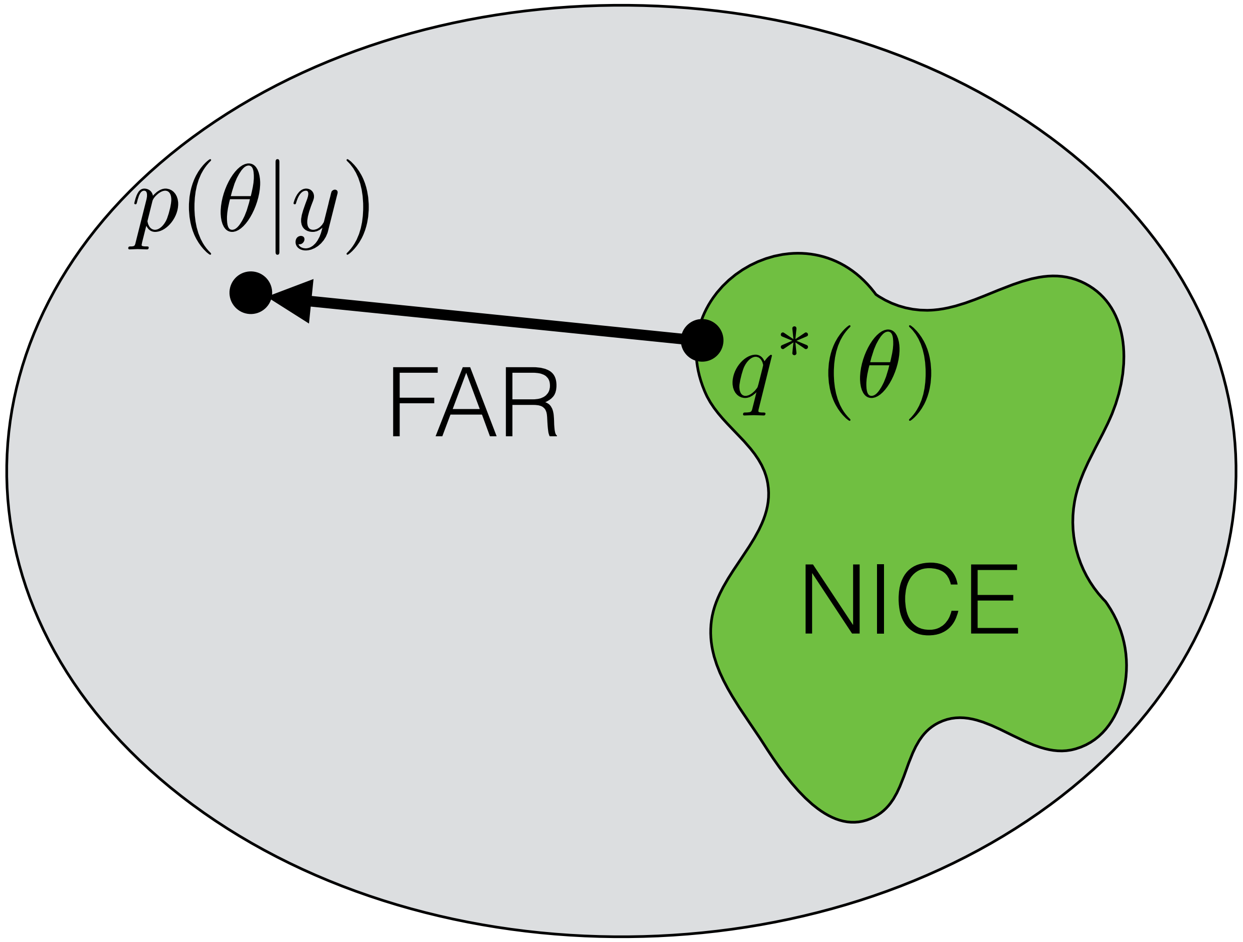
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

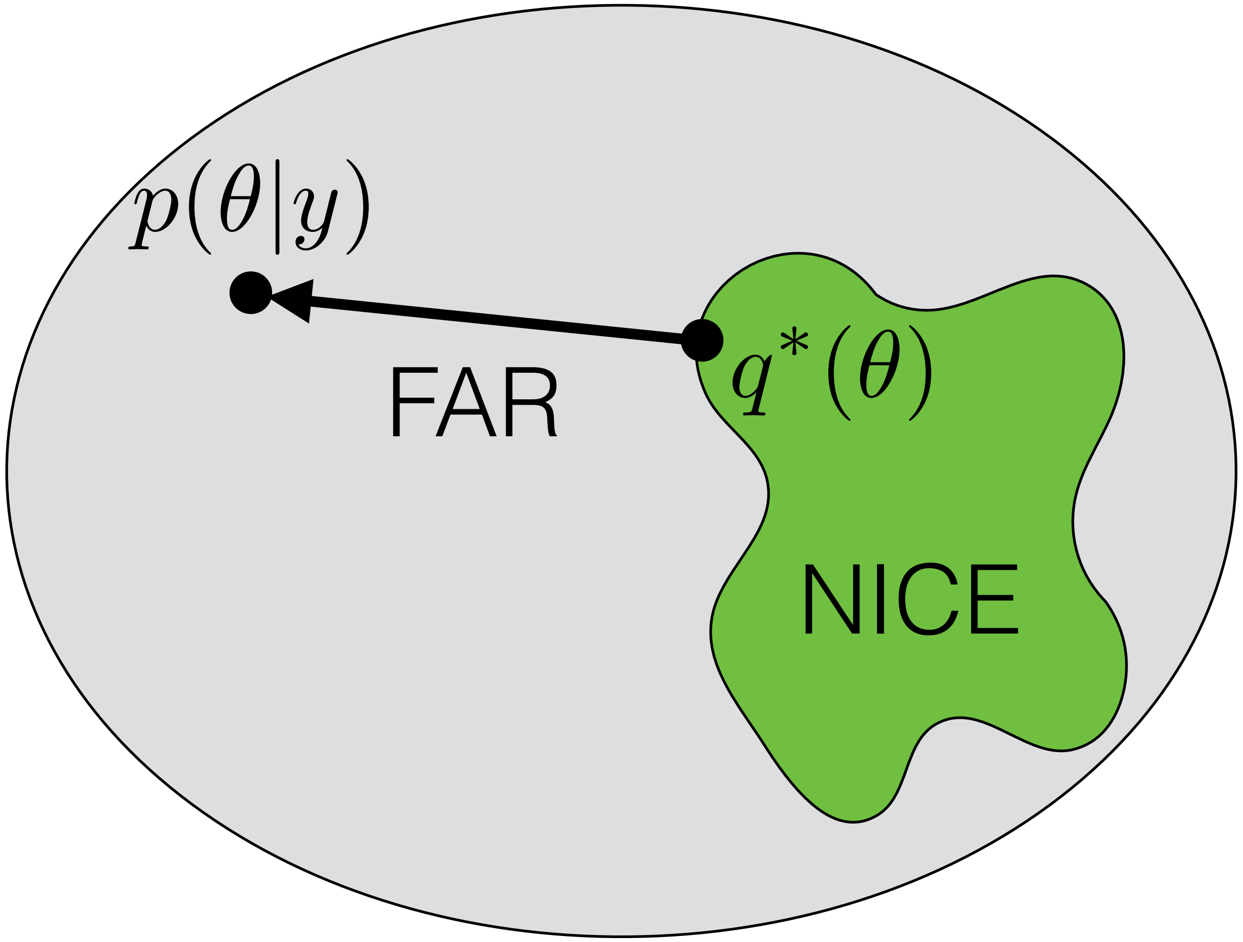
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

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

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

Why KL?

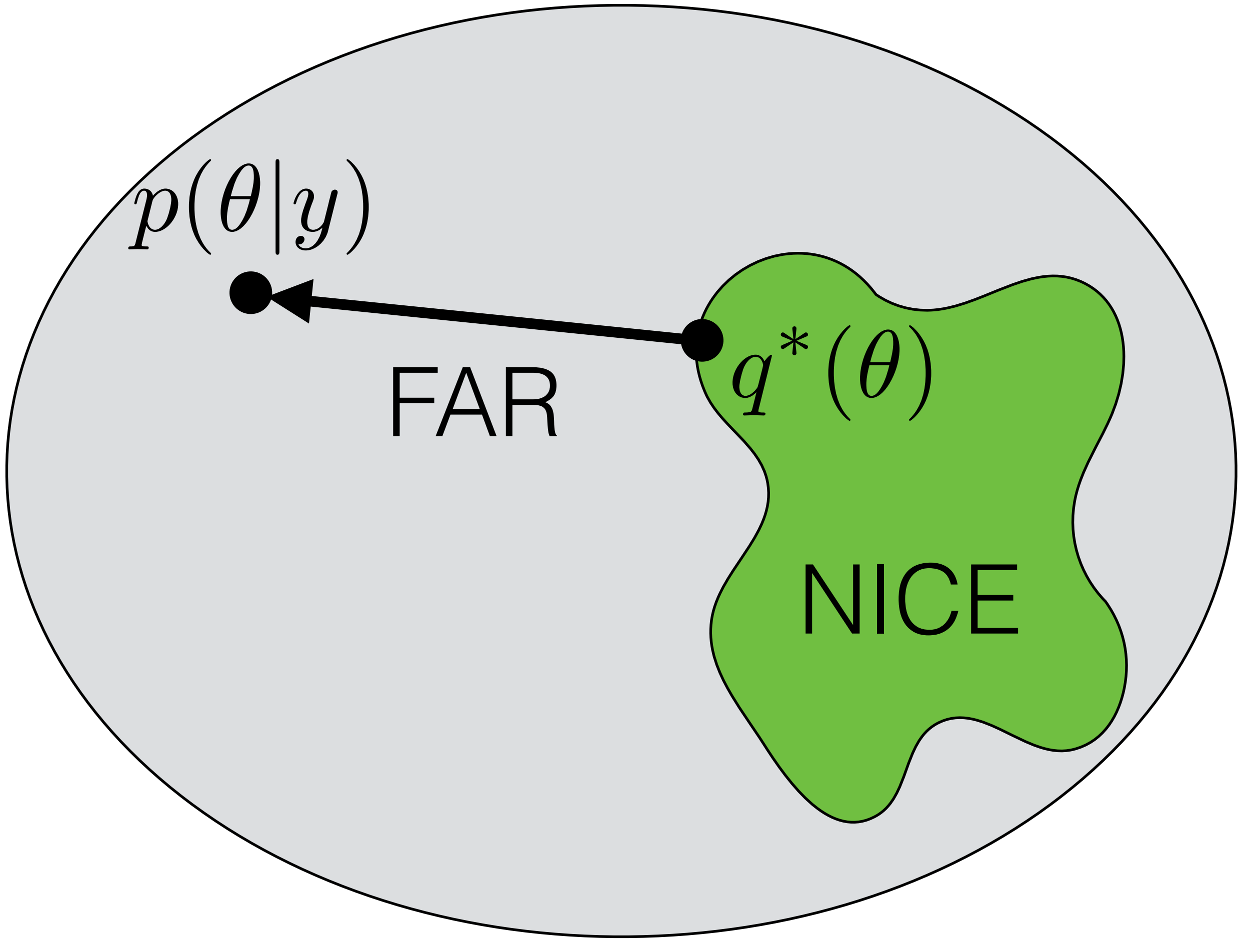
- Variational Bayes

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

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

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

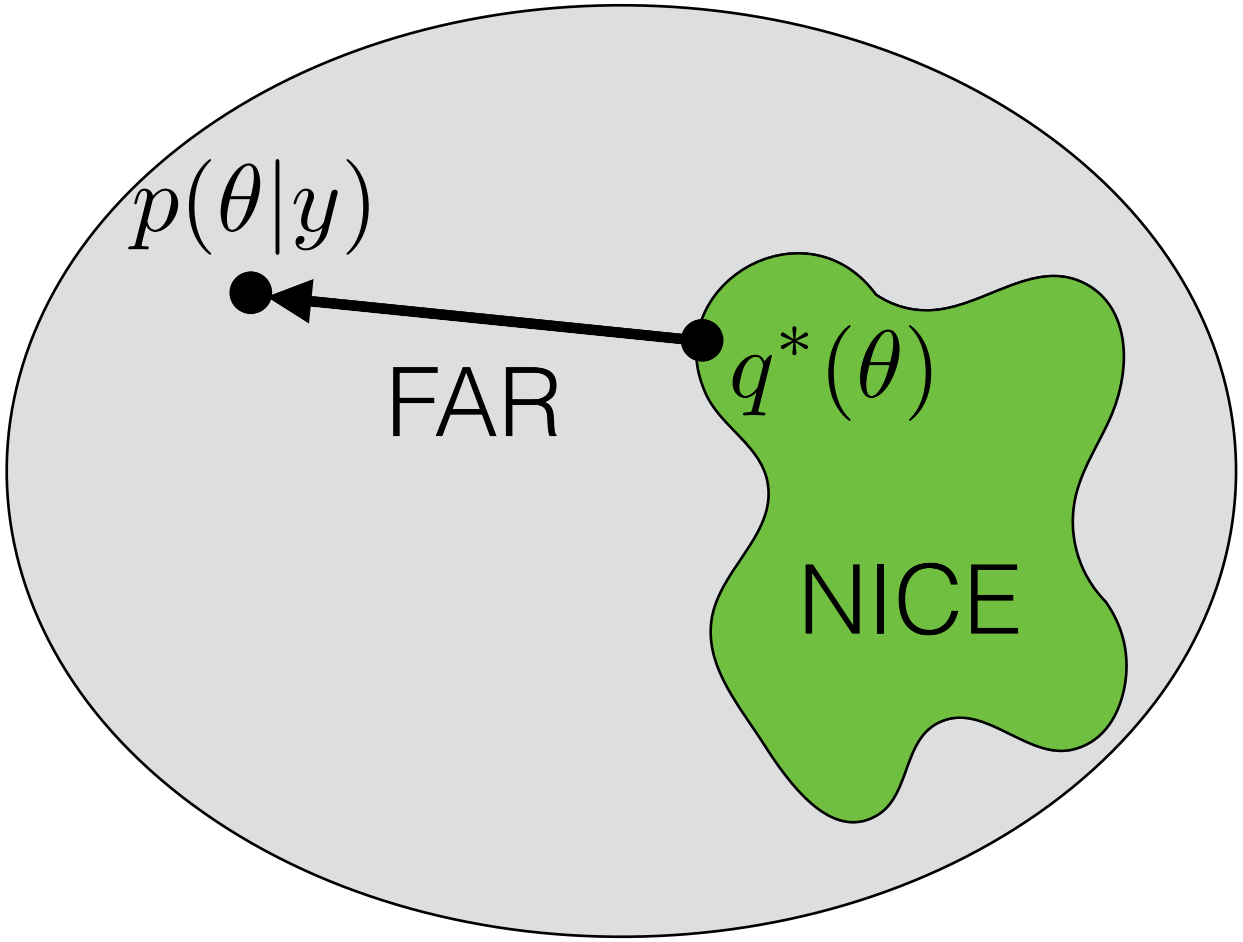
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta$$



Why KL?

- Variational Bayes

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



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

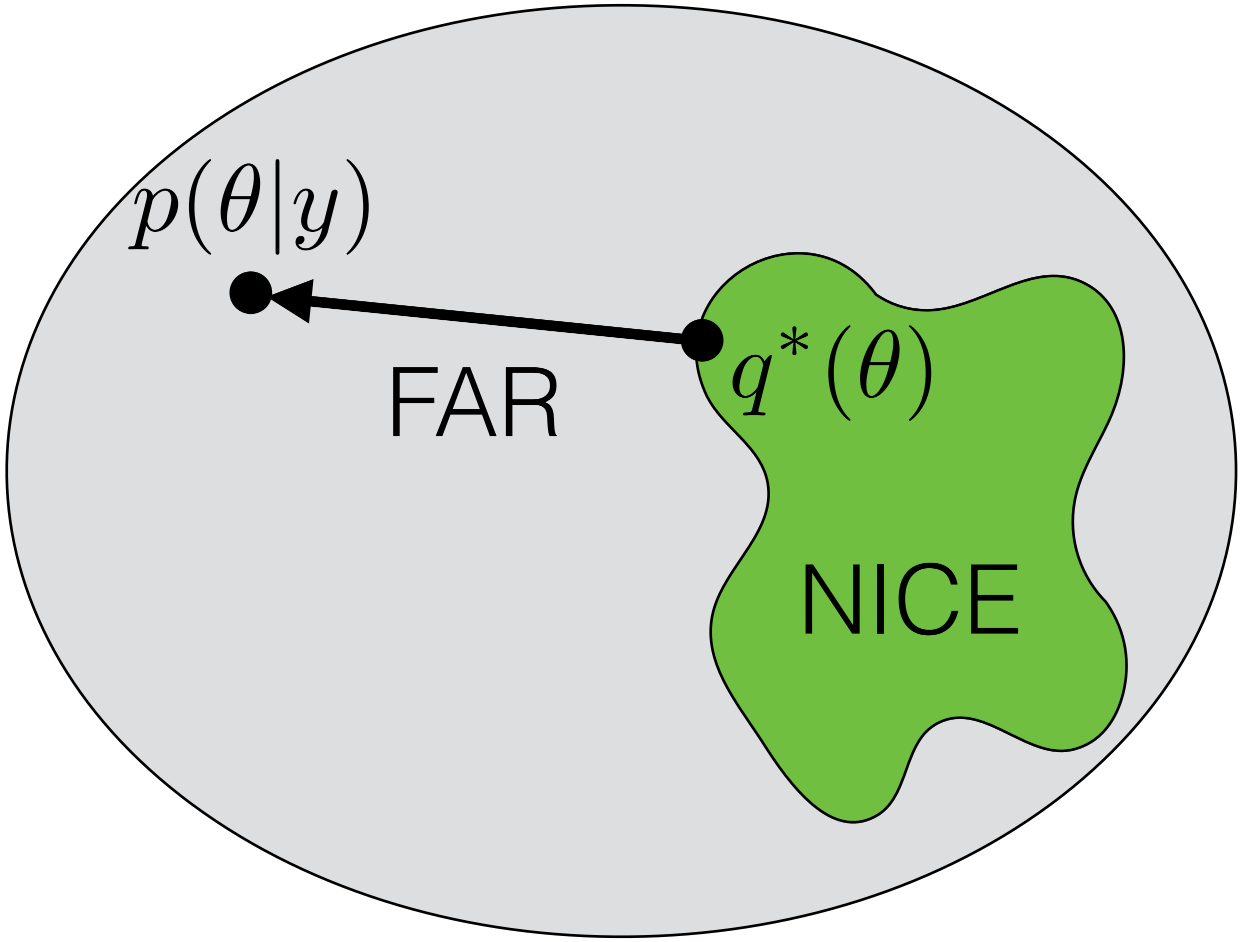
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta) p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[\log p(y) + \log \frac{q(\theta)}{p(\theta, y)} \right] d\theta$$

Why KL?

- Variational Bayes

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



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

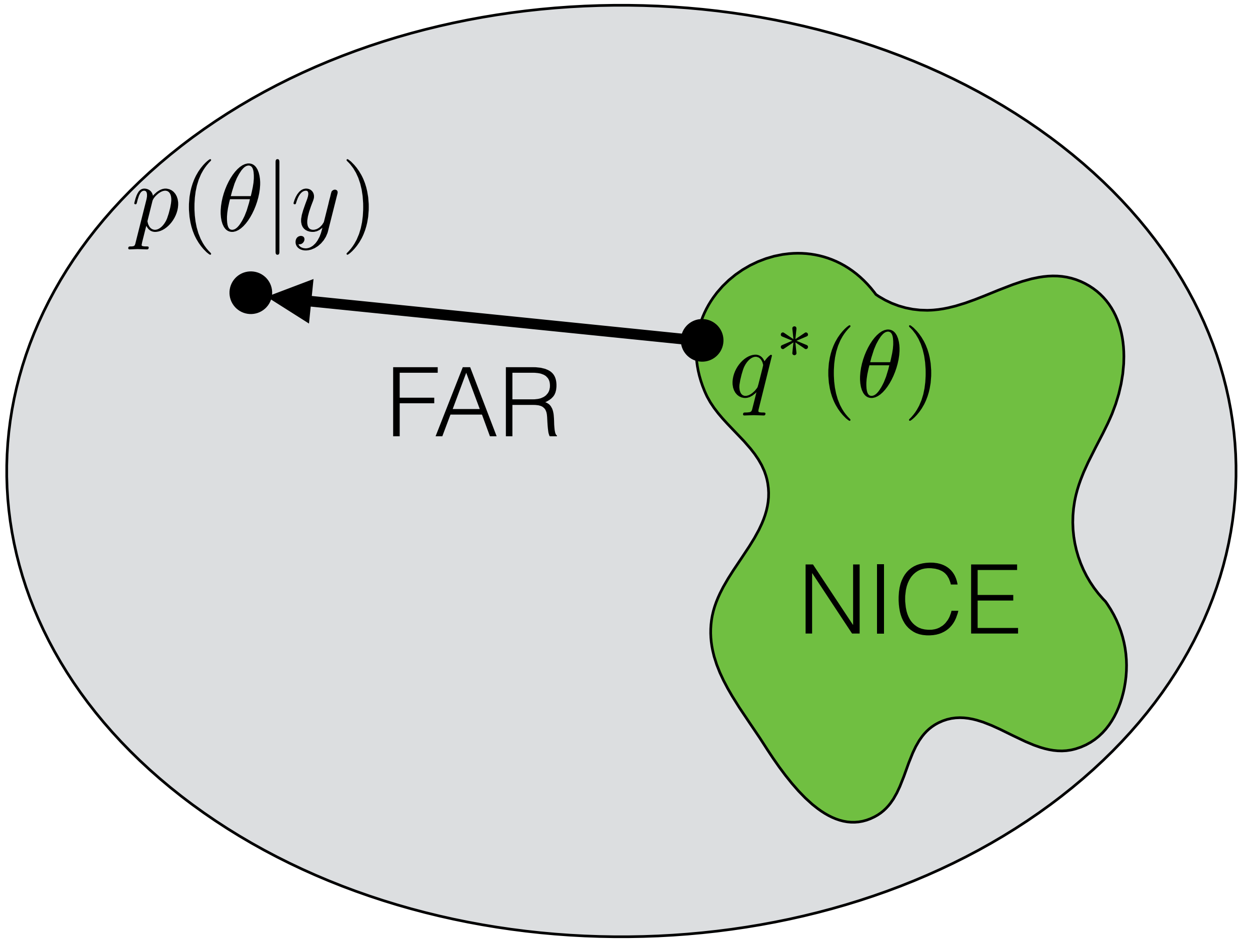
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[\log p(y) + \log \frac{q(\theta)}{p(\theta, y)}\right] d\theta$$

Why KL?

- Variational Bayes

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



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

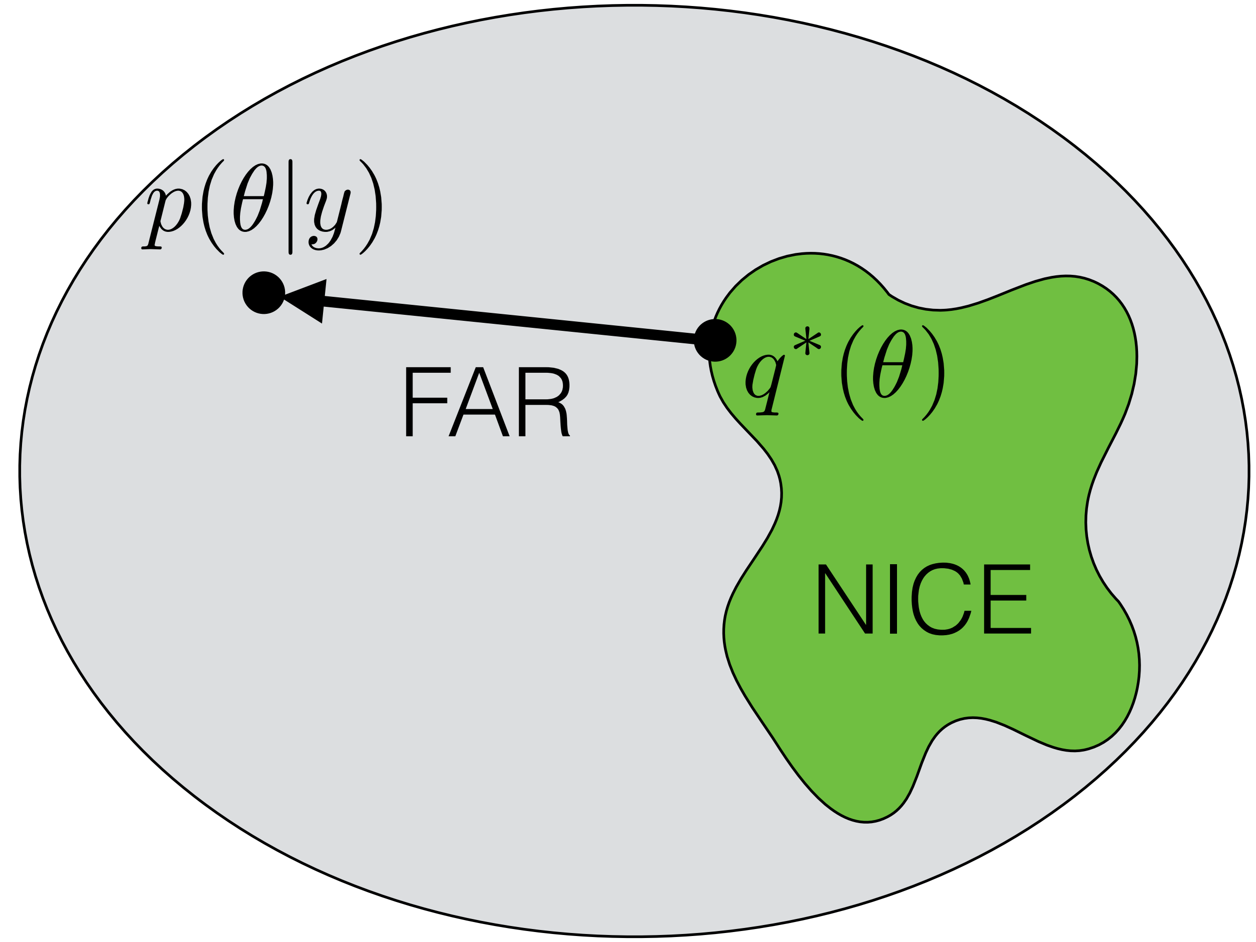
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta) p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[\log p(y) + \log \frac{q(\theta)}{p(\theta, y)}\right] d\theta$$

Why KL?

- Variational Bayes

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



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

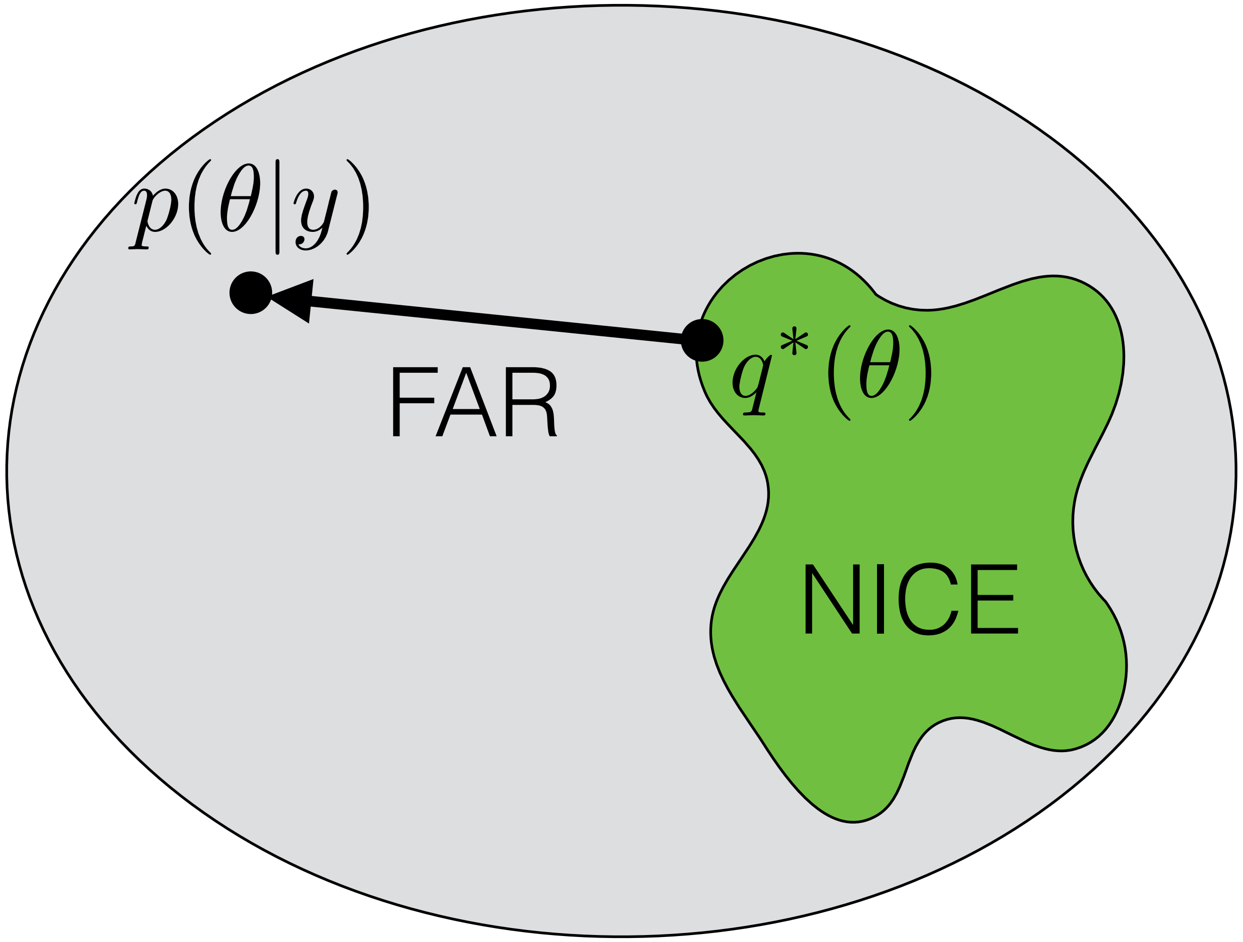
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[\log p(y) + \log \frac{q(\theta)}{p(\theta, y)}\right] d\theta$$

Why KL?

- Variational Bayes

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



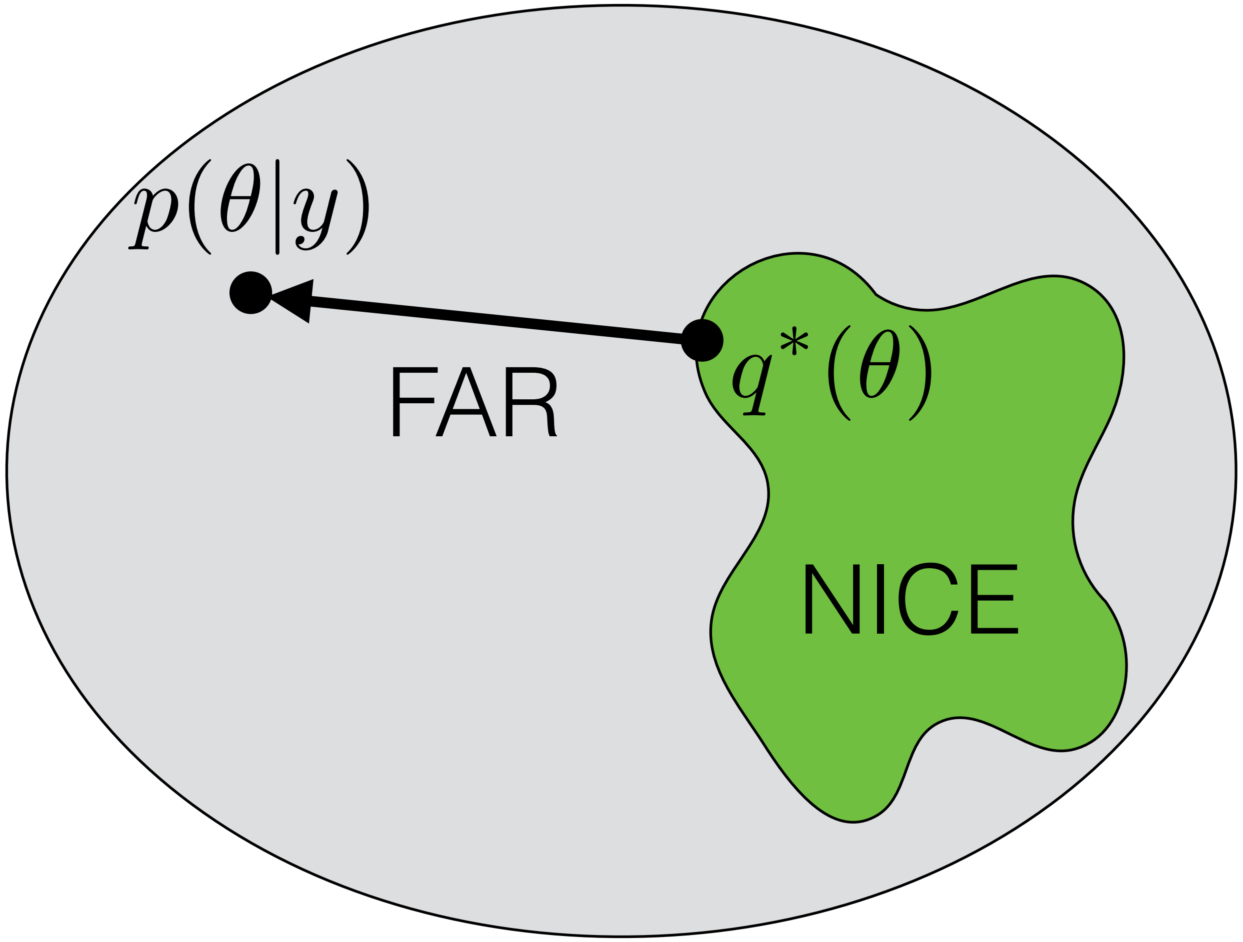
$$\text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

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

$$= \int q(\theta) \log \frac{q(\theta) p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[\log p(y) + \log \frac{q(\theta)}{p(\theta, y)}\right] d\theta$$

Why KL?

- Variational Bayes



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

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

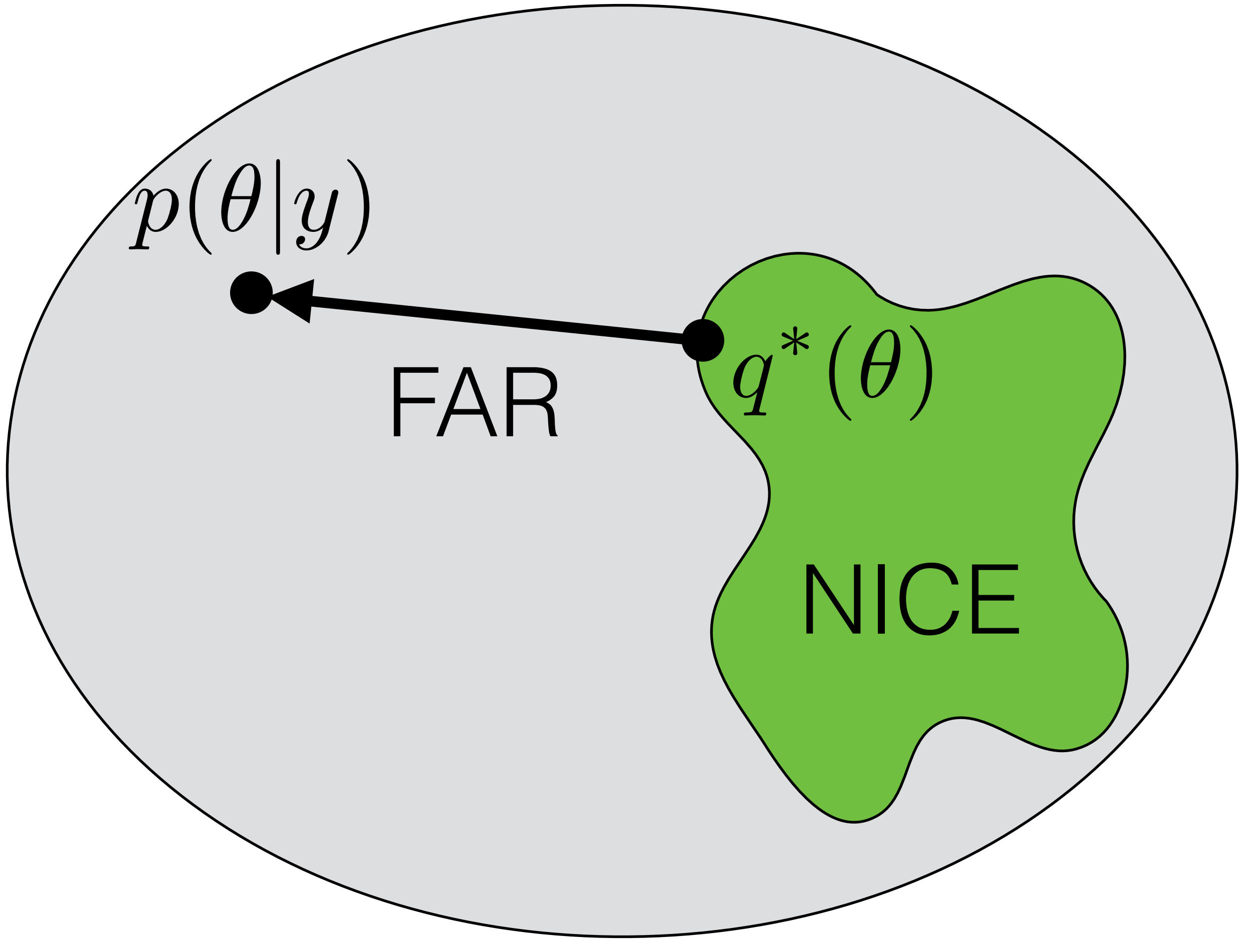
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[\log p(y) + \log \frac{q(\theta)}{p(\theta, y)}\right] d\theta$$

Why KL?

- Variational Bayes

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



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

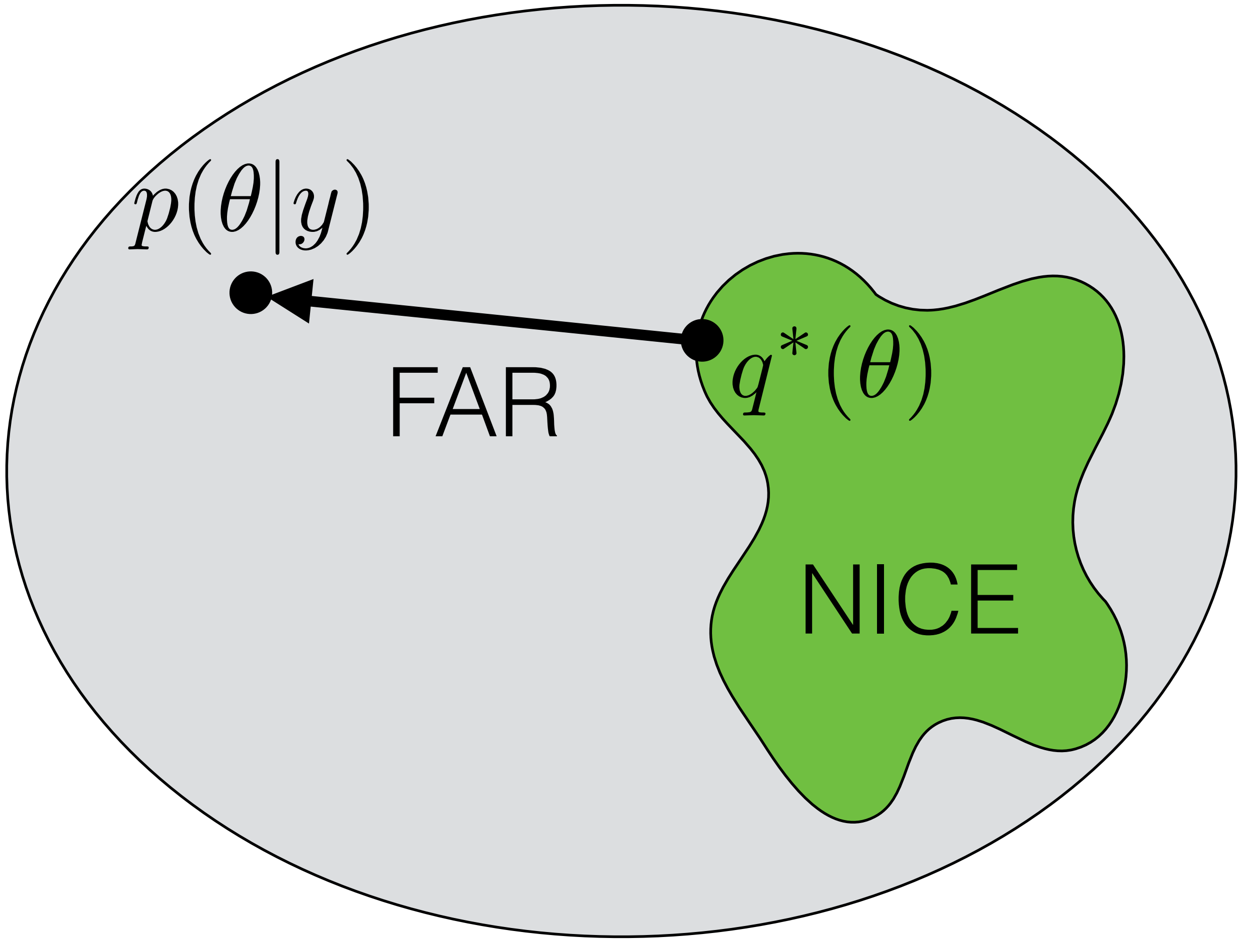
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[\log p(y) + \log \frac{q(\theta)}{p(\theta, y)} \right] d\theta$$

Why KL?

- Variational Bayes

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



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

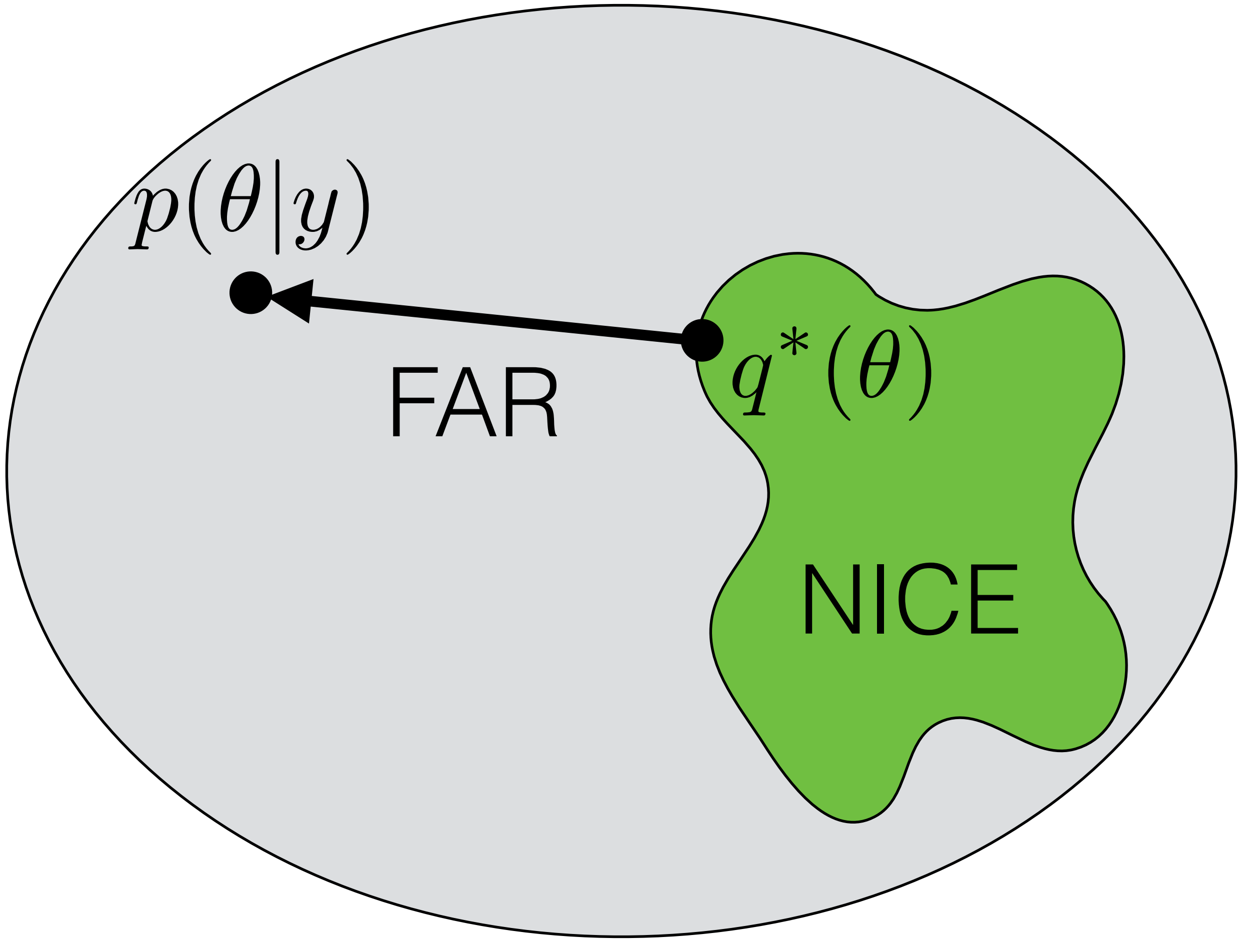
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

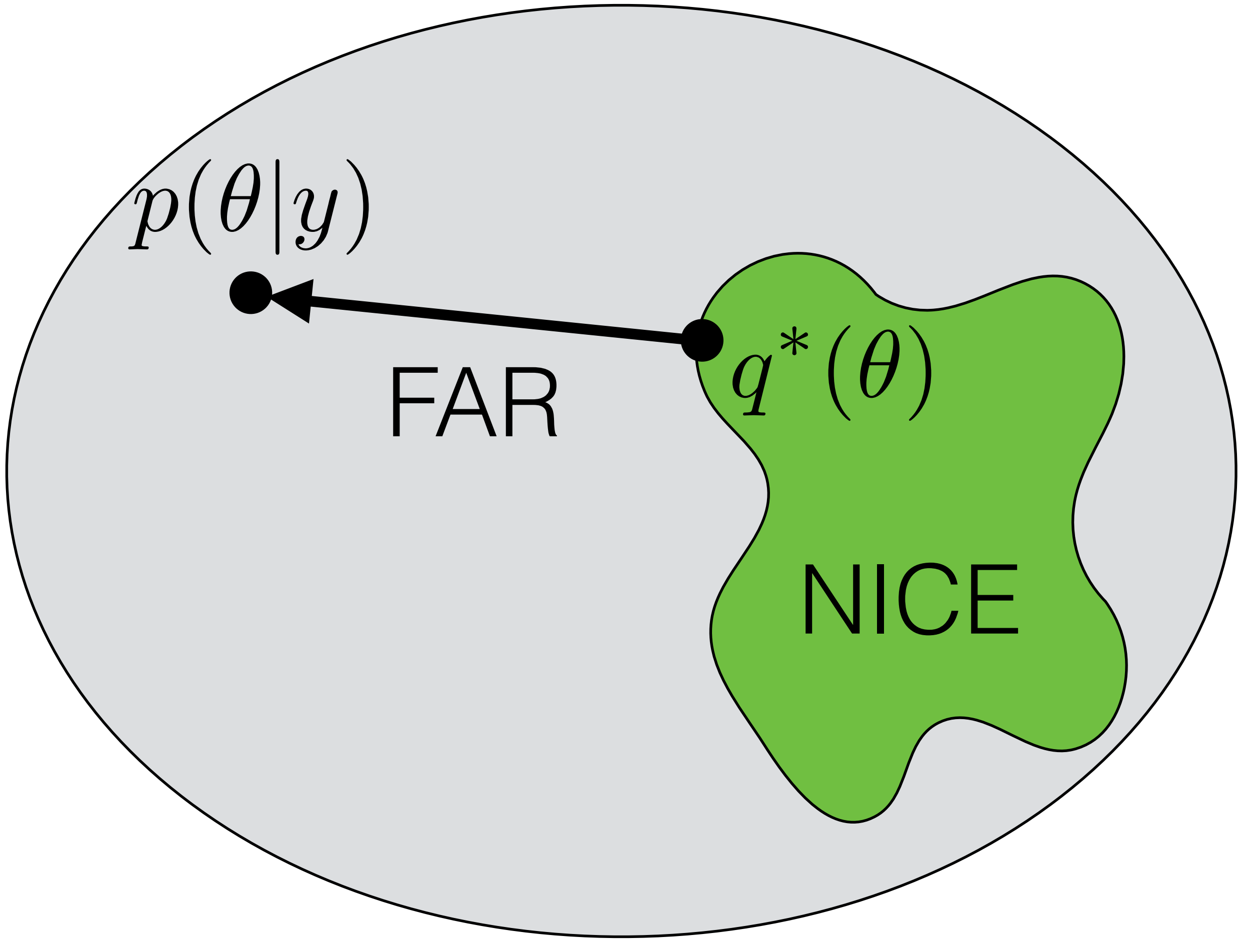
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

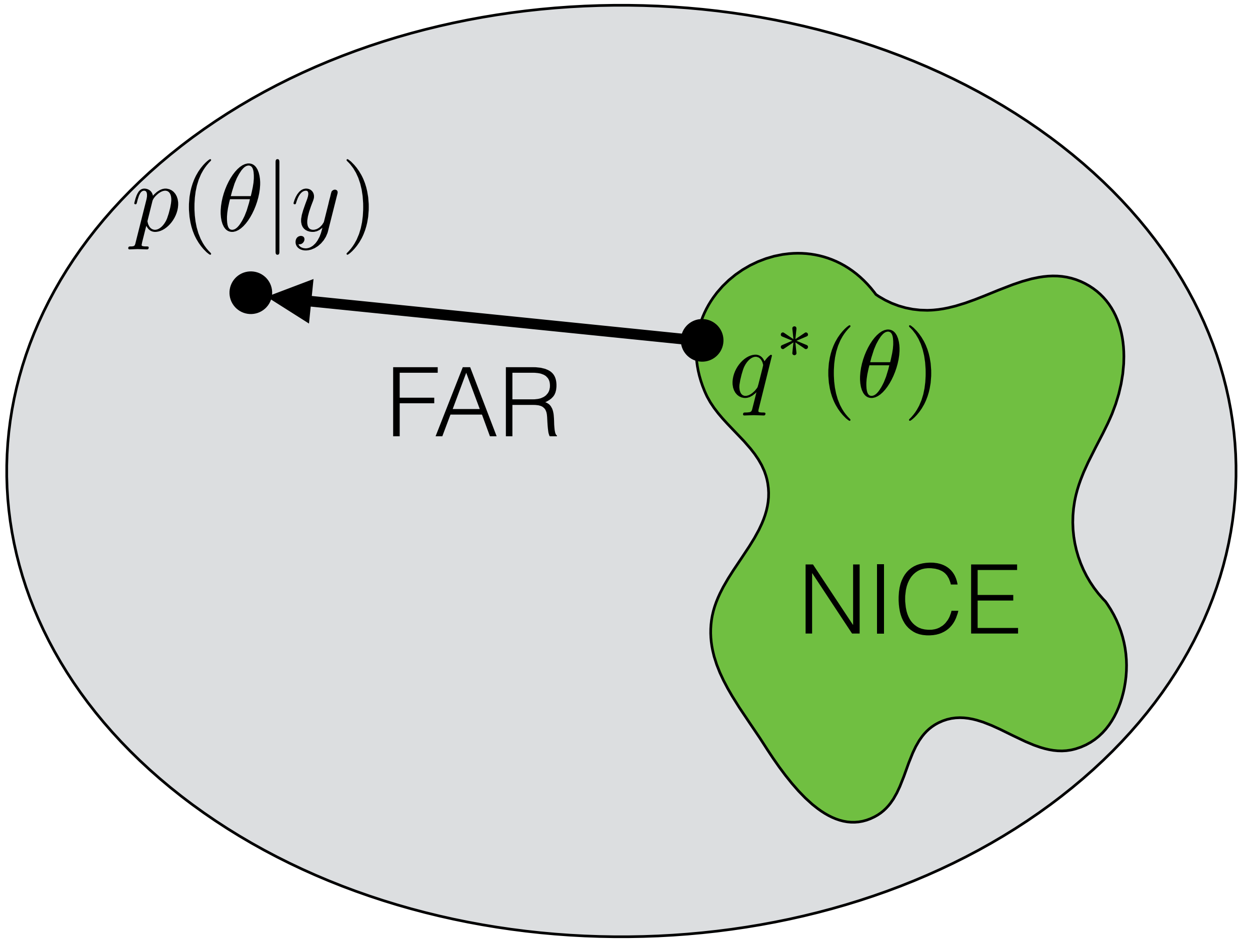
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

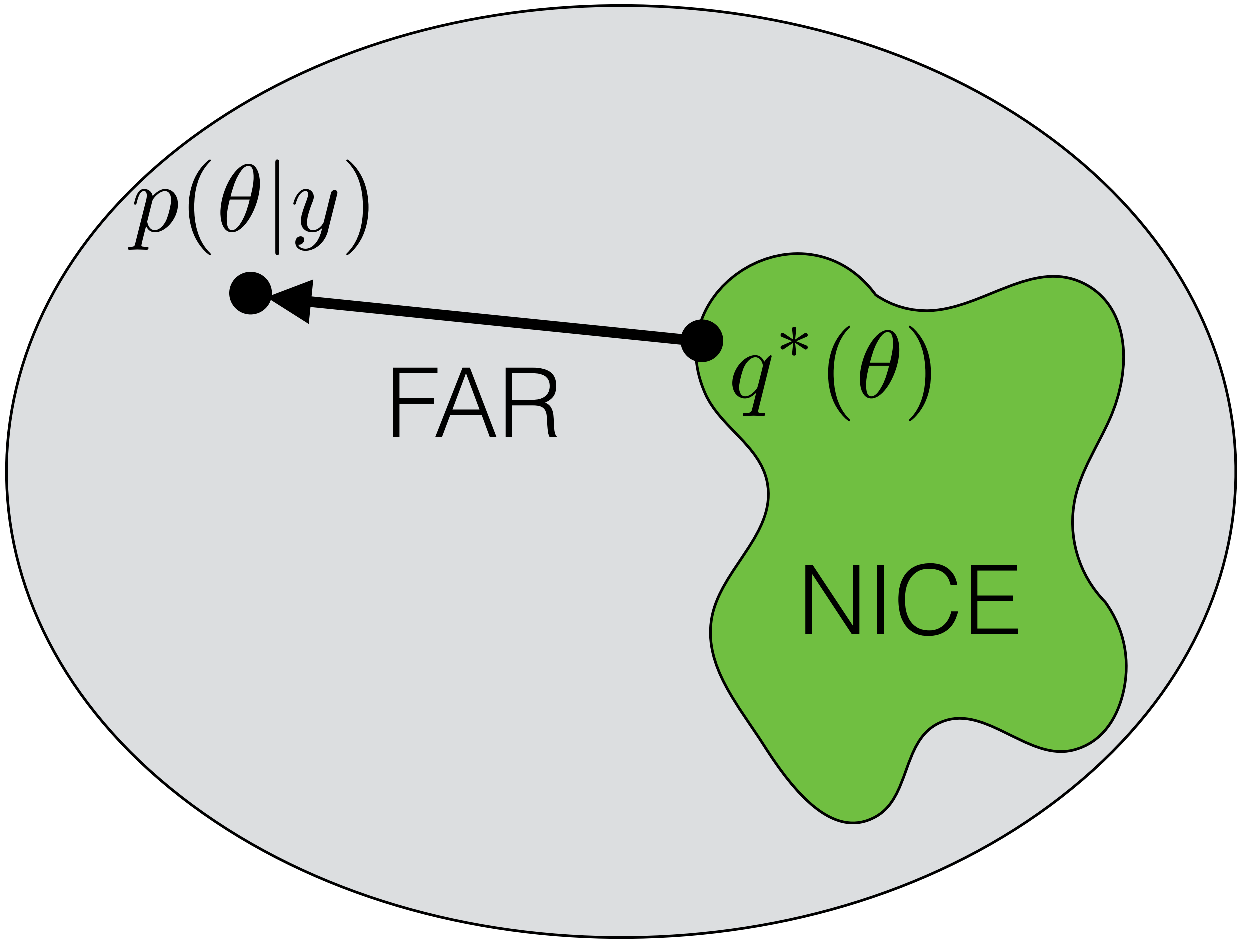
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

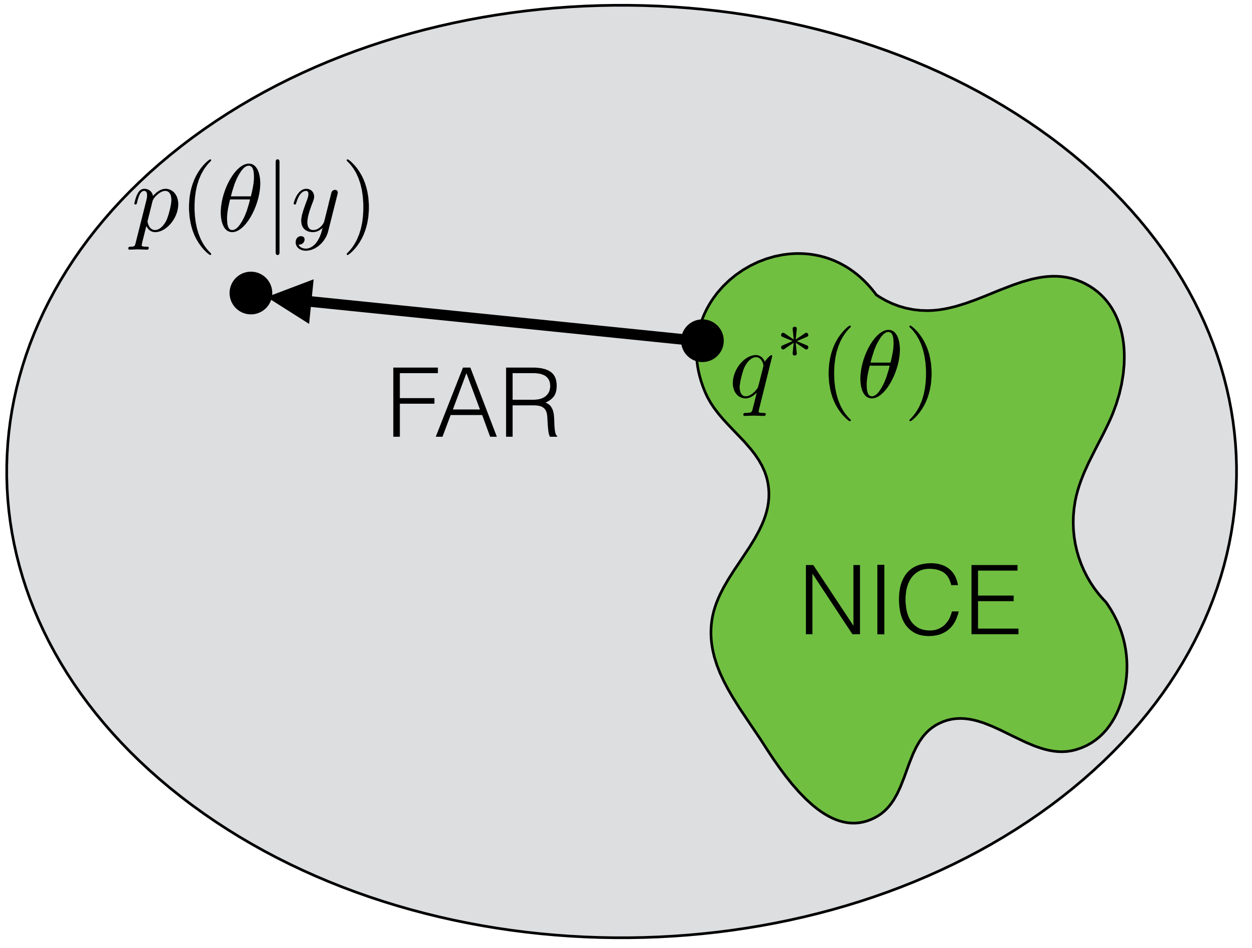
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

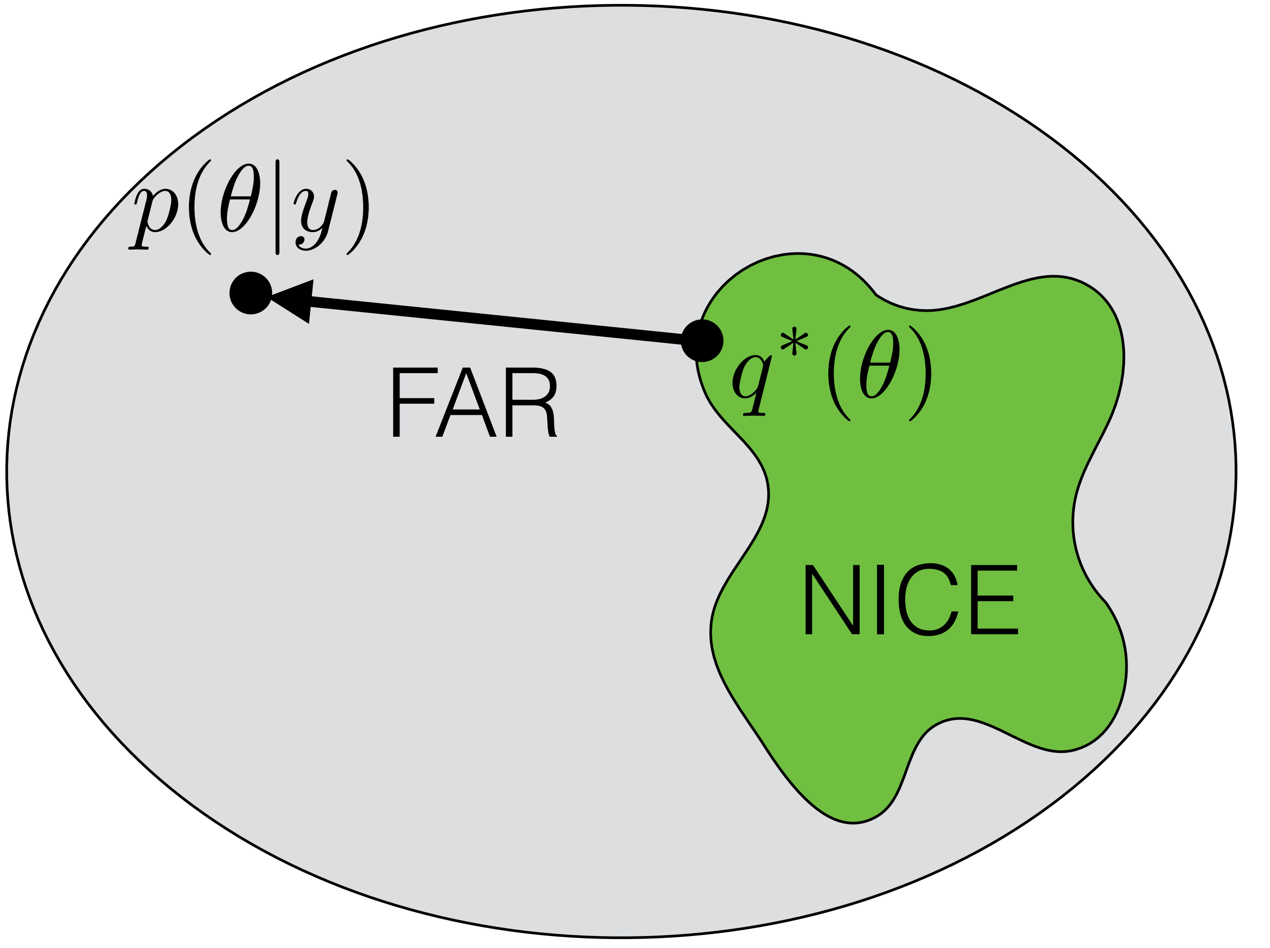
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

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

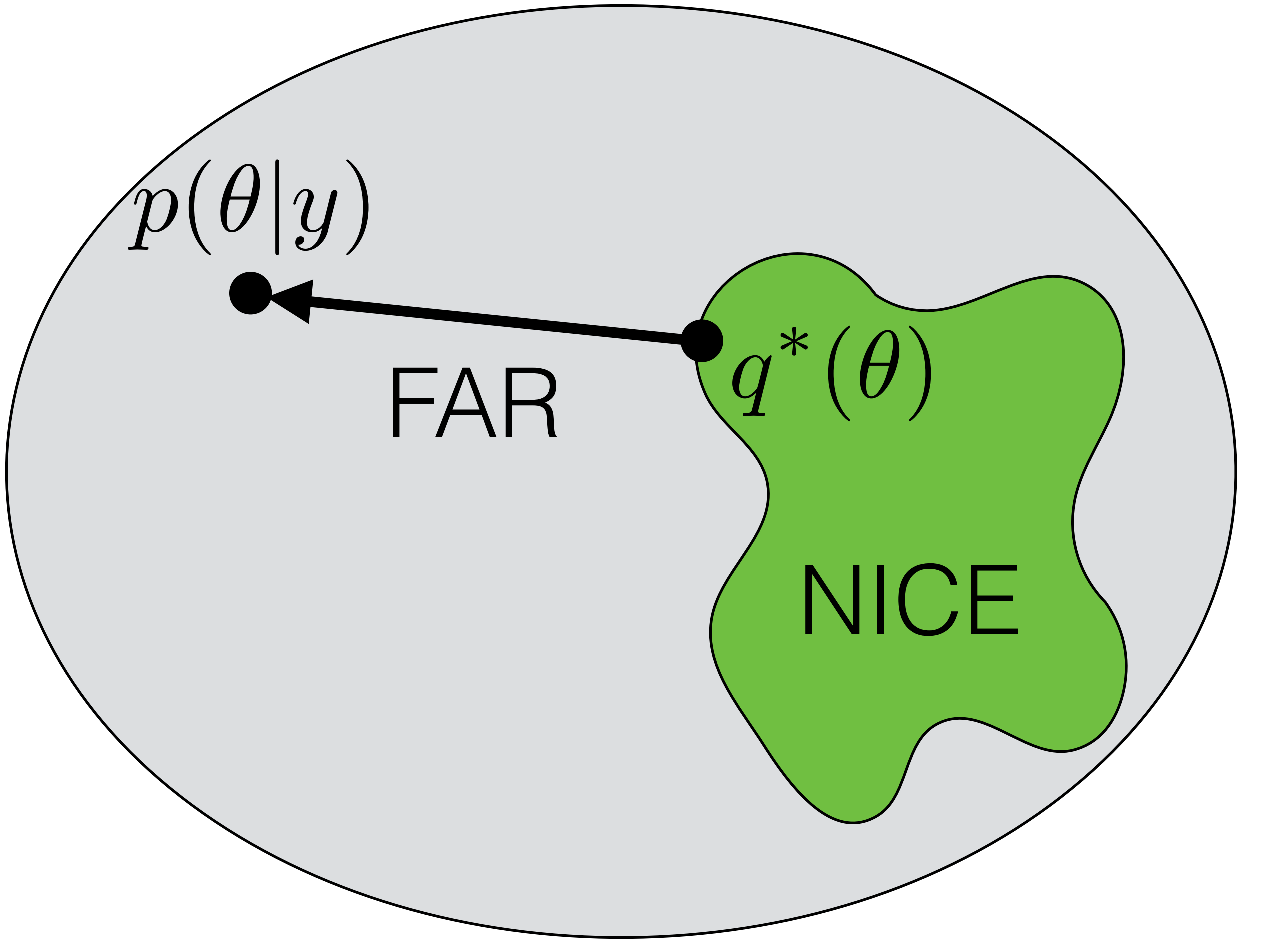
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta,y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta,y)}{q(\theta)} d\theta$$

"Evidence lower bound" (ELBO)

Why KL?

- Variational Bayes

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



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

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

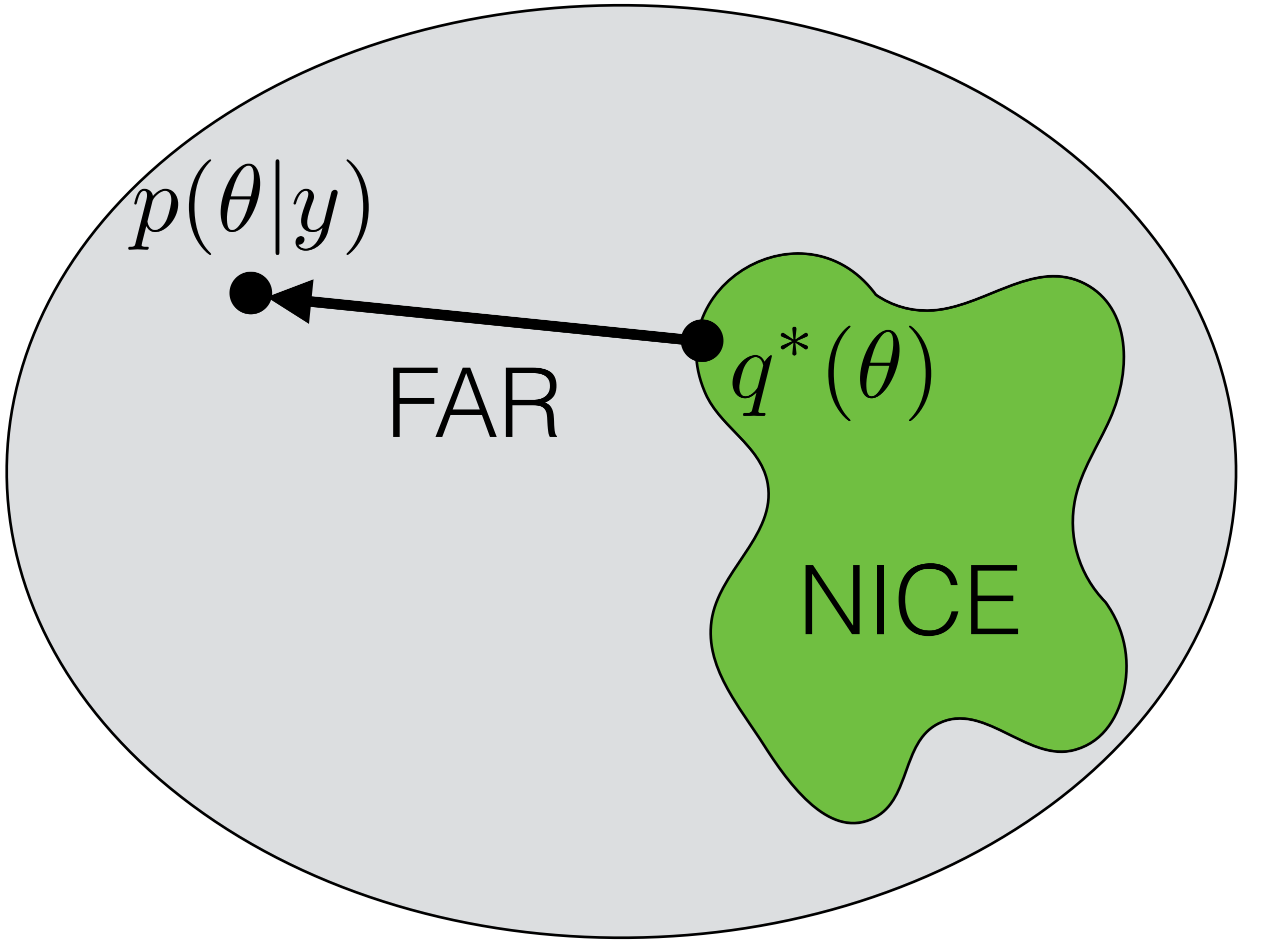
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

"Evidence lower bound" (ELBO)

Why KL?

- Variational Bayes

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



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

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

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

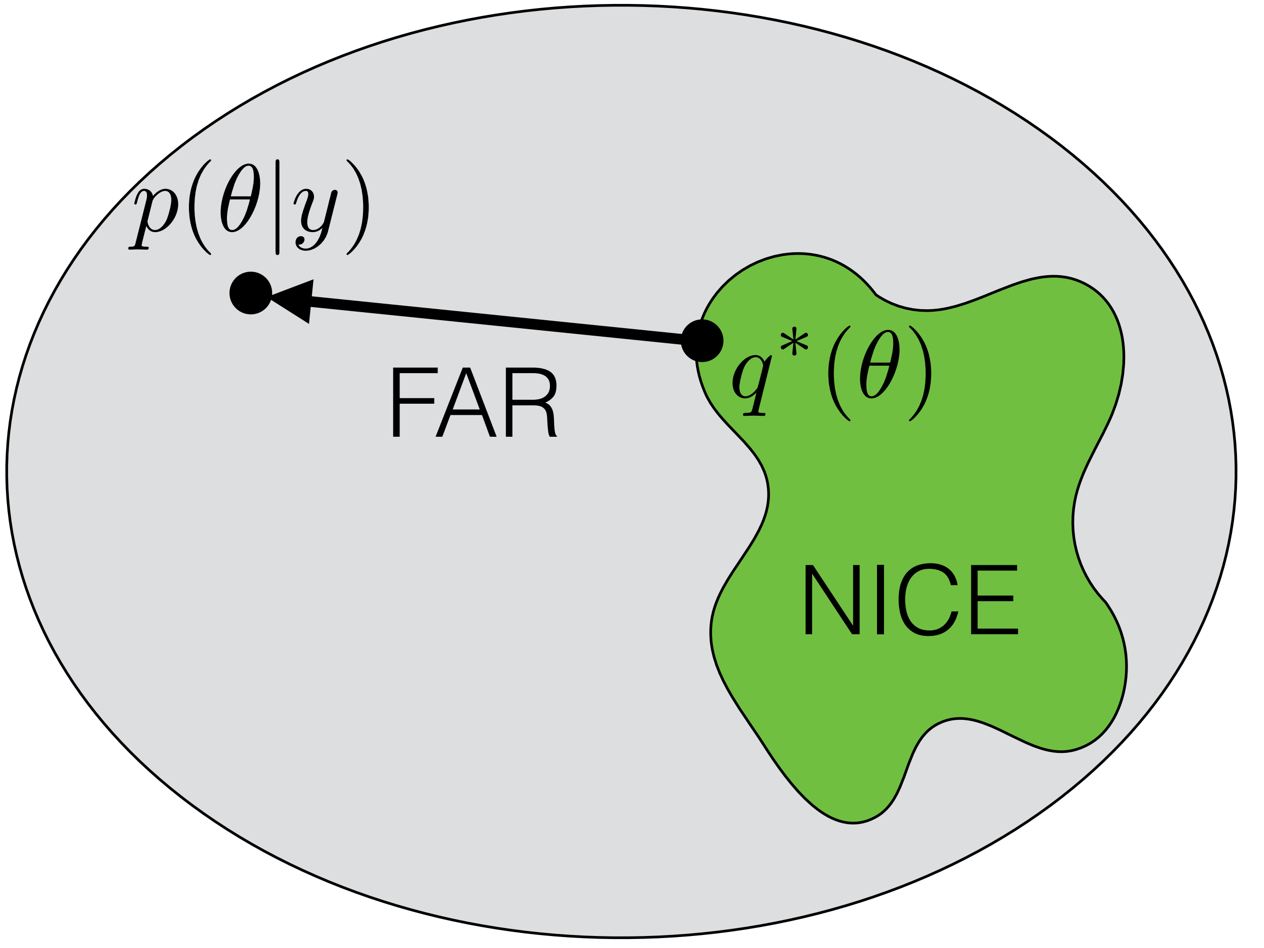
"Evidence lower bound" (ELBO)

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

Why KL?

- Variational Bayes

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



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

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

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

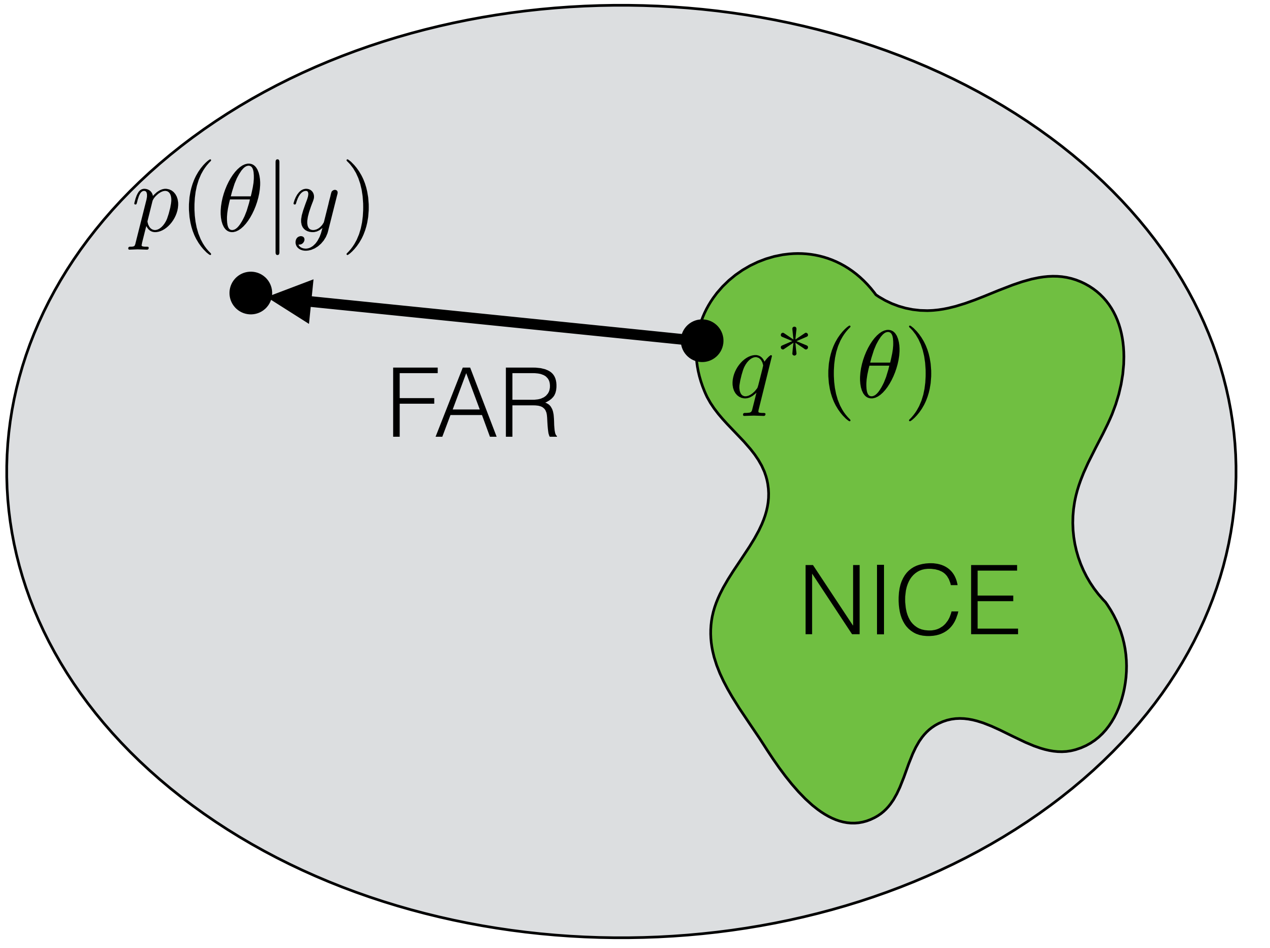
"Evidence lower bound" (ELBO)

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

Why KL?

- Variational Bayes

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



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

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

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

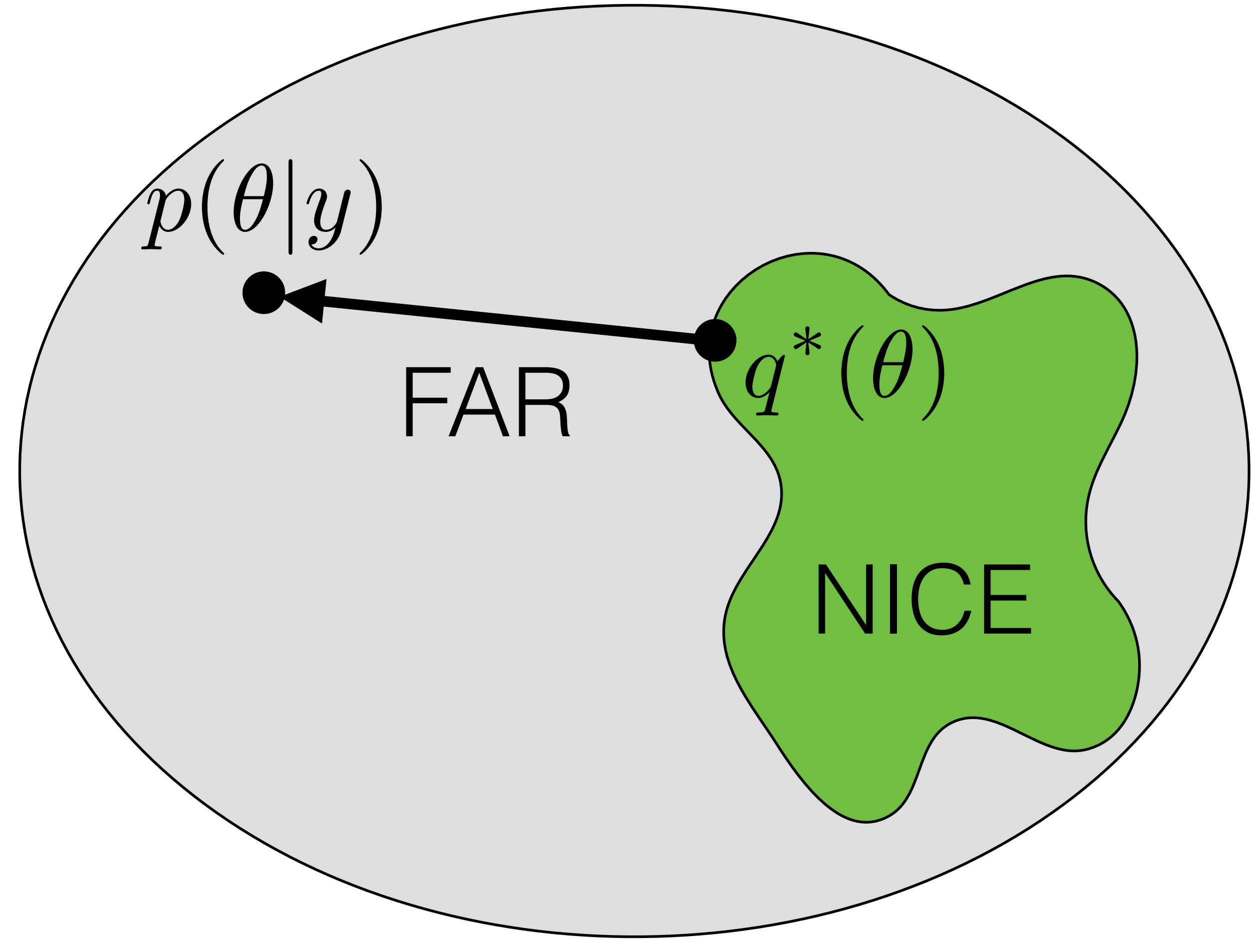
"Evidence lower bound" (ELBO)

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

Why KL?

- Variational Bayes

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



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

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

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

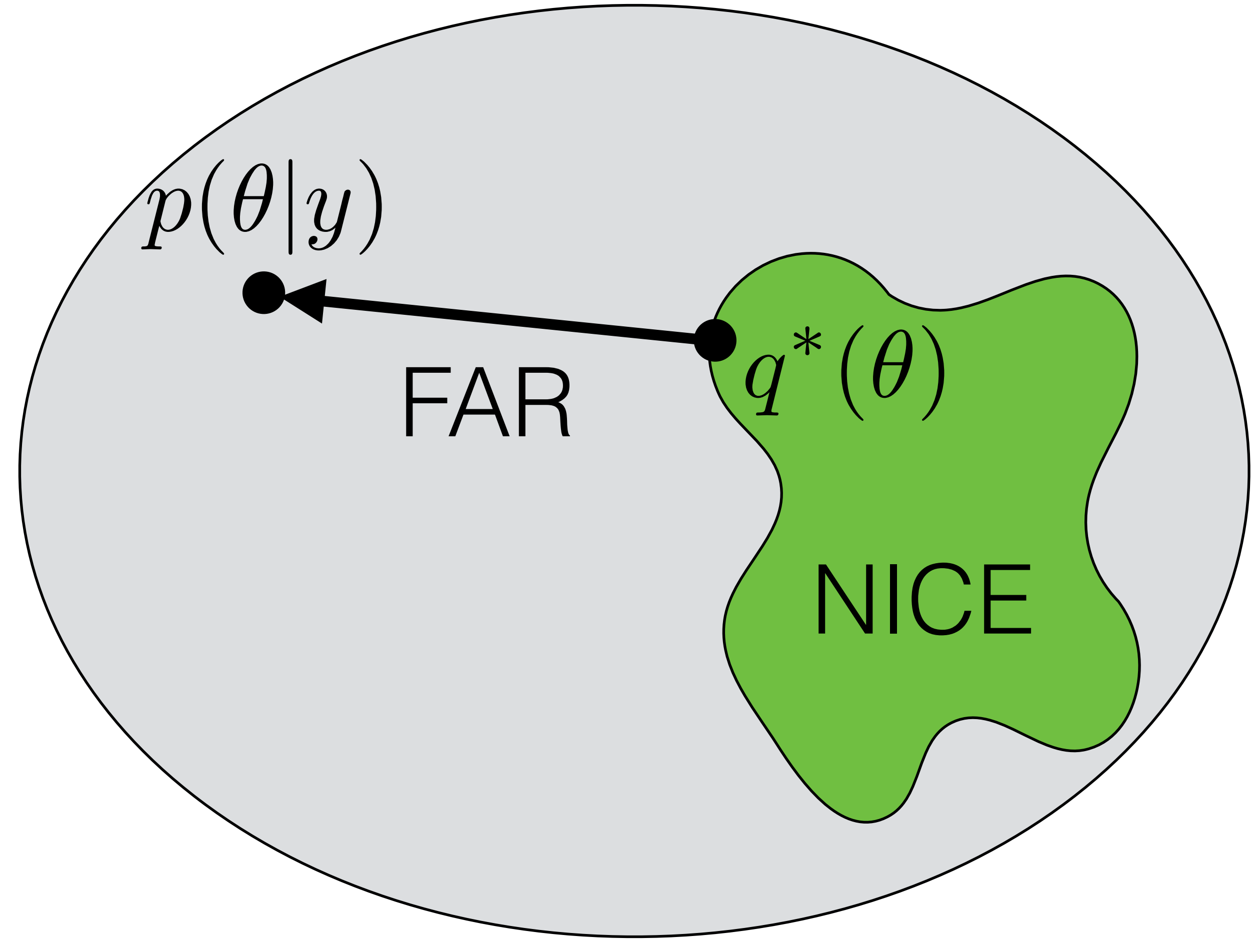
"Evidence lower bound" (ELBO)

- 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

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



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

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

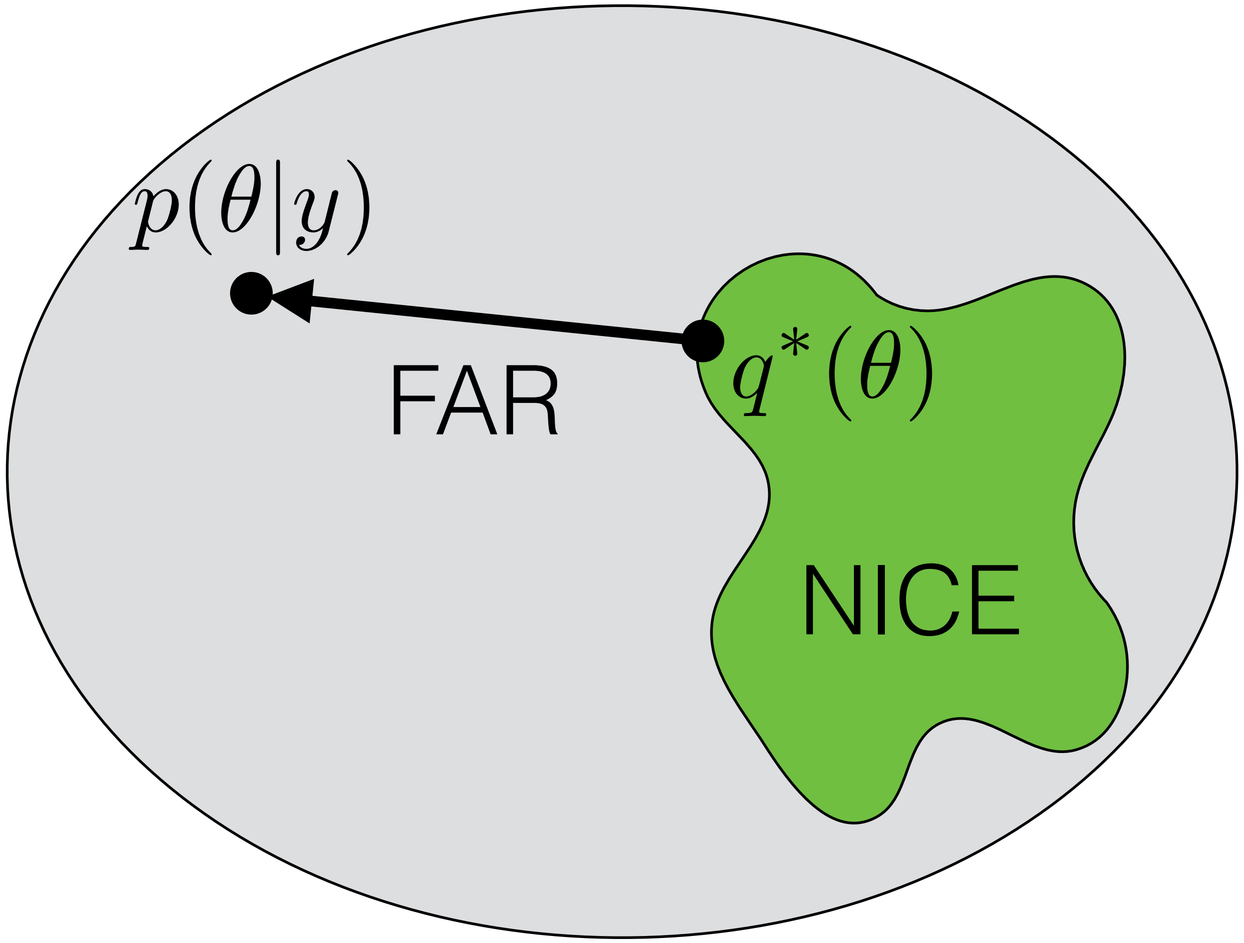
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

"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 (in this direction)?

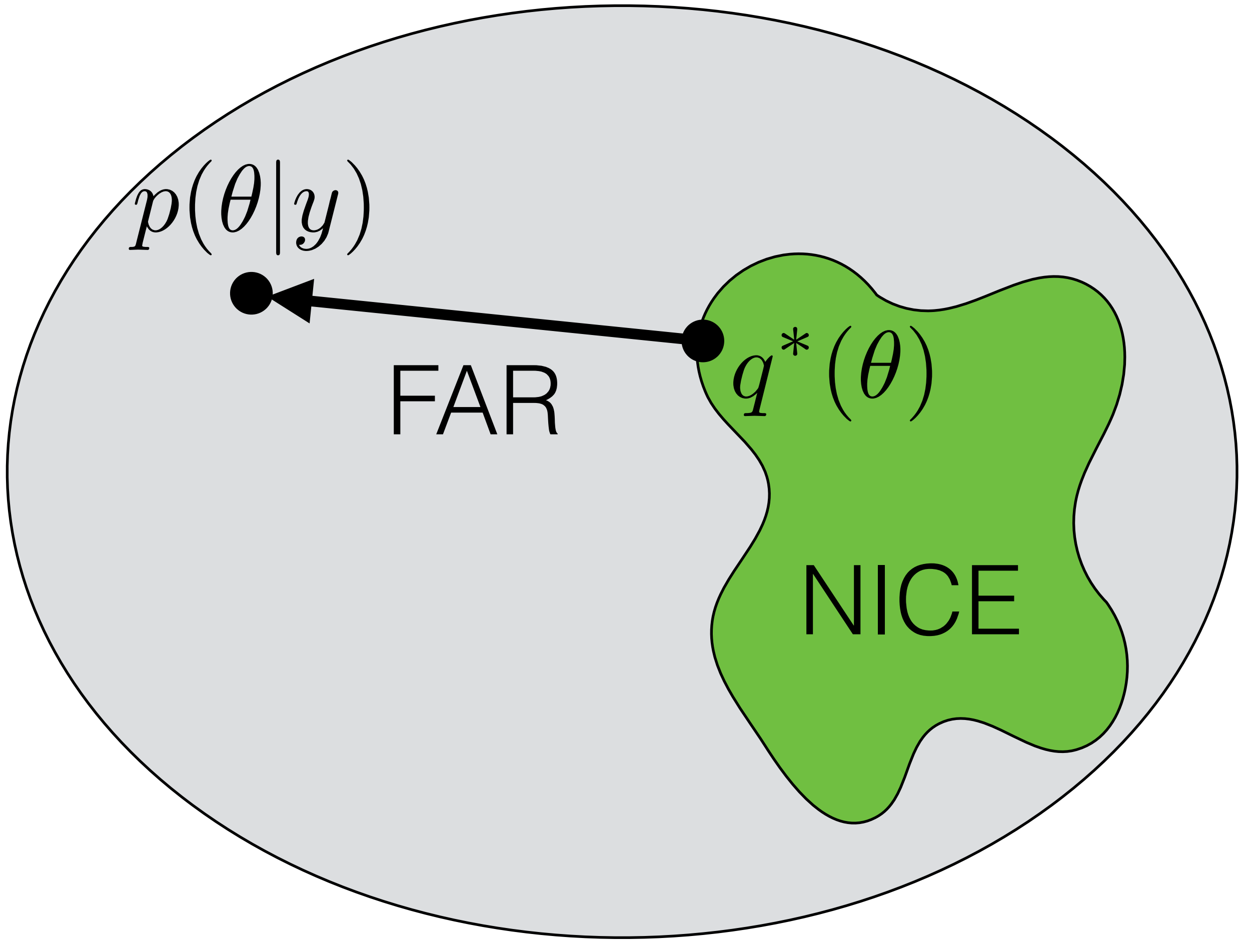
Variational Bayes

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



Variational Bayes

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

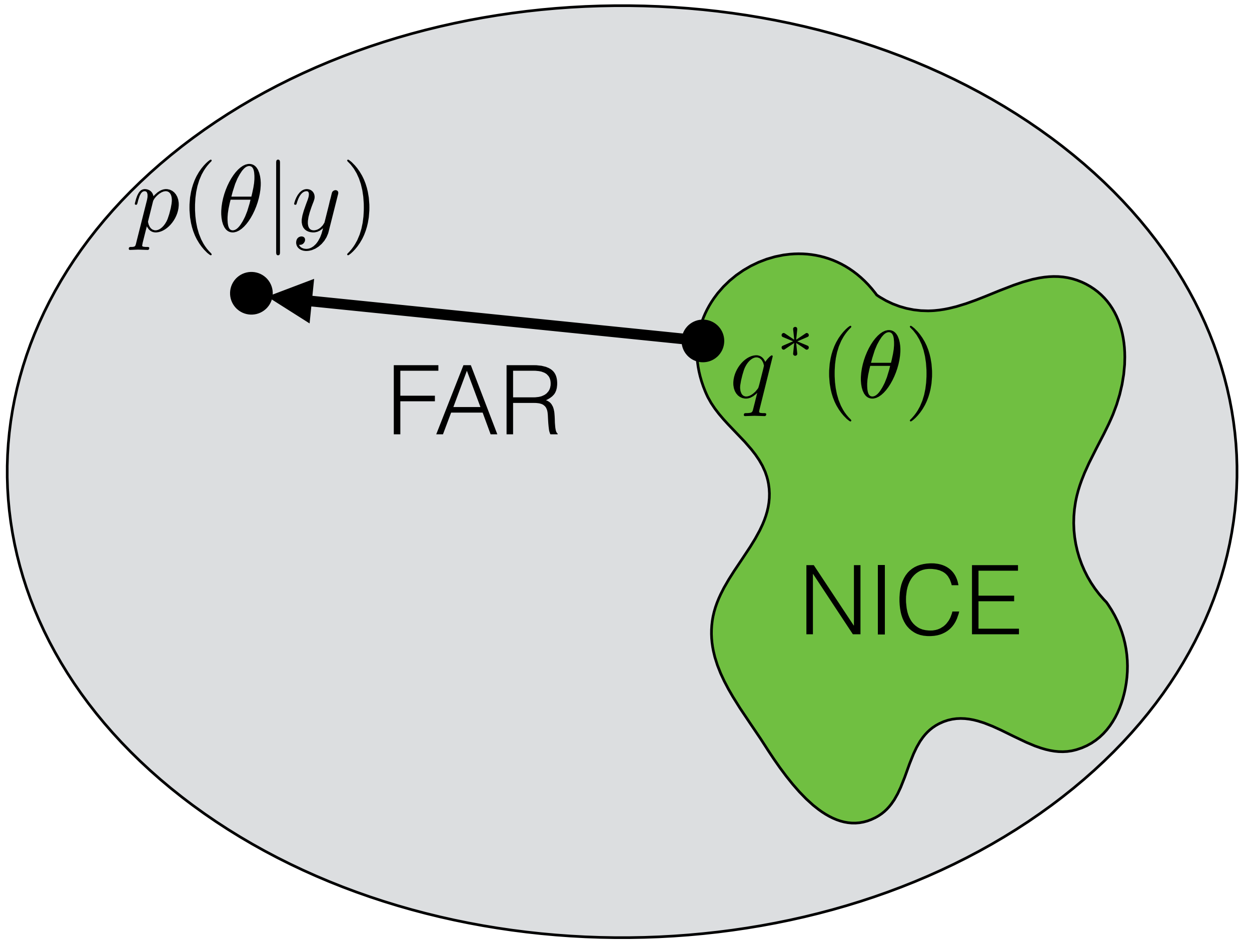


Choose "NICE" distributions

Variational Bayes

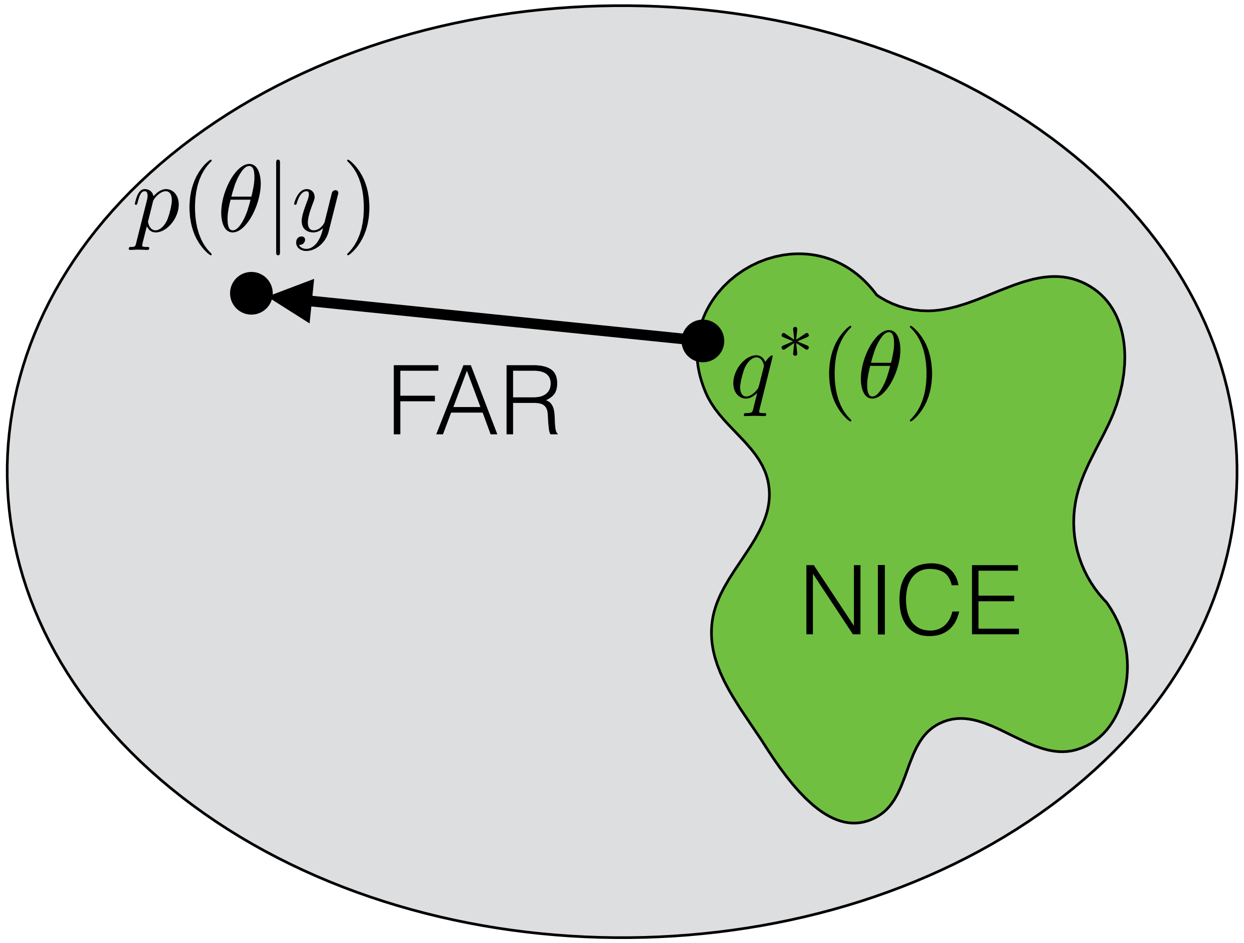
$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

Choose "NICE" distributions



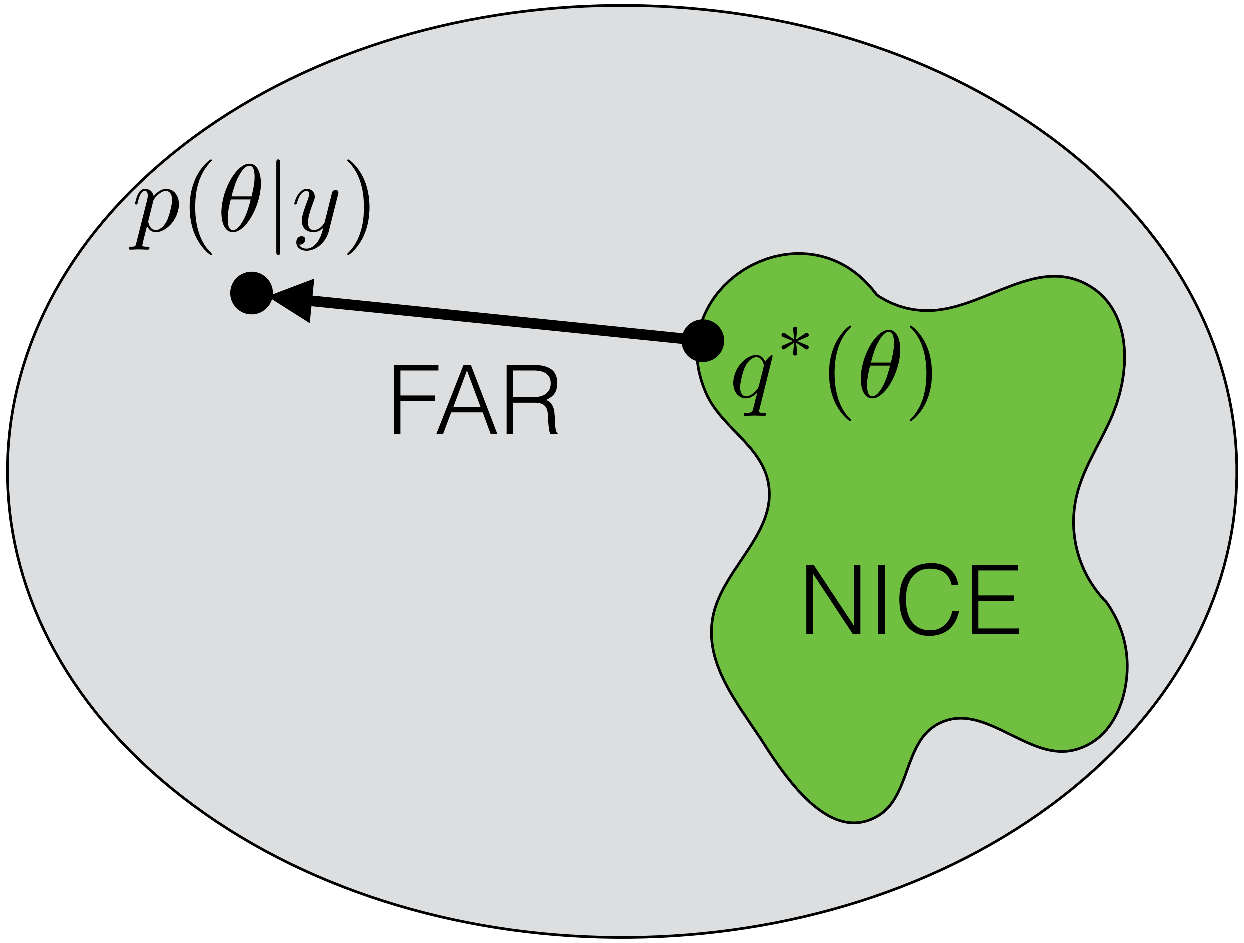
Variational Bayes

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



Choose "NICE" distributions

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



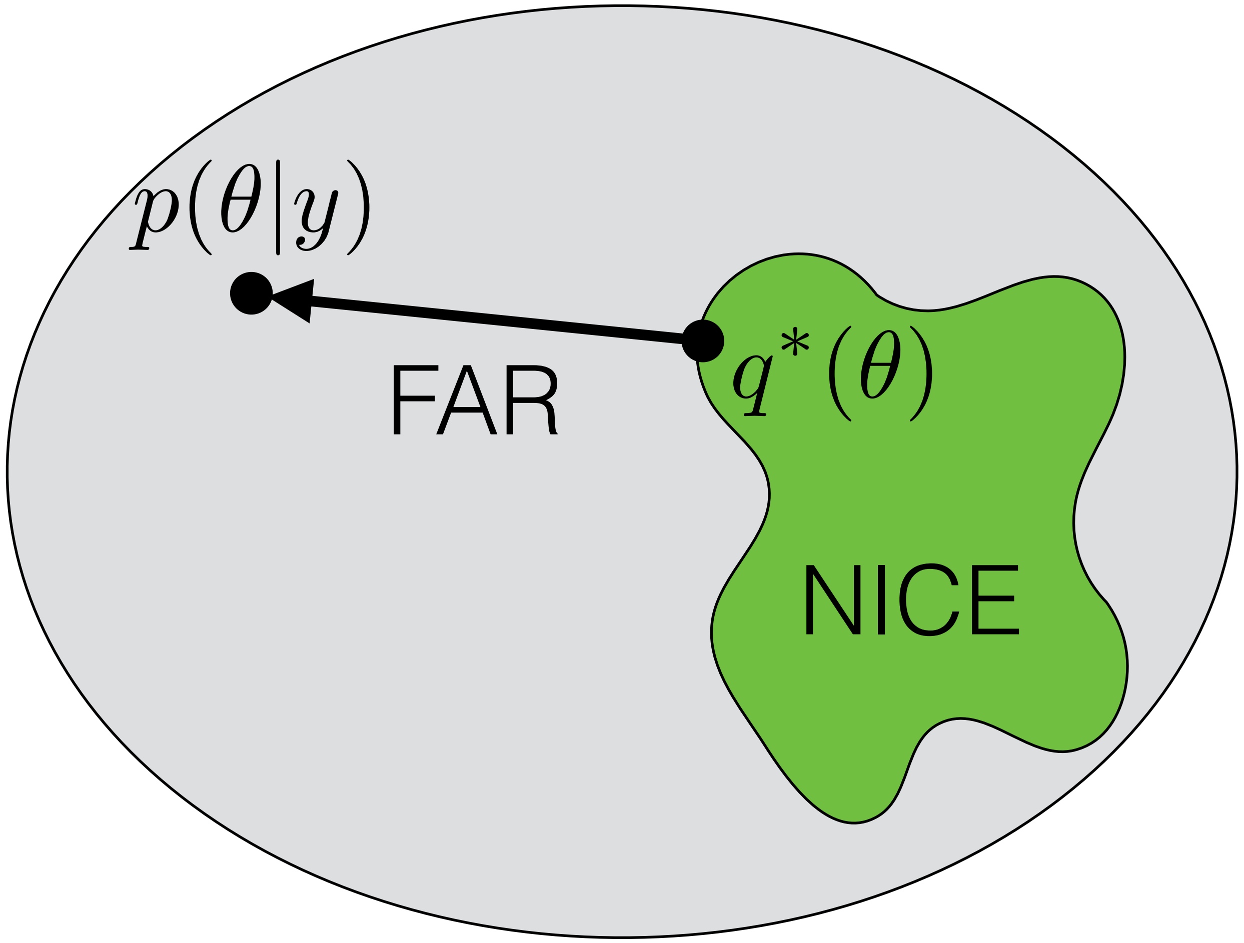
Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

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

Variational Bayes

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



Choose “NICE” distributions

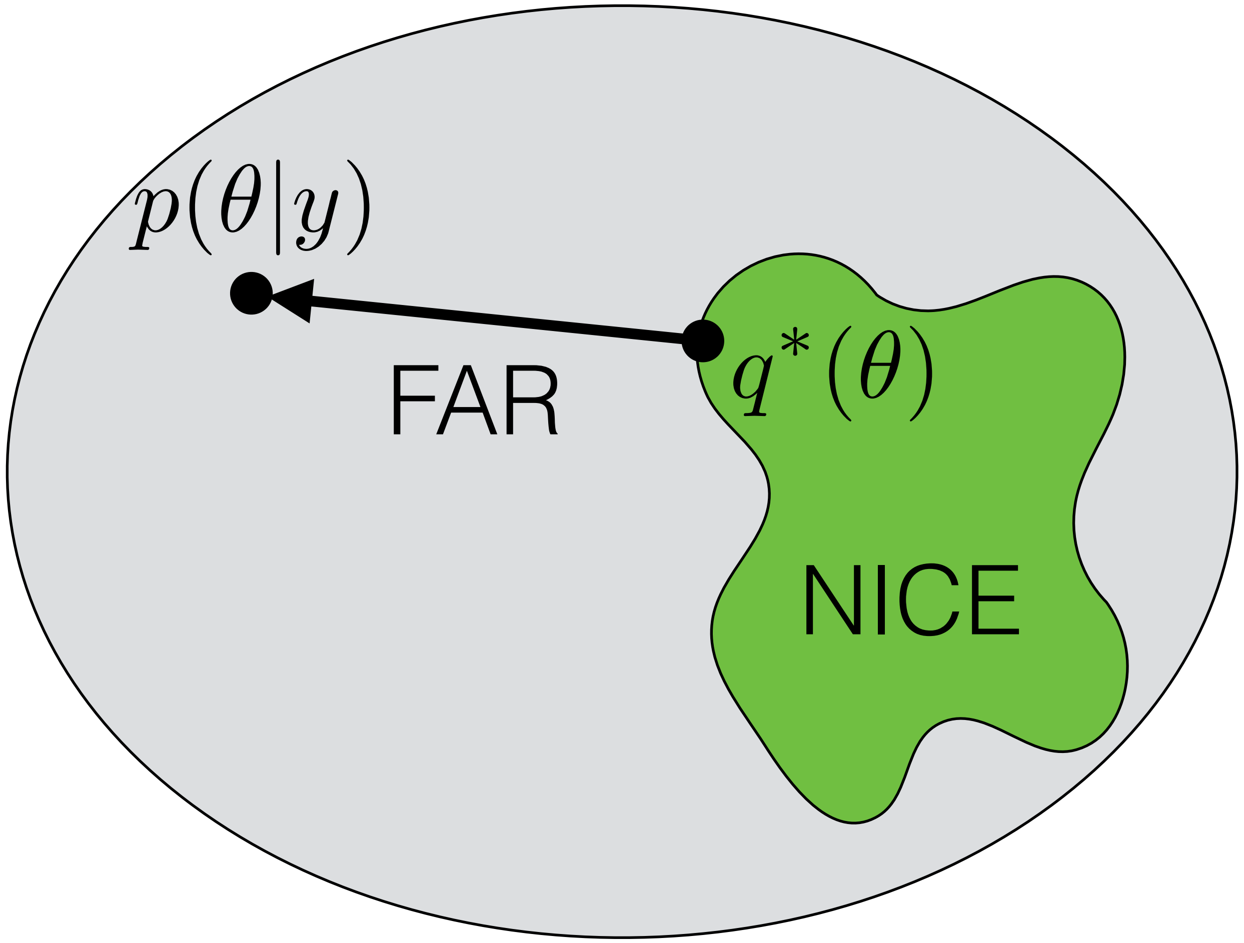
- Mean-field variational Bayes (MFVB)

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

- Often also exponential family

Variational Bayes

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



Choose "NICE" distributions

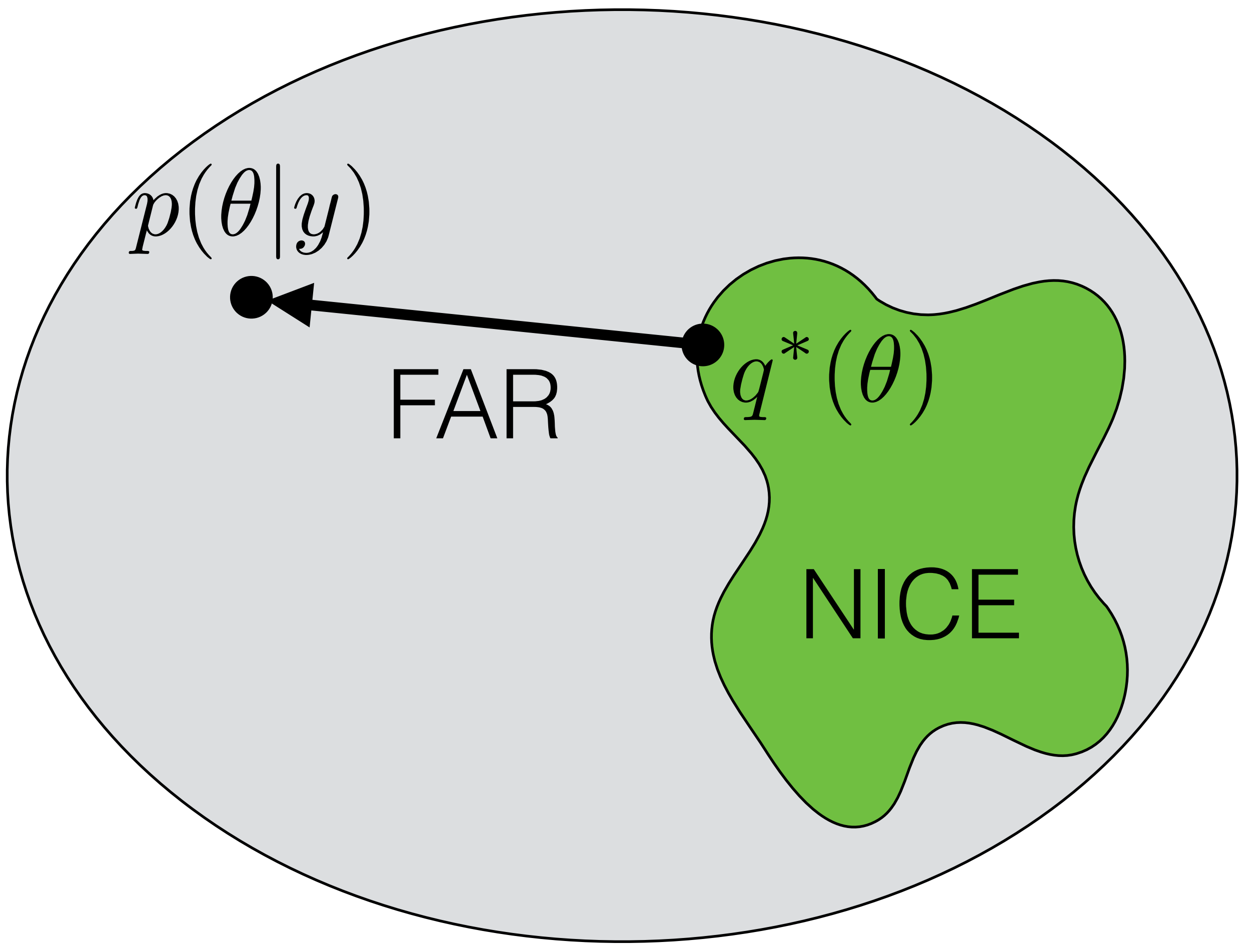
- Mean-field variational Bayes (MFVB)

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

- Often also exponential family
- *Not* a modeling assumption

Variational Bayes

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

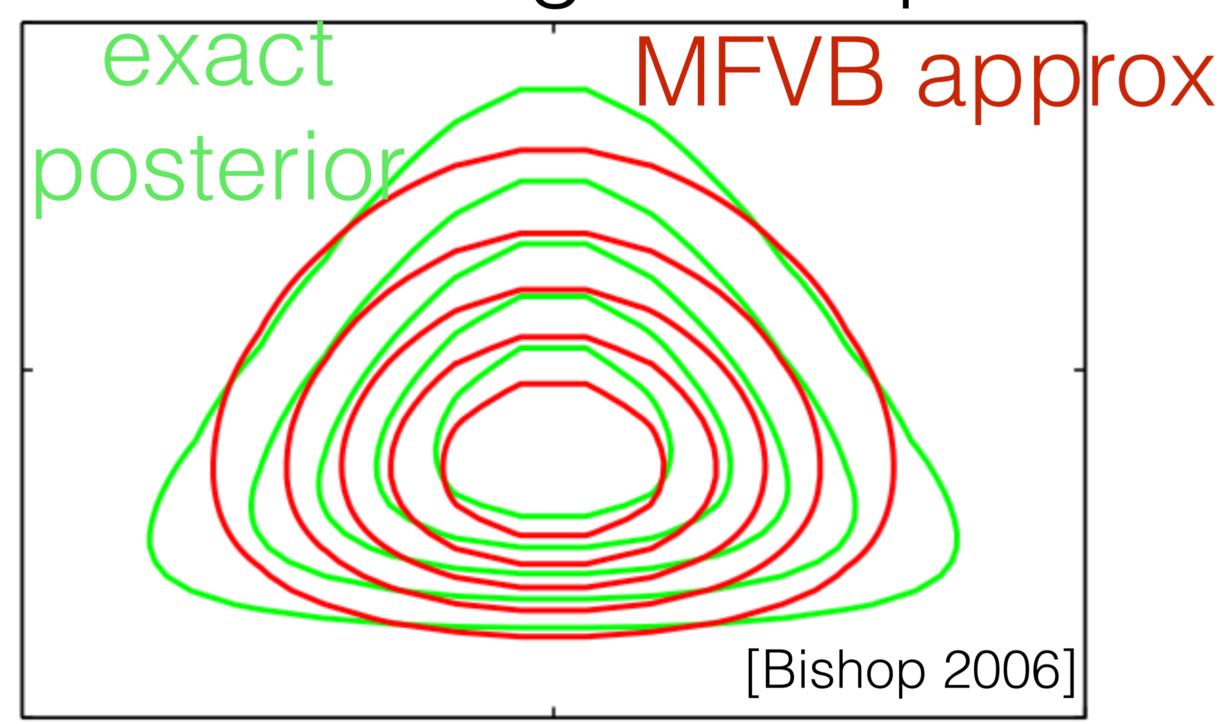


Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

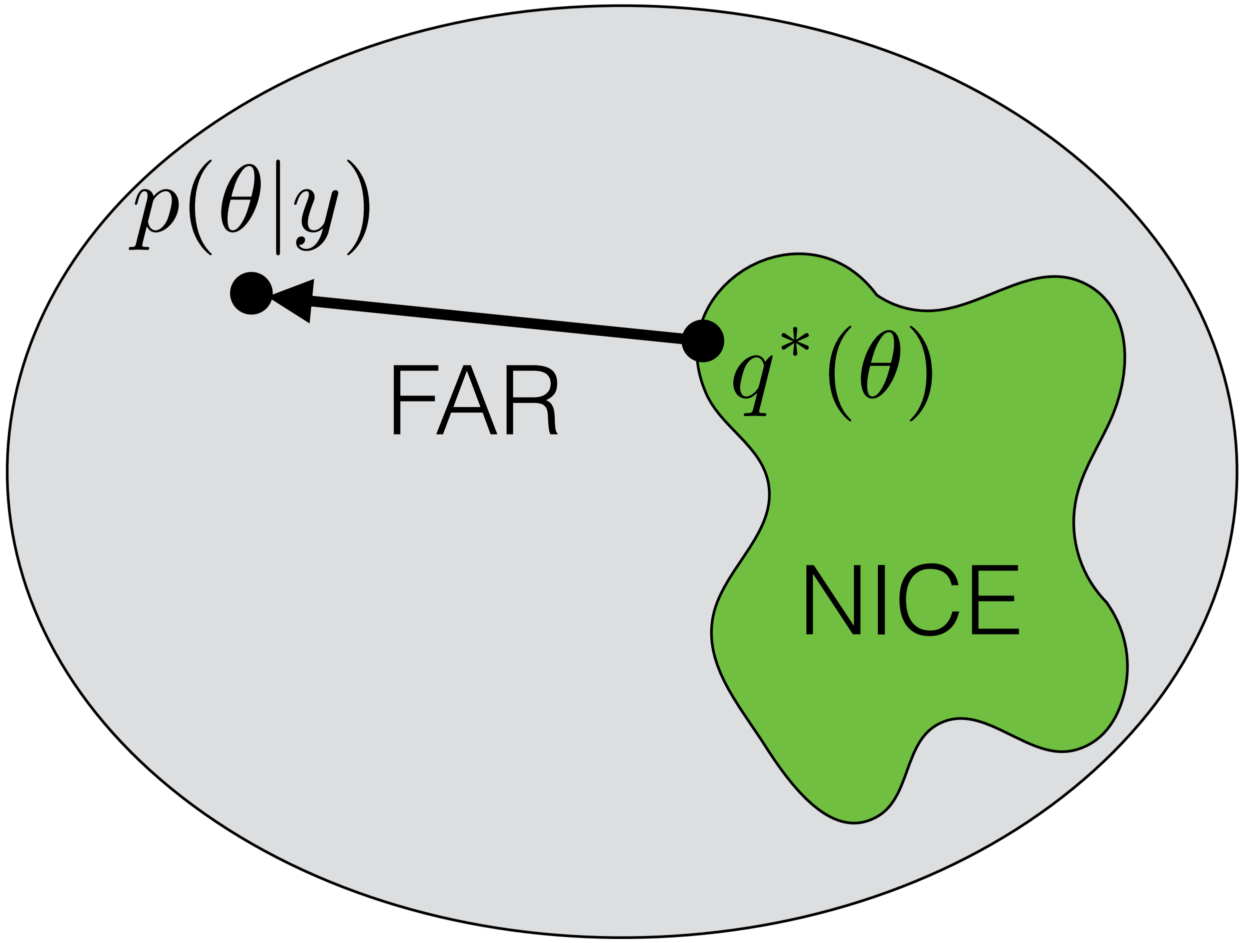
$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

- Often also exponential family
- *Not* a modeling assumption



Variational Bayes

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



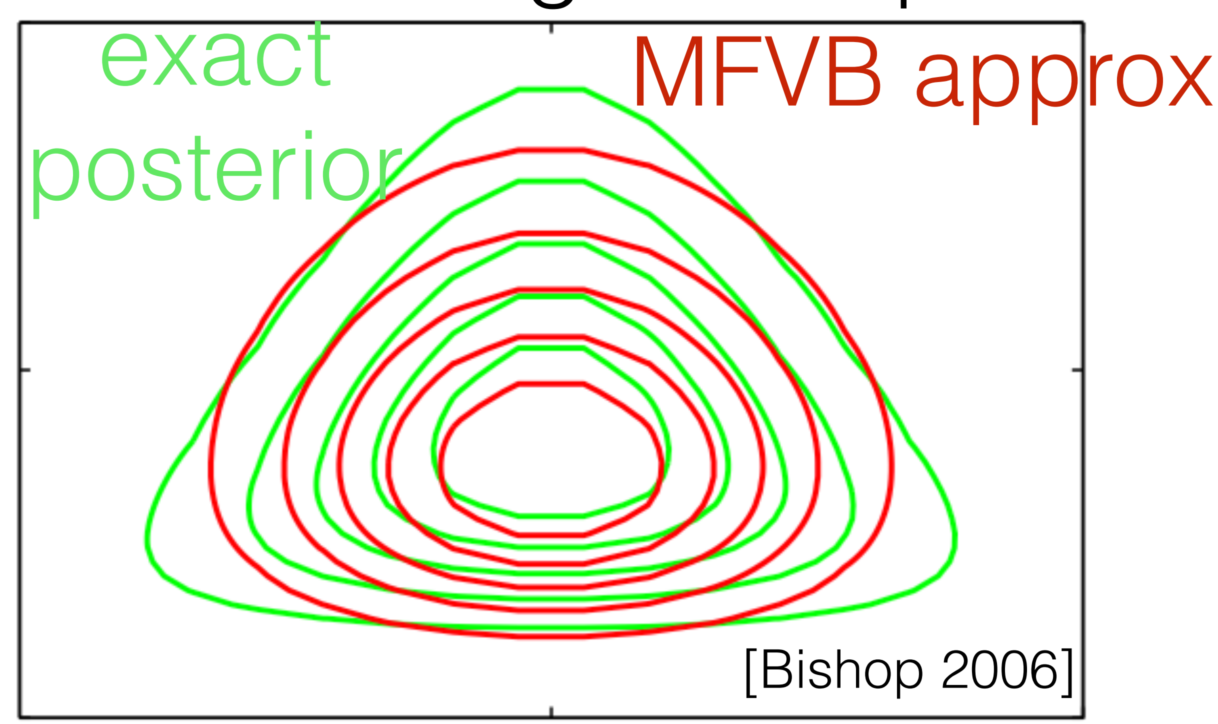
Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

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

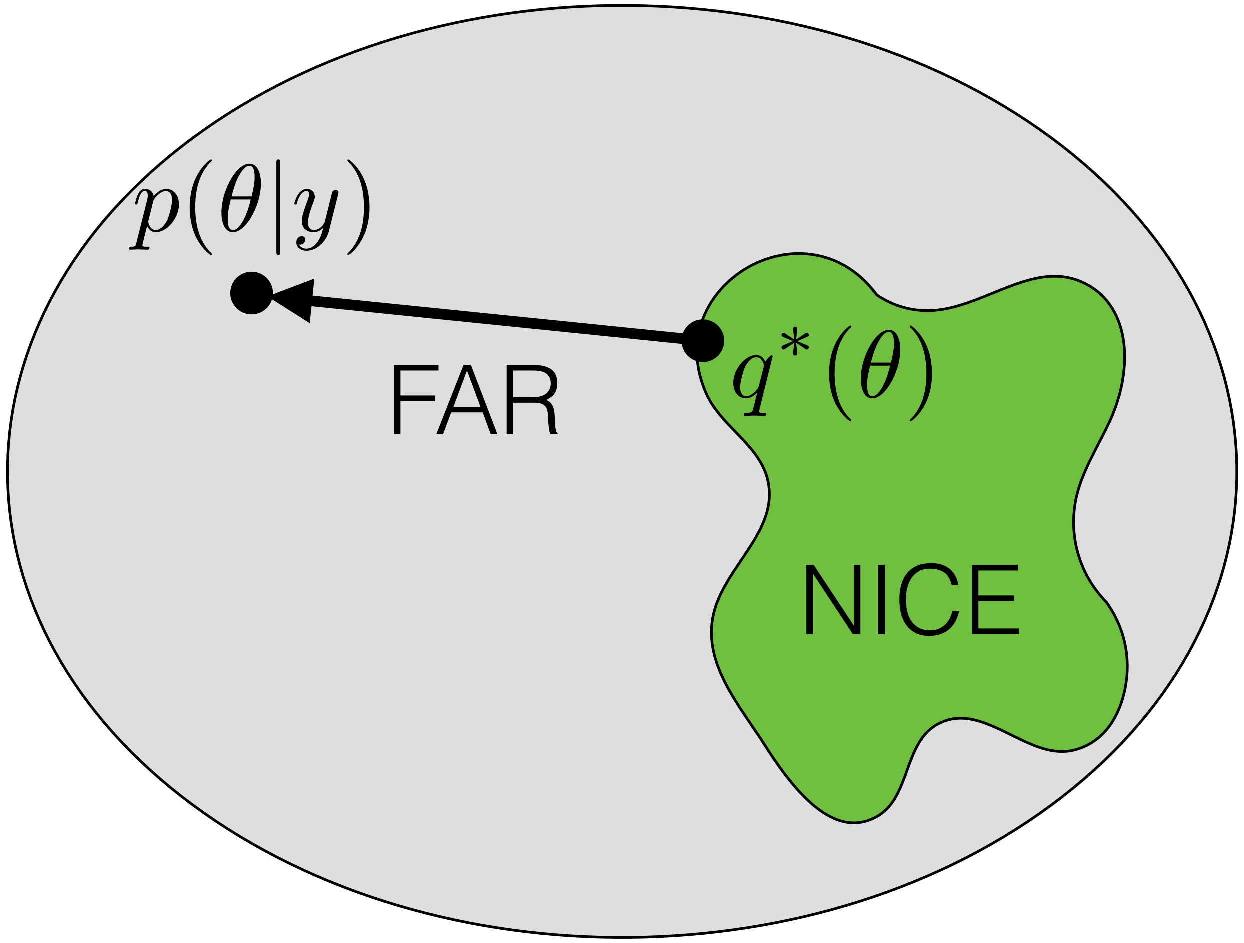
- Often also exponential family
- *Not* a modeling assumption

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



Variational Bayes

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

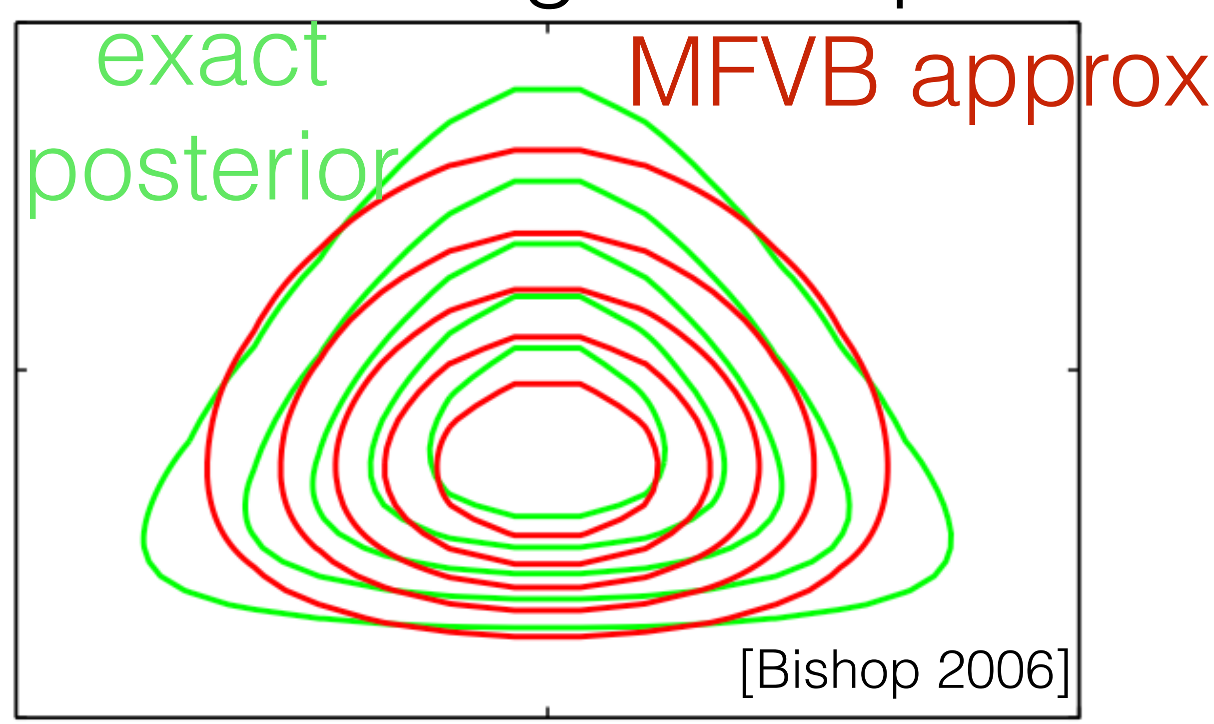


Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

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

- Often also exponential family
- *Not* a modeling assumption



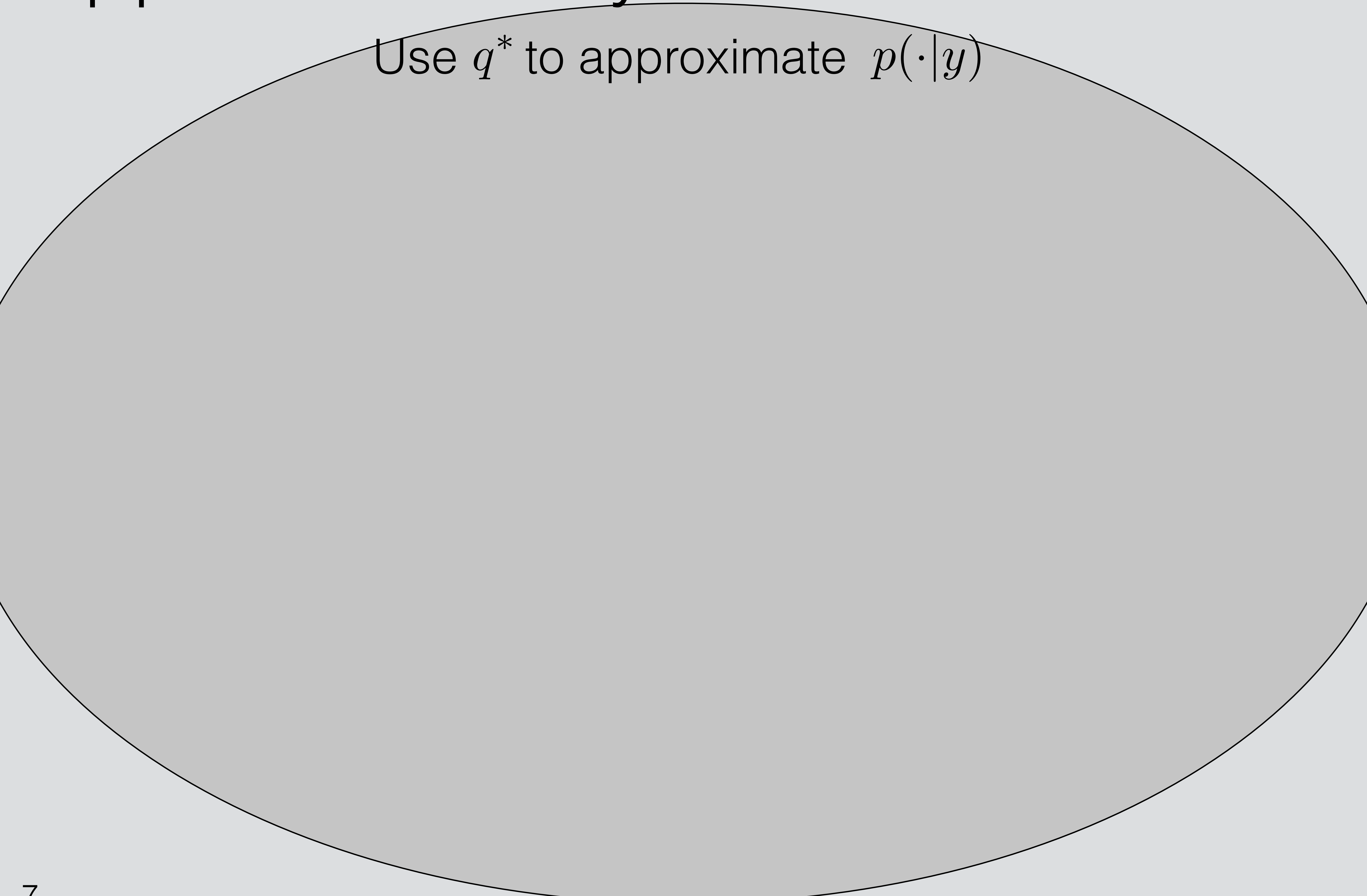
Now we have an optimization problem; how to solve it?

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

Approximate Bayesian inference

Approximate Bayesian inference

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



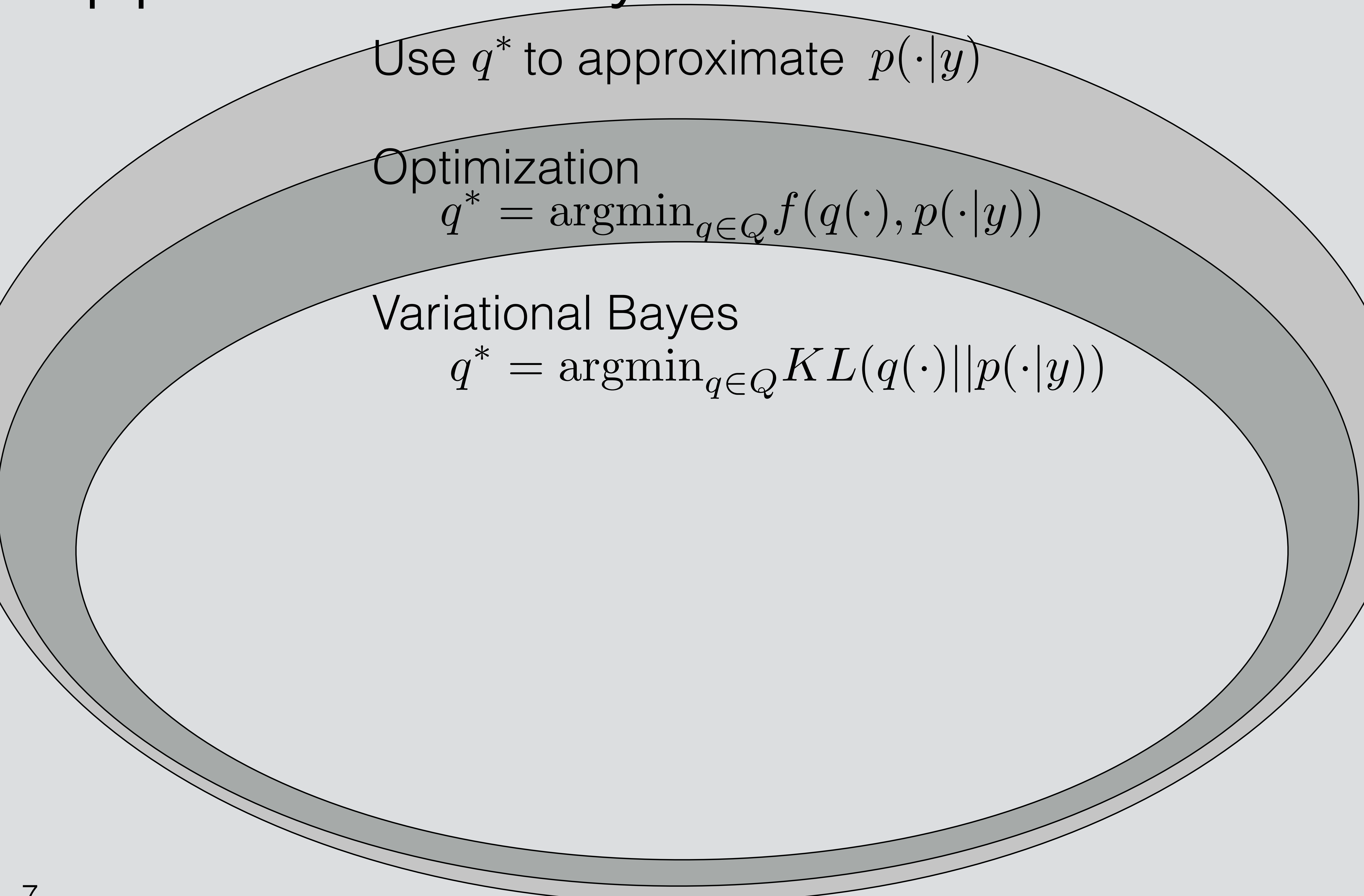
Approximate Bayesian inference

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

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

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

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

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

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

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

Variational Bayes

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

Mean-field variational Bayes

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

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

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al 2013]
- Newton trust region [used in e.g. Giordano et al 2018]

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al 2013]
- Newton trust region [used in e.g. Giordano et al 2018]

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]
- Newton trust region [used in e.g. Giordano et al 2018]

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]
- Newton trust region [used in e.g. Giordano et al 2018]

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB? Some VB successes (speed, accuracy)
- Some VB failure modes, and partial solutions
- Ease of use / automation
  - Automatic differentiation variational inference (ADVI) and beyond

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB? Some VB successes (speed, accuracy)
- Some VB failure modes, and partial solutions
- Ease of use / automation
  - Automatic differentiation variational inference (ADVI) and beyond

Air pollution: Particulate matter

[Krongut 2020]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$

[Krongut 2020]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$

- Model:

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and variance $\theta = (\mu, \sigma^2)$
- Model:

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and variance

$\theta = (\mu, \sigma^2)$

- Model:

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and variance

$$\theta = (\mu, \sigma^2)$$

- Model (conjugate prior):

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and variance

$\theta = (\mu, \sigma^2)$

- Model (conjugate prior): [Exercise: find the posterior]

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and variance $\theta = (\mu, \sigma^2)$
- Model (conjugate prior): [Exercise: find the posterior]

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision

$\theta = (\mu, \tau)$

- Model (conjugate prior): [Exercise: find the posterior]

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision

$$\theta = (\mu, \tau)$$

- Model (conjugate prior): [Exercise: find the posterior]

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision

$$\theta = (\mu, \tau)$$

- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}),$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})$$

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}),$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})$$

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}),$$

$$p(\theta): \quad \tau \sim \mathrm{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})$$

- Exercise: check

$$p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$$

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision

$$\theta = (\mu, \tau)$$

- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}),$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})$$

[Krongut 2020]

- Exercise: check

$$p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$$

- MFVB approximation:

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

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision

$$\theta = (\mu, \tau)$$

- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}),$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})$$

[Krongut 2020]

- Exercise: check

$$p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$$

- MFVB approximation:

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

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

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}),$$
$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$
$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})$$

[Krongut 2020]

- Exercise: check

$$p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$$

- MFVB approximation:

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

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

$$q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1})$$
$$q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$$

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}),$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})$$

[Krongut 2020]

- Exercise: check

$$p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$$

- MFVB approximation:

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

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

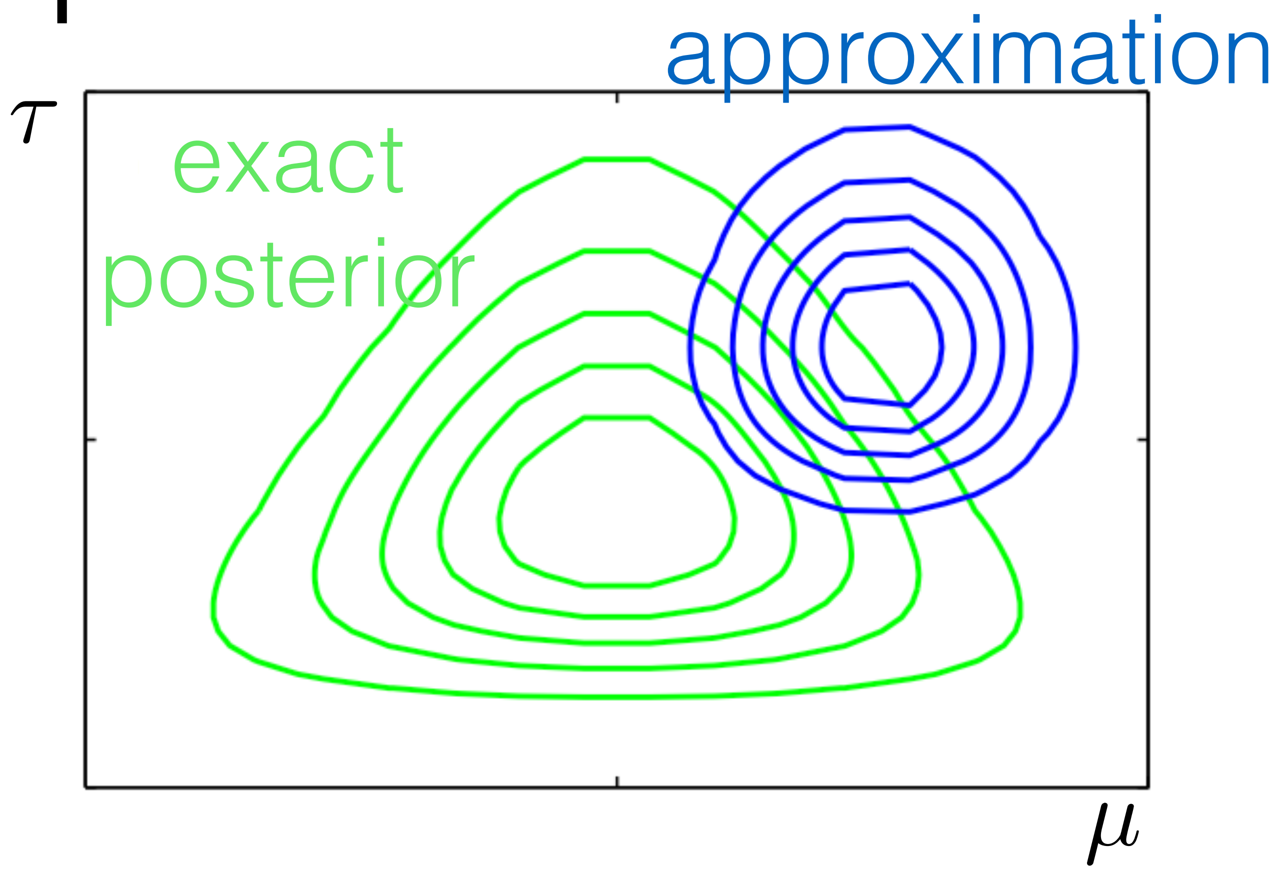
$$q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1})$$

$$q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$$

"variational parameters"

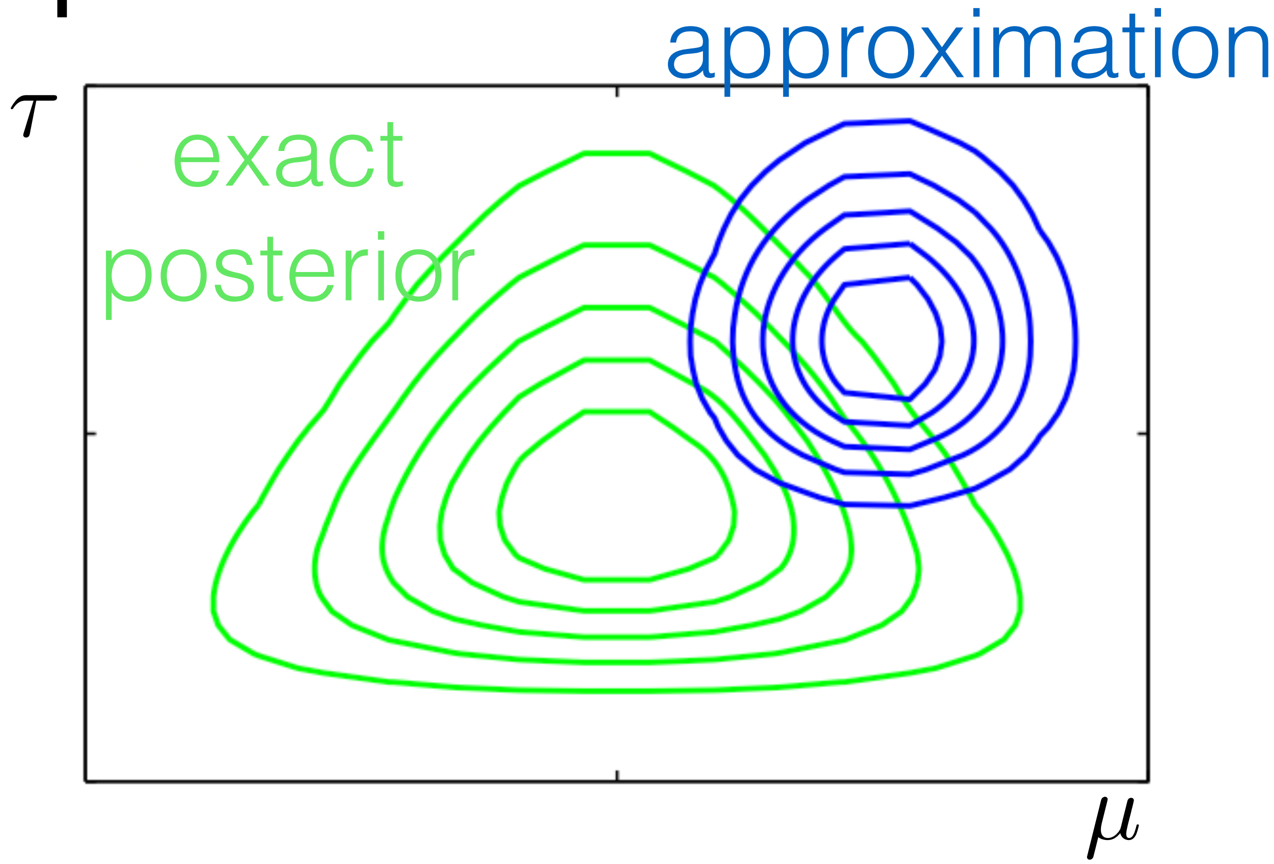
[MacKay 2003; Bishop 2006]

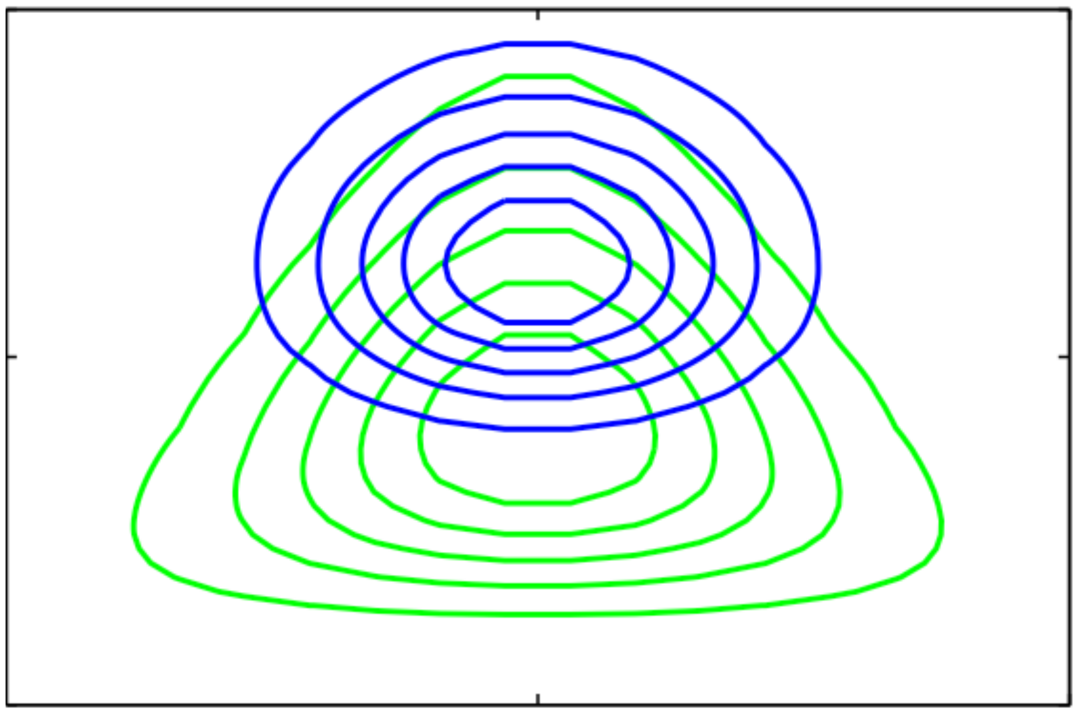
Air pollution: Particulate matter



[Bishop 2006]

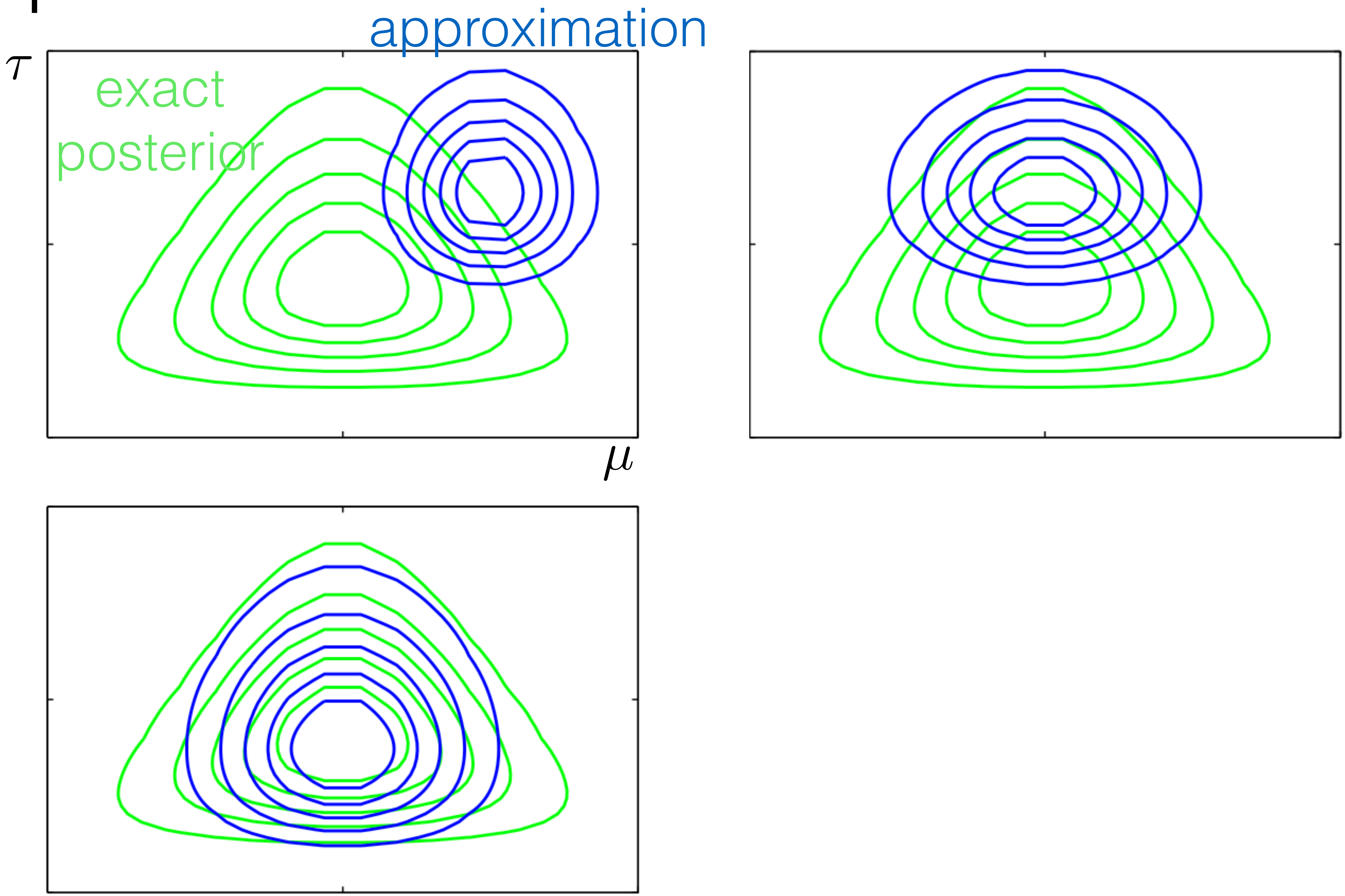
Air pollution: Particulate matter





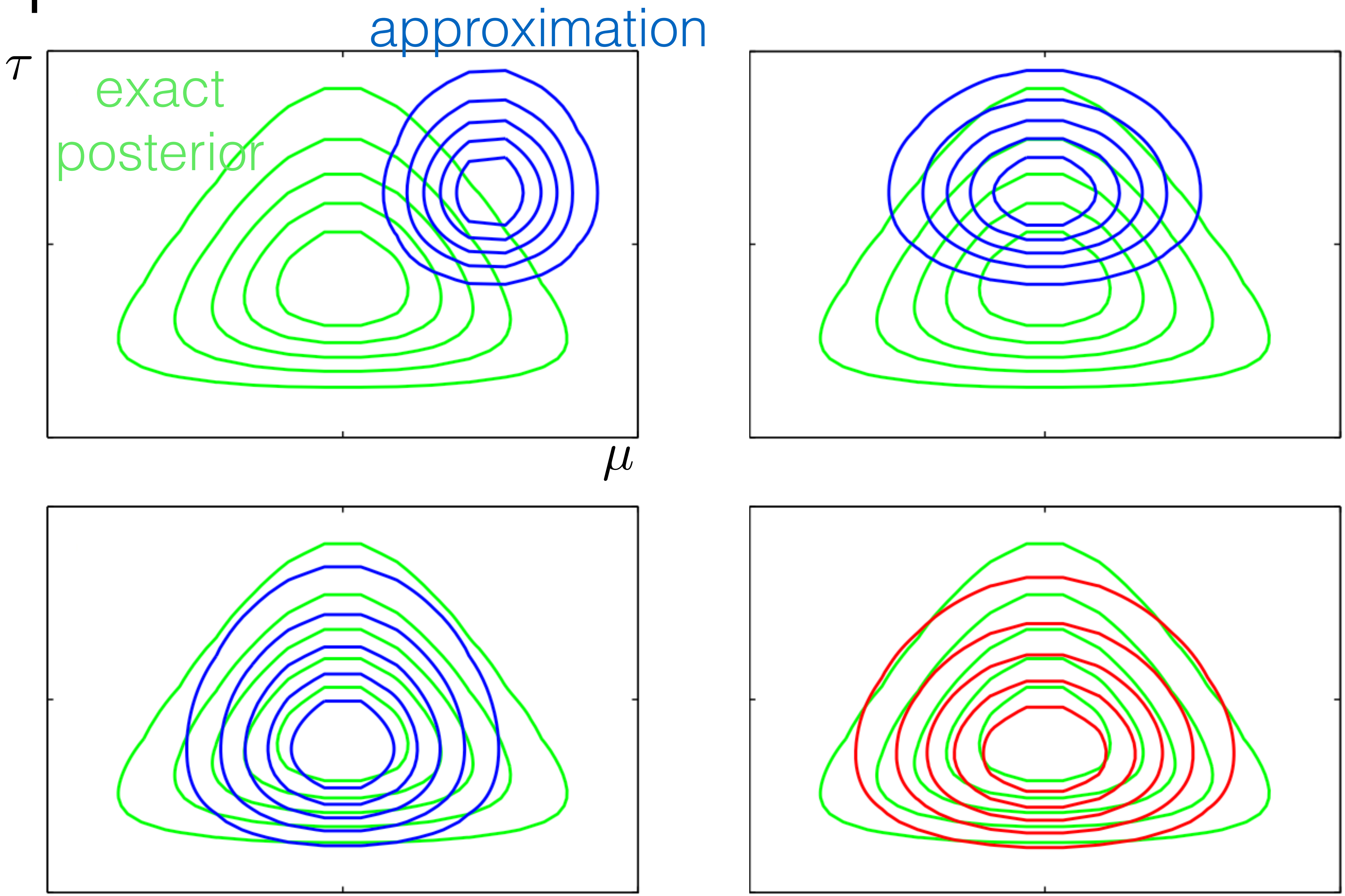
[Bishop 2006]

Air pollution: Particulate matter



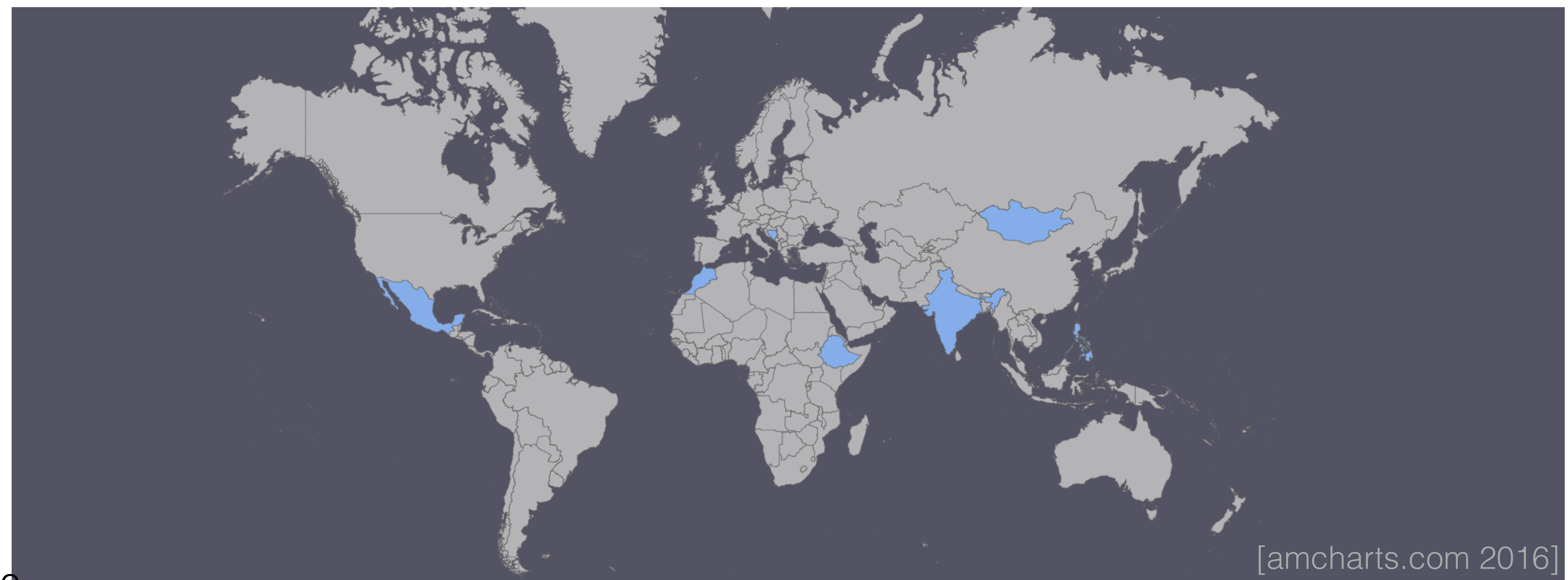
[Bishop 2006]

Air pollution: Particulate matter



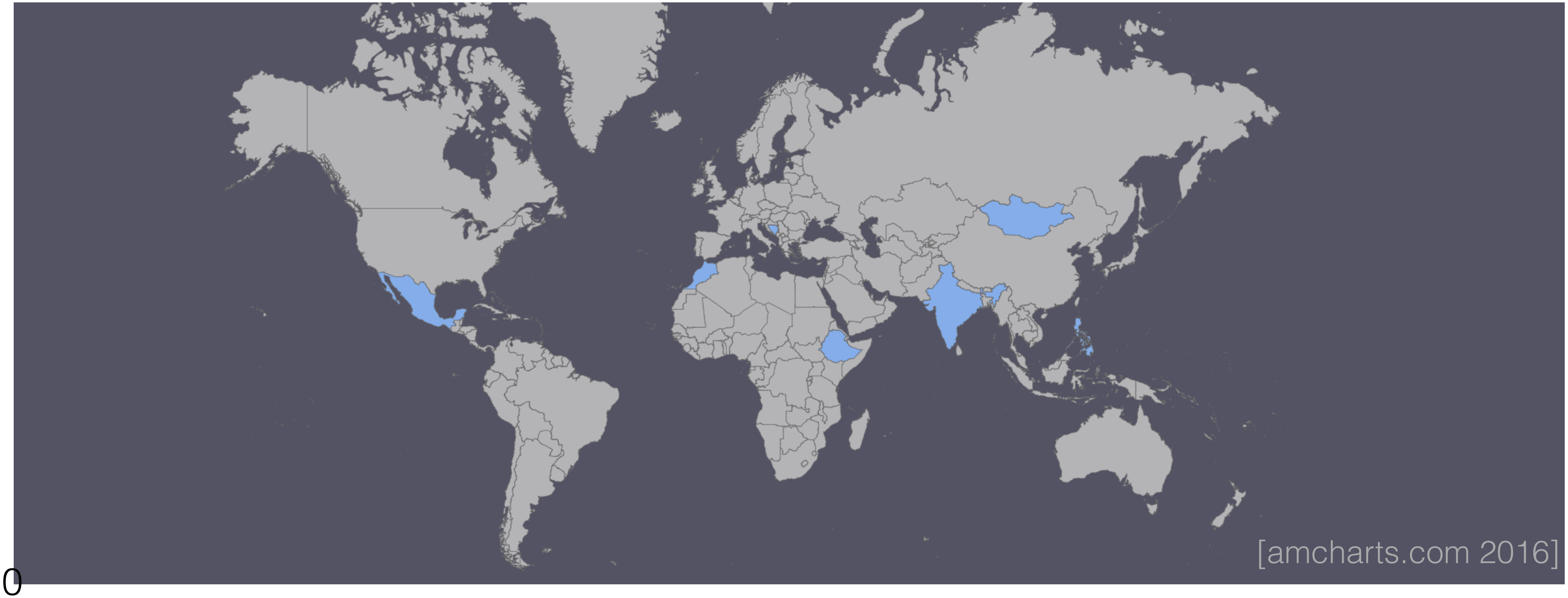
[Bishop 2006]

Microcredit Experiment



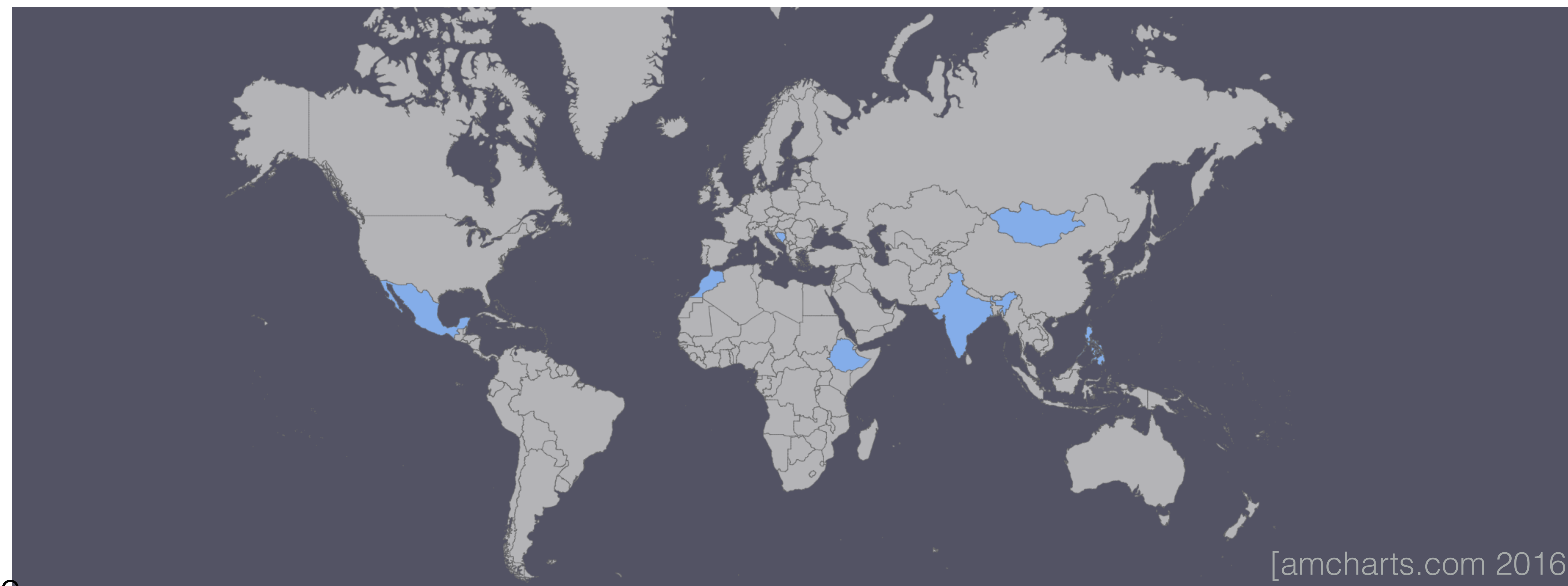
Microcredit Experiment

- Simplified from Meager (2019)



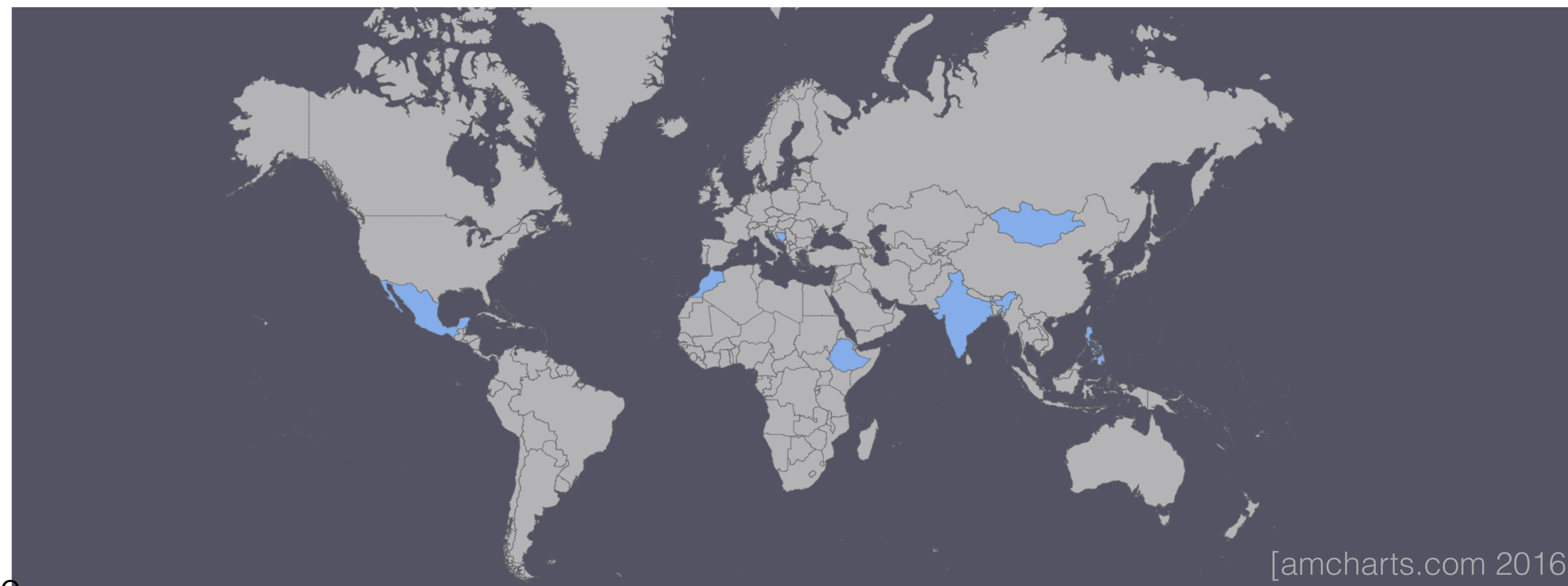
Microcredit Experiment

- Simplified from Meager (2019)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)



Microcredit Experiment

- Simplified from Meager (2019)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)



Microcredit Experiment

- Simplified from Meager (2019)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)

Microcredit Experiment

- Simplified from Meager (2019)
- K = 7 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 (2019)
- K = 7 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 (2019)
- K = 7 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

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

Microcredit Experiment

- Simplified from Meager (2019)
- K = 7 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

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

Microcredit Experiment

- Simplified from Meager (2019)
- K = 7 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

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

Microcredit Experiment

- Simplified from Meager (2019)
- K = 7 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 (2019)
- K = 7 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 (2019)
- K = 7 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 (2019)
- K = 7 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

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

Microcredit Experiment

- Simplified from Meager (2019)
- K = 7 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 (2019)
- K = 7 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:

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

Microcredit Experiment

- Simplified from Meager (2019)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

1 if microcredit

profit

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

- Priors and hyperpriors:

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

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

Microcredit Experiment

- Simplified from Meager (2019)
- K = 7 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}

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

- Priors and hyperpriors:

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

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

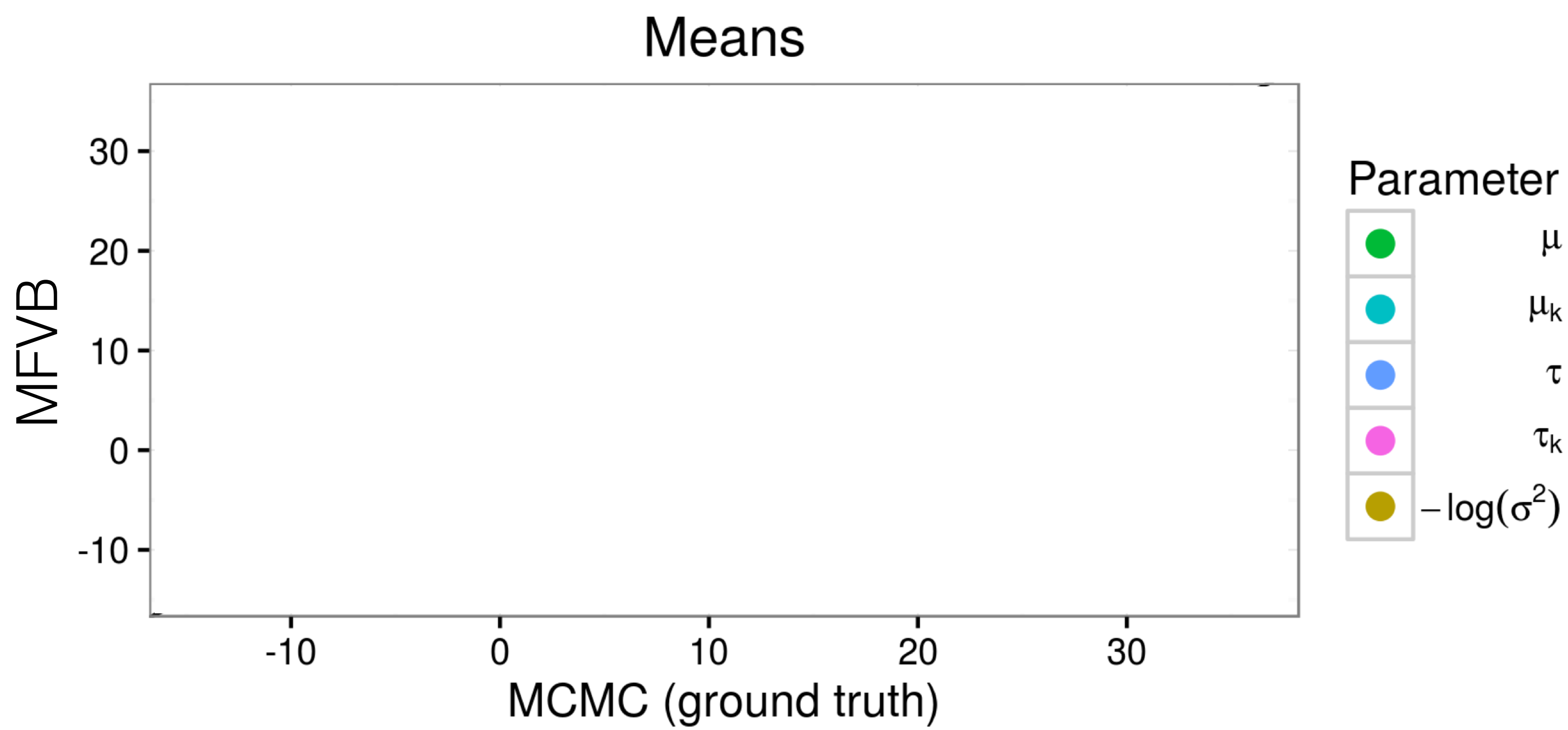
Microcredit

MFVB: Do we need to check the output?

Microcredit

MFVB: How will we know if it’s working?

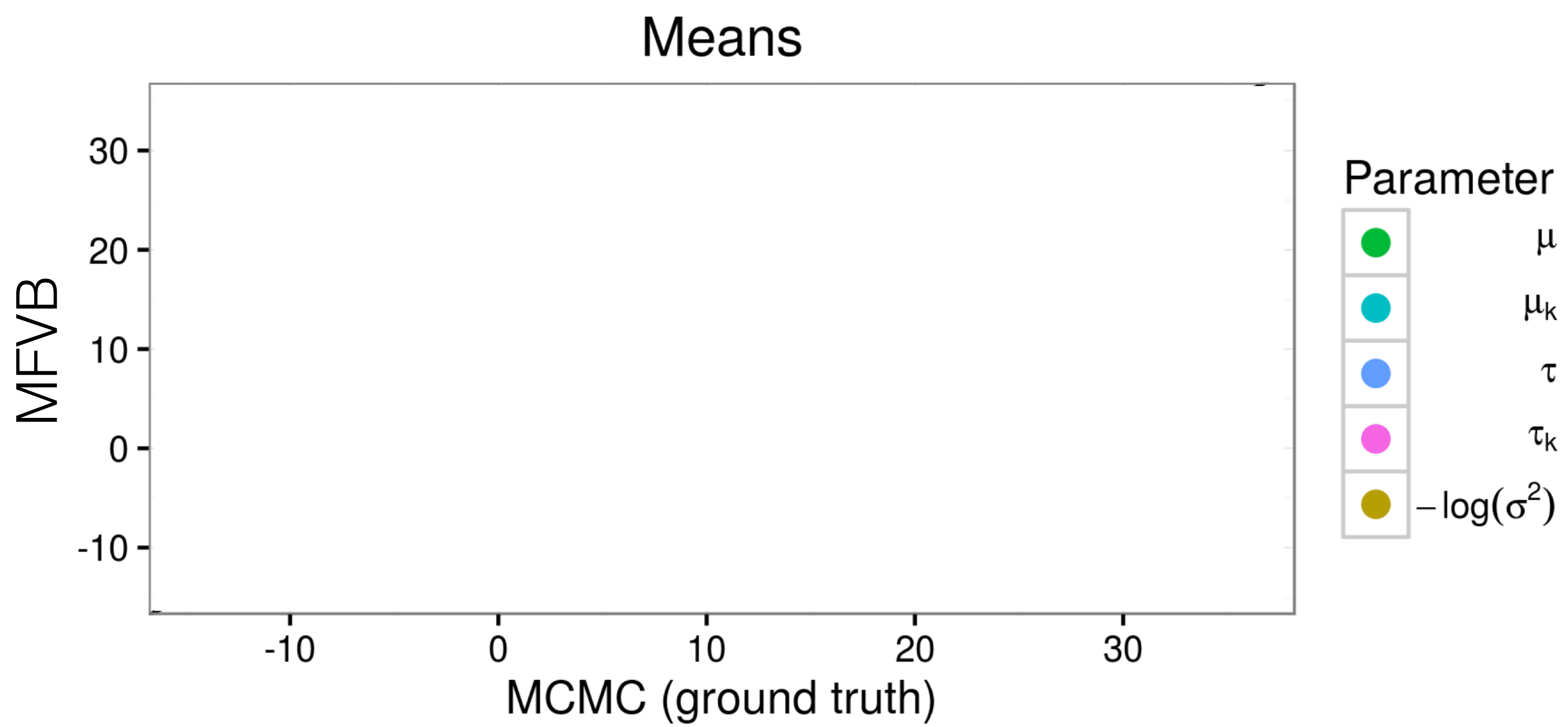
Microcredit



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

Microcredit

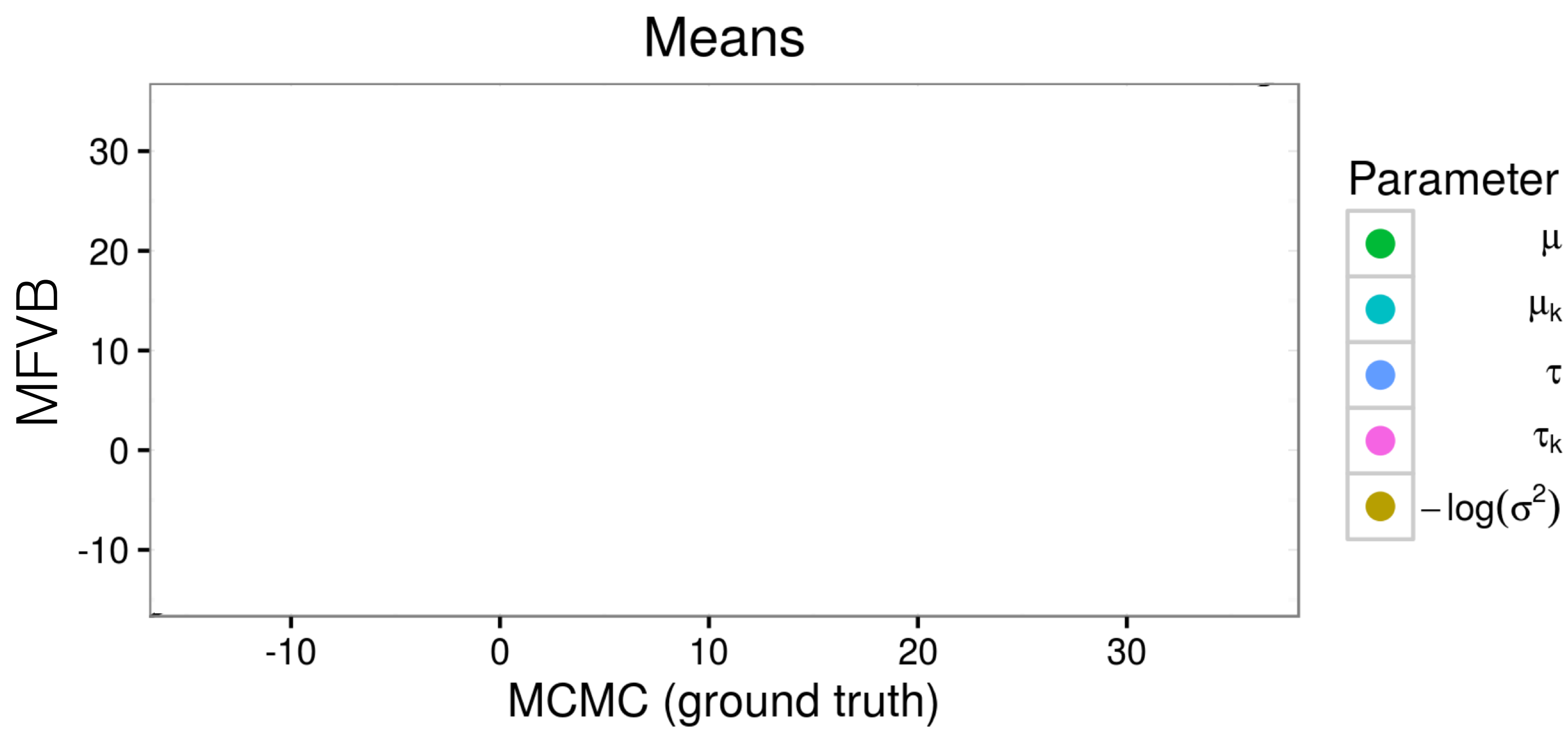
- *One set* of 2500 MCMC draws: **45 minutes**



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

Microcredit

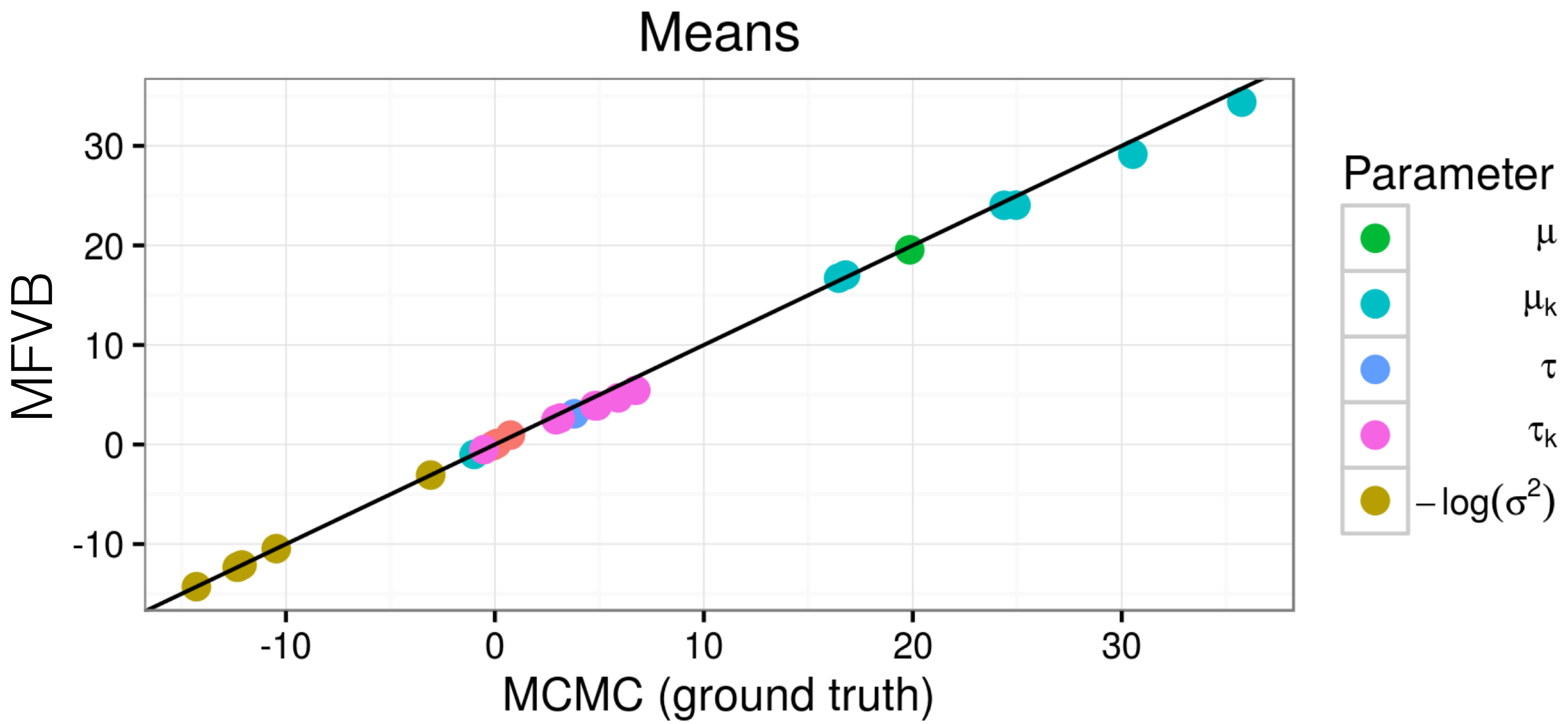
- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



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

Microcredit

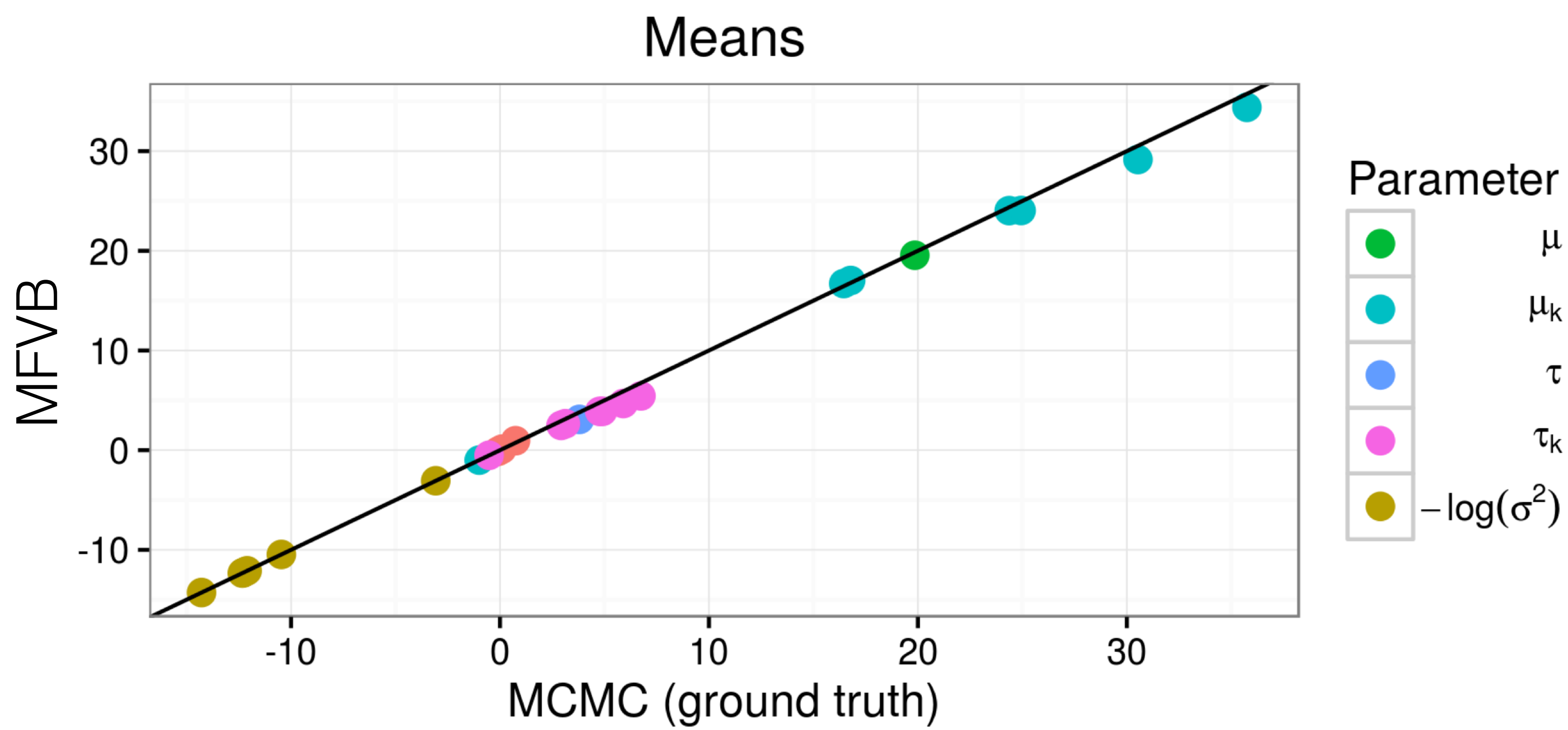
- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



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

Microcredit

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



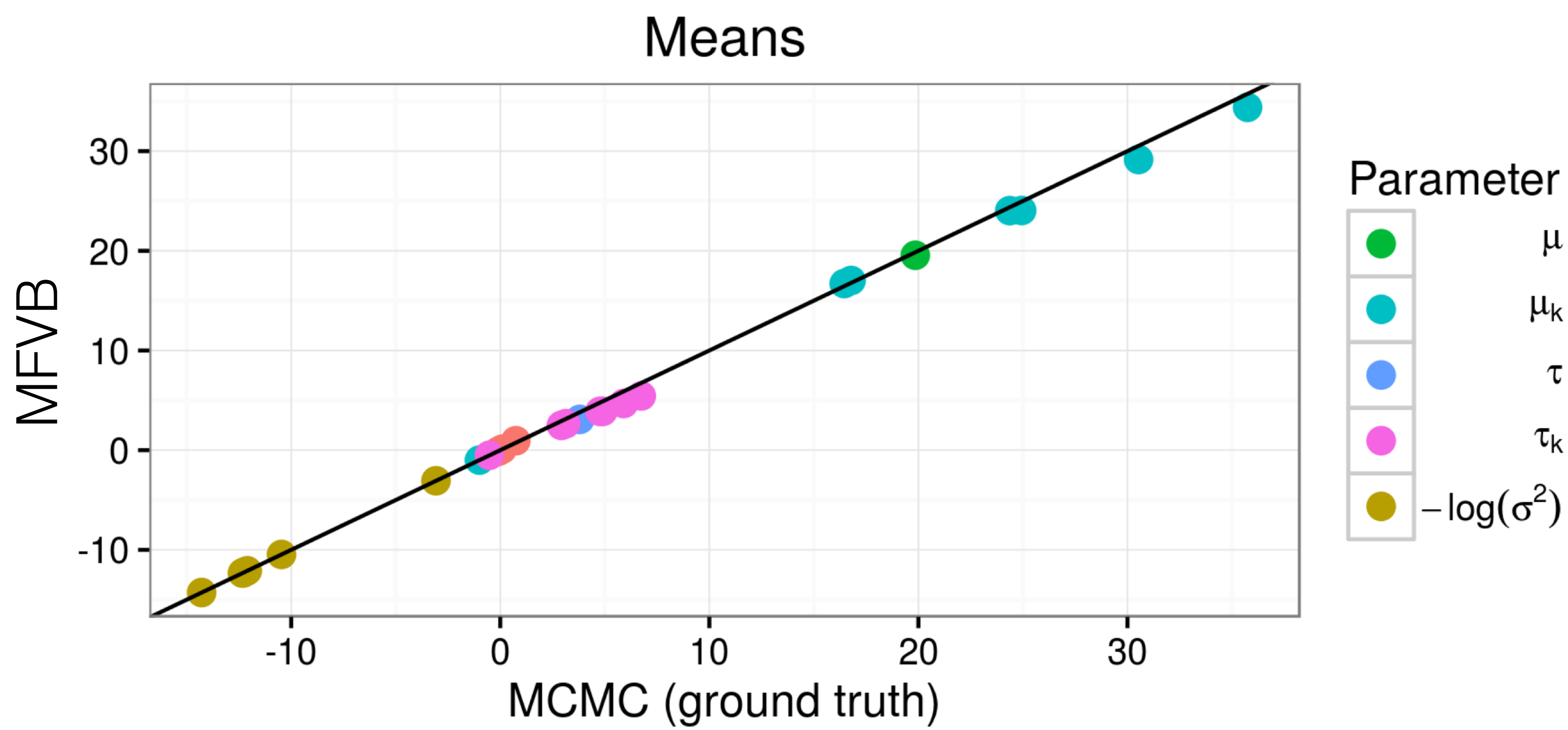
Criteo Online Ads Experiment

- 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**



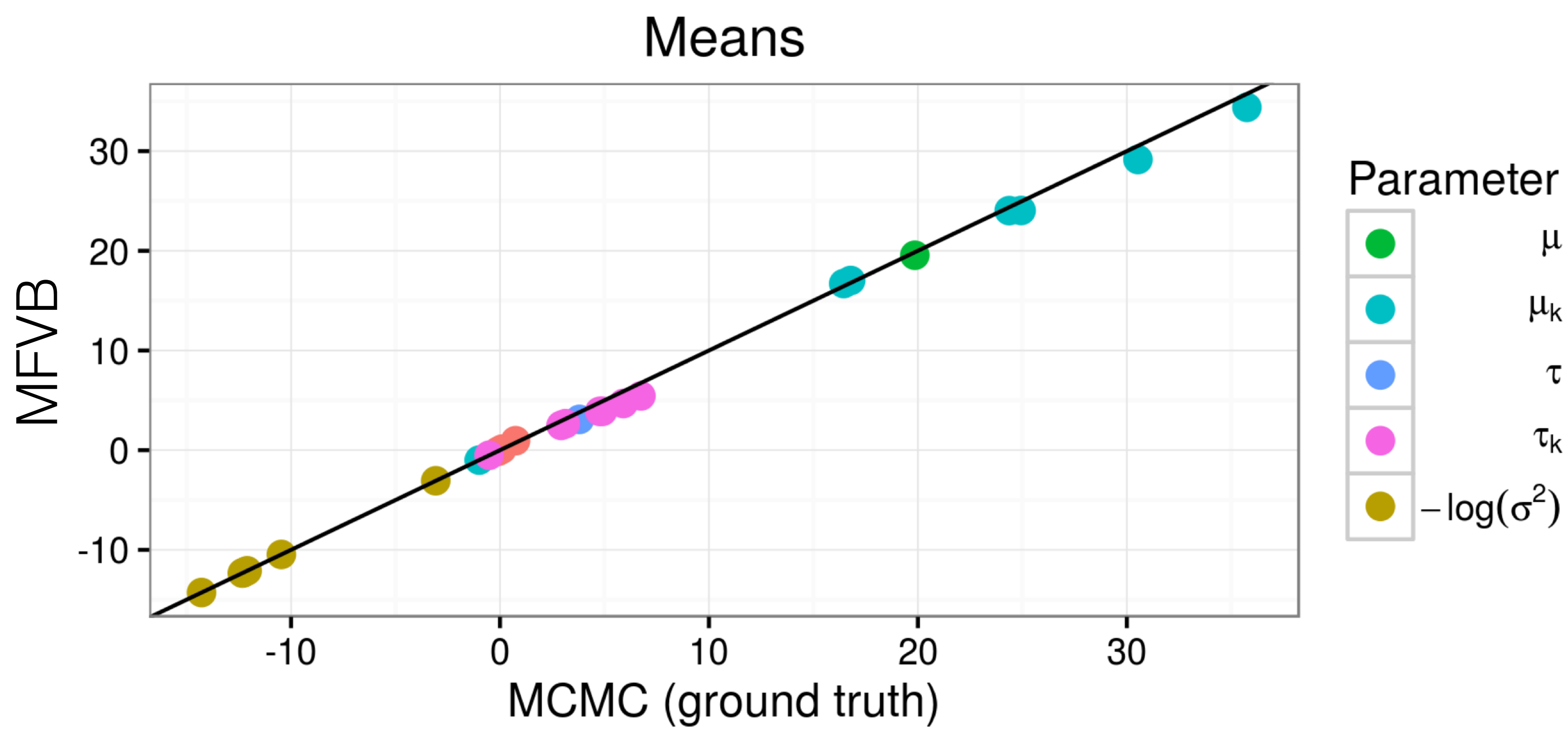
Criteo Online Ads Experiment

- 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**



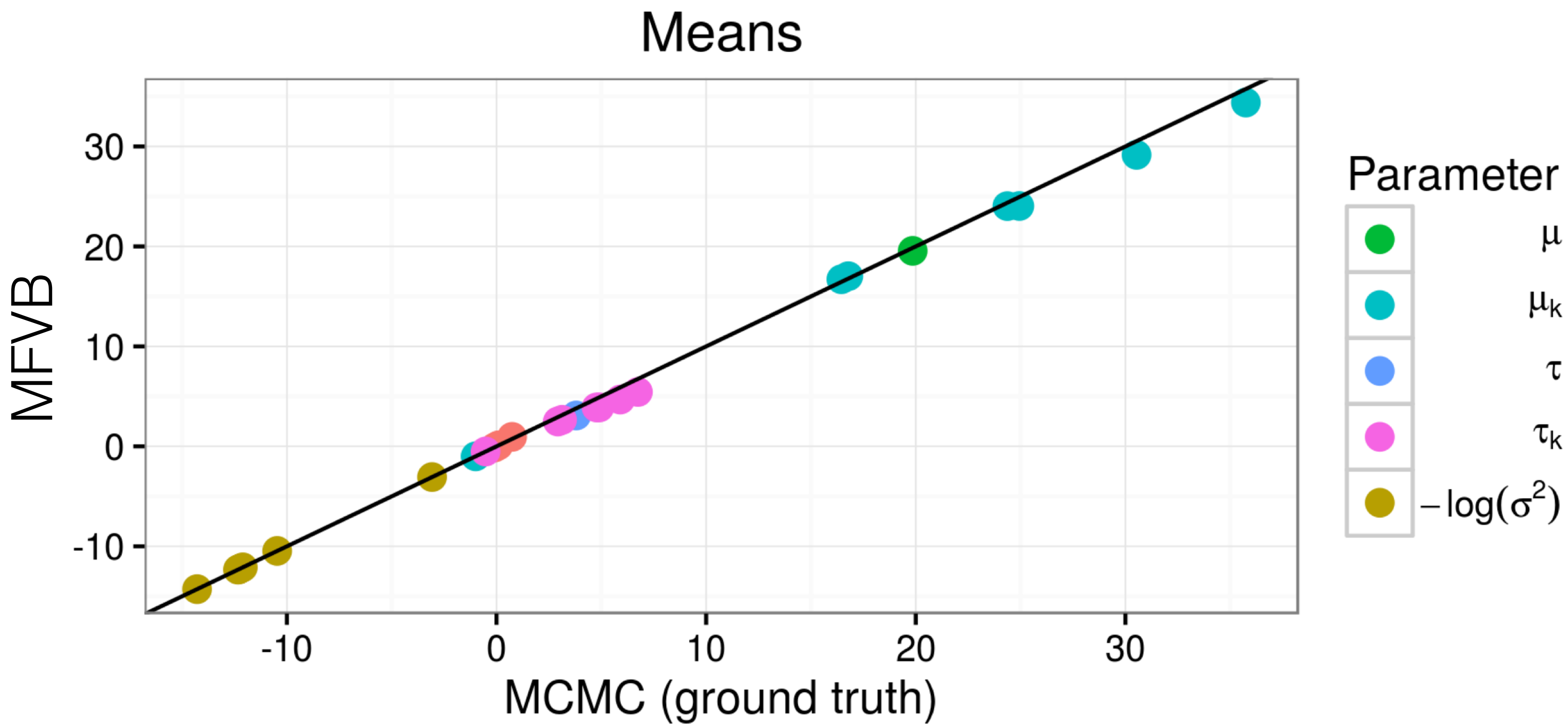
Criteo Online Ads Experiment

- 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

[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2018]

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 2018]

Criteo Online Ads Experiment

[Giordano, Broderick, Jordan 2018]

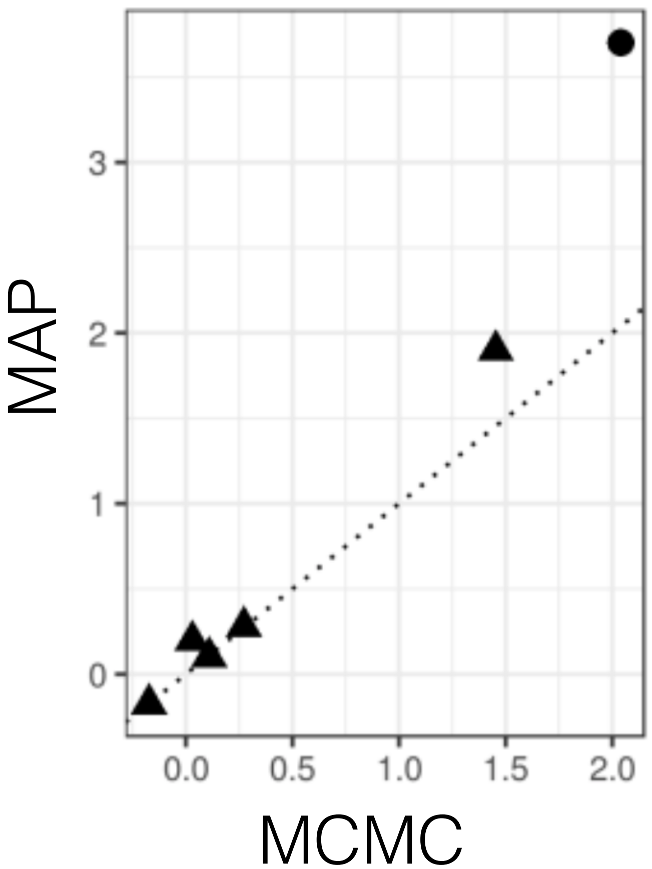
Criteo Online Ads Experiment

- MAP: **12 s**

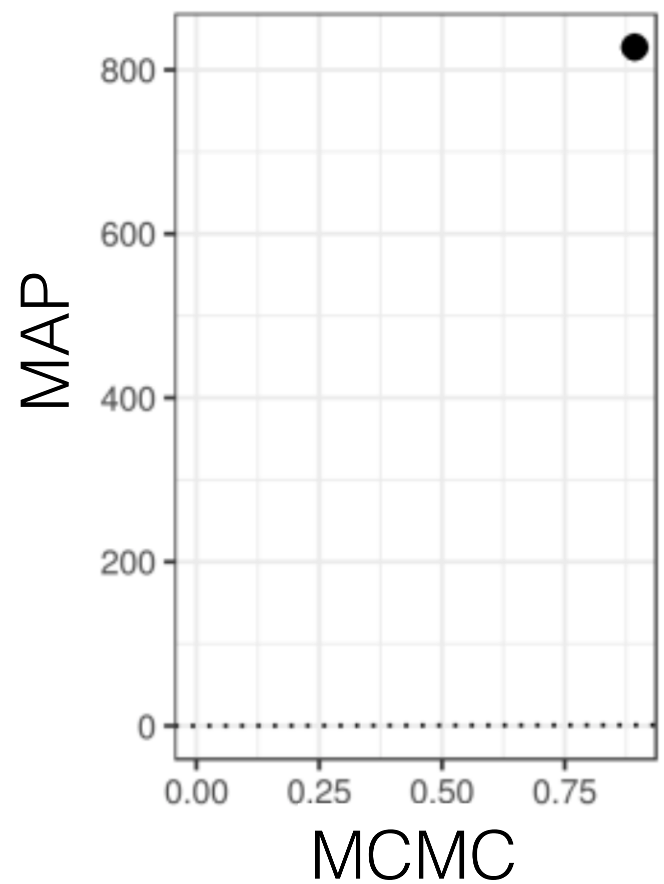
[Giordano, Broderick, Jordan 2018]

Criteo Online Ads Experiment

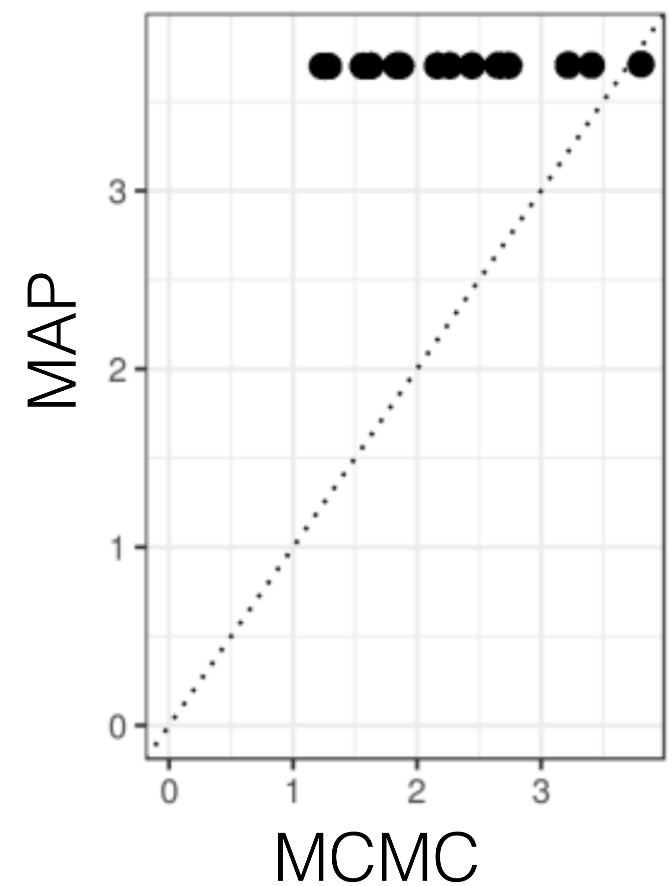
Global parameters (-τ)



Global parameter τ



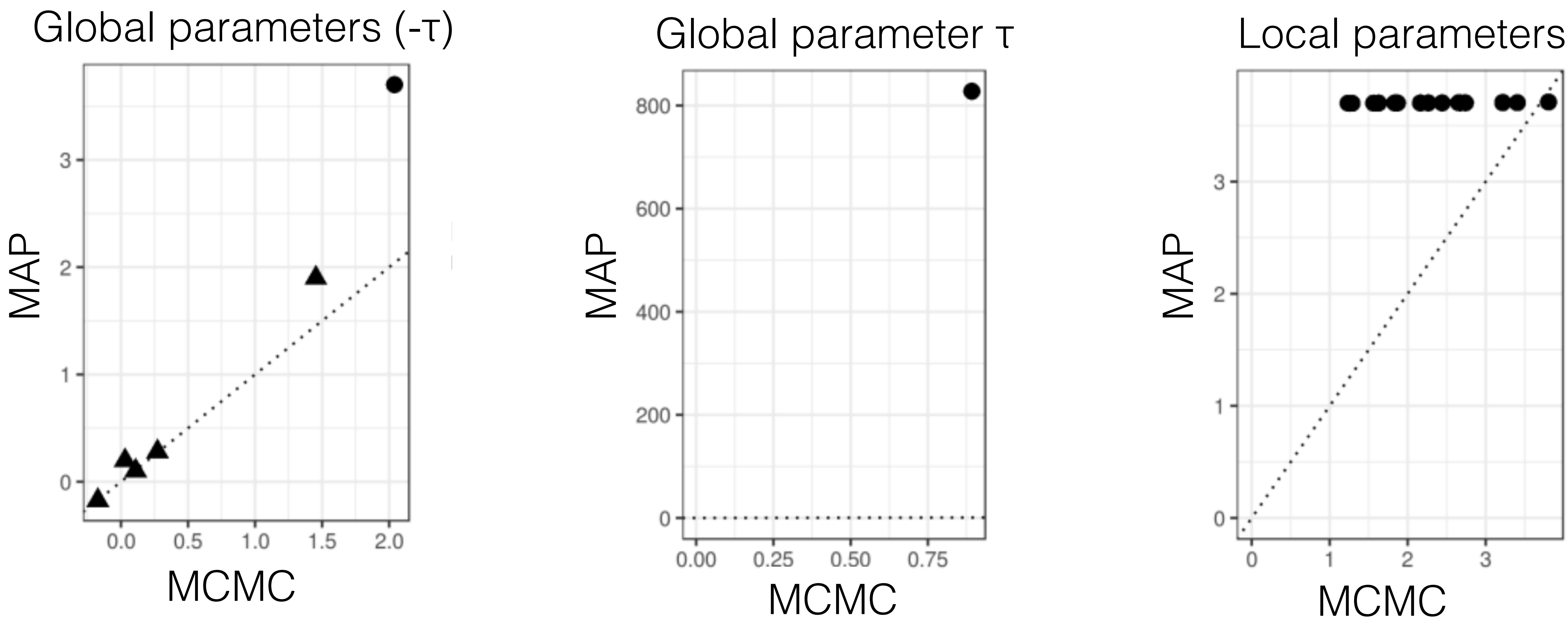
Local parameters



- MAP: **12 s**

[Giordano, Broderick, Jordan 2018]

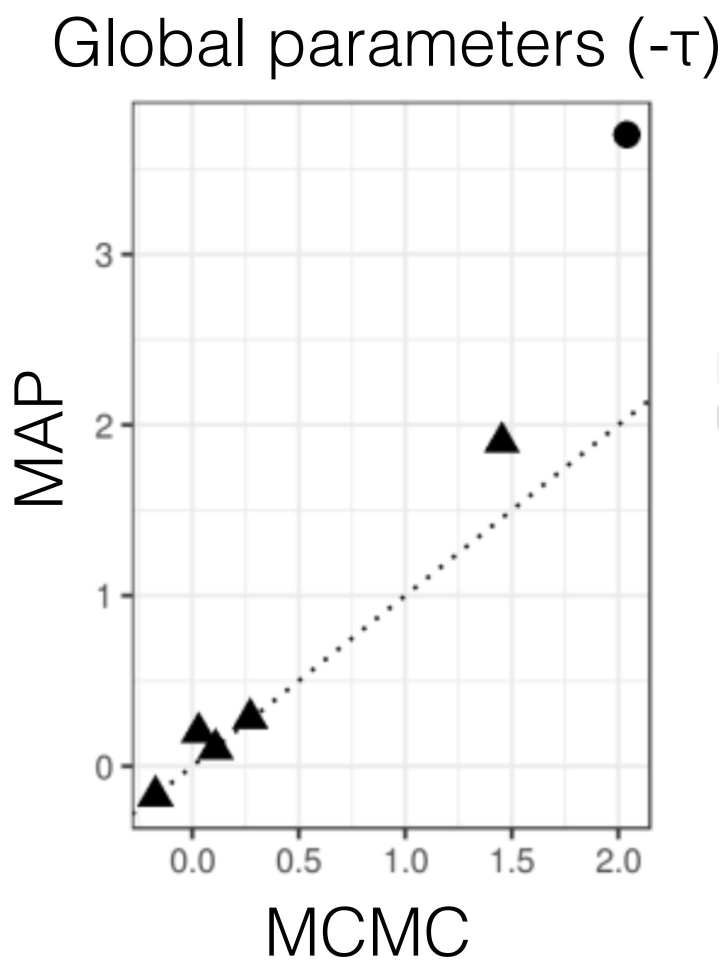
Criteo Online Ads Experiment

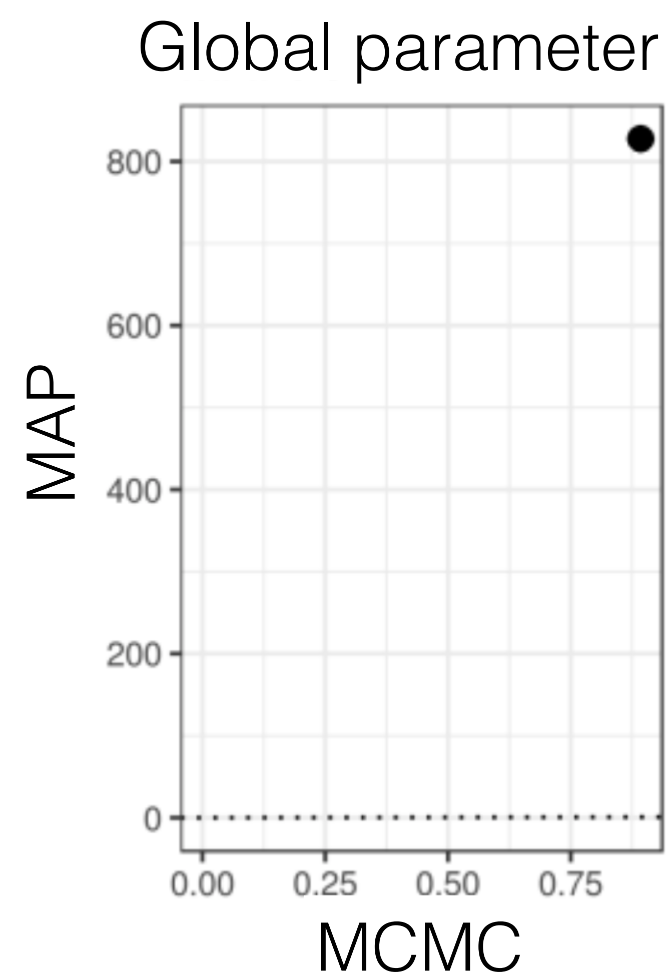


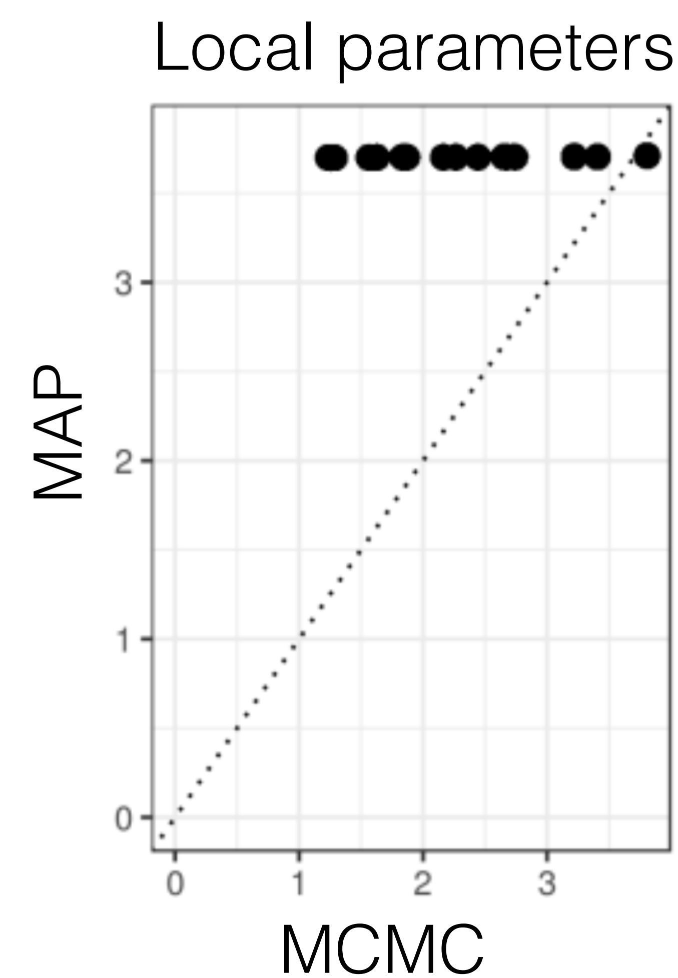
- MAP: **12 s**
- MFVB: **57 s**

[Giordano, Broderick, Jordan 2018]

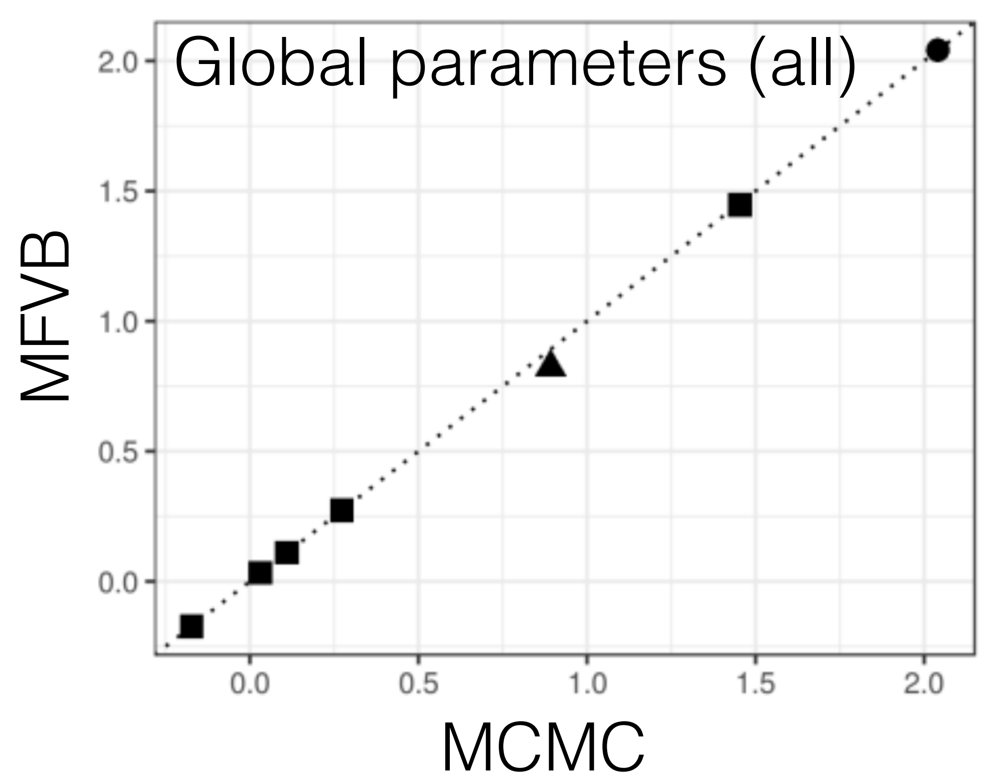
Criteo Online Ads Experiment

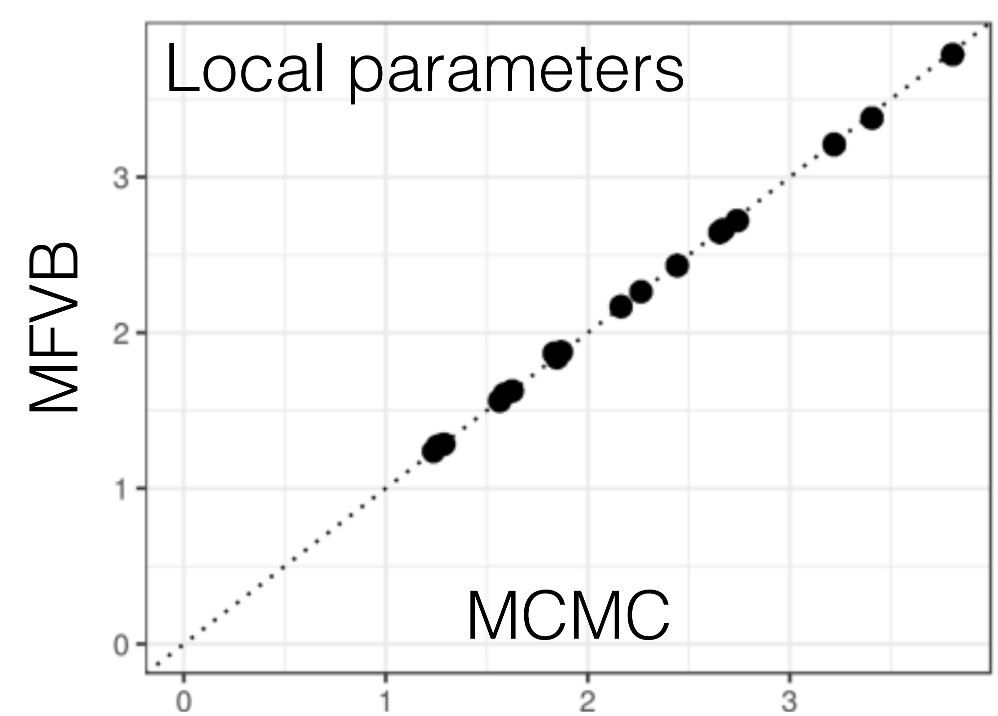






- MAP: **12 s**
- MFVB: **57 s**

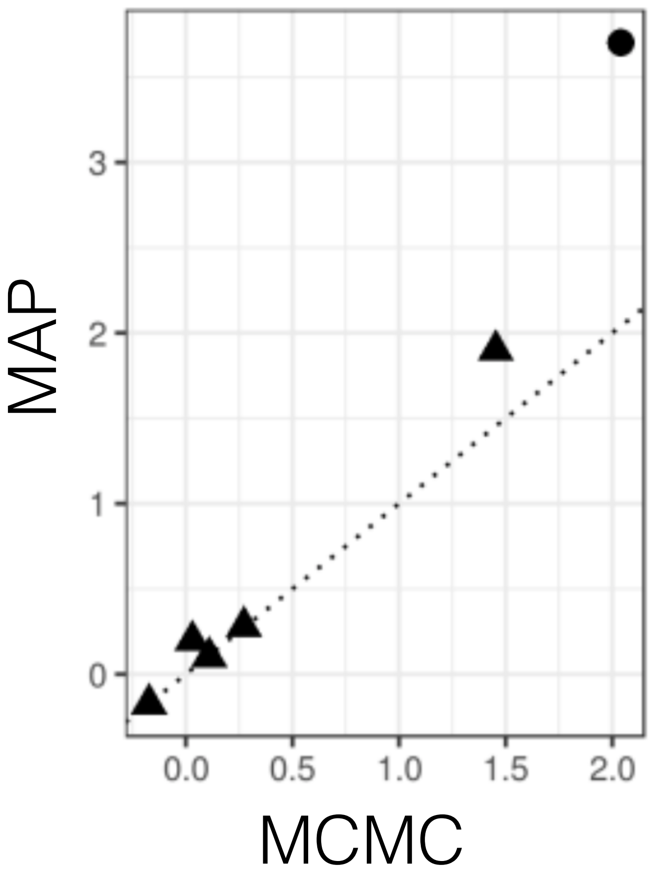




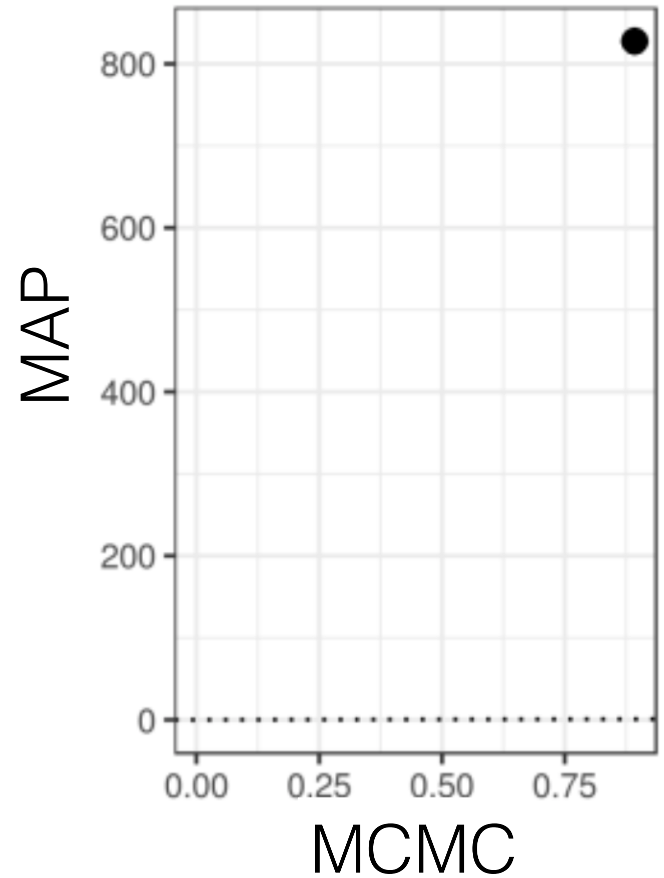
[Giordano, Broderick, Jordan 2018]

Criteo Online Ads Experiment

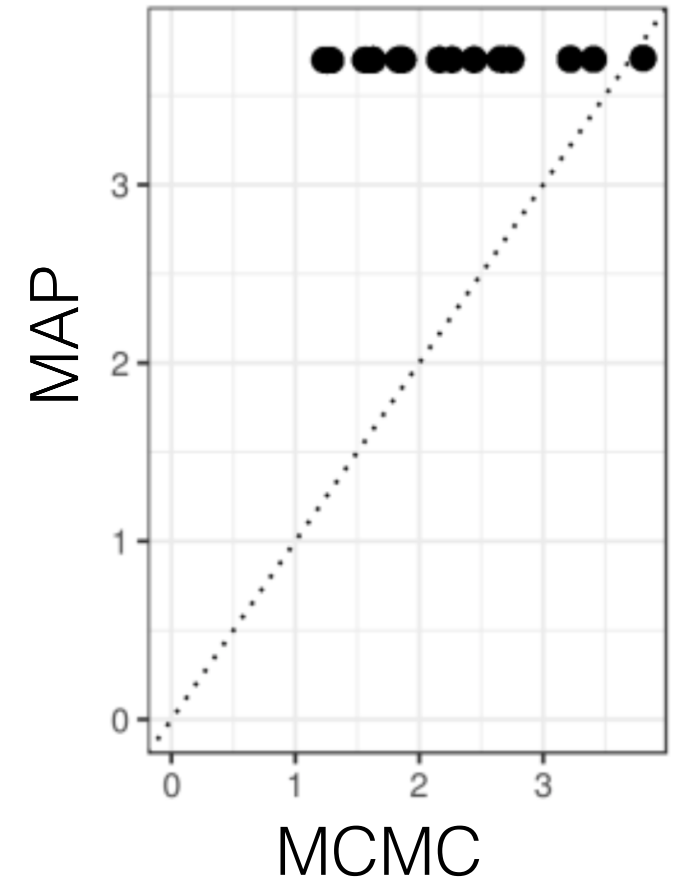
Global parameters (-τ)



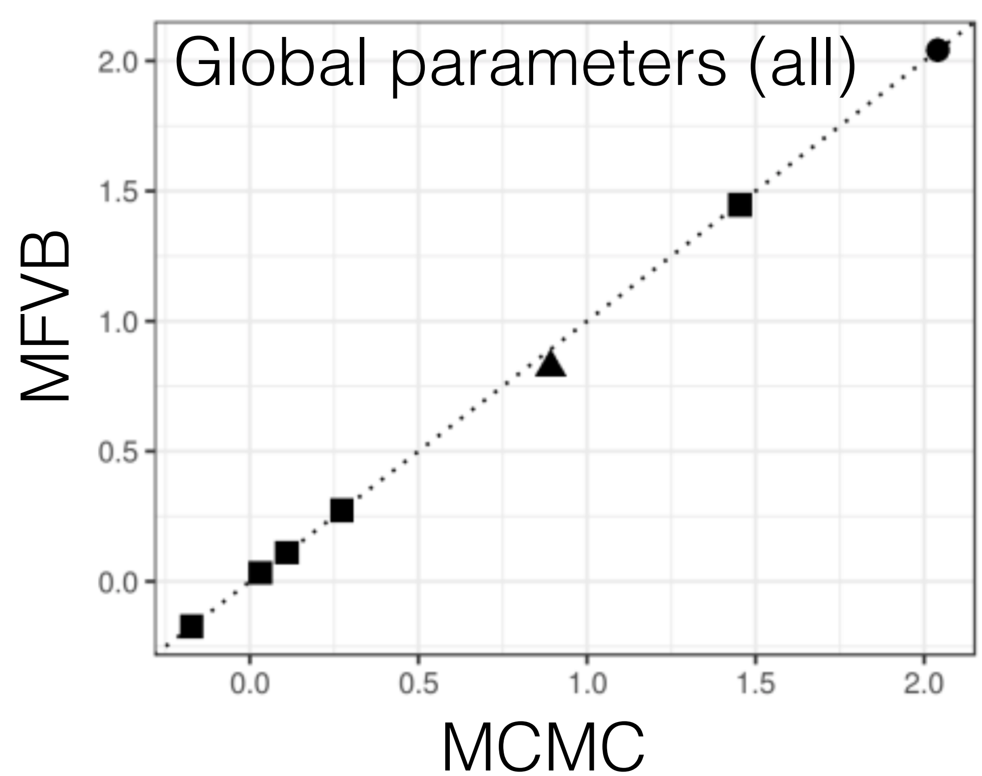
Global parameter τ

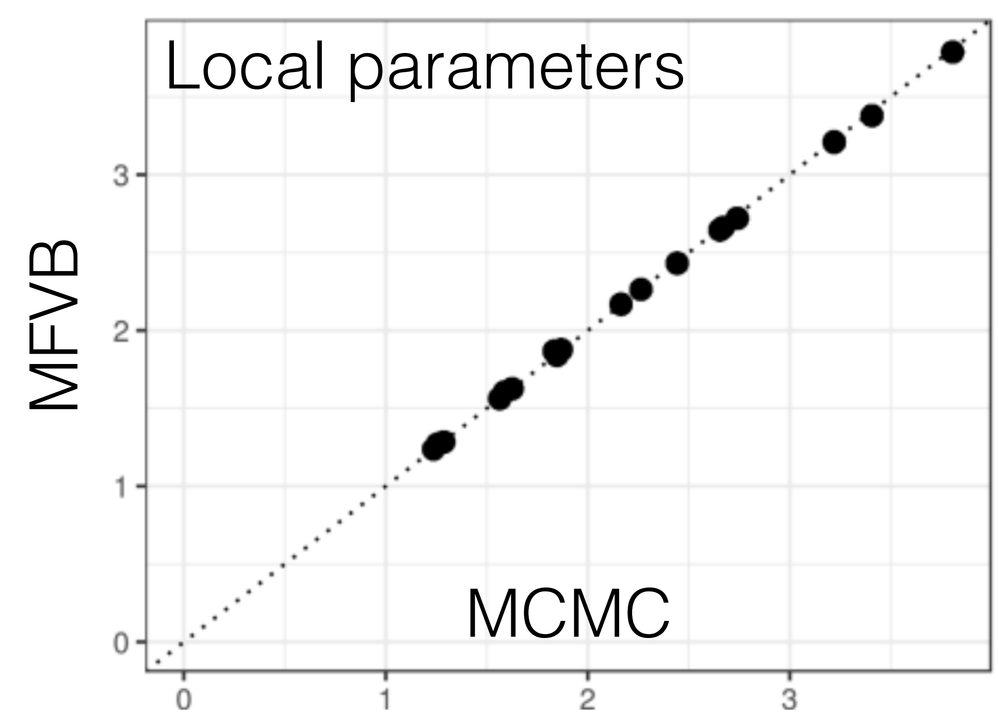


Local parameters



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





[Giordano, Broderick, Jordan 2018]

Why use MFVB?

- Topic discovery

"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]

Why use MFVB?

- Topic discovery
  - Latent Dirichlet allocation (LDA)

"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]

Why use MFVB?

- Topic discovery
  - Latent Dirichlet allocation (LDA): 52,700+ citations

"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]

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB? Some VB successes (speed, accuracy)
- Some VB failure modes, and partial solutions
- Ease of use / automation
  - Automatic differentiation variational inference (ADVI) and beyond

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB? Some VB successes (speed, accuracy)
  - Often fast and accurate for point estimates
- Some VB failure modes, and partial solutions
- Ease of use / automation
  - Automatic differentiation variational inference (ADVI) and beyond

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB? Some VB successes (speed, accuracy)
  - Often fast and accurate for point estimates
- Some VB failure modes, and partial solutions
- Ease of use / automation
  - Automatic differentiation variational inference (ADVI) and beyond

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB? Some VB successes (speed, accuracy)
  - Often fast and accurate for point estimates
- Some VB failure modes, and partial solutions
  - Issues with uncertainty and more
- Ease of use / automation
  - Automatic differentiation variational inference (ADVI) and beyond

What about uncertainty?

What about uncertainty?

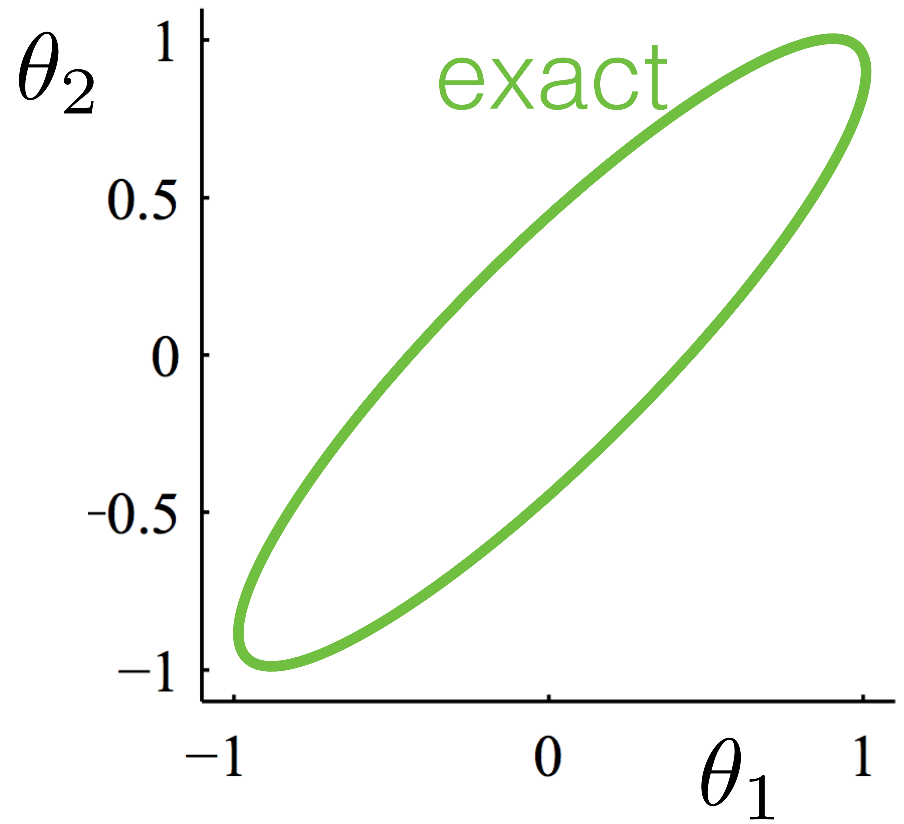
$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

What about uncertainty?

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

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

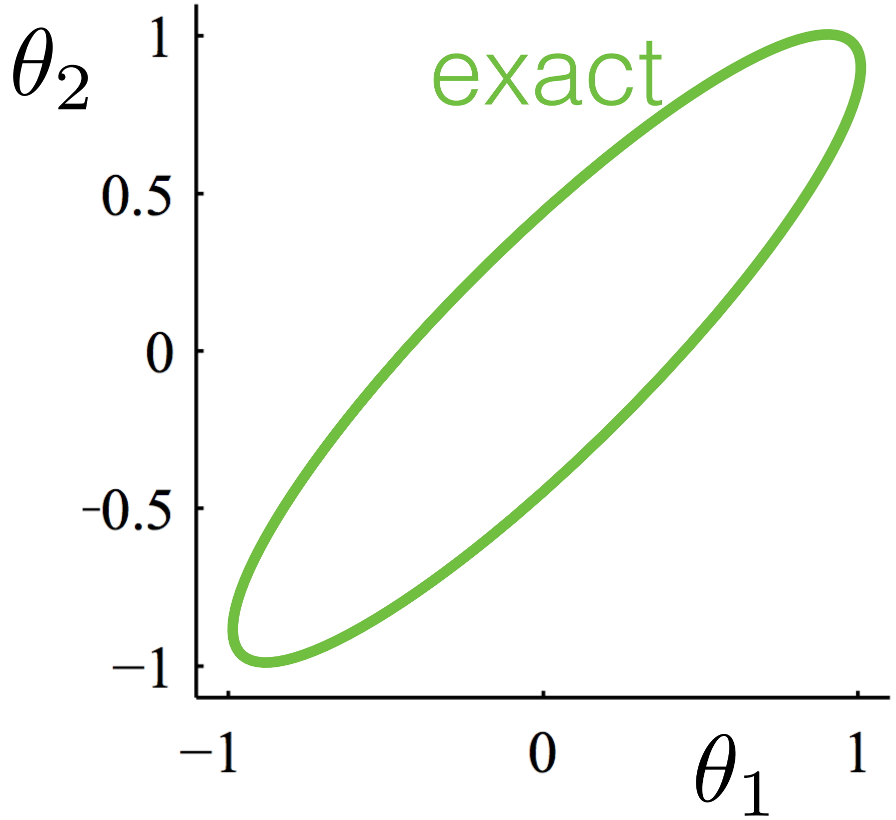


[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

What about uncertainty?

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

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



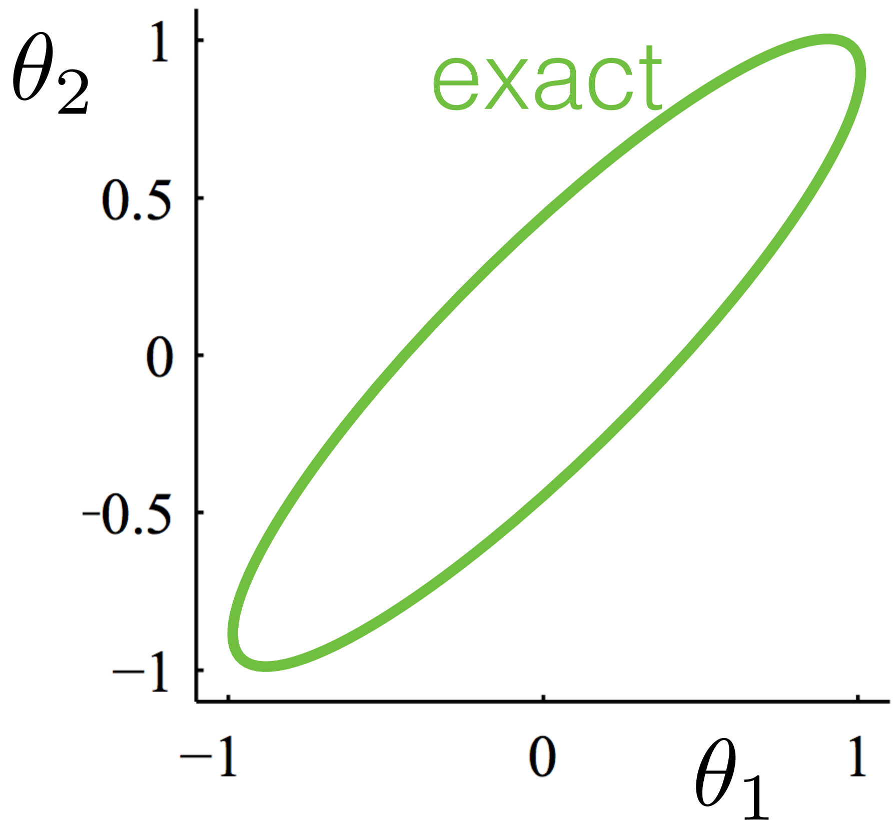
- Conjugate linear regression

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

What about uncertainty?

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

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

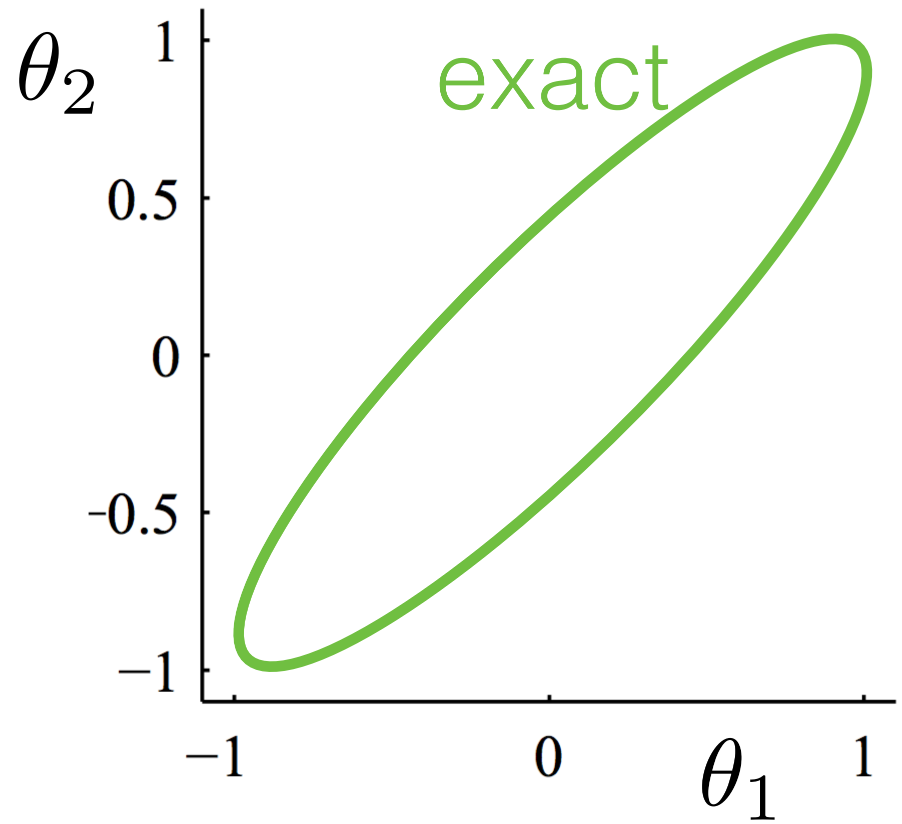


[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

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



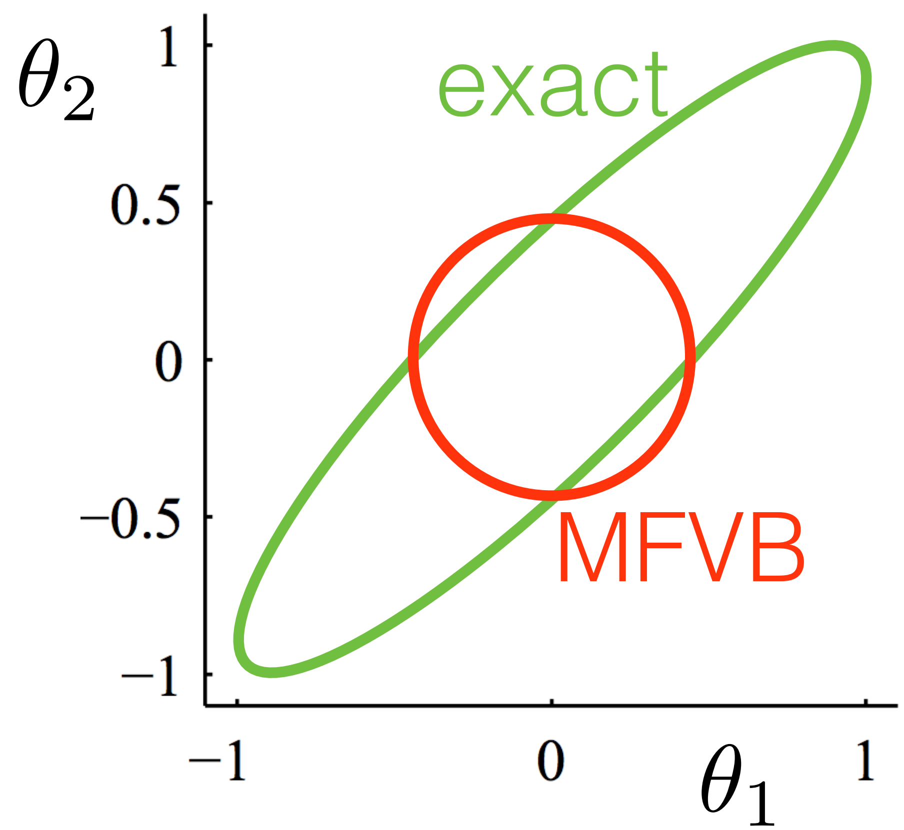
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

What about uncertainty?

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



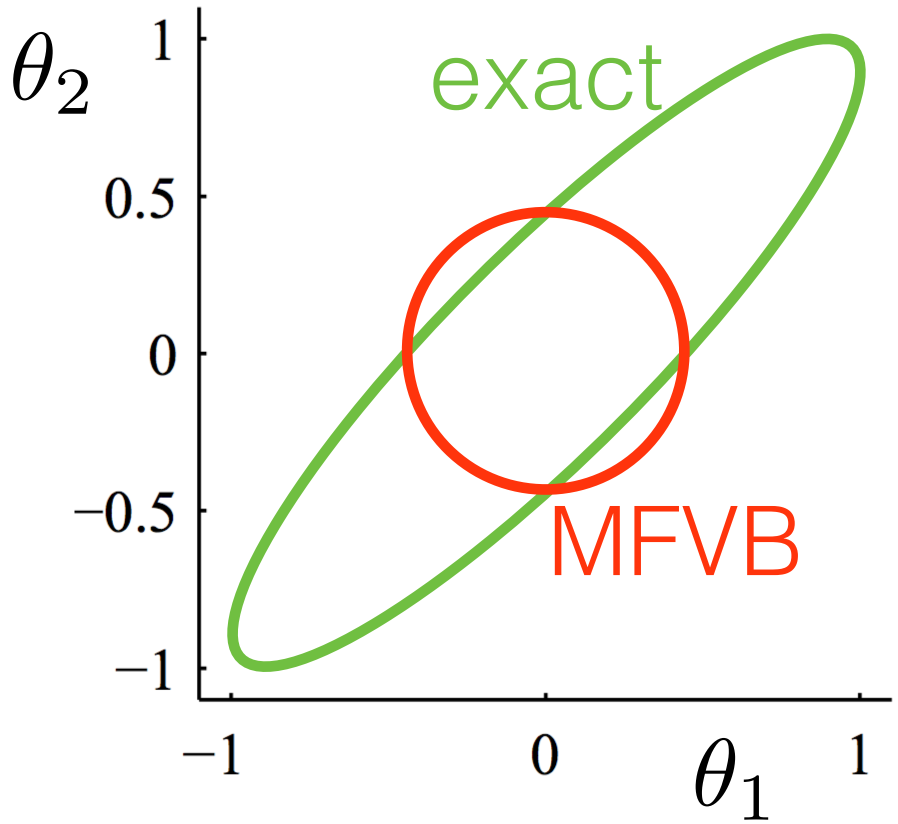
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

What about uncertainty?

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



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

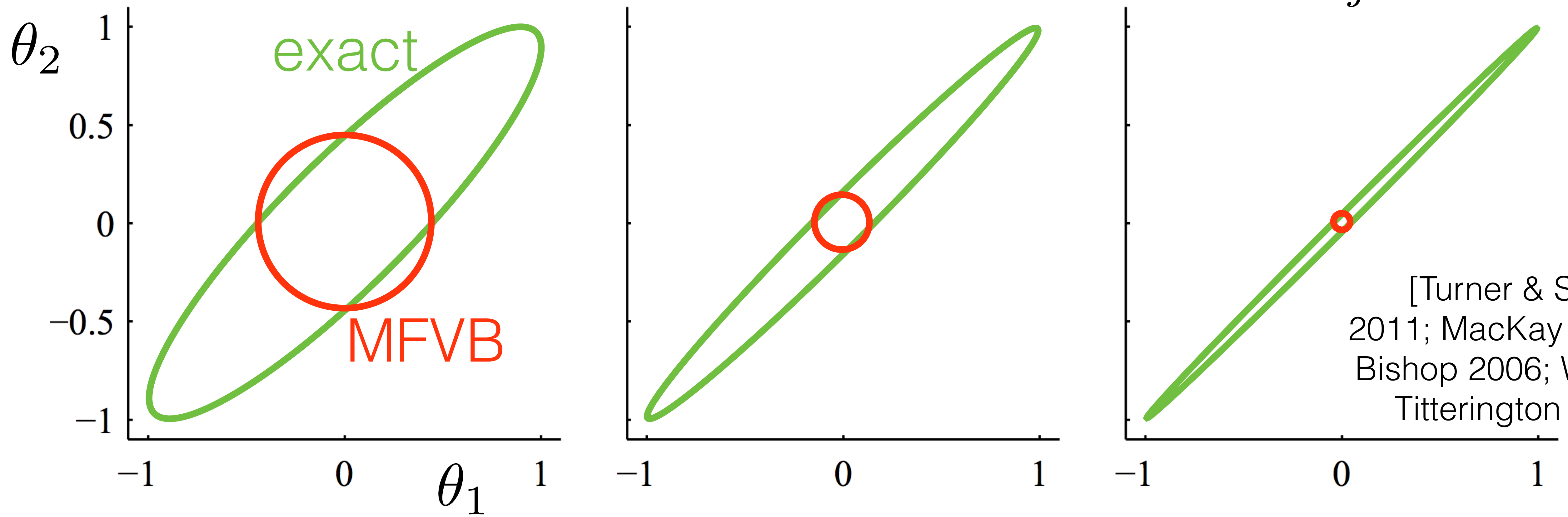
- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

- Misestimates variance (sometimes severely)

What about uncertainty?

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



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

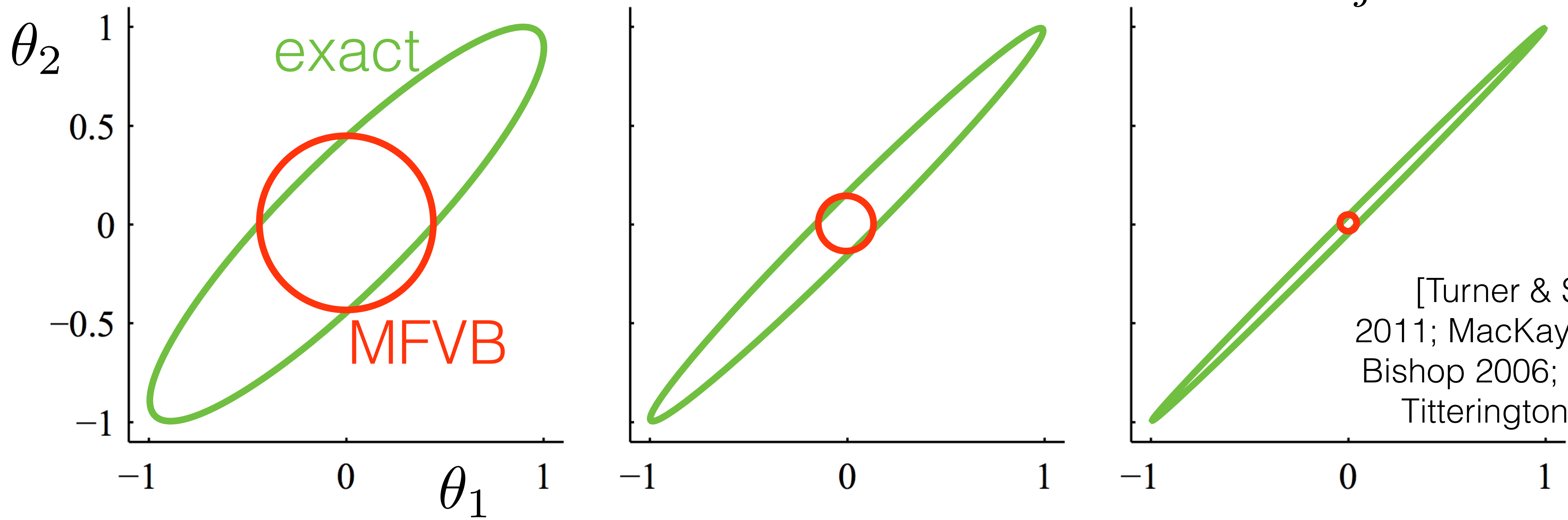
- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

- Misestimates variance (sometimes severely)

What about uncertainty?

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



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

- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

- Misestimates variance (sometimes severely)
- No covariance estimates

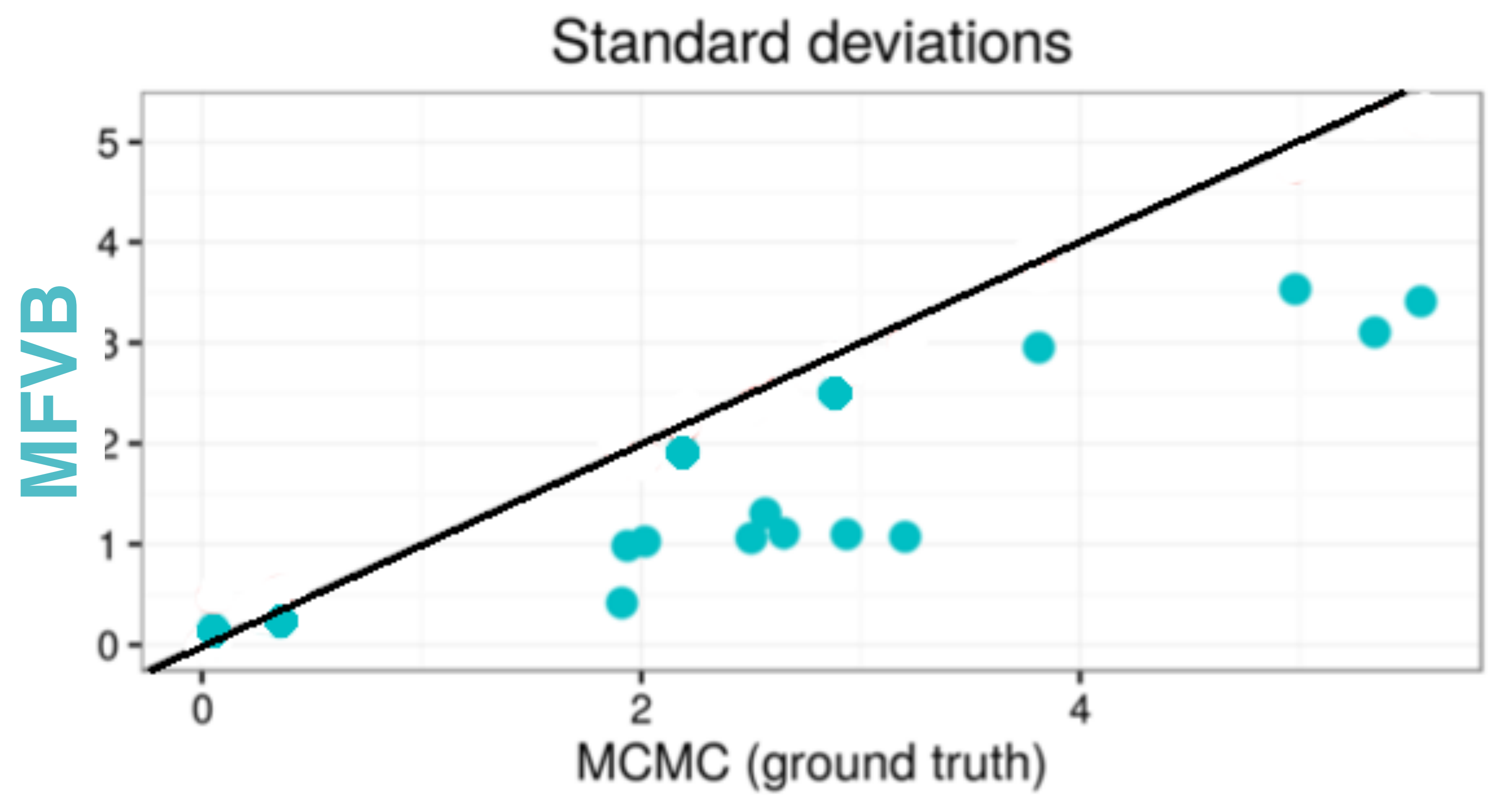
What about uncertainty?

- Microcredit

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

What about uncertainty?

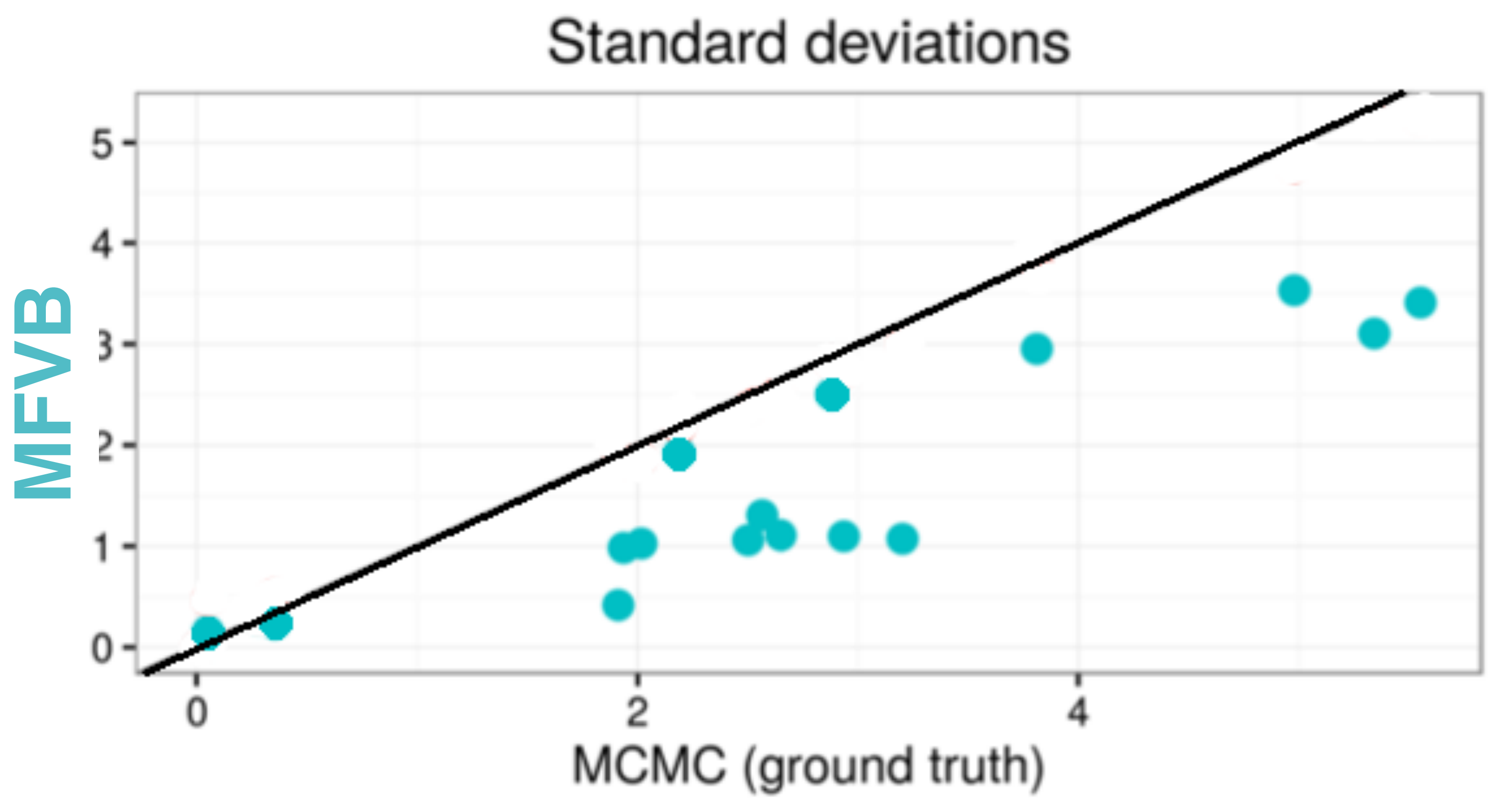
- Microcredit



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

What about uncertainty?

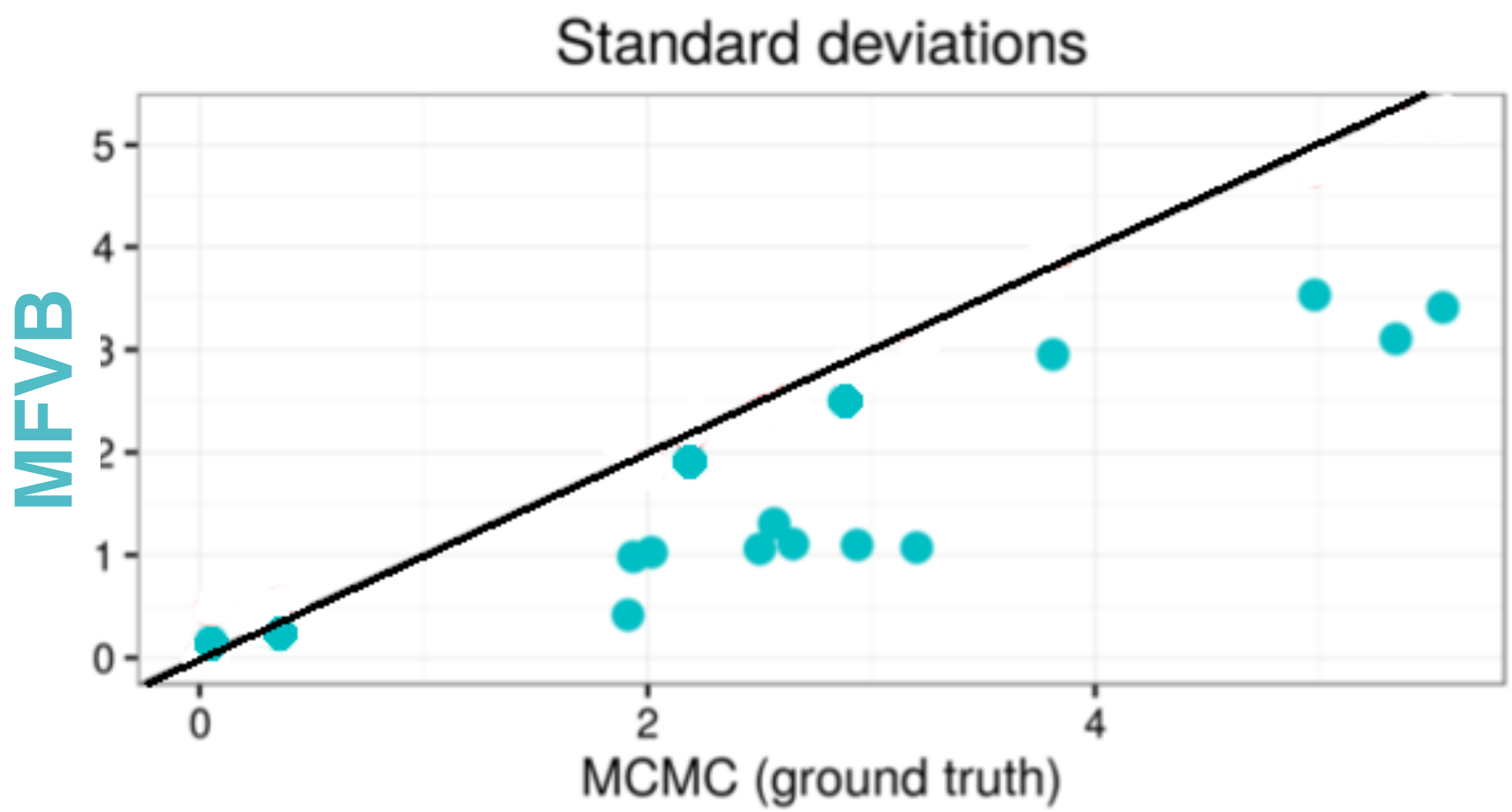
- Microcredit effect
- τ mean: 3.08 USD PPP



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

What about uncertainty?

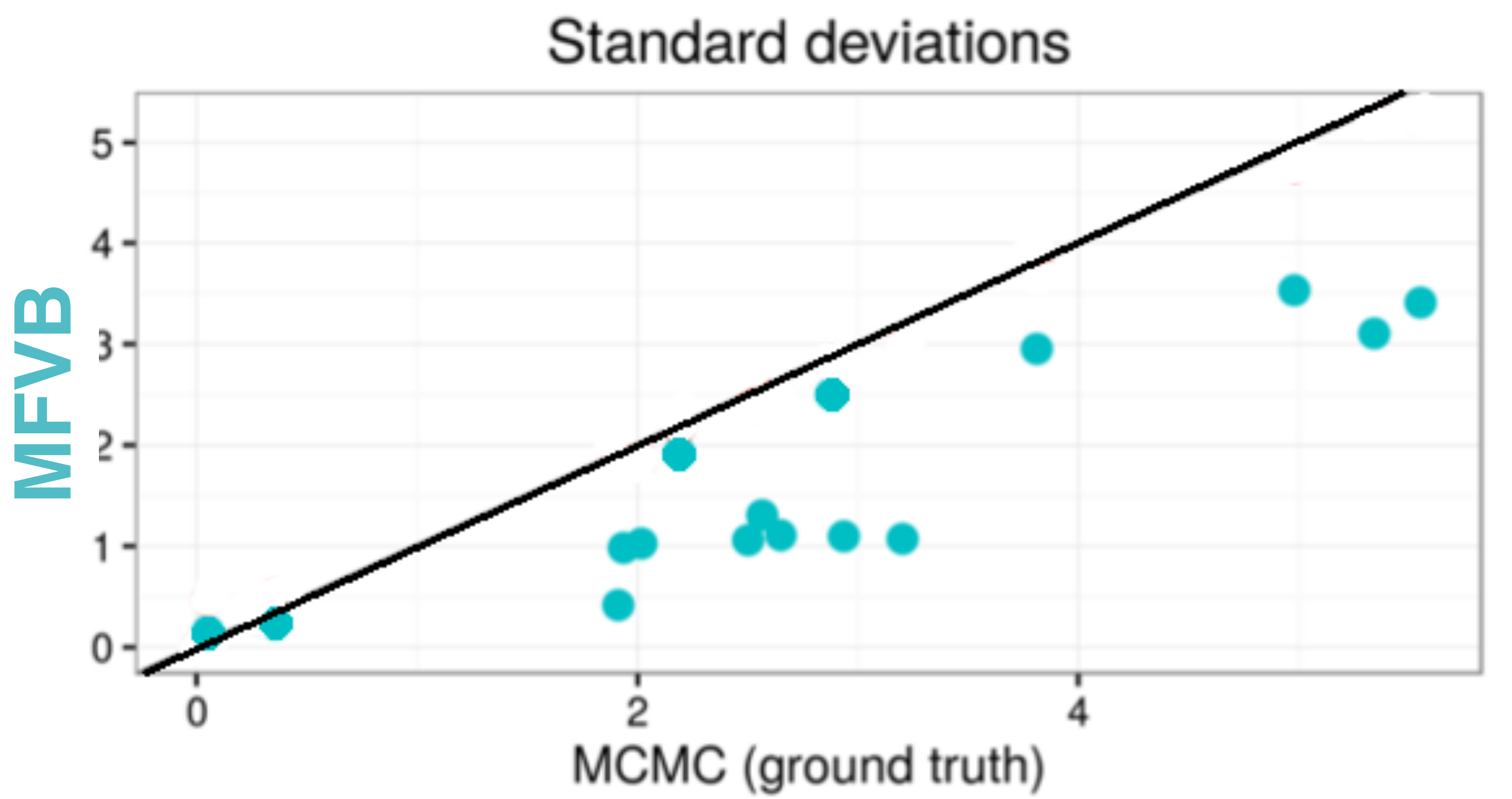
- Microcredit effect
- τ mean: 3.08 USD PPP
- τ std dev: 1.83 USD PPP



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

What about uncertainty?

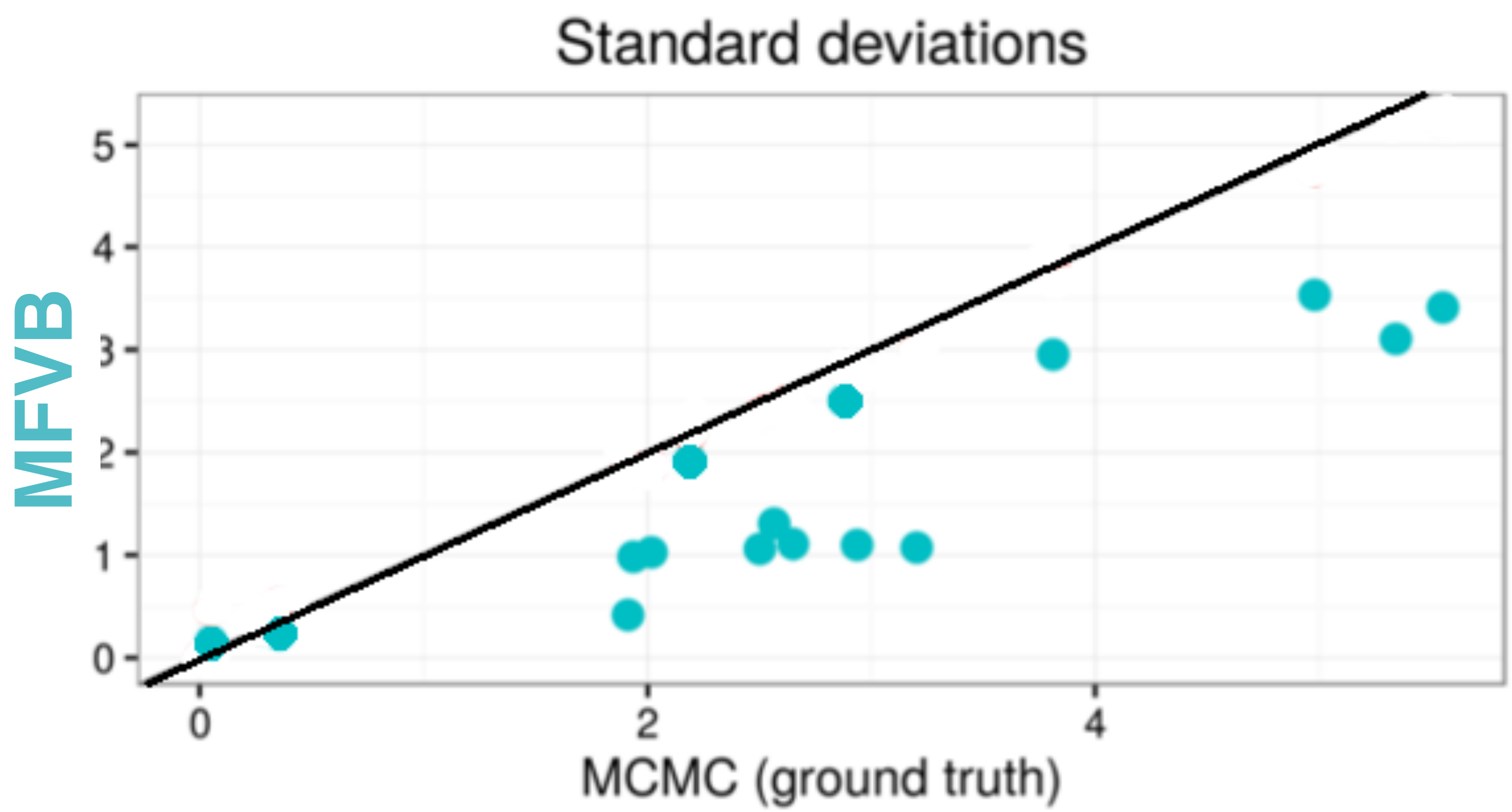
- Microcredit effect
- τ mean: 3.08 USD PPP
- τ std dev: 1.83 USD PPP
- Mean is 1.68 std dev from 0



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

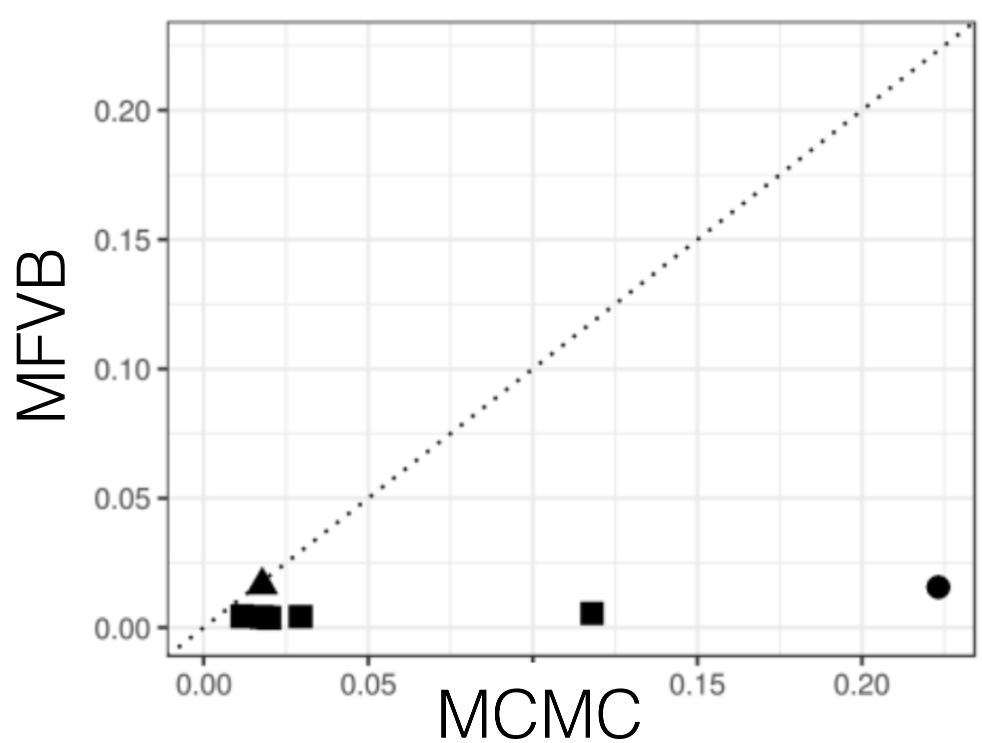
What about uncertainty?

- 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 (global parameters)

Posterior std dev estimates



[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2018]

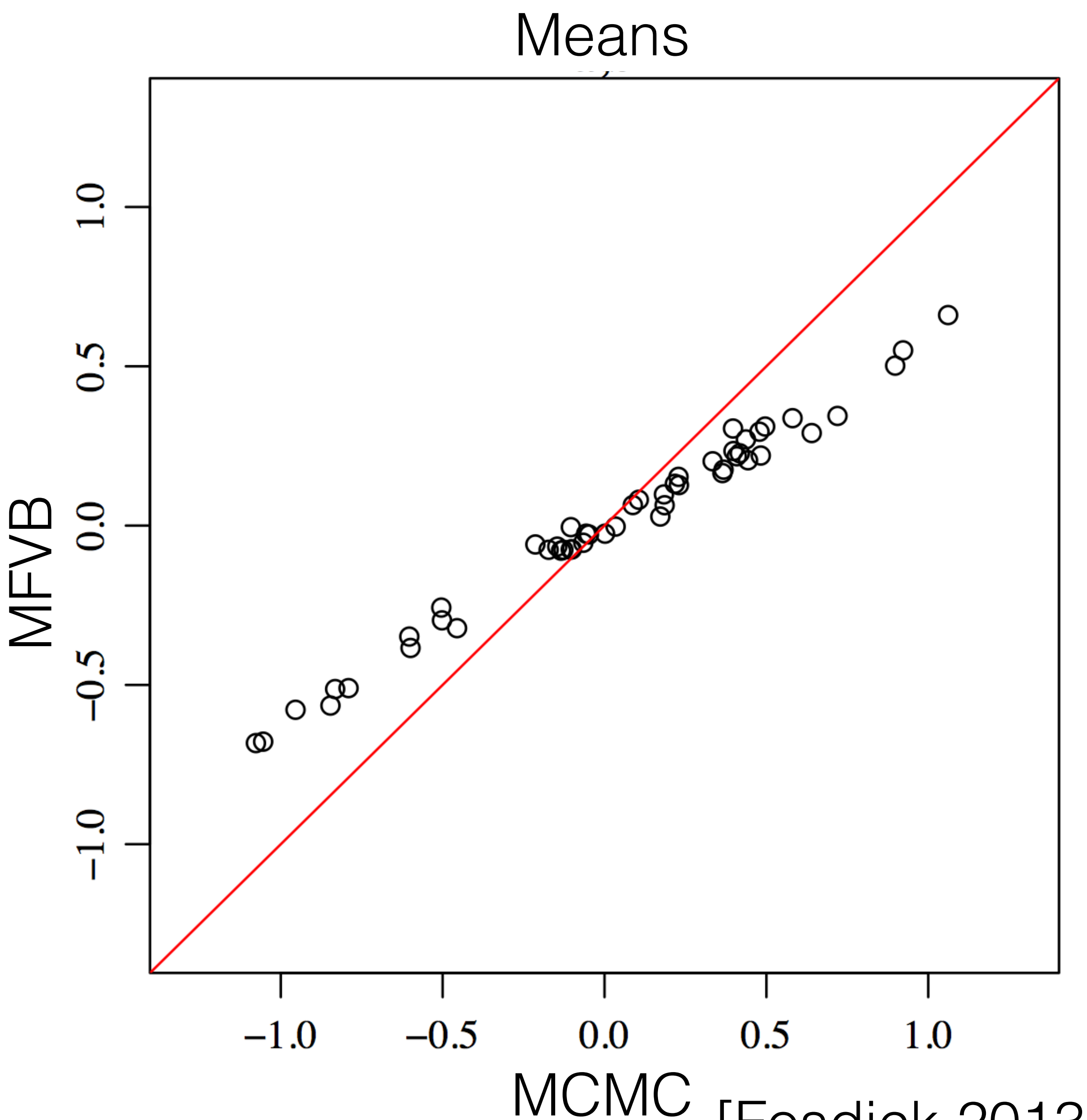
What about means?

What about means?

- Model for relational data with covariates
- When 1000+ nodes, MCMC > 1 day [Fosdick 2013, Ch 4]

What about means?

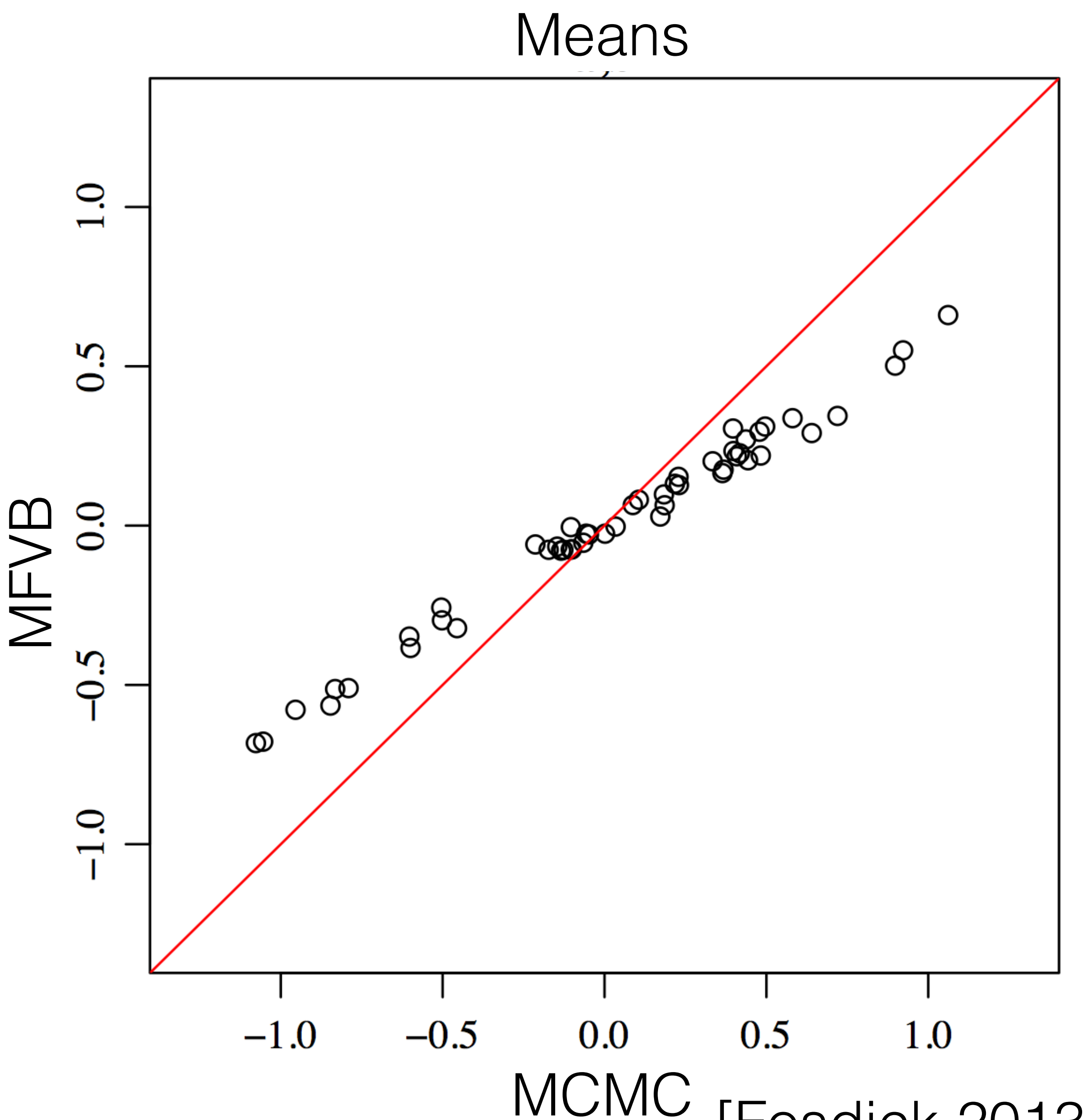
- Model for relational data with covariates
- When 1000+ nodes, MCMC > 1 day [Fosdick 2013, Ch 4]

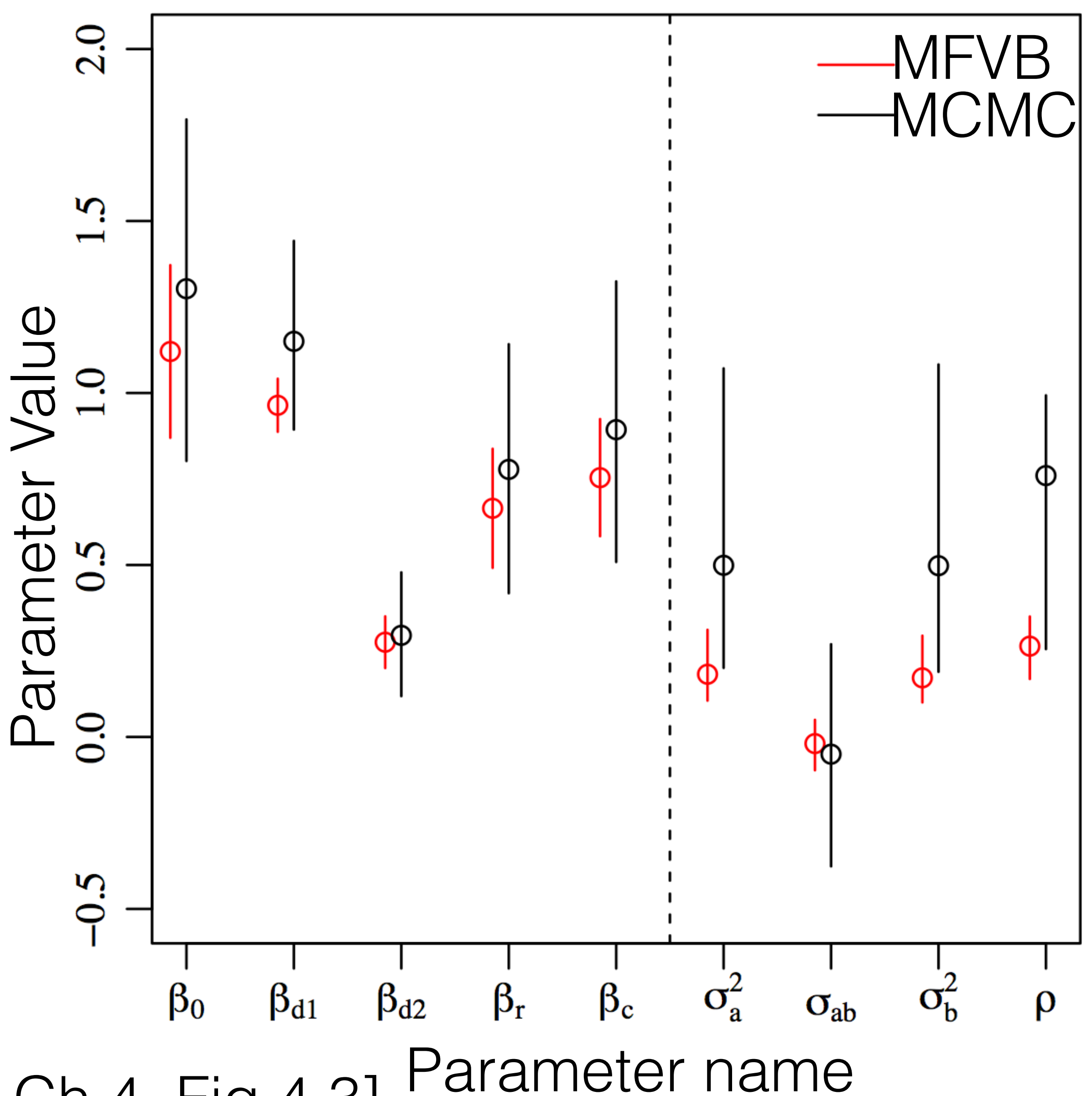


[Fosdick 2013, Ch 4, Fig 4.3]

What about means?

- Model for relational data with covariates
- When 1000+ nodes, MCMC > 1 day [Fosdick 2013, Ch 4]

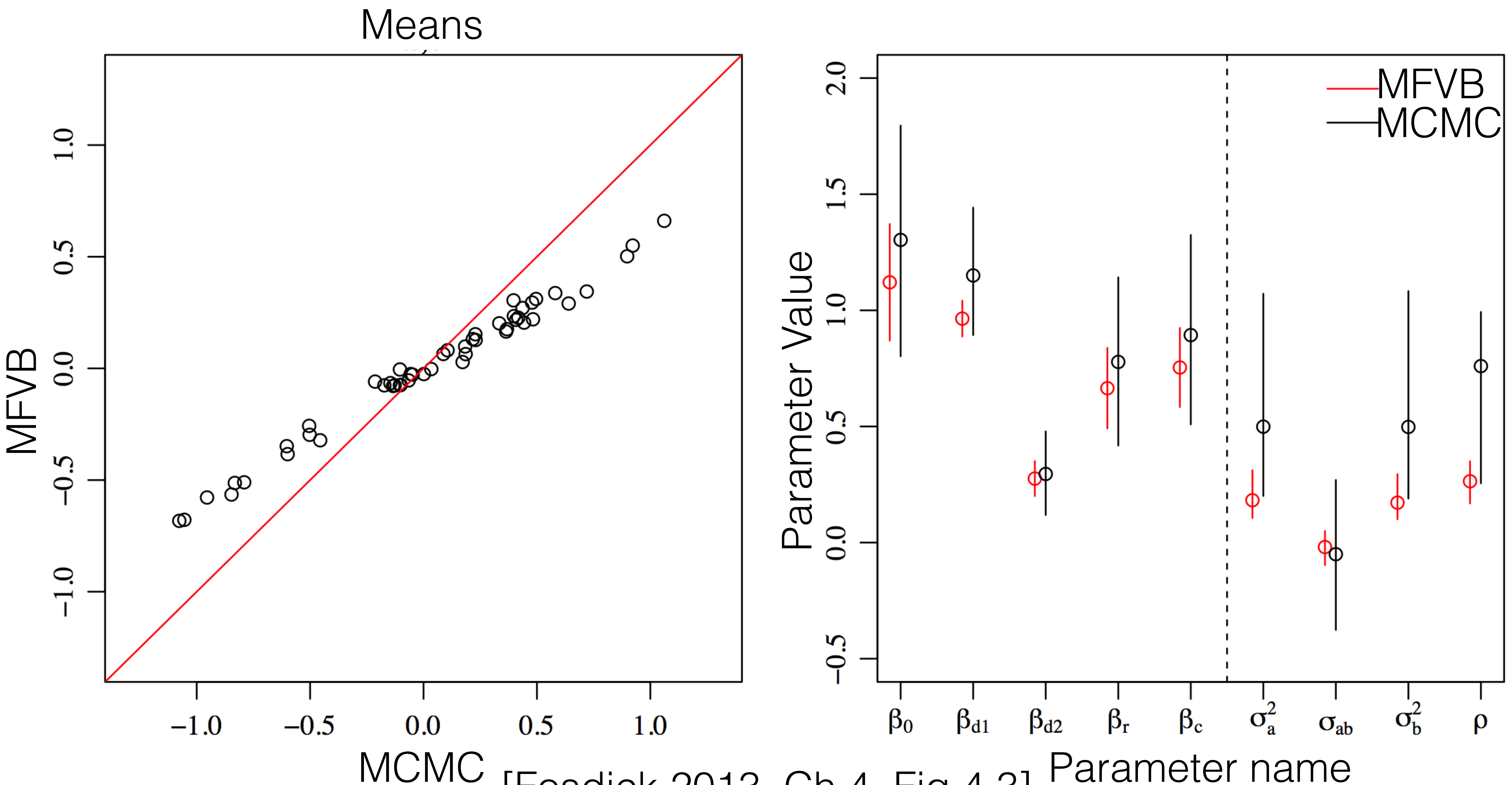




[Fosdick 2013, Ch 4, Fig 4.3]

What about means?

- Model for relational data with covariates
- When 1000+ nodes, MCMC > 1 day [Fosdick 2013, Ch 4]

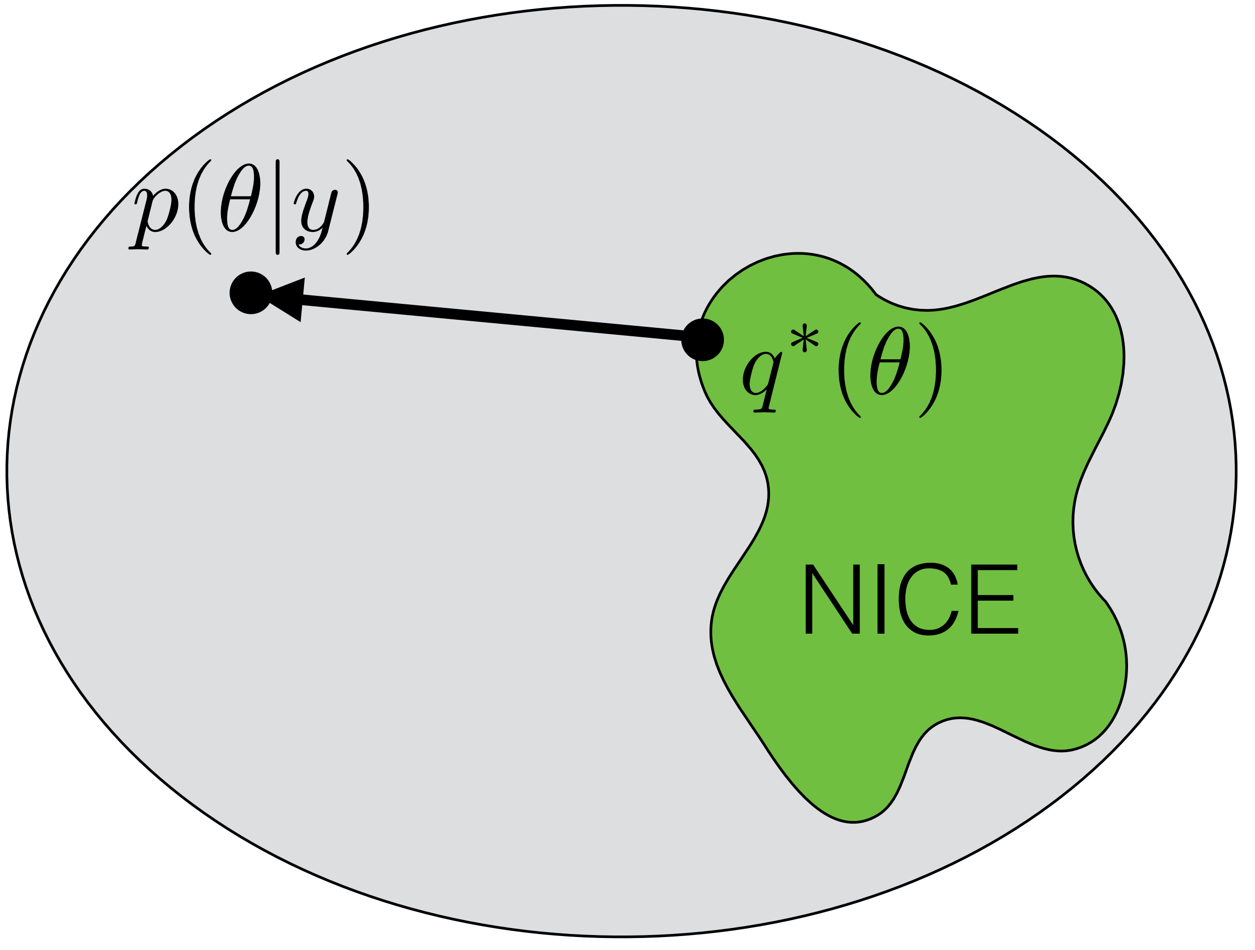


[Fosdick 2013, Ch 4, Fig 4.3]

- Can simulate data repeatedly from one model; sometimes estimates are good and sometimes not

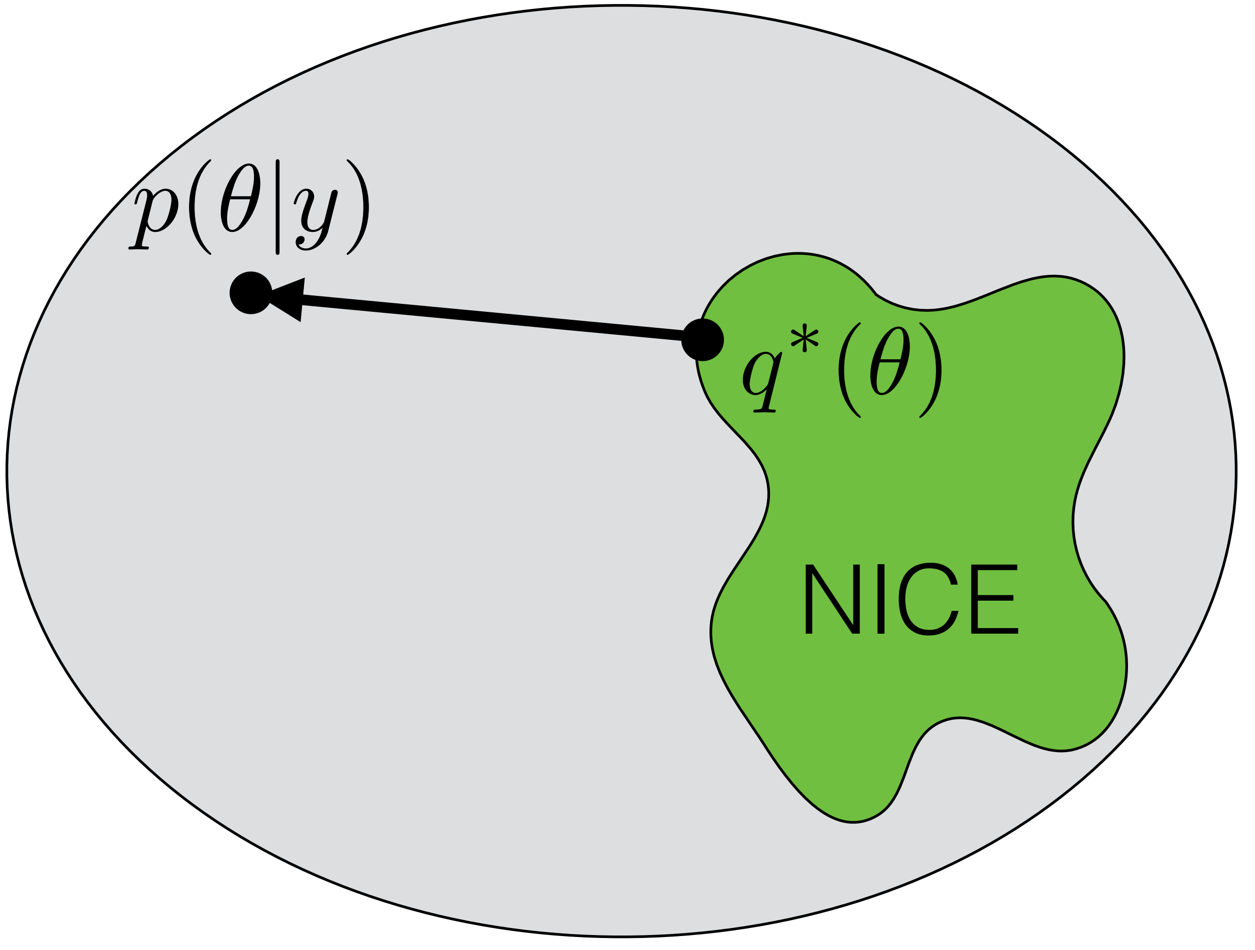
[Giordano, Broderick, Jordan 2015]

Can we fix the estimation problems?



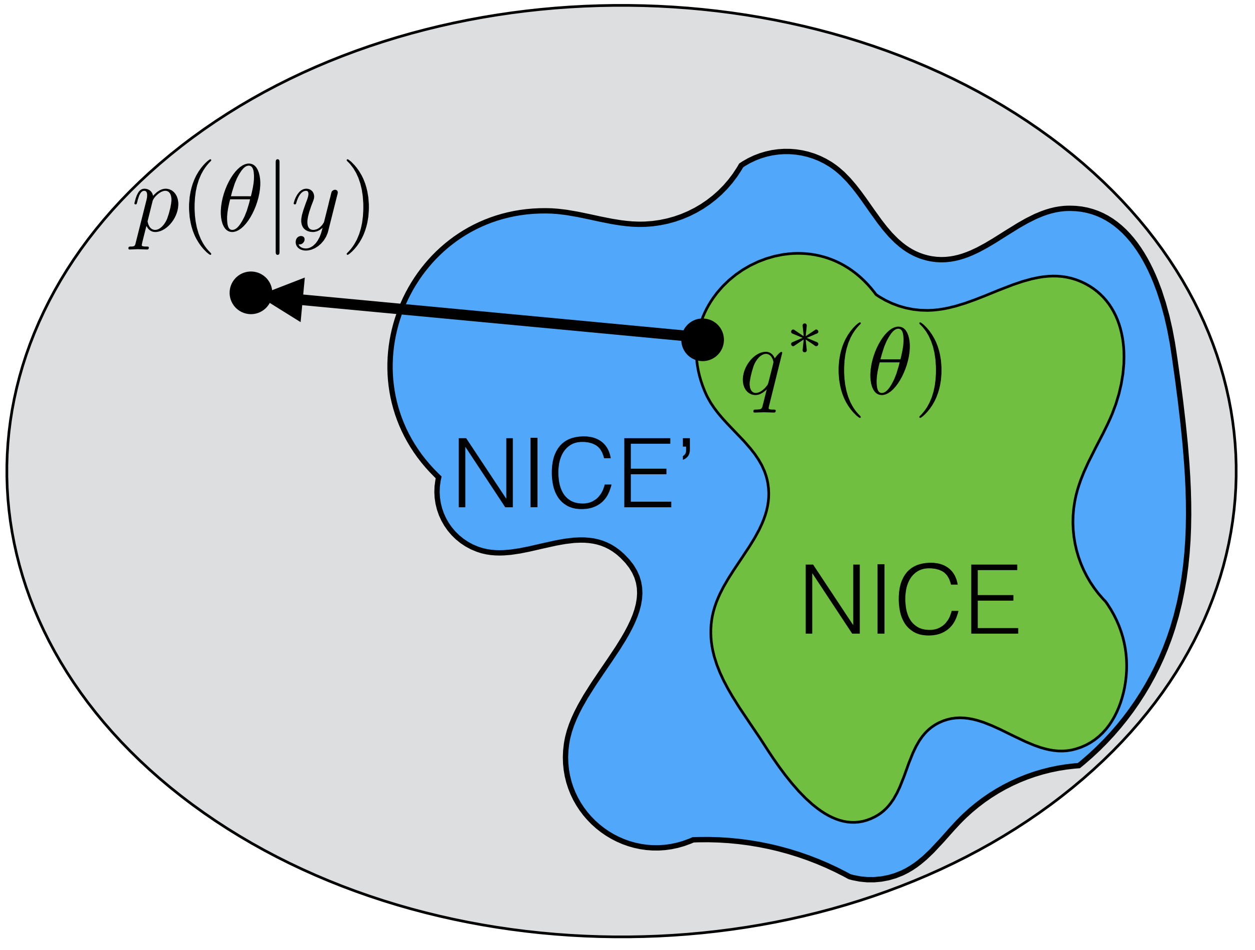
Can we fix the estimation problems?

- Common idea: use a richer set of NICE distributions



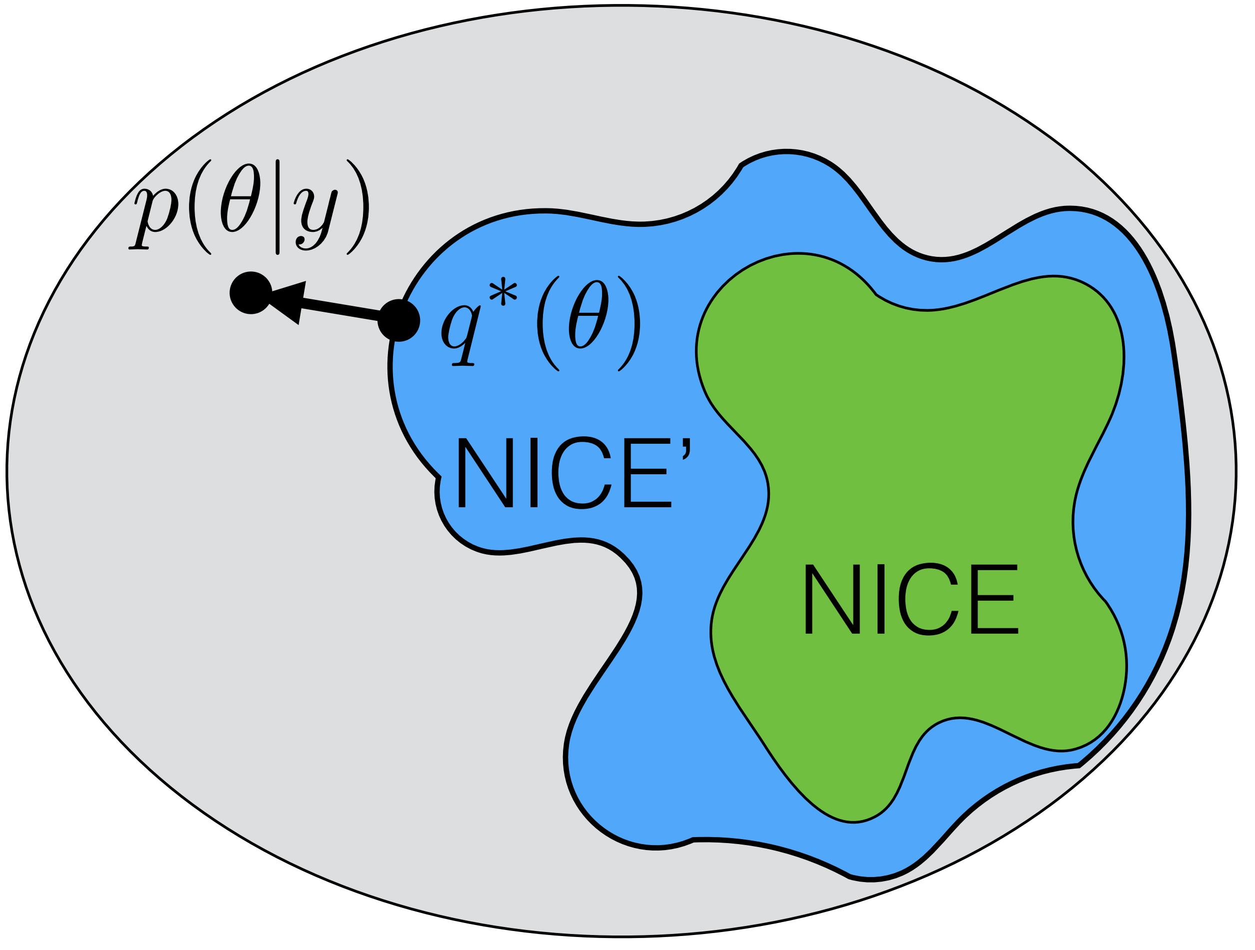
Can we fix the estimation problems?

- Common idea: use a richer set of NICE distributions



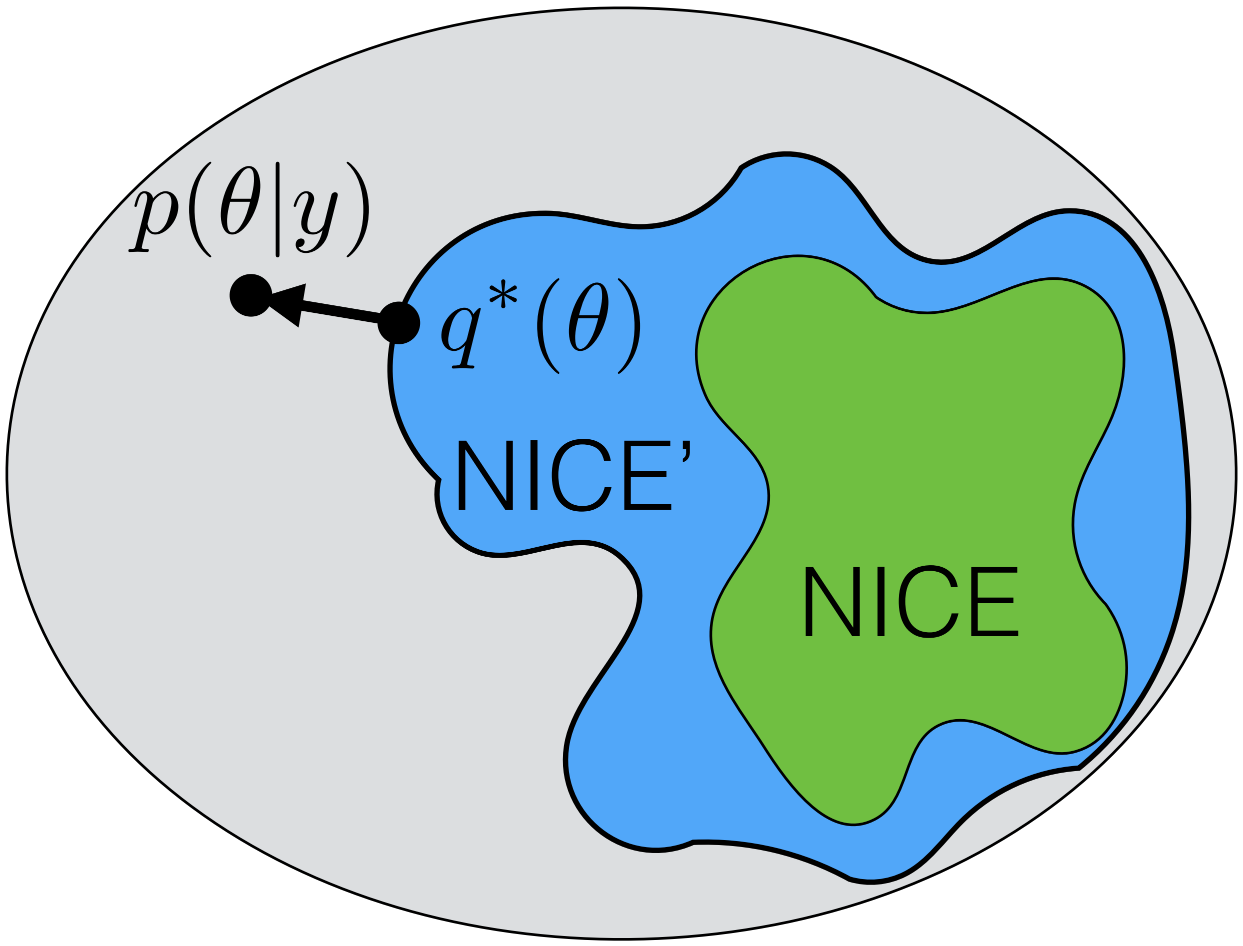
Can we fix the estimation problems?

- Common idea: use a richer set of NICE distributions



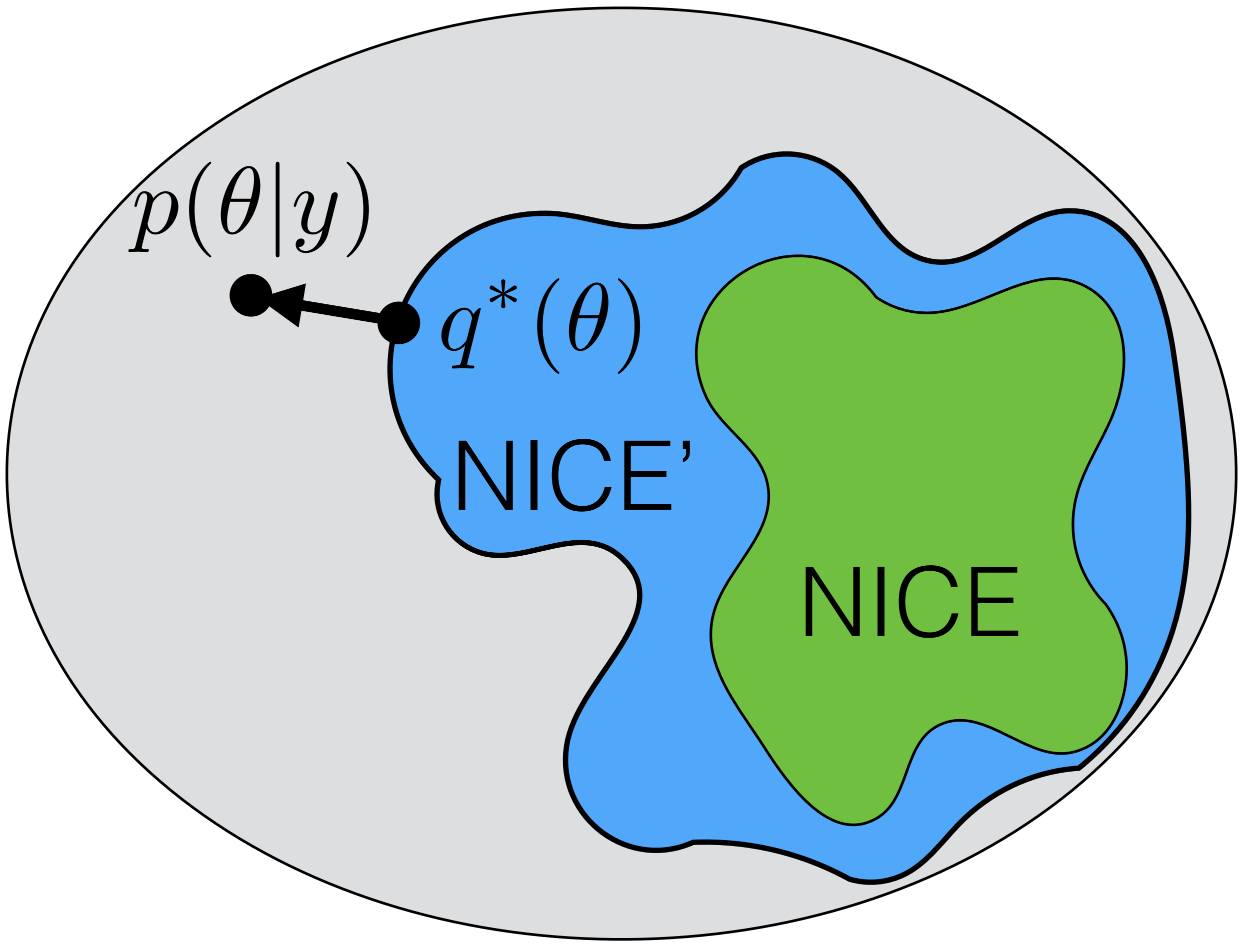
Can we fix the estimation problems?

- Common idea: use a richer set of NICE distributions
  - Often prohibitive computational expense



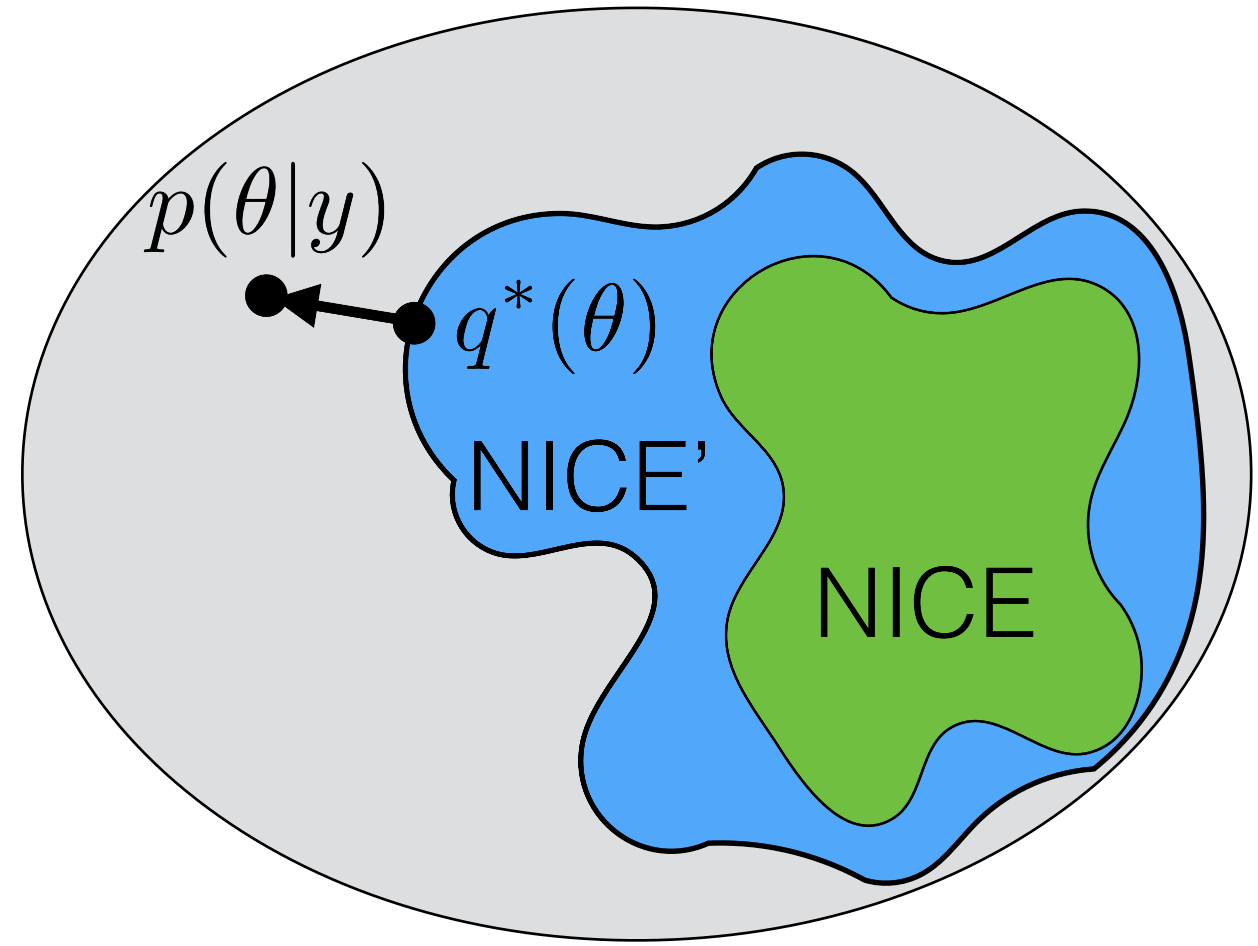
Can we fix the estimation problems?

- Common idea: use a richer set of NICE distributions
  - Often prohibitive computational expense
  - No guaranteed win

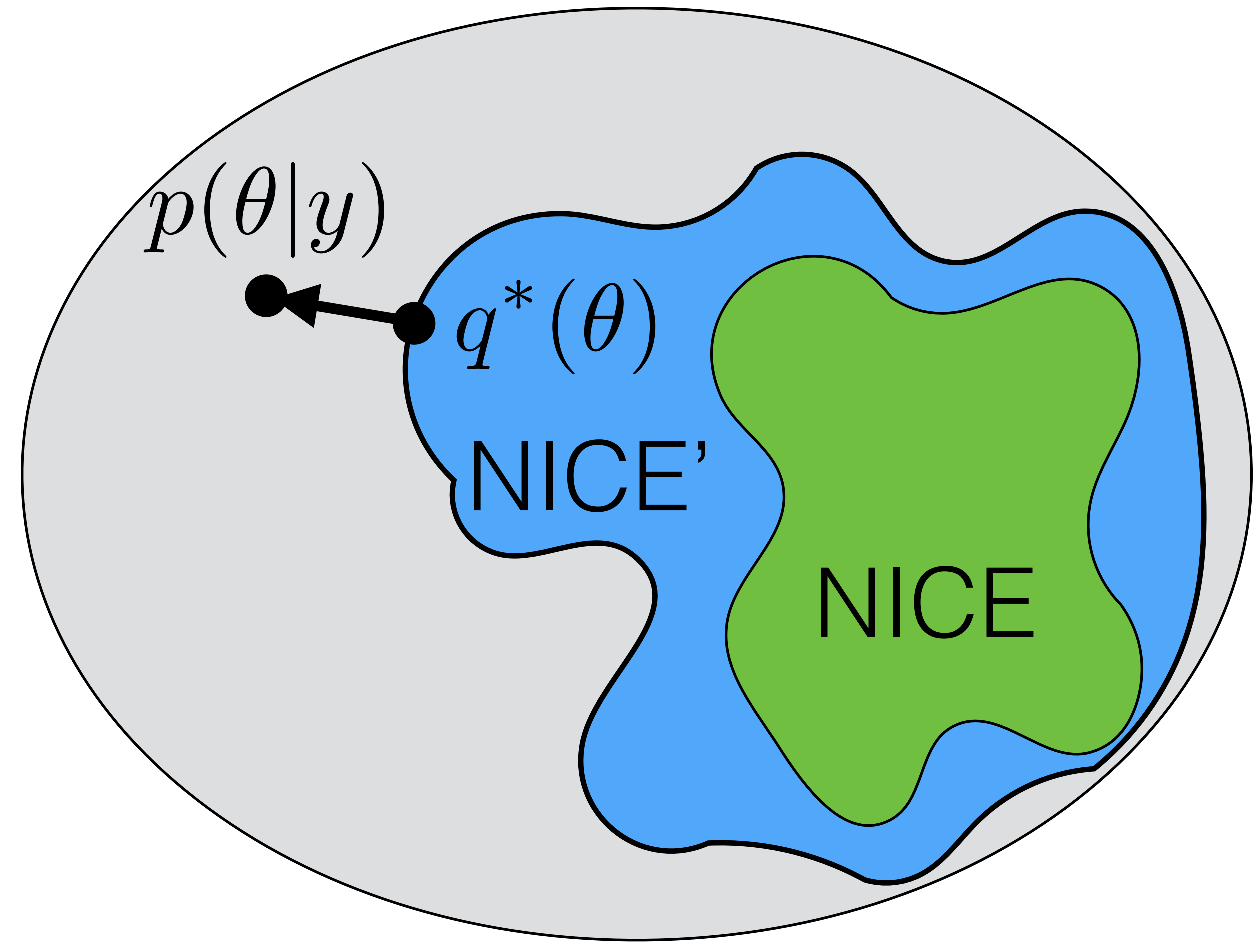


Can we fix the estimation problems?

- Common idea: use a richer set of NICE distributions
  - Often prohibitive computational expense
  - No guaranteed win
    - Can have larger NICE set but worse mean & variance estimates [Turner, Sahani 2011]



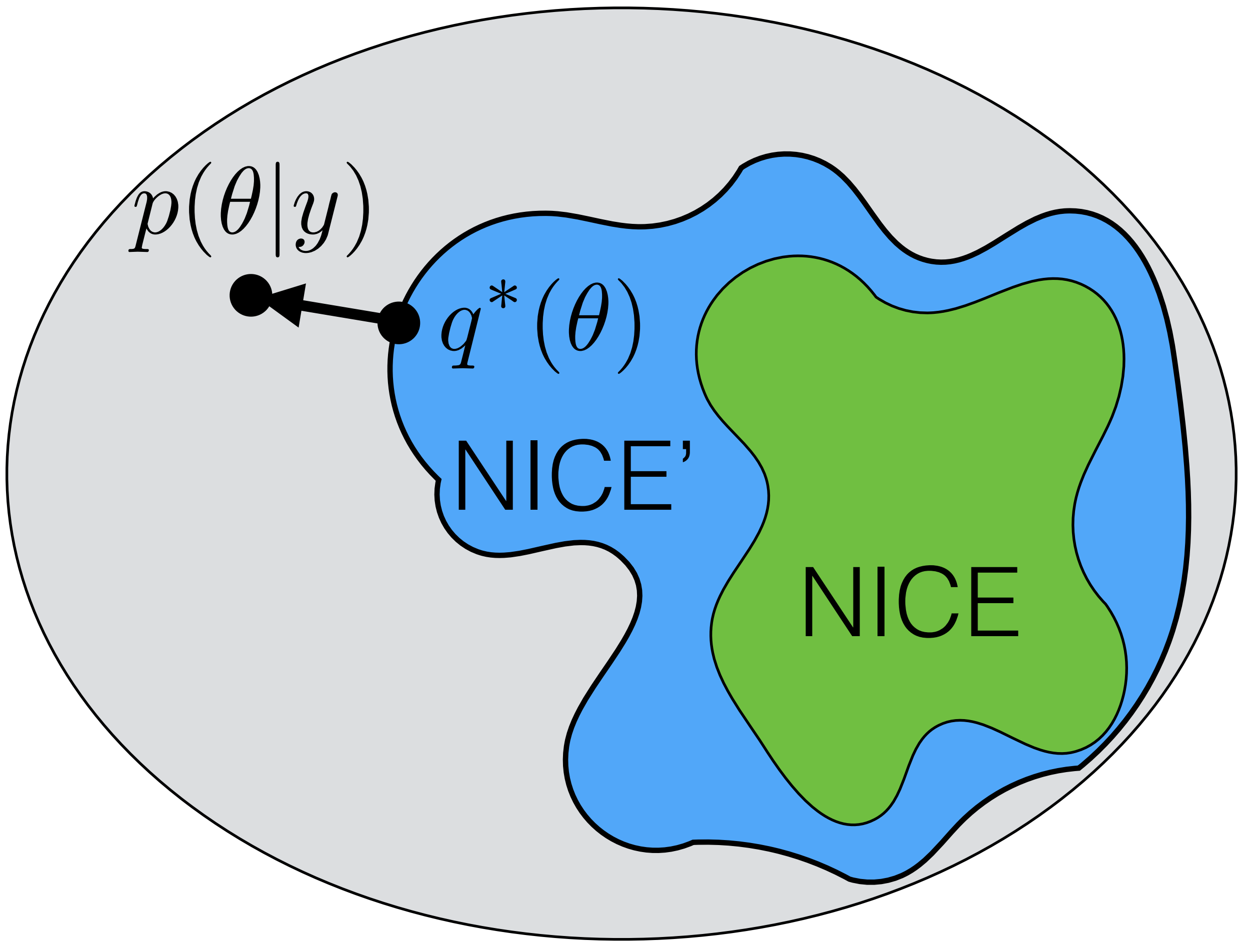
Can we fix the estimation problems?



- Common idea: use a richer set of NICE distributions
  - Often prohibitive computational expense
  - No guaranteed win
    - Can have larger NICE set but worse mean & variance estimates [Turner, Sahani 2011]
    - Even with unimodal 1D distributions, can have small KL but arbitrarily large mean or variance differences [Huggins, Karsprzak, Campbell, Broderick 2020]

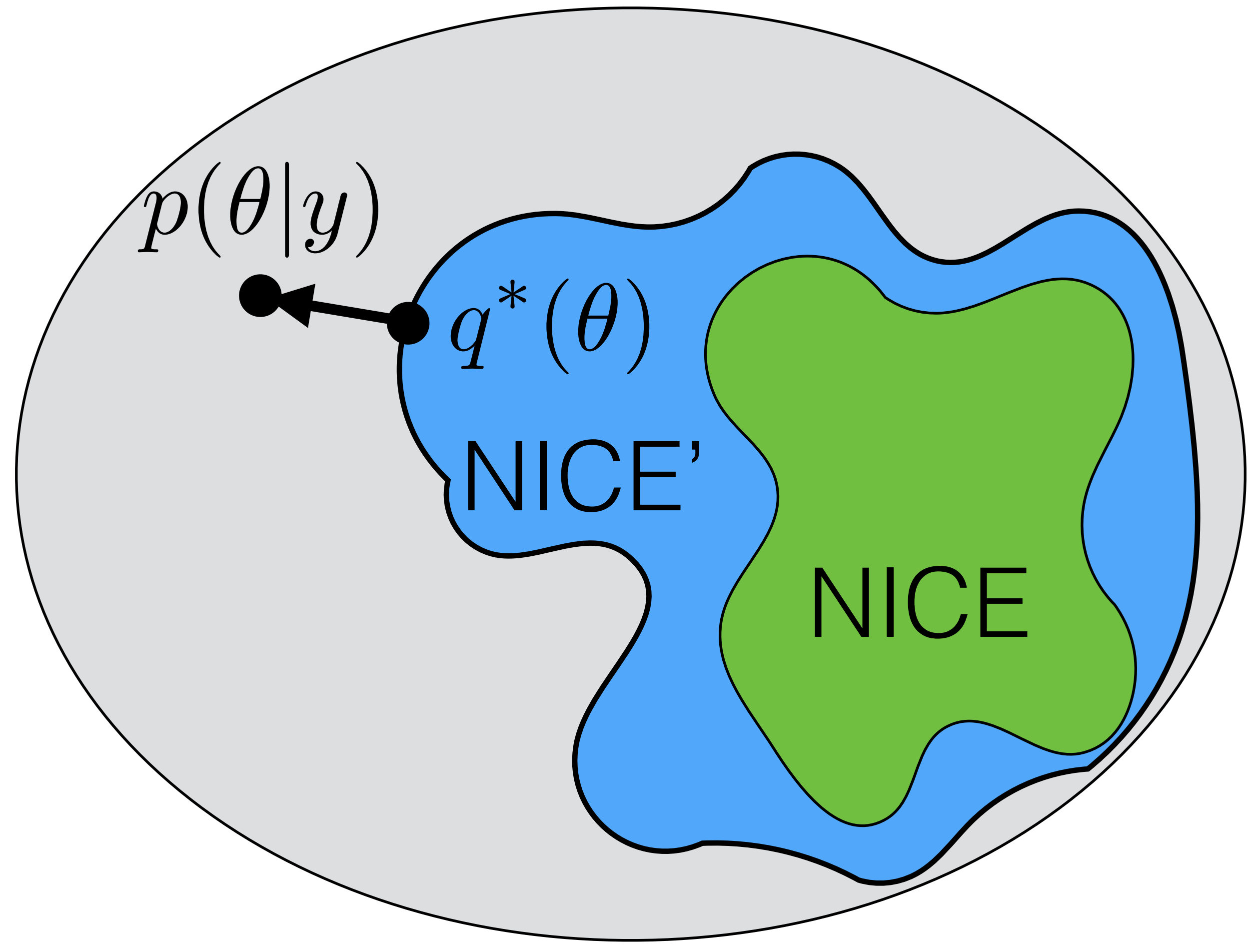
Can we fix the estimation problems?

- Common idea: use a richer set of NICE distributions
  - Often prohibitive computational expense
  - No guaranteed win
    - Can have larger NICE set but worse mean & variance estimates [Turner, Sahani 2011]



    - Even with unimodal 1D distributions, can have small KL but arbitrarily large mean or variance differences [Huggins, Karsprzak, Campbell, Broderick 2020]
  - Might still provide accuracy improvements "most" of the time in practice; see a systematic comparison shortly

Can we fix the estimation problems?



- Common idea: use a richer set of NICE distributions
  - Often prohibitive computational expense
  - No guaranteed win
    - Can have larger NICE set but worse mean & variance estimates [Turner, Sahani 2011]
    - Even with unimodal 1D distributions, can have small KL but arbitrarily large mean or variance differences [Huggins, Karsprzak, Campbell, Broderick 2020]
  - Might still provide accuracy improvements "most" of the time in practice; see a systematic comparison shortly
- Analogous challenges with changing the divergence

VB variance improvement

[Giordano, Broderick, Jordan 2015, 2018]

VB variance improvement

- Linear response variational Bayes (LRVB): A post hoc correction to a VB posterior covariance estimate

[Giordano, Broderick, Jordan 2015, 2018]

VB variance improvement

- Linear response variational Bayes (LRVB): A post hoc correction to a VB posterior covariance estimate
  - Works by applying a first-order Taylor expansion at the VB optimum (so requires being at the optimum)

[Giordano, Broderick, Jordan 2015, 2018]

VB variance improvement

- Linear response variational Bayes (LRVB): A post hoc correction to a VB posterior covariance estimate
  - Works by applying a first-order Taylor expansion at the VB optimum (so requires being at the optimum)
  - Exact when the posterior mean estimate is exact

[Giordano, Broderick, Jordan 2015, 2018]

VB variance improvement

- Linear response variational Bayes (LRVB): A post hoc correction to a VB posterior covariance estimate
  - Works by applying a first-order Taylor expansion at the VB optimum (so requires being at the optimum)
  - Exact when the posterior mean estimate is exact

  - Microcredit

[Giordano, Broderick, Jordan 2015, 2018]

VB variance improvement

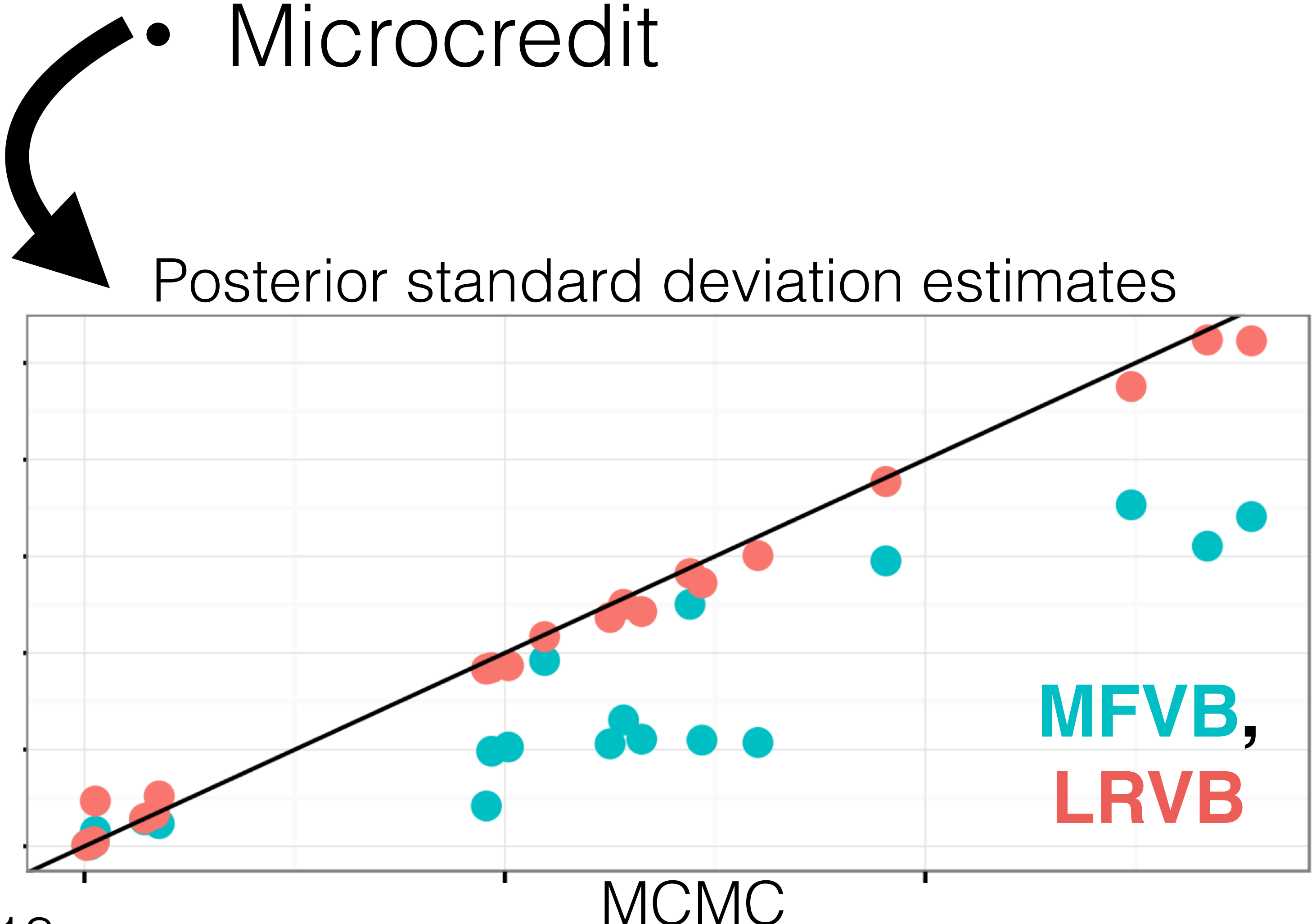
- Linear response variational Bayes (LRVB): A post hoc correction to a VB posterior covariance estimate
  - Works by applying a first-order Taylor expansion at the VB optimum (so requires being at the optimum)
  - Exact when the posterior mean estimate is exact

- Microcredit

[Giordano, Broderick, Jordan 2015, 2018]

VB variance improvement

- Linear response variational Bayes (LRVB): A post hoc correction to a VB posterior covariance estimate
  - Works by applying a first-order Taylor expansion at the VB optimum (so requires being at the optimum)
  - Exact when the posterior mean estimate is exact
  - Microcredit



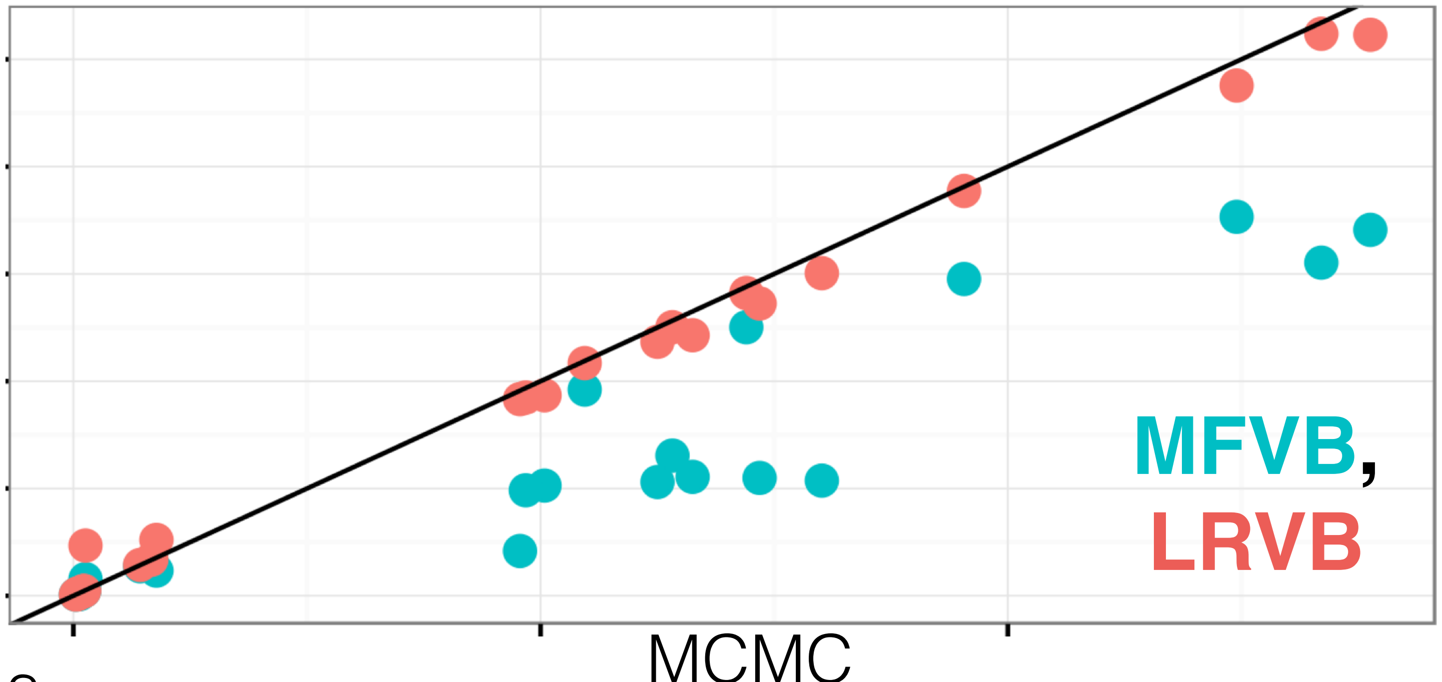
[Giordano, Broderick, Jordan 2015, 2018]

VB variance improvement

- Linear response variational Bayes (LRVB): A post hoc correction to a VB posterior covariance estimate
  - Works by applying a first-order Taylor expansion at the VB optimum (so requires being at the optimum)
  - Exact when the posterior mean estimate is exact

- Microcredit
- Criteo ads

Posterior standard deviation estimates



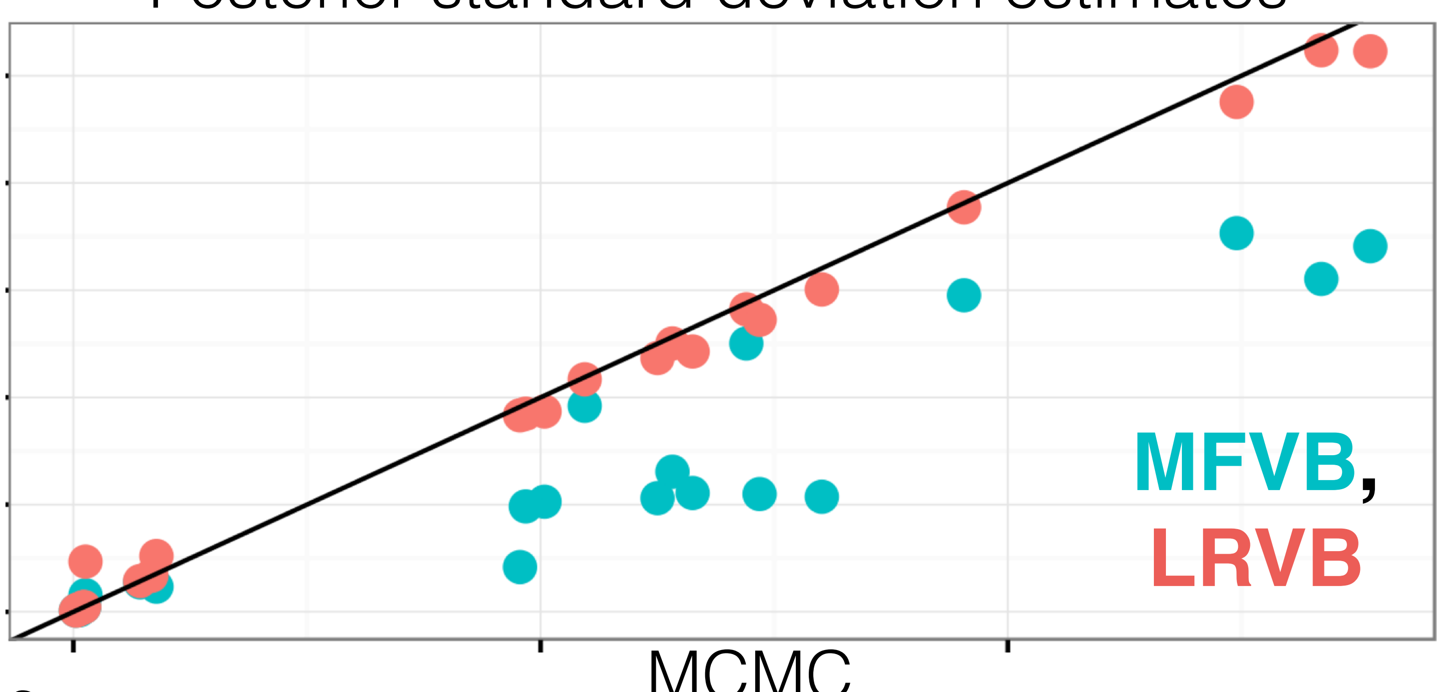
[Giordano, Broderick, Jordan 2015, 2018]

VB variance improvement

- Linear response variational Bayes (LRVB): A post hoc correction to a VB posterior covariance estimate
  - Works by applying a first-order Taylor expansion at the VB optimum (so requires being at the optimum)
  - Exact when the posterior mean estimate is exact

- Microcredit
- Criteo ads

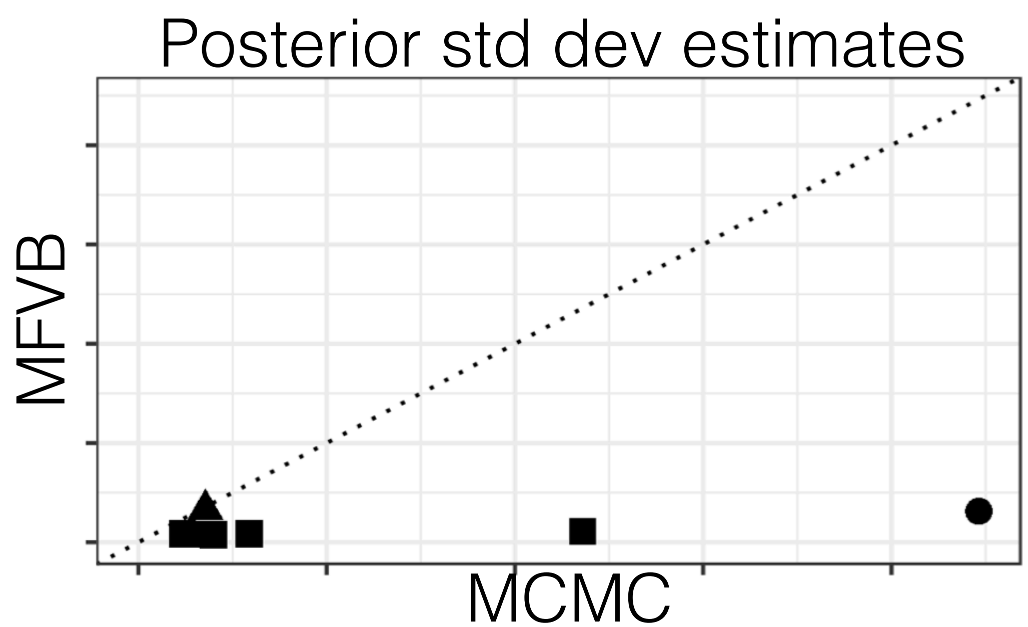
Posterior standard deviation estimates



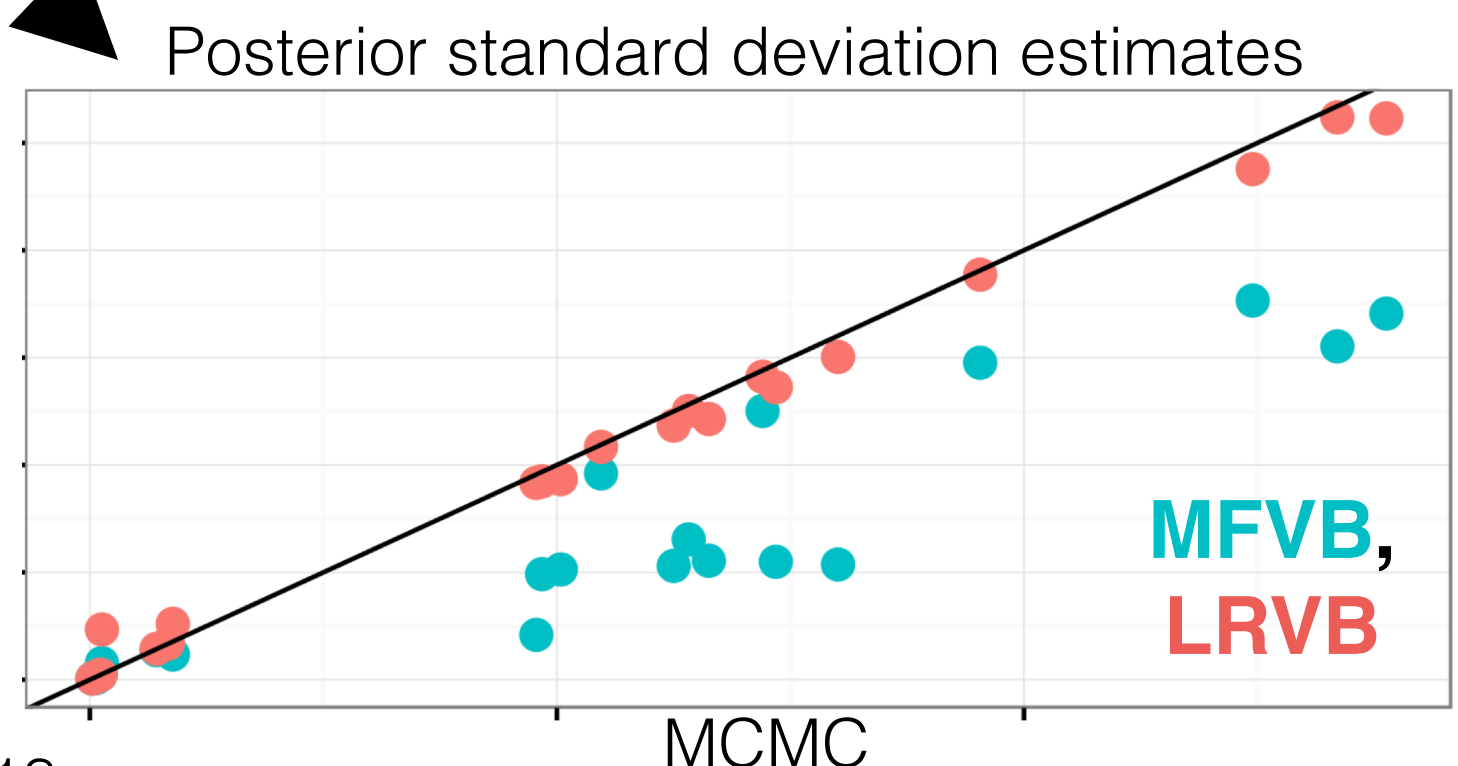
[Giordano, Broderick, Jordan 2015, 2018]

VB variance improvement

- Linear response variational Bayes (LRVB): A post hoc correction to a VB posterior covariance estimate
  - Works by applying a first-order Taylor expansion at the VB optimum (so requires being at the optimum)
  - Exact when the posterior mean estimate is exact



- Microcredit
- Criteo ads



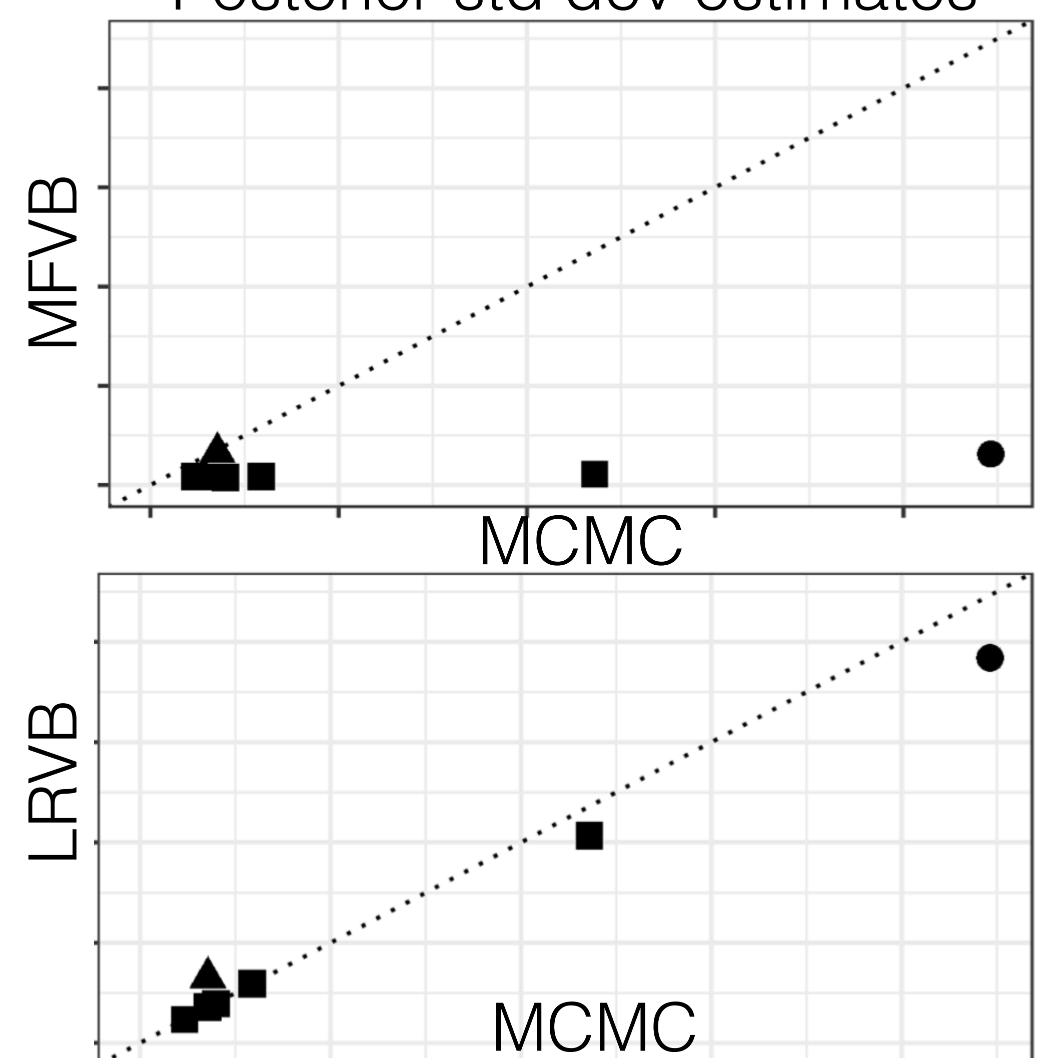
[Giordano, Broderick, Jordan 2015, 2018]

VB variance improvement

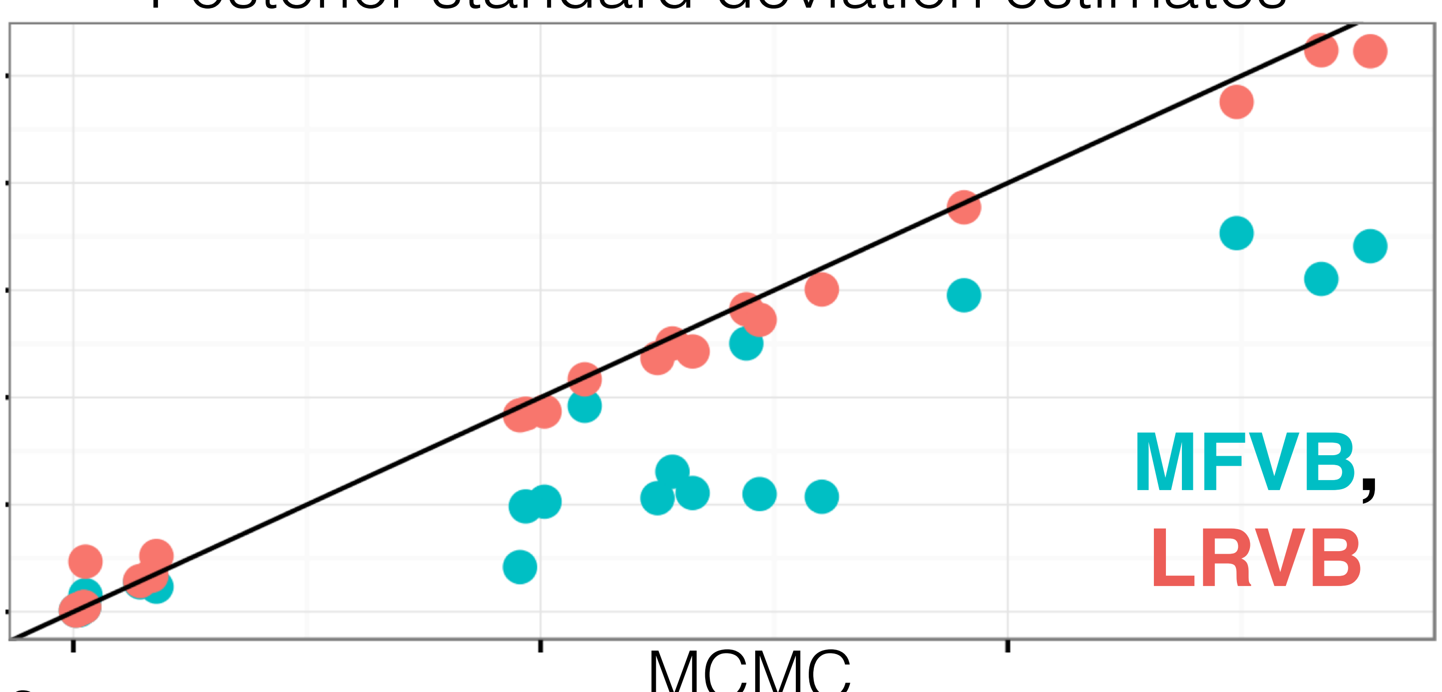
- Linear response variational Bayes (LRVB): A post hoc correction to a VB posterior covariance estimate
  - Works by applying a first-order Taylor expansion at the VB optimum (so requires being at the optimum)
  - Exact when the posterior mean estimate is exact

- Microcredit
- Criteo ads

Posterior std dev estimates



Posterior standard deviation estimates



[Giordano, Broderick, Jordan 2015, 2018]

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB? Some VB successes (speed, accuracy)
  - Often fast and accurate for point estimates
- Some VB failure modes, and partial solutions
  - Issues with uncertainty and more
- Ease of use / automation
  - Automatic differentiation variational inference (ADVI) and beyond

Roadmap

- Bayes & Approximate Bayes setup
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB? Some VB successes (speed, accuracy)
  - Often fast and accurate for point estimates
- Some VB failure modes, and partial solutions
  - Issues with uncertainty and more
- Ease of use / automation
  - Automatic differentiation variational inference (ADVI) and beyond

How much can be automated?

How much can be automated?

- Problem: no one has time to derive update equations!

How much can be automated?

- Problem: no one has time to derive update equations!
- Ideal “black-box” Bayesian approximation: user needs to specify only the model (i.e. prior, likelihood) and data

How much can be automated?

- Problem: no one has time to derive update equations!
- Ideal "black-box" Bayesian approximation: user needs to specify only the model (i.e. prior, likelihood) and data
  - The method should output an estimate of the posterior summaries of interest (e.g. posterior mean and variance)

How much can be automated?

- Problem: no one has time to derive update equations!
- Ideal “black-box” Bayesian approximation: user needs to specify only the model (i.e. prior, likelihood) and data
  - The method should output an estimate of the posterior summaries of interest (e.g. posterior mean and variance)
- Modern automated MCMC methods are commonly in the lineage of the No-U-Turn Sampler (Stan, PyMC, etc.)

How much can be automated?

- Problem: no one has time to derive update equations!
- Ideal "black-box" Bayesian approximation: user needs to specify only the model (i.e. prior, likelihood) and data
  - The method should output an estimate of the posterior summaries of interest (e.g. posterior mean and variance)
- Modern automated MCMC methods are commonly in the lineage of the No-U-Turn Sampler (Stan, PyMC, etc.)
  - But often too slow to run

How much can be automated?

- Problem: no one has time to derive update equations!
- Ideal “black-box” Bayesian approximation: user needs to specify only the model (i.e. prior, likelihood) and data
  - The method should output an estimate of the posterior summaries of interest (e.g. posterior mean and variance)
- Modern automated MCMC methods are commonly in the lineage of the No-U-Turn Sampler (Stan, PyMC, etc.)
  - But often too slow to run
- Dominant black-box variational method: “Automatic differentiation variational inference” (ADVI)

How much can be automated?

- Problem: no one has time to derive update equations!
- Ideal “black-box” Bayesian approximation: user needs to specify only the model (i.e. prior, likelihood) and data
  - The method should output an estimate of the posterior summaries of interest (e.g. posterior mean and variance)
- Modern automated MCMC methods are commonly in the lineage of the No-U-Turn Sampler (Stan, PyMC, etc.)
  - But often too slow to run
- Dominant black-box variational method: “Automatic differentiation variational inference” (ADVI)
  - In the major probabilistic programming languages: Stan, PyMC, Edward, etc.

How much can be automated?

- Problem: no one has time to derive update equations!
- Ideal “black-box” Bayesian approximation: user needs to specify only the model (i.e. prior, likelihood) and data
  - The method should output an estimate of the posterior summaries of interest (e.g. posterior mean and variance)
- Modern automated MCMC methods are commonly in the lineage of the No-U-Turn Sampler (Stan, PyMC, etc.)
  - But often too slow to run
- Dominant black-box variational method: “Automatic differentiation variational inference” (ADVI)
  - In the major probabilistic programming languages: Stan, PyMC, Edward, etc.
  - What exactly counts as being automated? Is ADVI faster than MCMC? Is ADVI accurate?

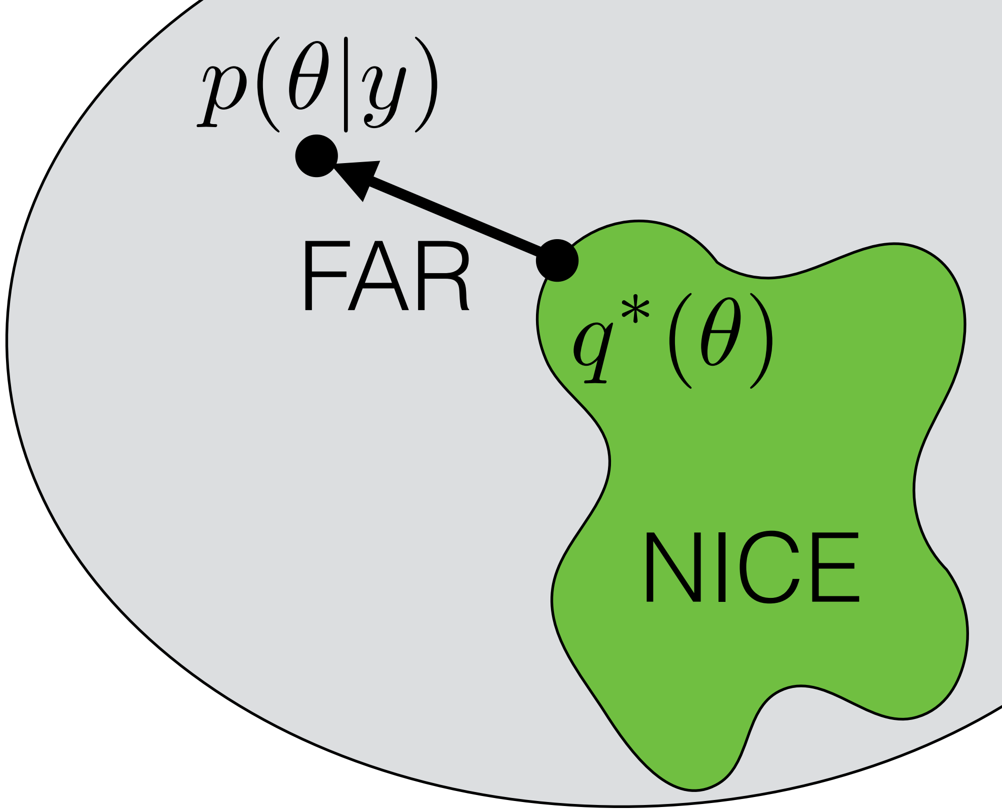
Automating VB

Automating VB

- Recall: Data y, parameters θ

Automating VB

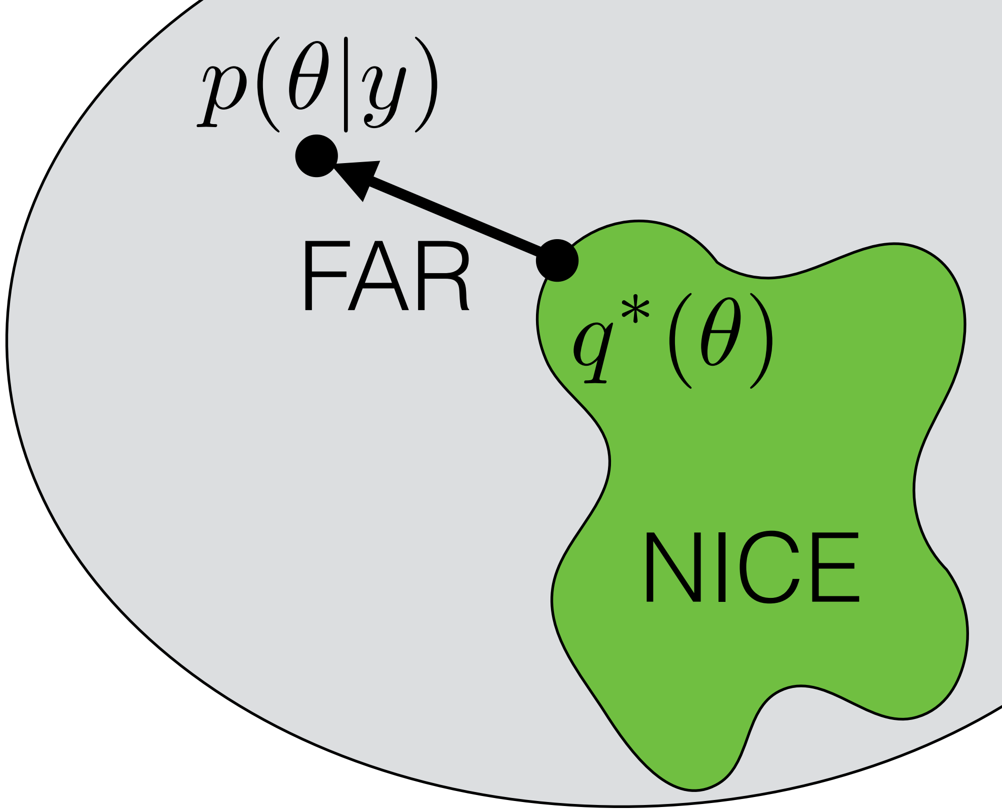
- Recall: Data y, parameters θ



Automating VB

- Recall: Data y, parameters θ
- Variational Bayes:

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

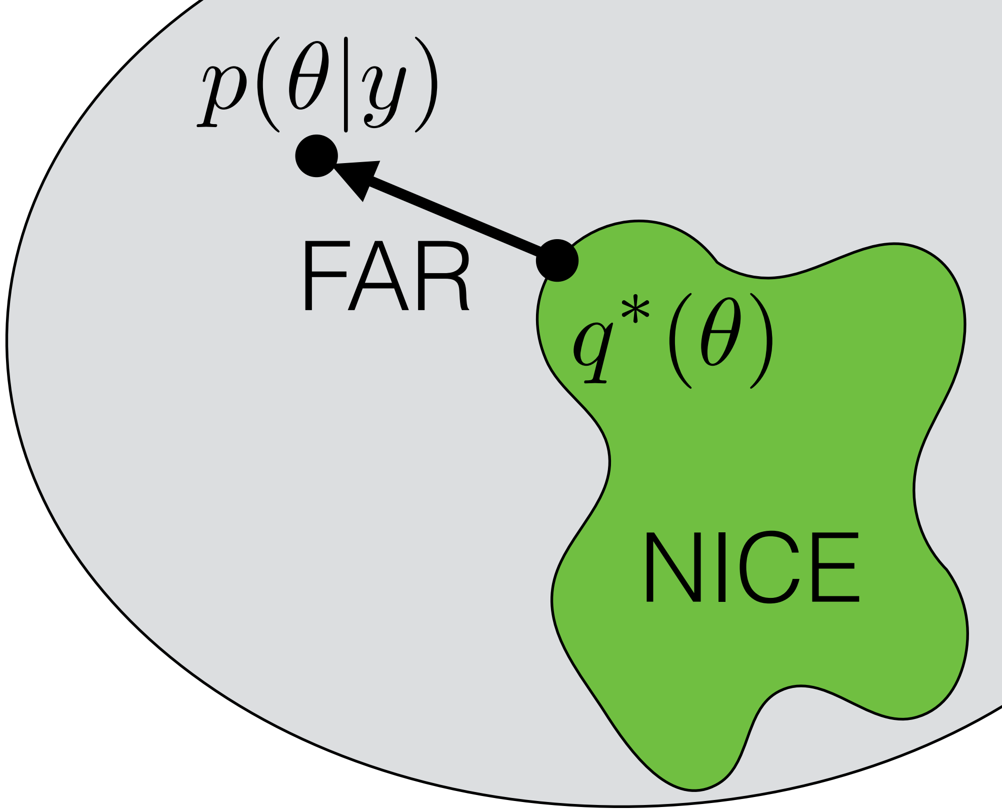


Automating VB

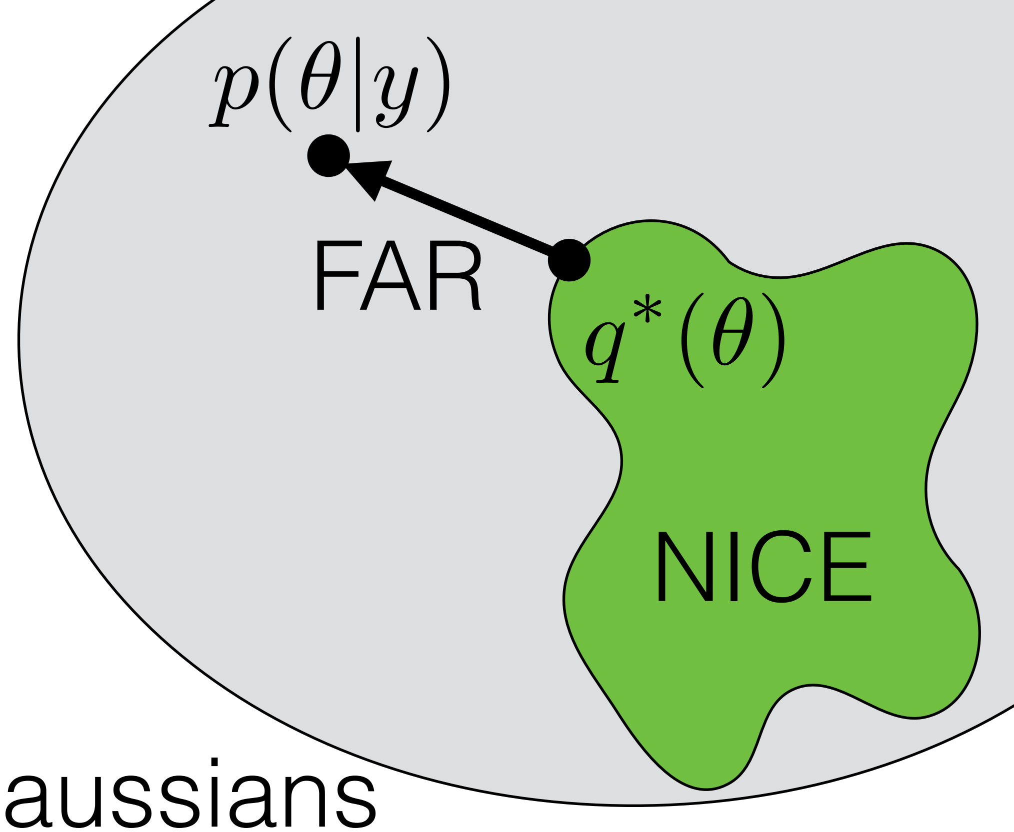
- Recall: Data y, parameters θ
- Variational Bayes:

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

- Auto diff variational inference (ADVI):



Automating VB

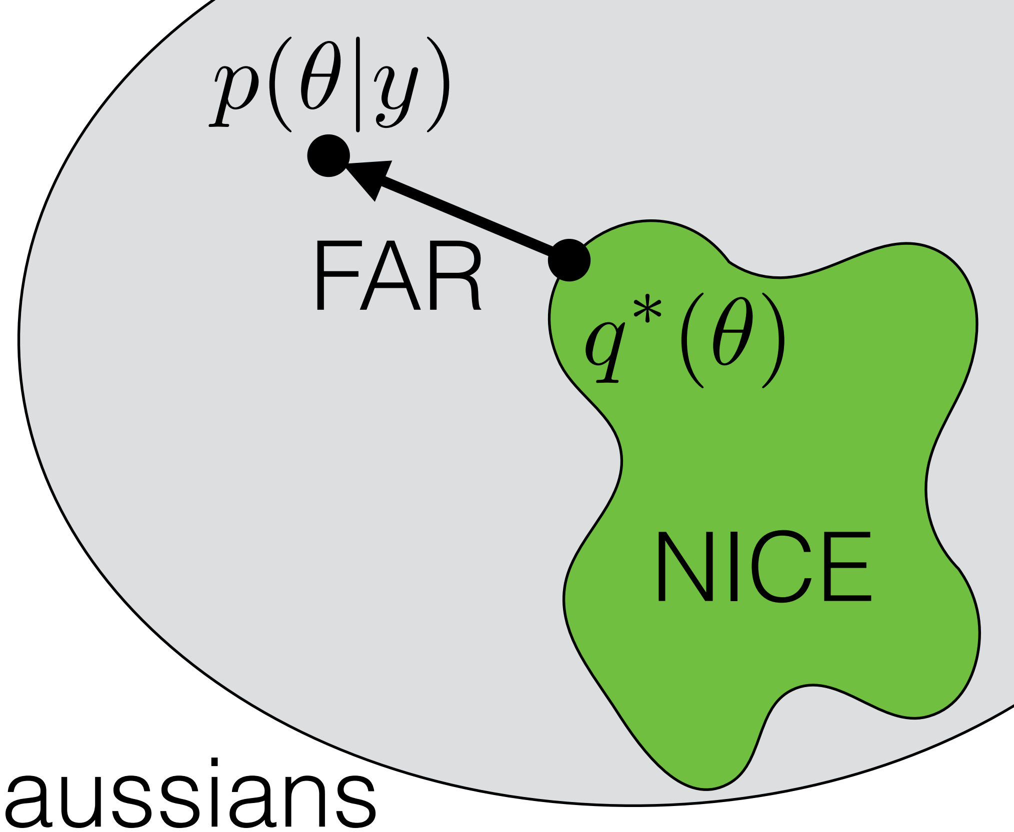


- Recall: Data y, parameters θ
- Variational Bayes:

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

- Auto diff variational inference (ADVI):
  - “Nice” distributions chosen to be Gaussians (“mean field” or “full rank”): $q(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$

Automating VB



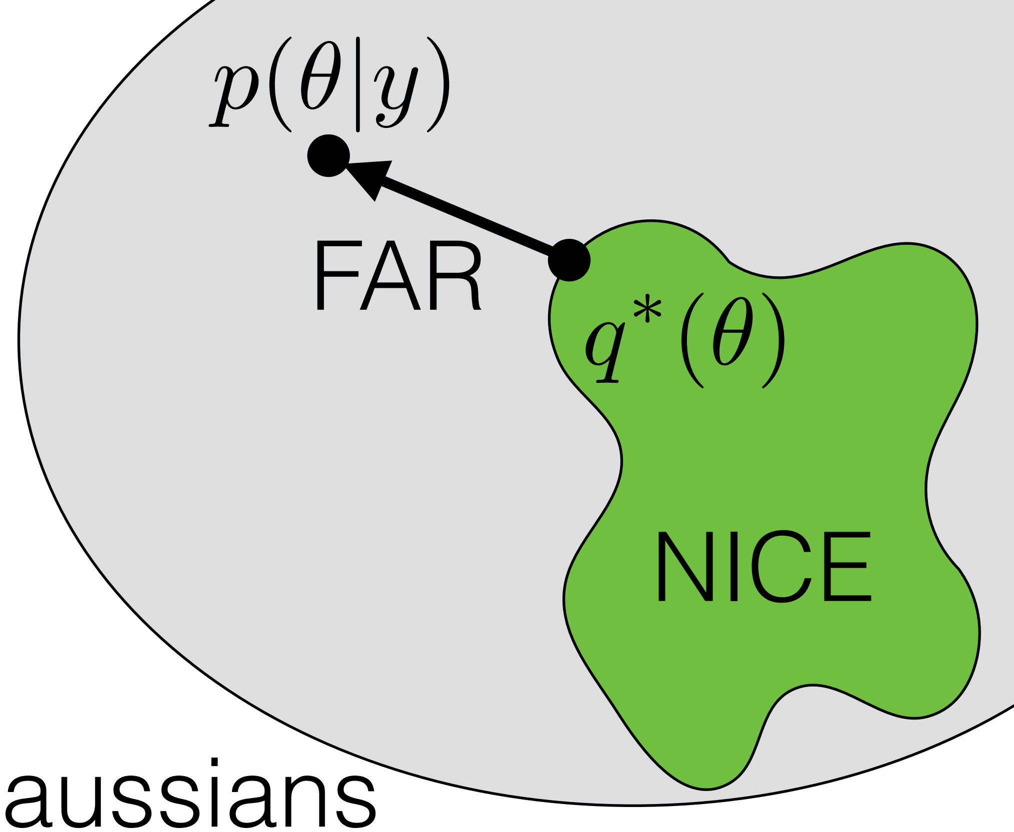
- Recall: Data y, parameters θ
- Variational Bayes:

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

- Auto diff variational inference (ADVI):
  - “Nice” distributions chosen to be Gaussians (“mean field” or “full rank”): $q(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$

$$\mu^*, \Sigma^* = \mathrm{argmin}_{\mu,\Sigma} g(\mu, \Sigma) + \int_z \mathcal{N}(z; 0, I) h(z, \mu, \Sigma) \, dz$$

Automating VB



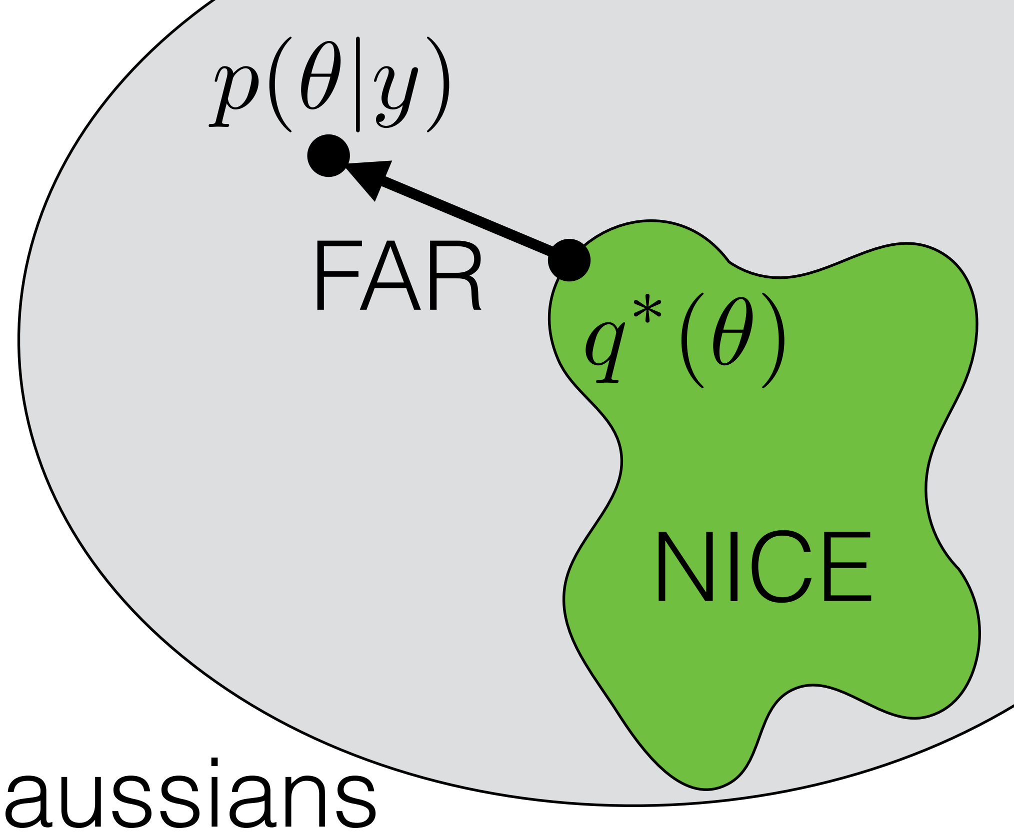
- Recall: Data y, parameters θ
- Variational Bayes:

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

- Auto diff variational inference (ADVI):
  - “Nice” distributions chosen to be Gaussians (“mean field” or “full rank”): $q(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$

$$\mu^*, \Sigma^* = \mathrm{argmin}_{\mu, \Sigma} \underbrace{g(\mu, \Sigma)}_{\text{easy}} + \int_z \mathcal{N}(z; 0, I) h(z, \mu, \Sigma)\, dz$$

Automating VB



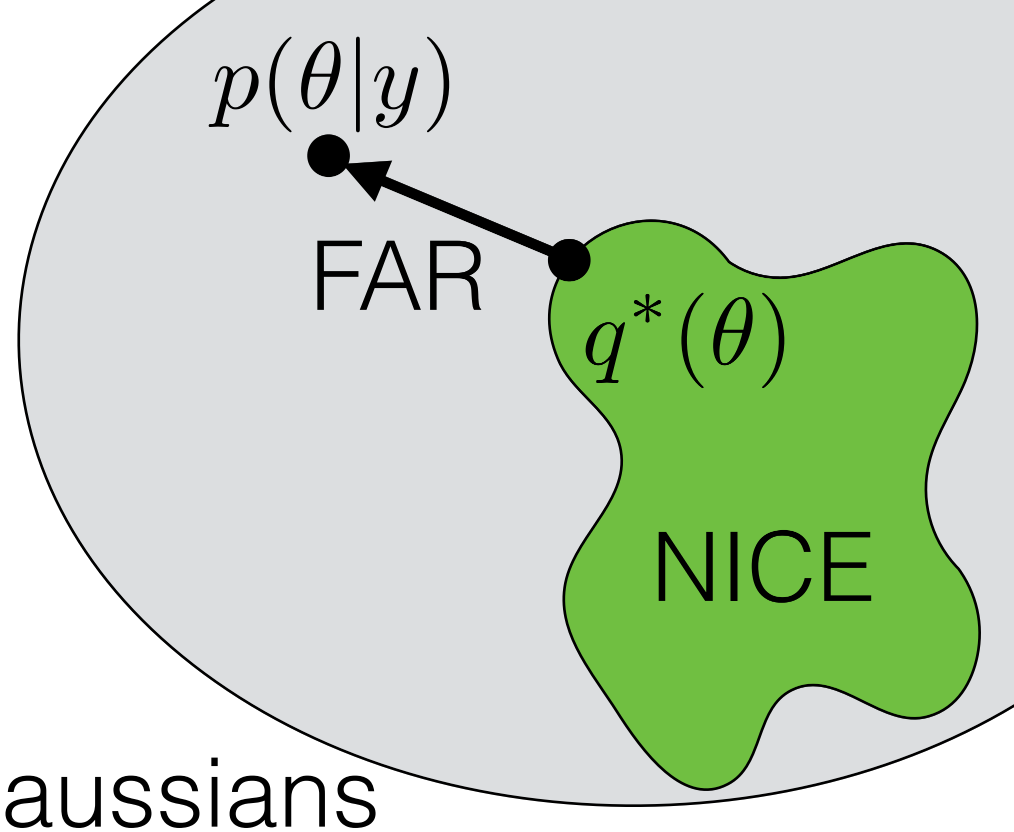
- Recall: Data y, parameters θ
- Variational Bayes:

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

- Auto diff variational inference (ADVI):
  - "Nice" distributions chosen to be Gaussians ("mean field" or "full rank"): $q(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$

$$\mu^*, \Sigma^* = \text{argmin}_{\mu, \Sigma} \underbrace{g(\mu, \Sigma)}_{\text{easy}} + \underbrace{\int_z \mathcal{N}(z; 0, I) h(z, \mu, \Sigma)\, dz}_{\text{integral: requires approximation}}$$

Automating VB



- Recall: Data y, parameters θ
- Variational Bayes:

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

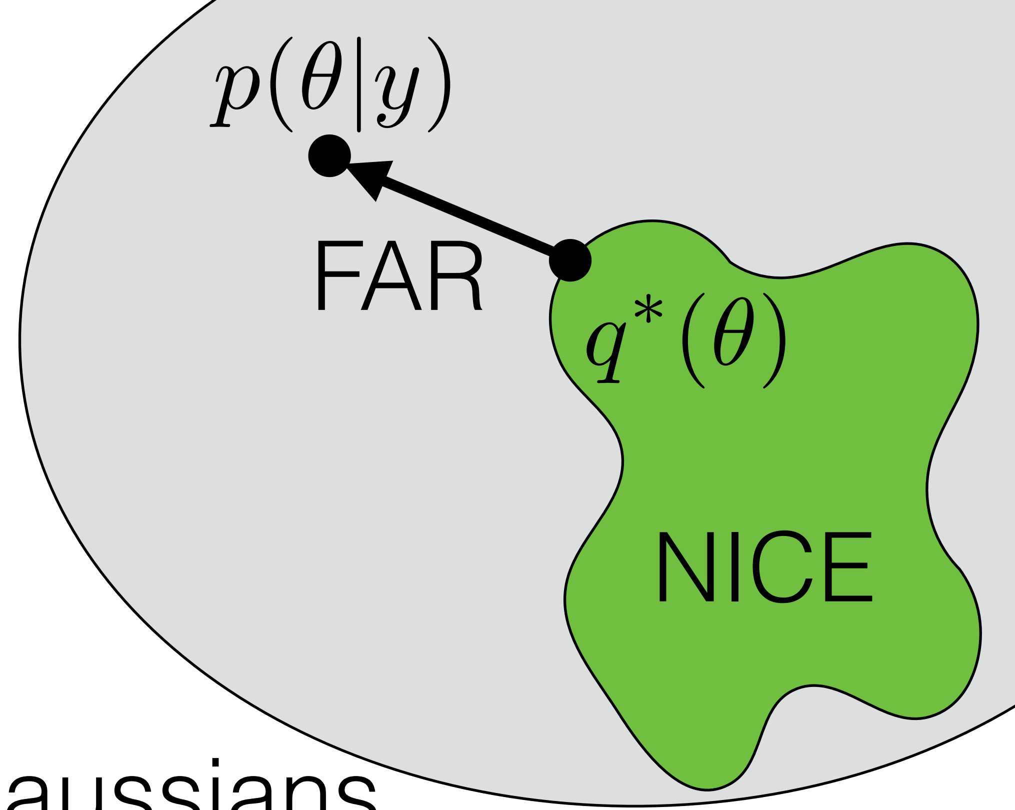
- Auto diff variational inference (ADVI):
  - "Nice" distributions chosen to be Gaussians ("mean field" or "full rank"): $q(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$

$$\mu^*, \Sigma^* = \text{argmin}_{\mu, \Sigma} \underbrace{g(\mu, \Sigma)}_{\text{easy}} + \underbrace{\int_z \mathcal{N}(z; 0, I) h(z, \mu, \Sigma)\, dz}_{\text{integral: requires approximation}}$$

- Stochastic gradient descent (SGD)→original ADVI

[Kucukelbir et al 2015, 2017]

Automating VB



- Recall: Data y, parameters θ
- Variational Bayes:

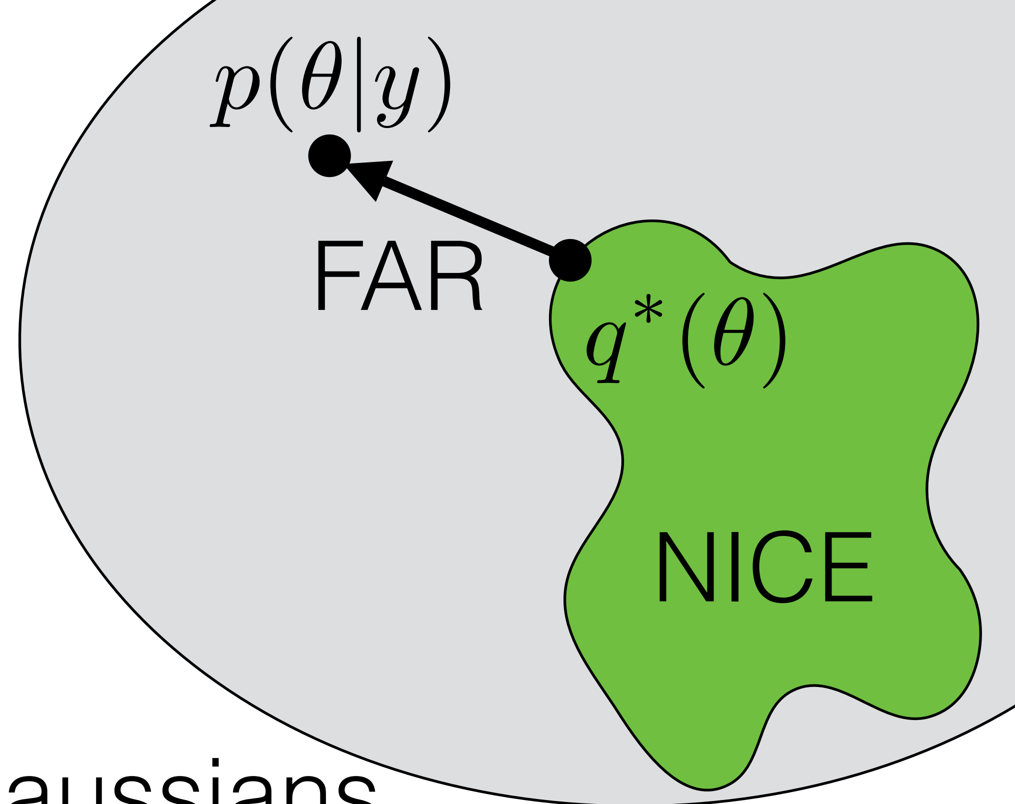
$$q^* = \text{argmin}_{q \in \text{NICE}} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

- Auto diff variational inference (ADVI):
  - "Nice" distributions chosen to be Gaussians ("mean field" or "full rank"): $q(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$

$$\mu^*, \Sigma^* = \text{argmin}_{\mu,\Sigma} \underbrace{g(\mu, \Sigma)}_{\text{easy}} + \underbrace{\int_z \mathcal{N}(z; 0, I) h(z, \mu, \Sigma)\, dz}_{\text{integral: requires approximation}}$$

- Stochastic gradient descent (SGD)→original ADVI [Kucukelbir et al 2015, 2017]
  - Want a step in the direction of the gradient (for μ, Σ)
  - Approximate the gradient with Monte Carlo

Automating VB



- Recall: Data y, parameters θ
- Variational Bayes:

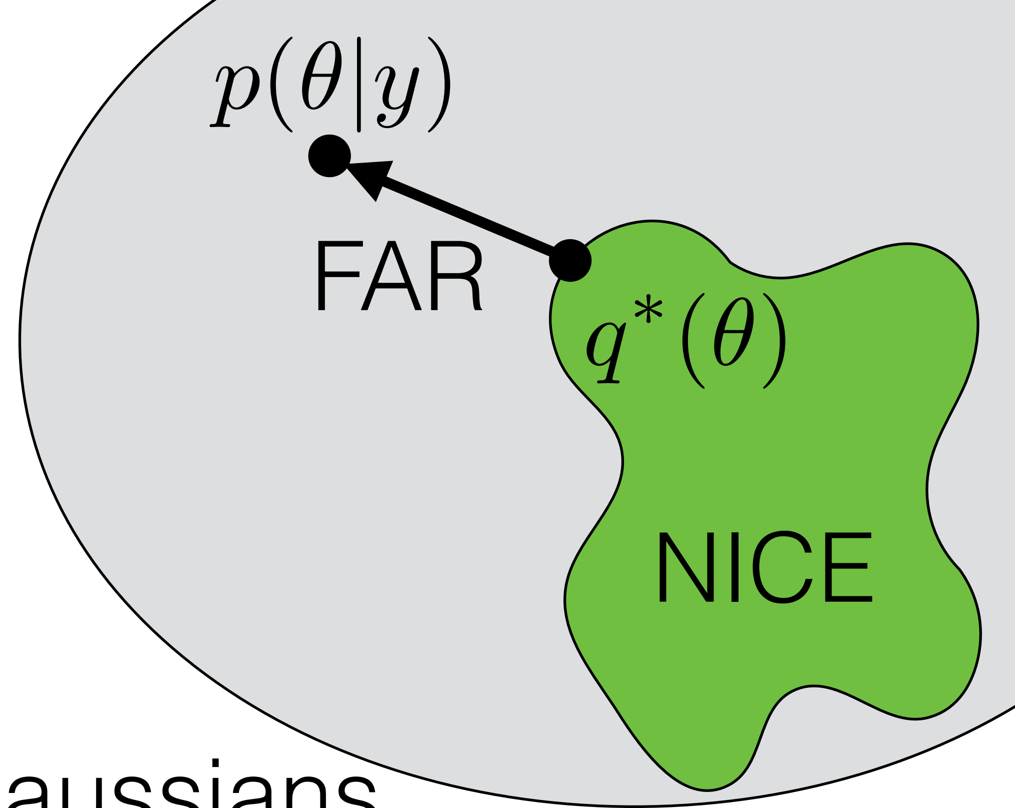
$$q^* = \mathrm{argmin}_{q \in \mathrm{NICE}} \mathrm{KL}\left(q(\cdot)||p(\cdot|y)\right)$$

- Auto diff variational inference (ADVI):
  - "Nice" distributions chosen to be Gaussians ("mean field" or "full rank"): $q(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$

$$\mu^*, \Sigma^* = \mathrm{argmin}_{\mu, \Sigma} \underbrace{g(\mu, \Sigma)}_{\text{easy}} + \underbrace{\int_z \mathcal{N}(z; 0, I) h(z, \mu, \Sigma)\, dz}_{\text{integral: requires approximation}}$$

- Stochastic gradient descent (SGD)→original ADVI [Kucukelbir et al 2015, 2017]
  - Want a step in the direction of the gradient (for μ, Σ)
  - Approximate the gradient with Monte Carlo
- Sample average approx (SAA) [Giordano et al 2023; Burroni et al 2023]

Automating VB



- Recall: Data y, parameters θ
- Variational Bayes:

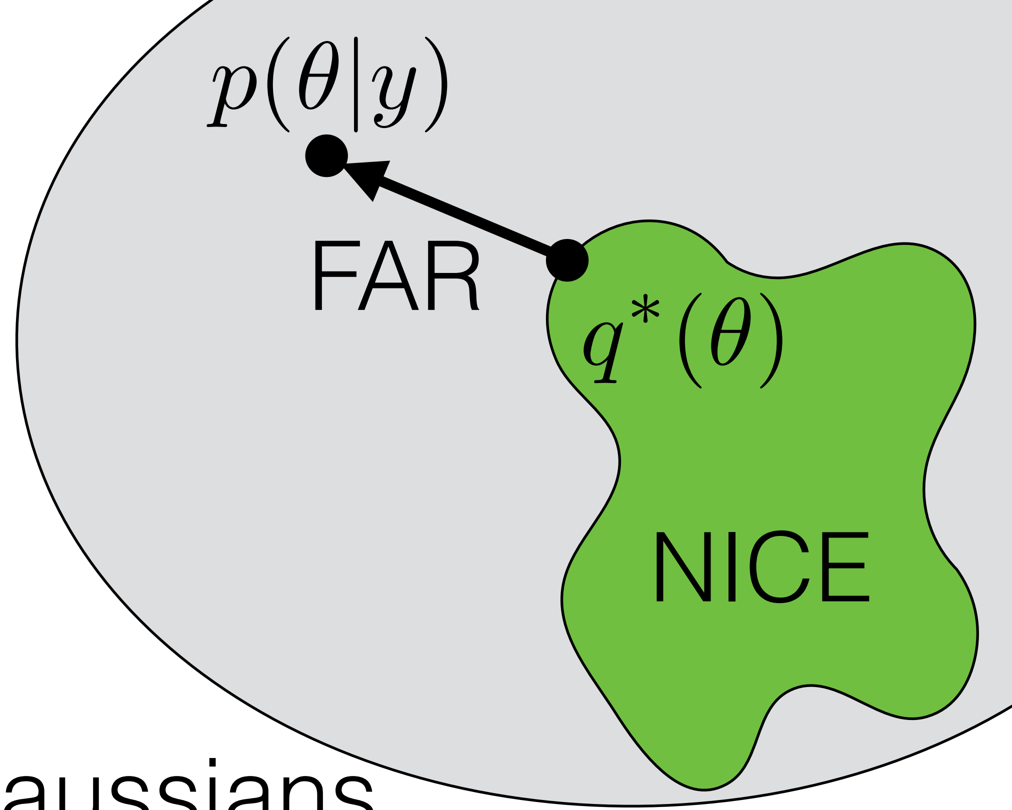
$$q^* = \mathrm{argmin}_{q \in \mathrm{NICE}} \mathrm{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

- Auto diff variational inference (ADVI):
  - "Nice" distributions chosen to be Gaussians ("mean field" or "full rank"): $q(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$

$$\mu^*, \Sigma^* = \mathrm{argmin}_{\mu, \Sigma} \underbrace{g(\mu, \Sigma)}_{\text{easy}} + \underbrace{\int_z \mathcal{N}(z; 0, I) h(z, \mu, \Sigma)\, dz}_{\text{integral: requires approximation}}$$

- Stochastic gradient descent (SGD)→original ADVI [Kucukelbir et al 2015, 2017]
  - Want a step in the direction of the gradient (for μ, Σ)
  - Approximate the gradient with Monte Carlo
- Sample average approx (SAA)→*deterministic ADVI* (DADVI) [Giordano et al 2023; Burroni et al 2023]

Automating VB



- Recall: Data y, parameters θ
- Variational Bayes:

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

- Auto diff variational inference (ADVI):
  - "Nice" distributions chosen to be Gaussians ("mean field" or "full rank"): $q(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$

$$\mu^*, \Sigma^* = \text{argmin}_{\mu,\Sigma} \underbrace{g(\mu, \Sigma)}_{\text{easy}} + \underbrace{\int_z \mathcal{N}(z; 0, I) h(z, \mu, \Sigma)\, dz}_{\text{integral: requires approximation}}$$

- Stochastic gradient descent (SGD)→original ADVI [Kucukelbir et al 2015, 2017]
  - Want a step in the direction of the gradient (for μ, Σ)
  - Approximate the gradient with Monte Carlo
- Sample average approx (SAA)→*deterministic ADVI* (DADVI) [Giordano et al 2023; Burroni et al 2023]
  - Approximate the *objective* with Monte Carlo
  - (Deterministically) optimize the *approximate objective*

ADVI & Deterministic ADVI (DADVI)

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)
- **ADVI**:

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)
- **ADVI**: ✅ Already in major prob programming languages

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)
- **ADVI**: ✅ Already in major prob programming languages
  ❌ Need to choose a step size schedule (typically involves tuning parameters)

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)
- **ADVI**: ✅ Already in major prob programming languages
  ❌ Need to choose a step size schedule (typically involves tuning parameters)
  ❌ No clear convergence criterion (unlike variational Bayes with no Monte Carlo noise)

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)
- **ADVI**: ✅ Already in major prob programming languages
  ❌ Need to choose a step size schedule (typically involves tuning parameters)
  ❌ No clear convergence criterion (unlike variational Bayes with no Monte Carlo noise)
  ❌ Cannot apply linear response methods to correct the posterior (co)variances [Giordano, Broderick, Jordan 2015, 2018]

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)
- **ADVI**: ✅ Already in major prob programming languages
  ❌ Need to choose a step size schedule (typically involves tuning parameters)
  ❌ No clear convergence criterion (unlike variational Bayes with no Monte Carlo noise)
  ❌ Cannot apply linear response methods to correct the posterior (co)variances [Giordano, Broderick, Jordan 2015, 2018]
- **DADVI**:

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)
- **ADVI**: ✅ Already in major prob programming languages
  ❌ Need to choose a step size schedule (typically involves tuning parameters)
  ❌ No clear convergence criterion (unlike variational Bayes with no Monte Carlo noise)
  ❌ Cannot apply linear response methods to correct the posterior (co)variances [Giordano, Broderick, Jordan 2015, 2018]
- **DADVI**: 🟨 Code, but not yet in major prob prog langs

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)
- **ADVI**: ✅ Already in major prob programming languages
  ❌ Need to choose a step size schedule (typically involves tuning parameters)
  ❌ No clear convergence criterion (unlike variational Bayes with no Monte Carlo noise)
  ❌ Cannot apply linear response methods to correct the posterior (co)variances [Giordano, Broderick, Jordan 2015, 2018]
- **DADVI**: 🟨 Code, but not yet in major prob prog langs
  ✅ Can use more reliable optimization tools. We use trust-region Newton conjugate gradient (trust-ncg in scipy)

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)
- **ADVI**: ✅ Already in major prob programming languages
  ❌ Need to choose a step size schedule (typically involves tuning parameters)
  ❌ No clear convergence criterion (unlike variational Bayes with no Monte Carlo noise)
  ❌ Cannot apply linear response methods to correct the posterior (co)variances [Giordano, Broderick, Jordan 2015, 2018]
- **DADVI**: 🟨 Code, but not yet in major prob prog langs
  ✅ Can use more reliable optimization tools. We use trust-region Newton conjugate gradient (trust-ncg in scipy)
  ✅ No tuning parameters thanks to second-order methods

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)
- **ADVI**: ✅ Already in major prob programming languages
  ❌ Need to choose a step size schedule (typically involves tuning parameters)
  ❌ No clear convergence criterion (unlike variational Bayes with no Monte Carlo noise)
  ❌ Cannot apply linear response methods to correct the posterior (co)variances [Giordano, Broderick, Jordan 2015, 2018]
- **DADVI**: 🟨 Code, but not yet in major prob prog langs
  ✅ Can use more reliable optimization tools. We use trust-region Newton conjugate gradient (trust-ncg in scipy)
  ✅ No tuning parameters thanks to second-order methods
  ✅ Clear convergence criterion: gradient norm near zero

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)
- **ADVI**: ✅ Already in major prob programming languages
  ❌ Need to choose a step size schedule (typically involves tuning parameters)
  ❌ No clear convergence criterion (unlike variational Bayes with no Monte Carlo noise)
  ❌ Cannot apply linear response methods to correct the posterior (co)variances [Giordano, Broderick, Jordan 2015, 2018]
- **DADVI**: 🟨 Code, but not yet in major prob prog langs
  ✅ Can use more reliable optimization tools. We use trust-region Newton conjugate gradient (trust-ncg in scipy)
  ✅ No tuning parameters thanks to second-order methods
  ✅ Clear convergence criterion: gradient norm near zero
  ✅ Can correct posterior variances with linear response

ADVI & Deterministic ADVI (DADVI)

- **Both**: ✅ Avoid manual derivations of update steps
  ❌ Require all parameters to be continuous (like NUTS)
- **ADVI**: ✅ Already in major prob programming languages
  ❌ Need to choose a step size schedule (typically involves tuning parameters)
  ❌ No clear convergence criterion (unlike variational Bayes with no Monte Carlo noise)
  ❌ Cannot apply linear response methods to correct the posterior (co)variances [Giordano, Broderick, Jordan 2015, 2018]
- **DADVI**: 🟨 Code, but not yet in major prob prog langs
  ✅ Can use more reliable optimization tools. We use trust-region Newton conjugate gradient (trust-ncg in scipy)
  ✅ No tuning parameters thanks to second-order methods
  ✅ Clear convergence criterion: gradient norm near zero
  ✅ Can correct posterior variances with linear response
  ✅ More functionality: e.g. estimate of sampling variability

Experiments: runtime

Experiments: runtime

- 53 Applied Regression Modeling + 4 “bigger” problems

Experiments: runtime

- 53 Applied Regression Modeling + 4 “bigger” problems

“Bigger” models

microcredit
ecology
elections
tennis

Experiments: runtime

- 53 Applied Regression Modeling + 4 “bigger” problems
- Parameter dim from 2 to 15K

“Bigger” models

microcredit
ecology
elections
tennis

Experiments: runtime

- 53 Applied Regression Modeling + 4 “bigger” problems
- Parameter dim from 2 to 15K; max MCMC time >6h

Experiments: runtime

- 53 Applied Regression Modeling + 4 “bigger” problems
- Parameter dim from 2 to 15K; max MCMC time >6h
- Run to convergence (or max 100,000 iterations)

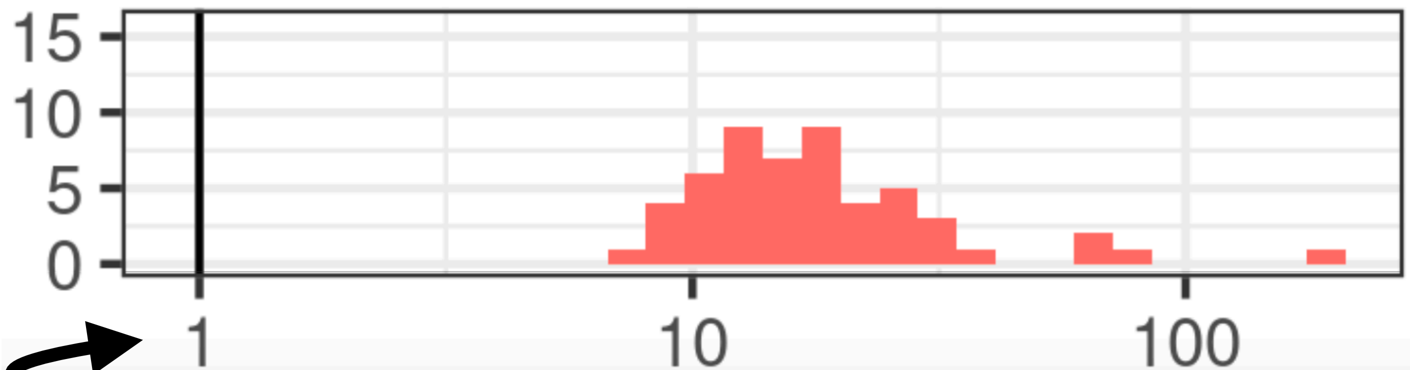
“Bigger” models

microcredit
ecology
elections
tennis

Experiments: runtime

- 53 Applied Regression Modeling + 4 “bigger” problems
- Parameter dim from 2 to 15K; max MCMC time >6h
- Run to convergence (or max 100,000 iterations)

ARM models



runtime / (DADVI’s runtime)

“Bigger” models

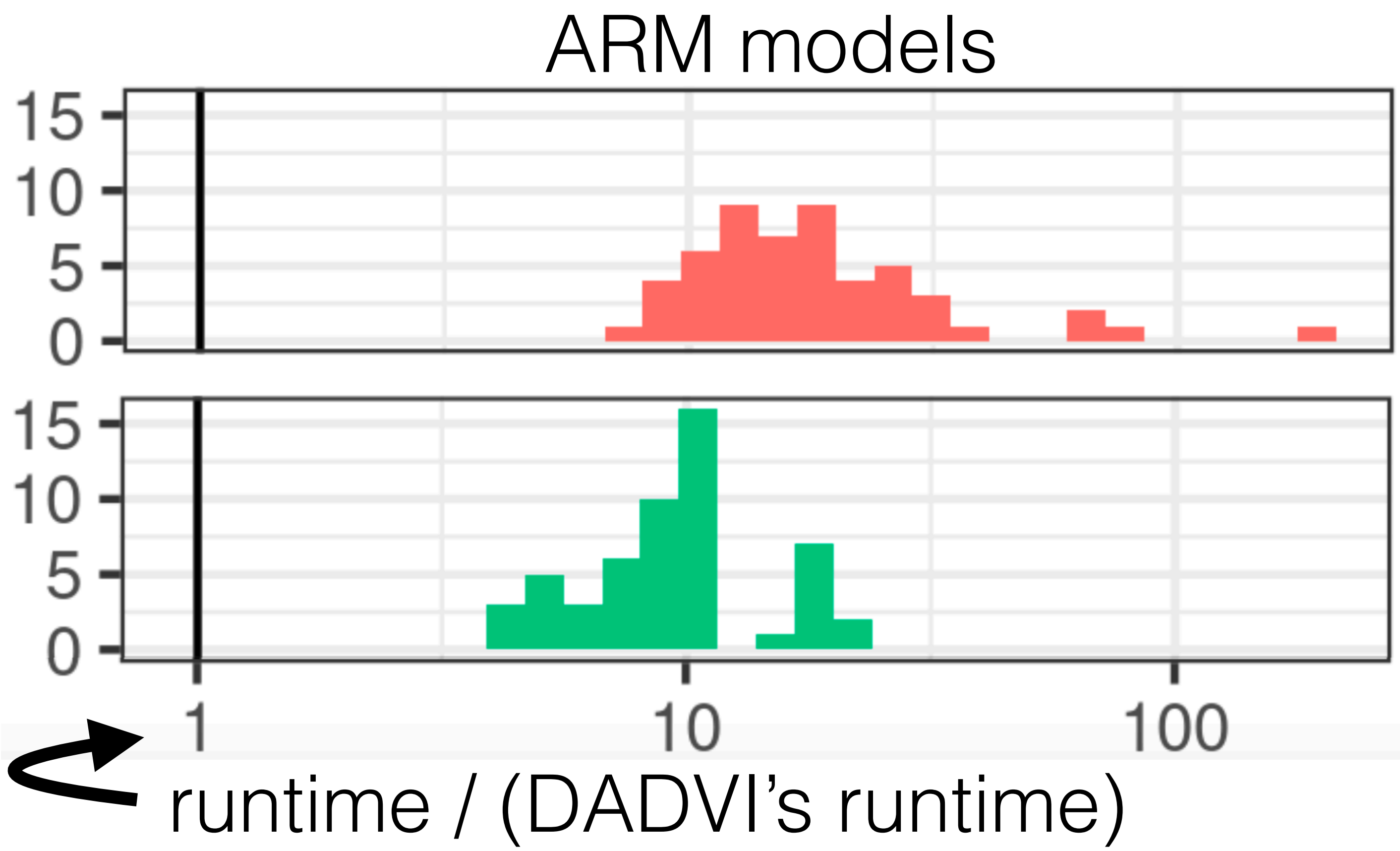
microcredit
ecology
elections
tennis

[i.e. full cost of DADVI + linear response correction]

MCMC (PyMC NUTS)

Experiments: runtime

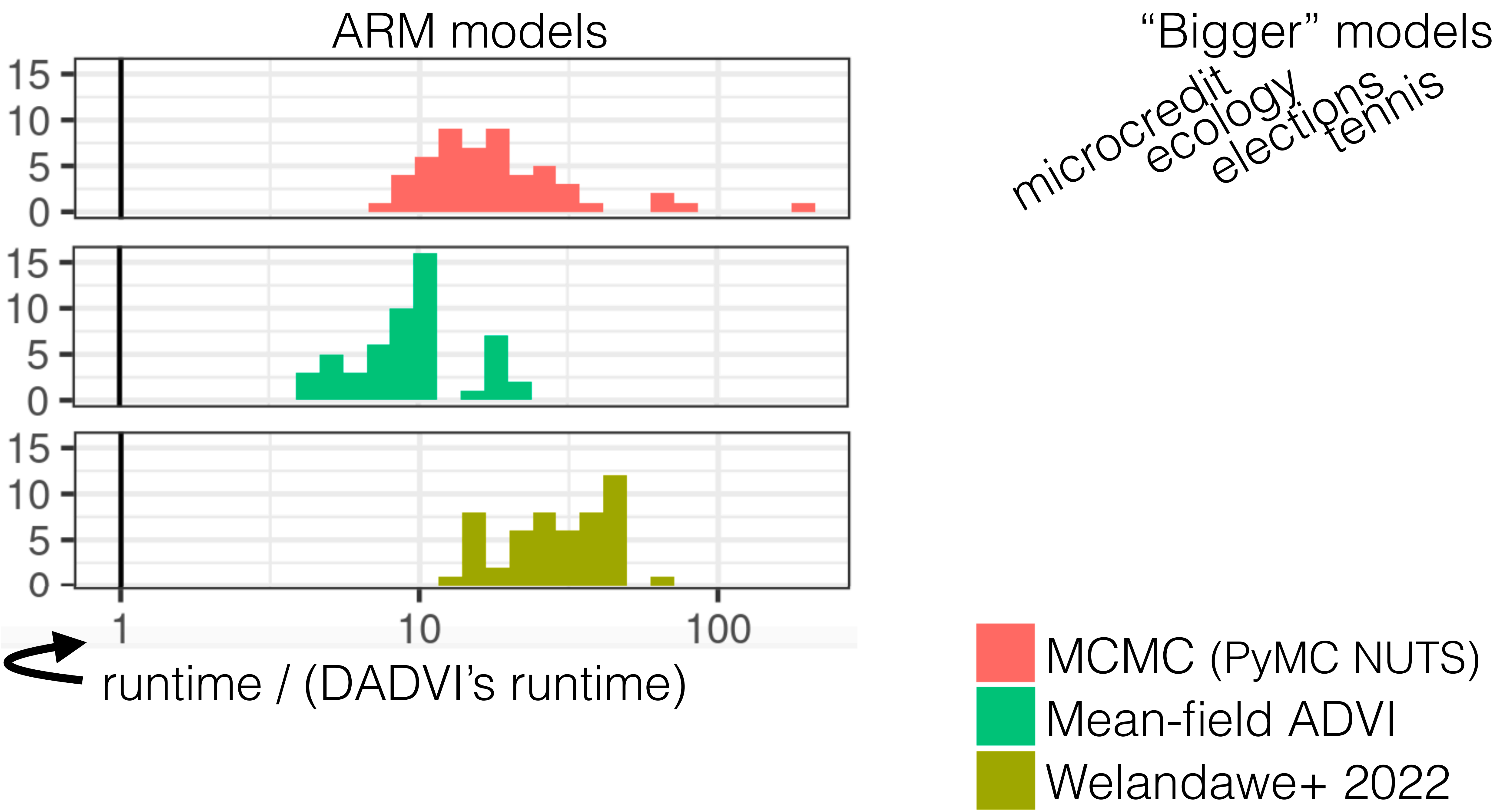
- 53 Applied Regression Modeling + 4 “bigger” problems
- Parameter dim from 2 to 15K; max MCMC time >6h
- Run to convergence (or max 100,000 iterations)



MCMC (PyMC NUTS)
Mean-field ADVI

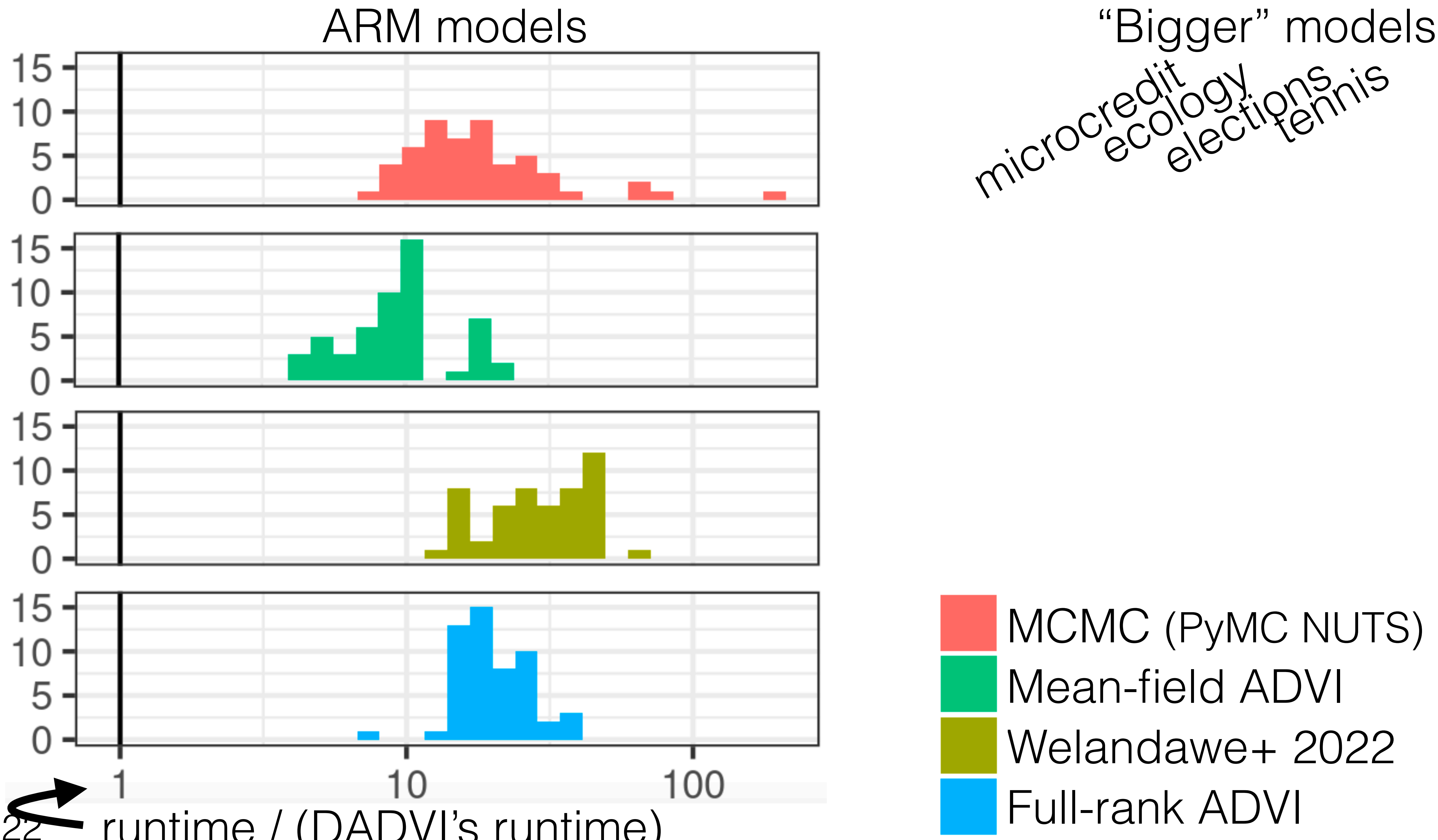
Experiments: runtime

- 53 Applied Regression Modeling + 4 “bigger” problems
- Parameter dim from 2 to 15K; max MCMC time >6h
- Run to convergence (or max 100,000 iterations)



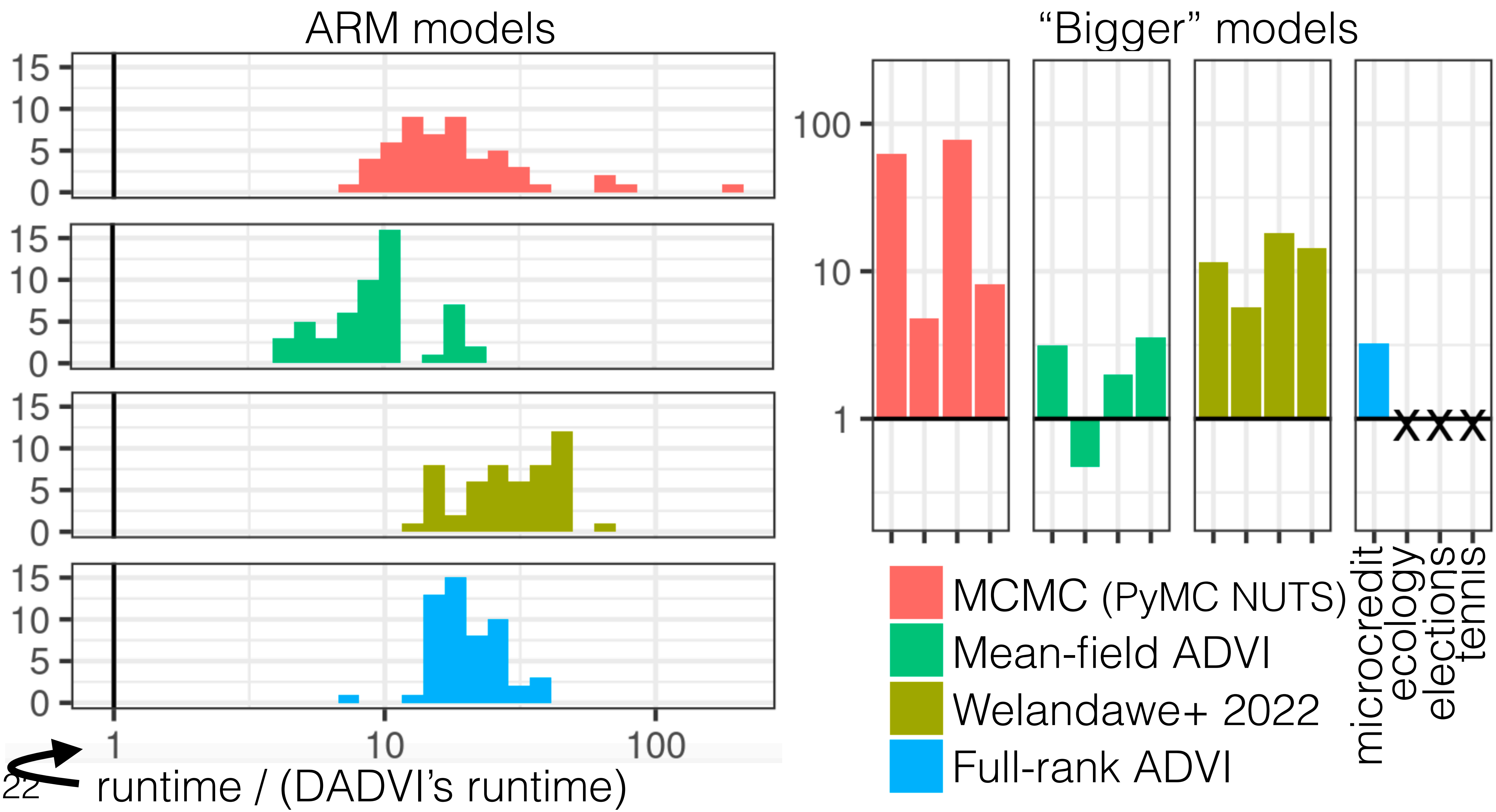
Experiments: runtime

- 53 Applied Regression Modeling + 4 “bigger” problems
- Parameter dim from 2 to 15K; max MCMC time >6h
- Run to convergence (or max 100,000 iterations)



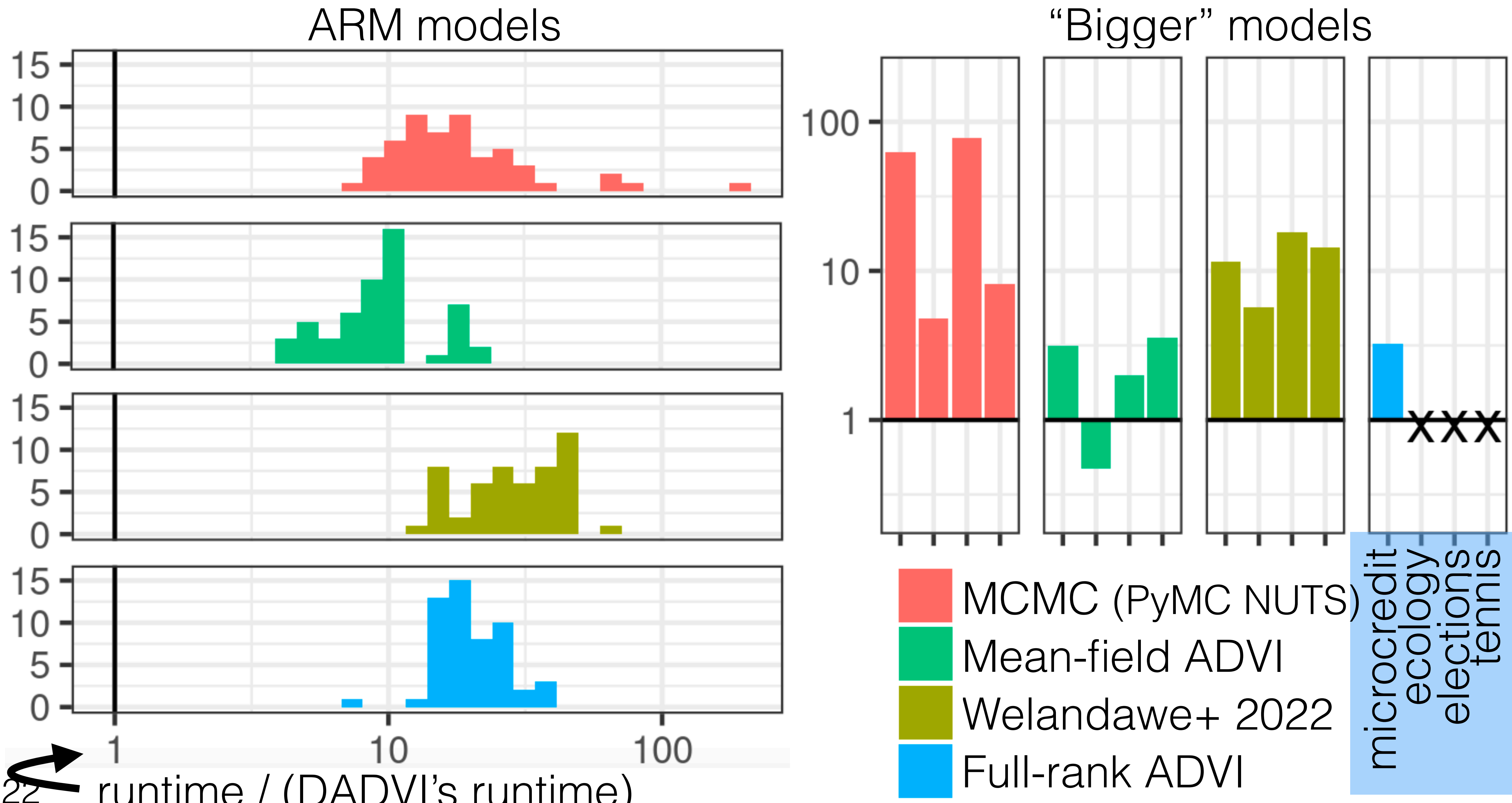
Experiments: runtime

- 53 Applied Regression Modeling + 4 “bigger” problems
- Parameter dim from 2 to 15K; max MCMC time >6h
- Run to convergence (or max 100,000 iterations)



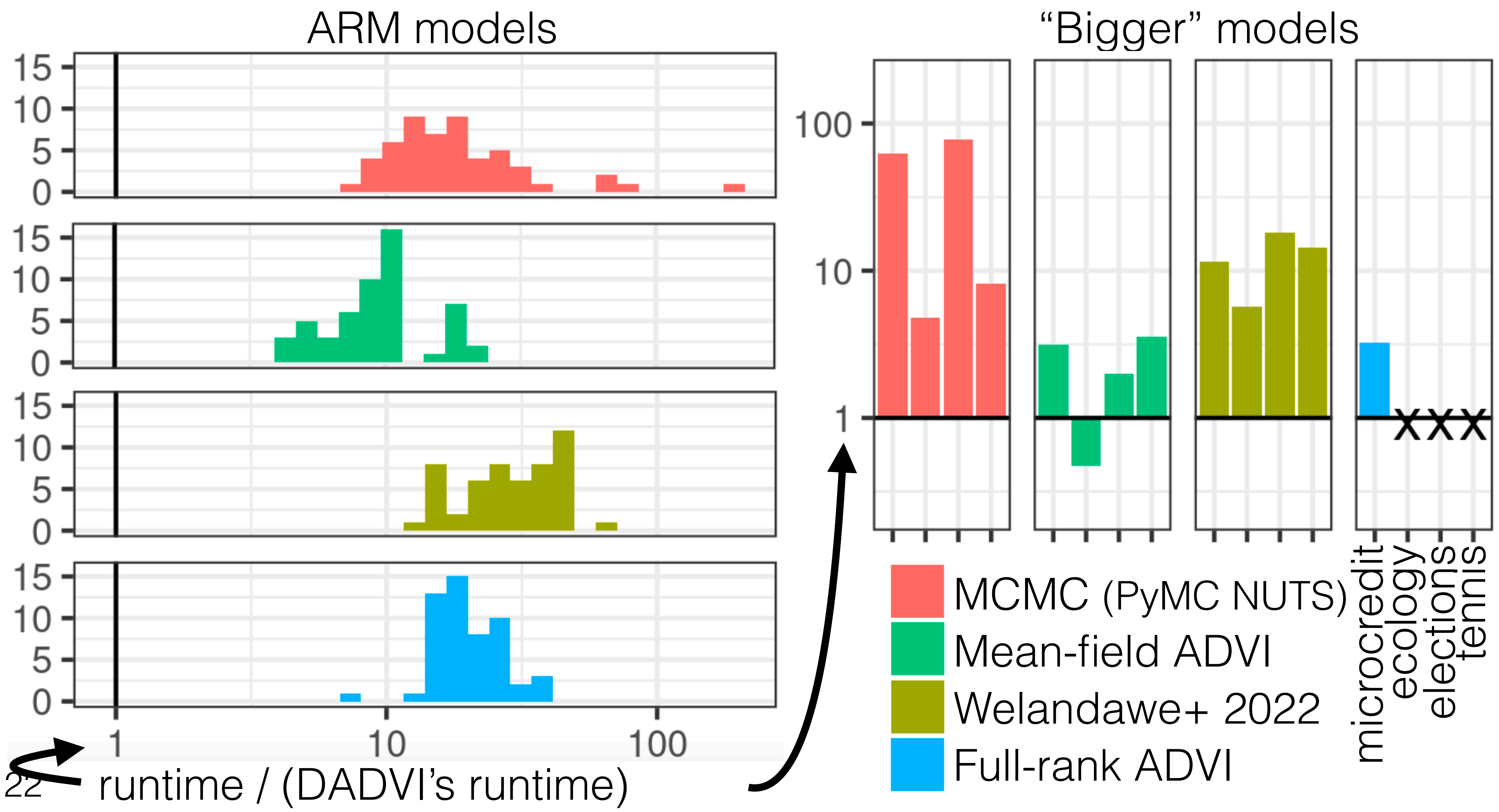
Experiments: runtime

- 53 Applied Regression Modeling + 4 “bigger” problems
- Parameter dim from 2 to 15K; max MCMC time >6h
- Run to convergence (or max 100,000 iterations)



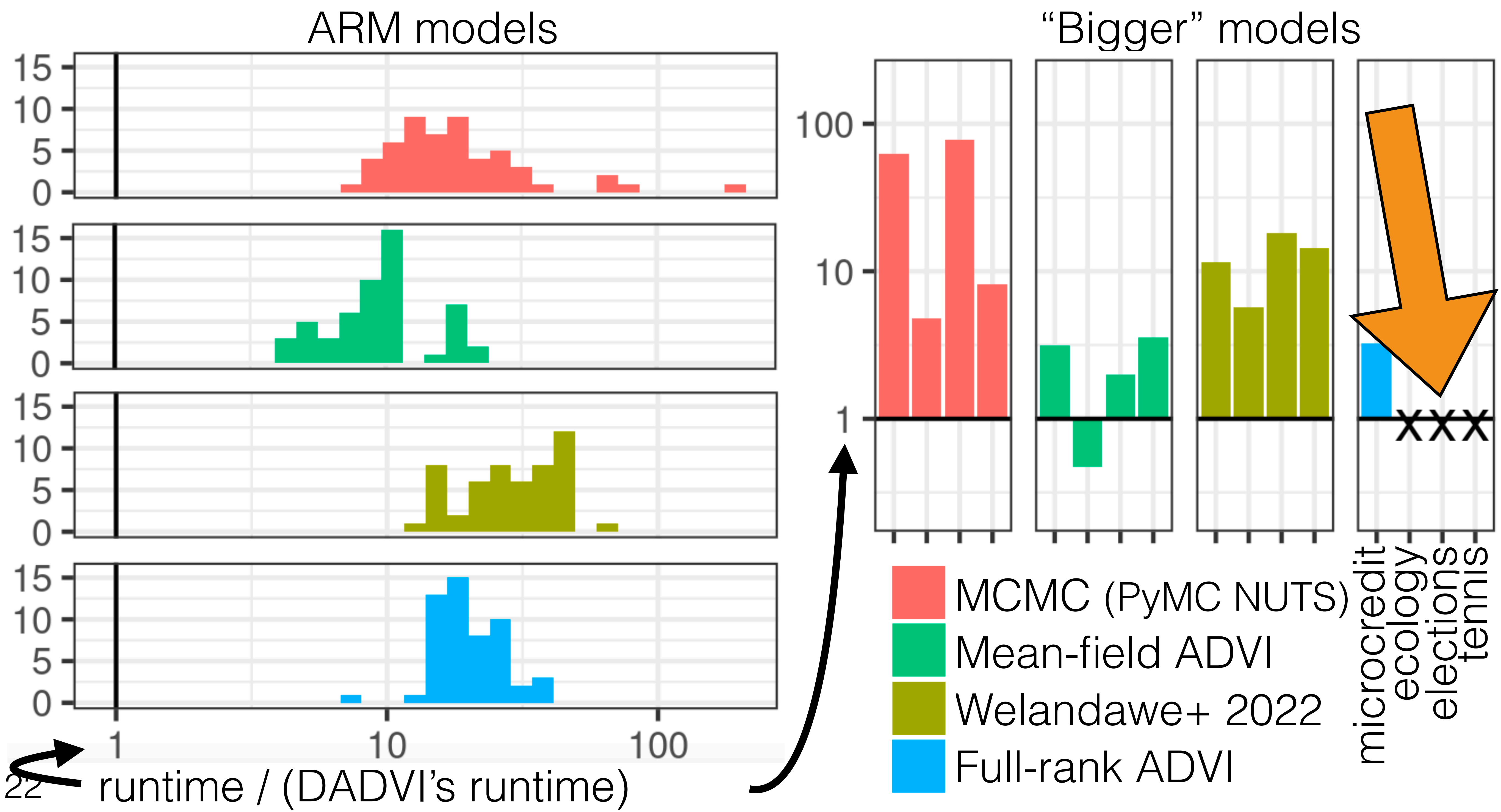
Experiments: runtime

- 53 Applied Regression Modeling + 4 “bigger” problems
- Parameter dim from 2 to 15K; max MCMC time >6h
- Run to convergence (or max 100,000 iterations)



Experiments: runtime

- 53 Applied Regression Modeling + 4 “bigger” problems
- Parameter dim from 2 to 15K; max MCMC time >6h
- Run to convergence (or max 100,000 iterations)



Quality of posterior variance estimates

Quality of posterior variance estimates

- DADVI allows a linear response (LR) correction

Quality of posterior variance estimates

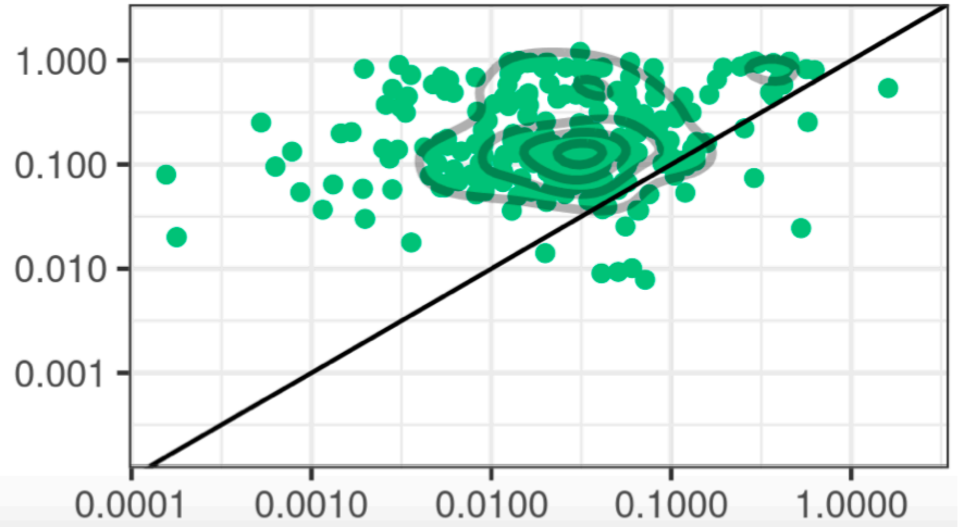
- DADVI allows a linear response (LR) correction

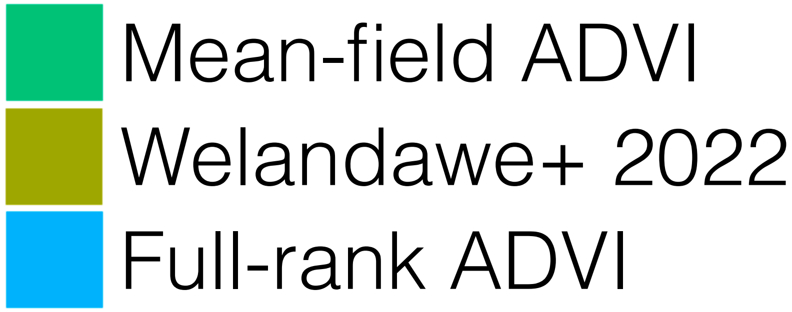
Mean-field ADVI
Welandawe+ 2022
Full-rank ADVI

Quality of posterior variance estimates

- DADVI allows a linear response (LR) correction

ARM models

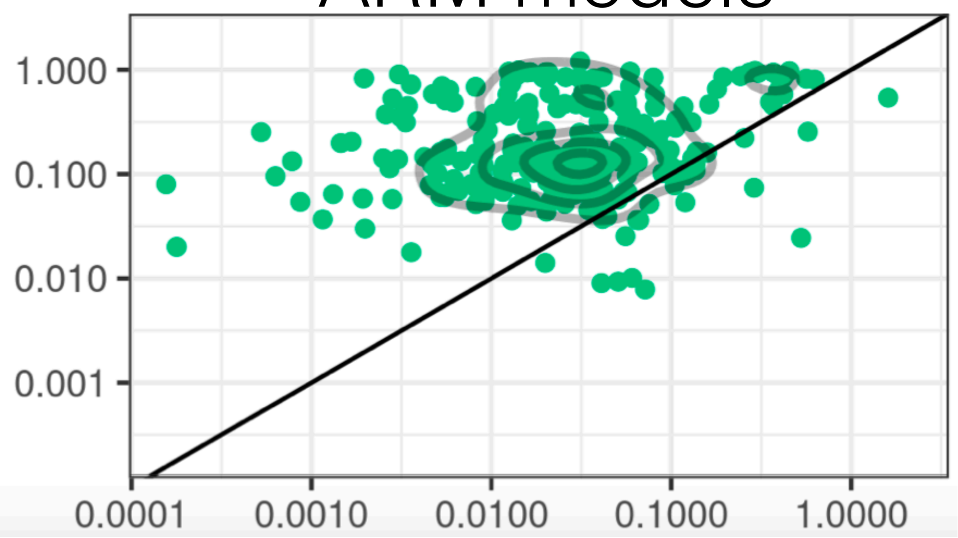




Quality of posterior variance estimates

- DADVI allows a linear response (LR) correction

ARM models

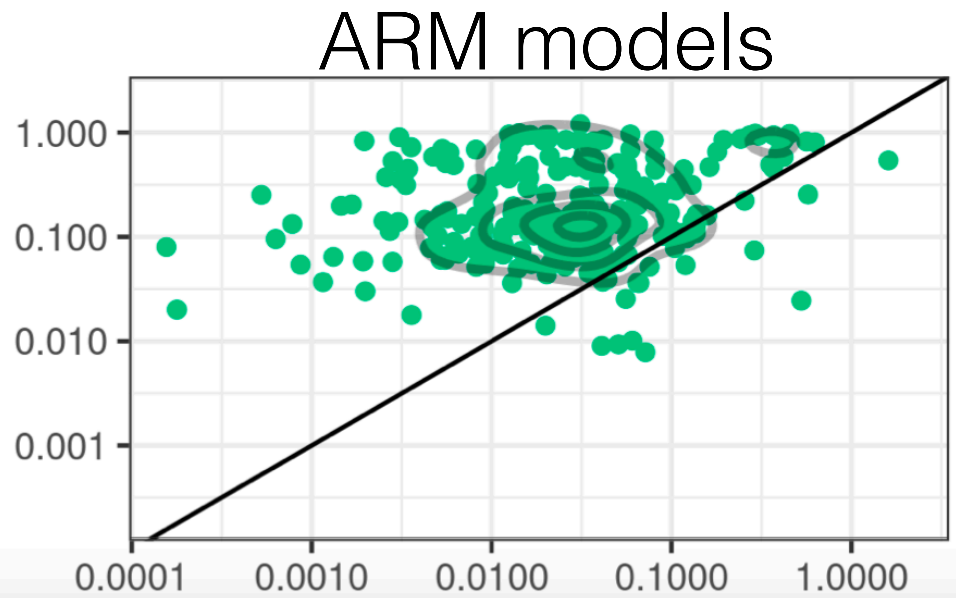


Mean-field ADVI
Welandawe+ 2022
Full-rank ADVI

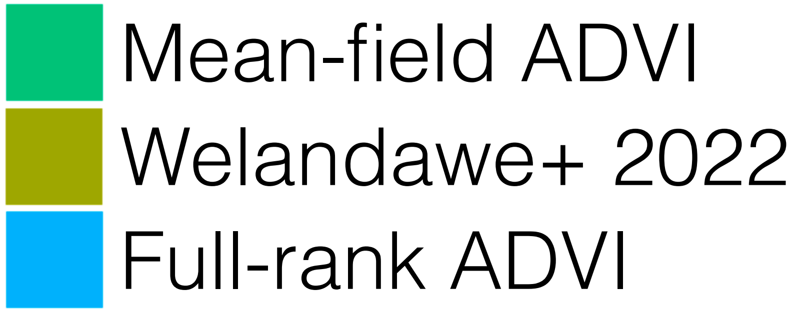
- Point for each parameter (can be a vector)

Quality of posterior variance estimates

- DADVI allows a linear response (LR) correction



DADVI's relative error for posterior standard deviation estimates

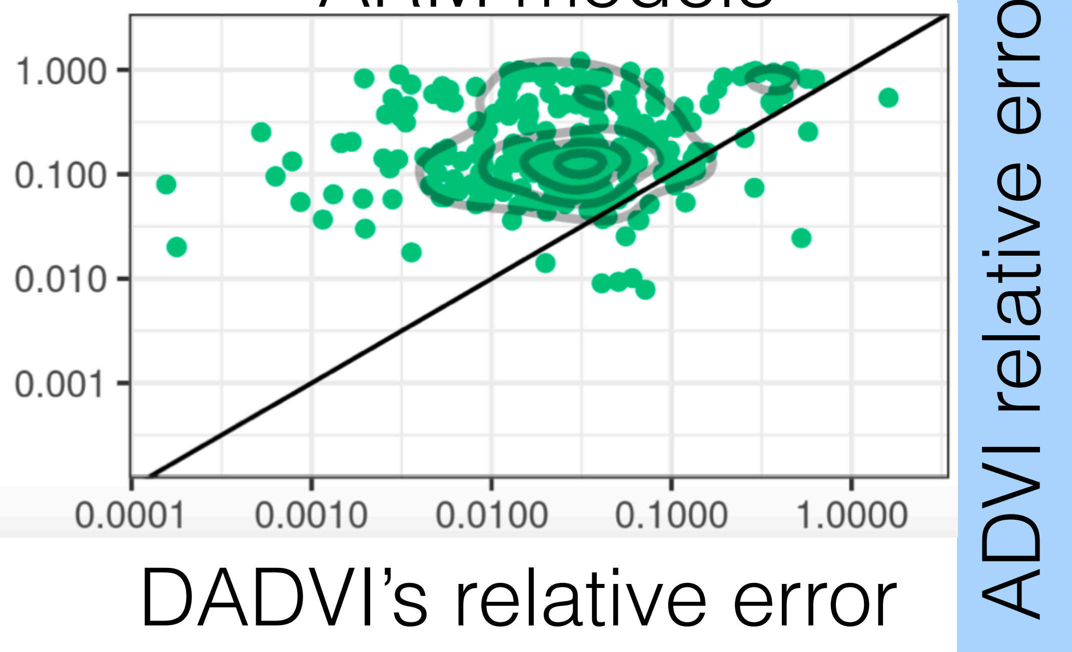


- Point for each parameter (can be a vector)

Quality of posterior variance estimates

- DADVI allows a linear response (LR) correction

ARM models



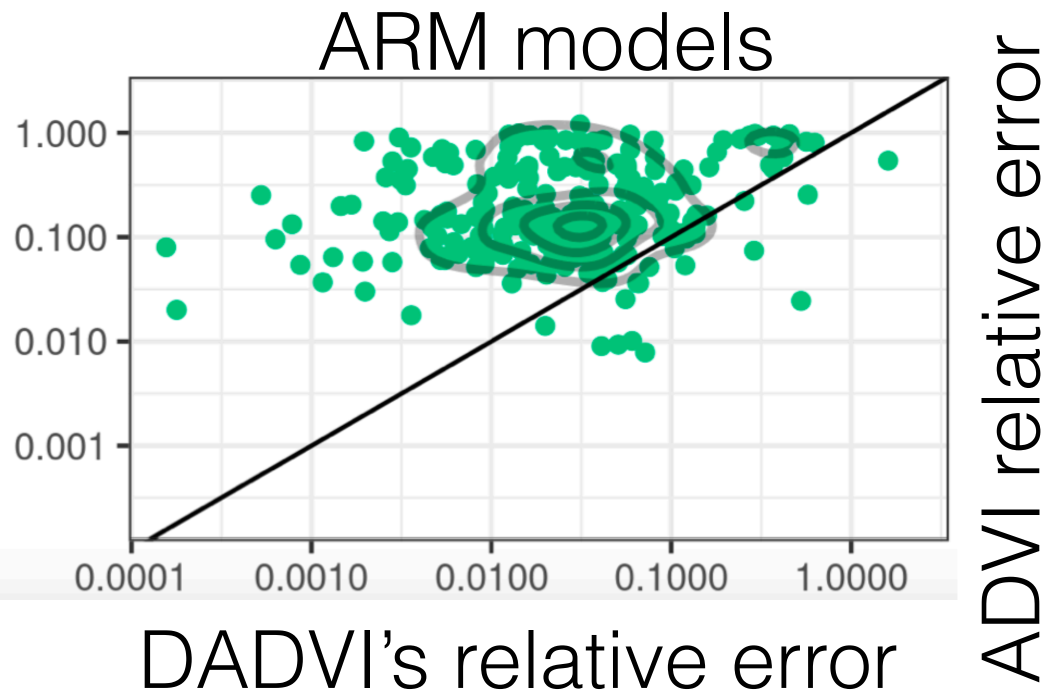
DADVI's relative error for posterior standard deviation estimates

Mean-field ADVI
Welandawe+ 2022
Full-rank ADVI

- Point for each parameter (can be a vector)

Quality of posterior variance estimates

- DADVI allows a linear response (LR) correction



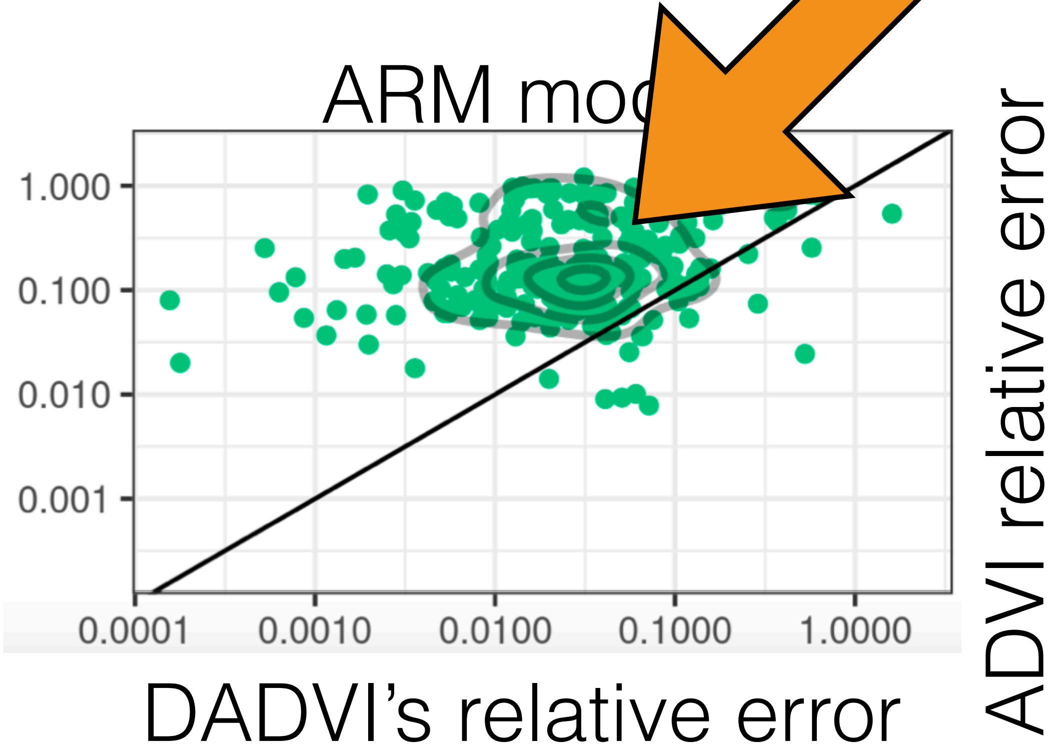
DADVI's relative error for posterior standard deviation estimates

Mean-field ADVI
Welandawe+ 2022
Full-rank ADVI

- Point for each parameter (can be a vector)
- If a point is above the diagonal line, using DADVI + linear response is better

Quality of posterior variance estimates

- DADVI allows a li[...]ear response (LR) correction



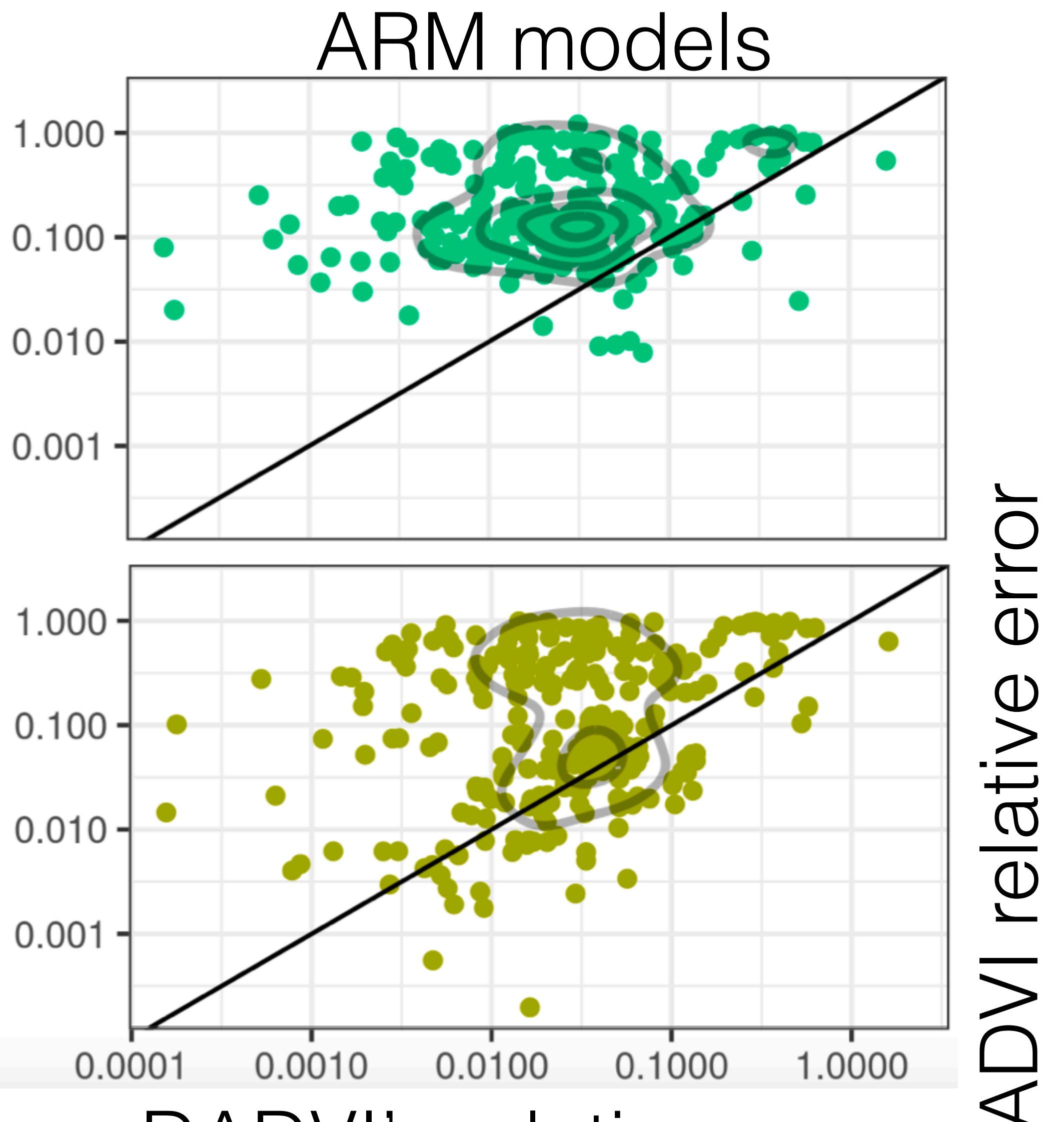
DADVI's relative error for posterior standard deviation estimates

Mean-field ADVI
Welandawe+ 2022
Full-rank ADVI

- Point for each parameter (can be a vector)
- If a point is above the diagonal line, using DADVI + linear response is better

Quality of posterior variance estimates

- DADVI allows a linear response (LR) correction

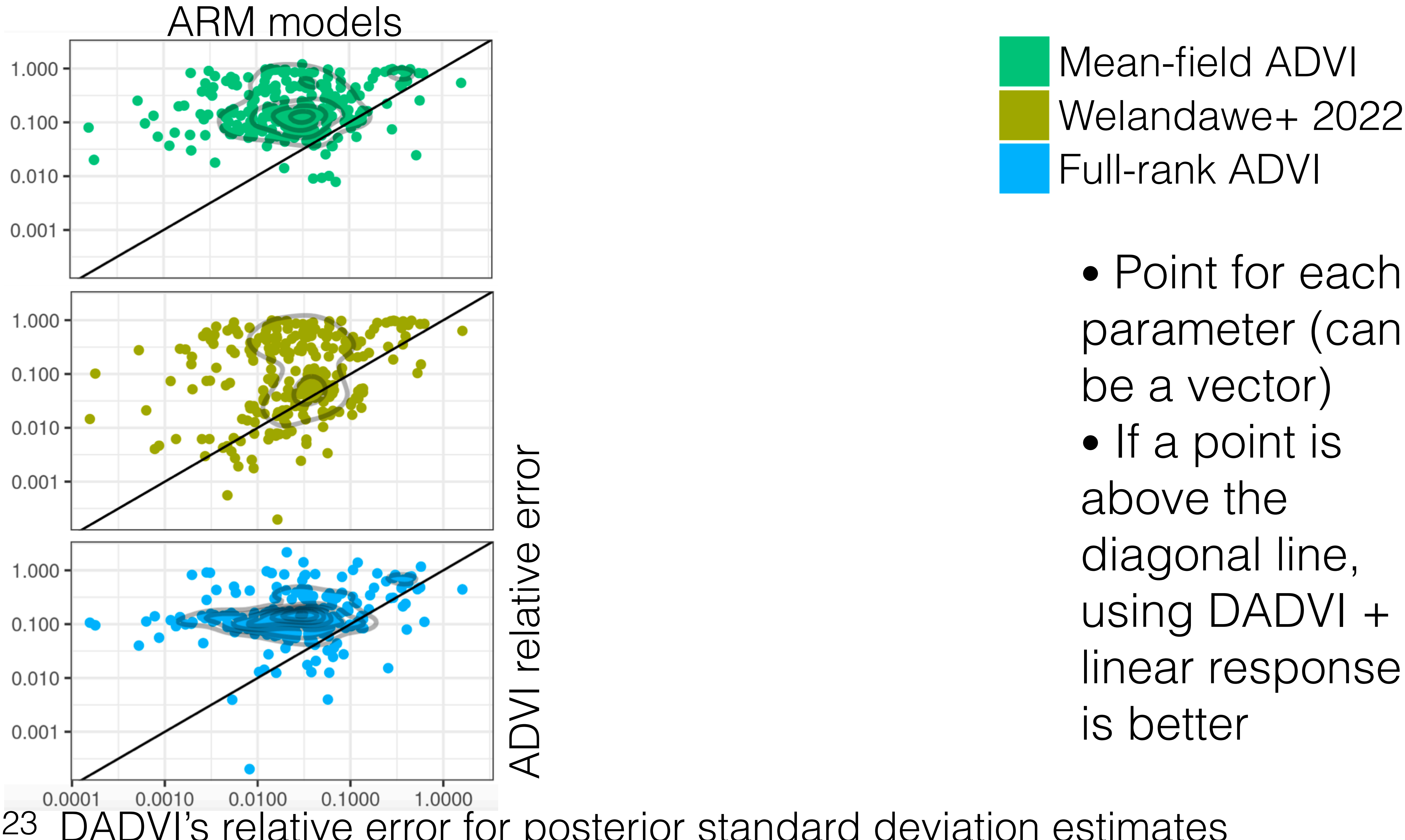


DADVI's relative error for posterior standard deviation estimates

- Point for each parameter (can be a vector)
- If a point is above the diagonal line, using DADVI + linear response is better

Quality of posterior variance estimates

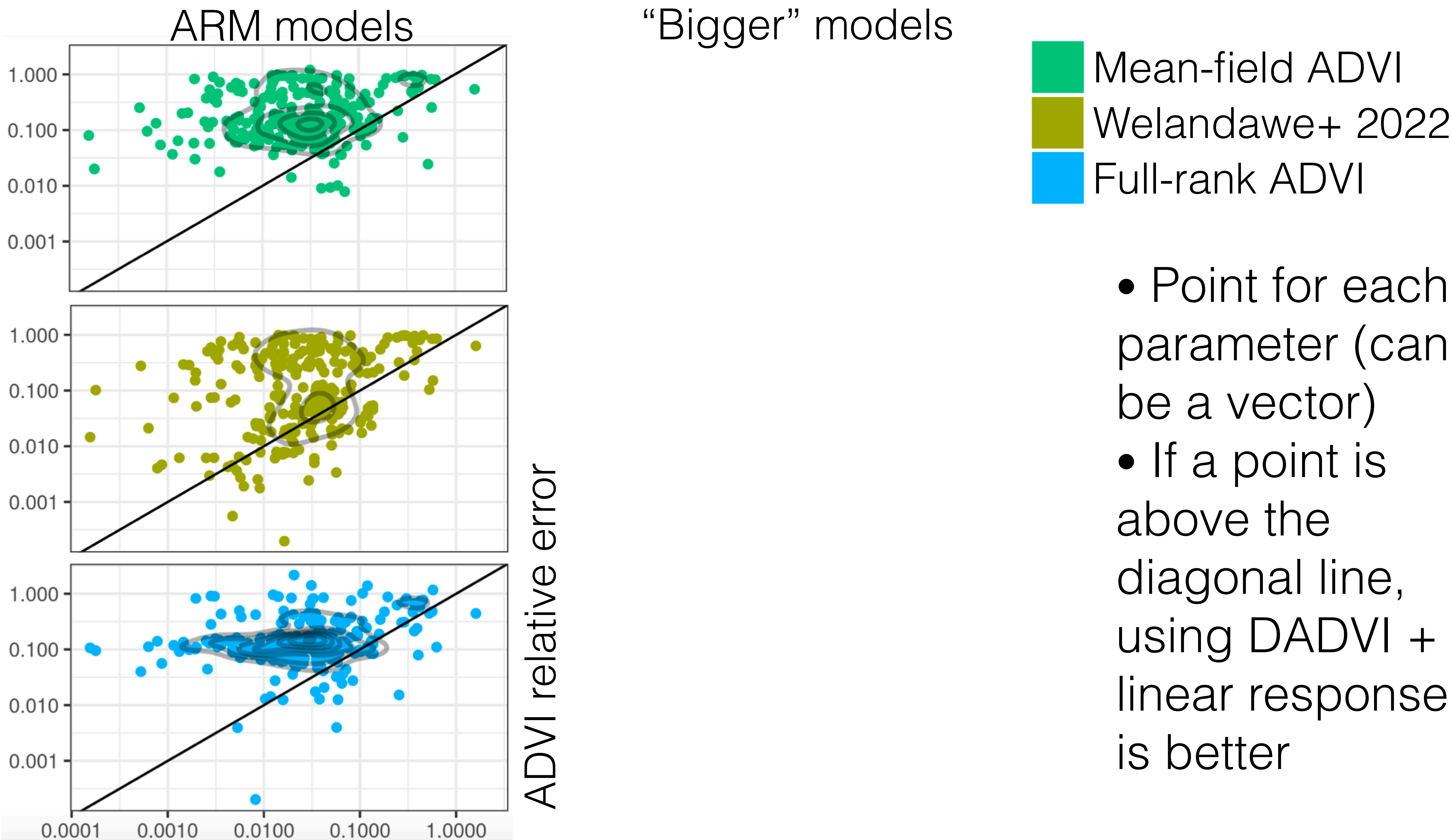
- DADVI allows a linear response (LR) correction



- Point for each parameter (can be a vector)
- If a point is above the diagonal line, using DADVI + linear response is better

Quality of posterior variance estimates

- DADVI allows a linear response (LR) correction

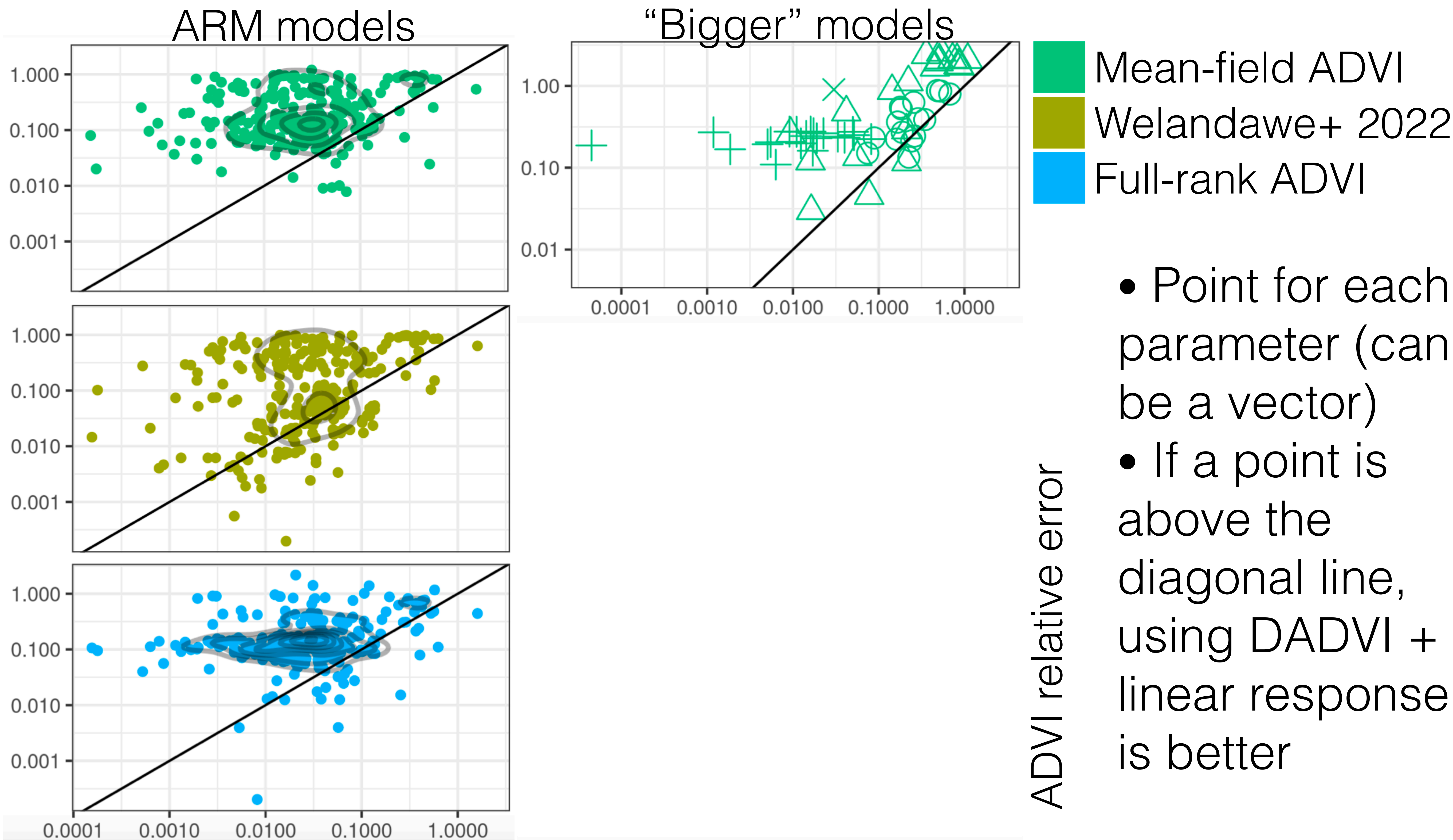


DADVI’s relative error for posterior standard deviation estimates

- Point for each parameter (can be a vector)
- If a point is above the diagonal line, using DADVI + linear response is better

Quality of posterior variance estimates

- DADVI allows a linear response (LR) correction

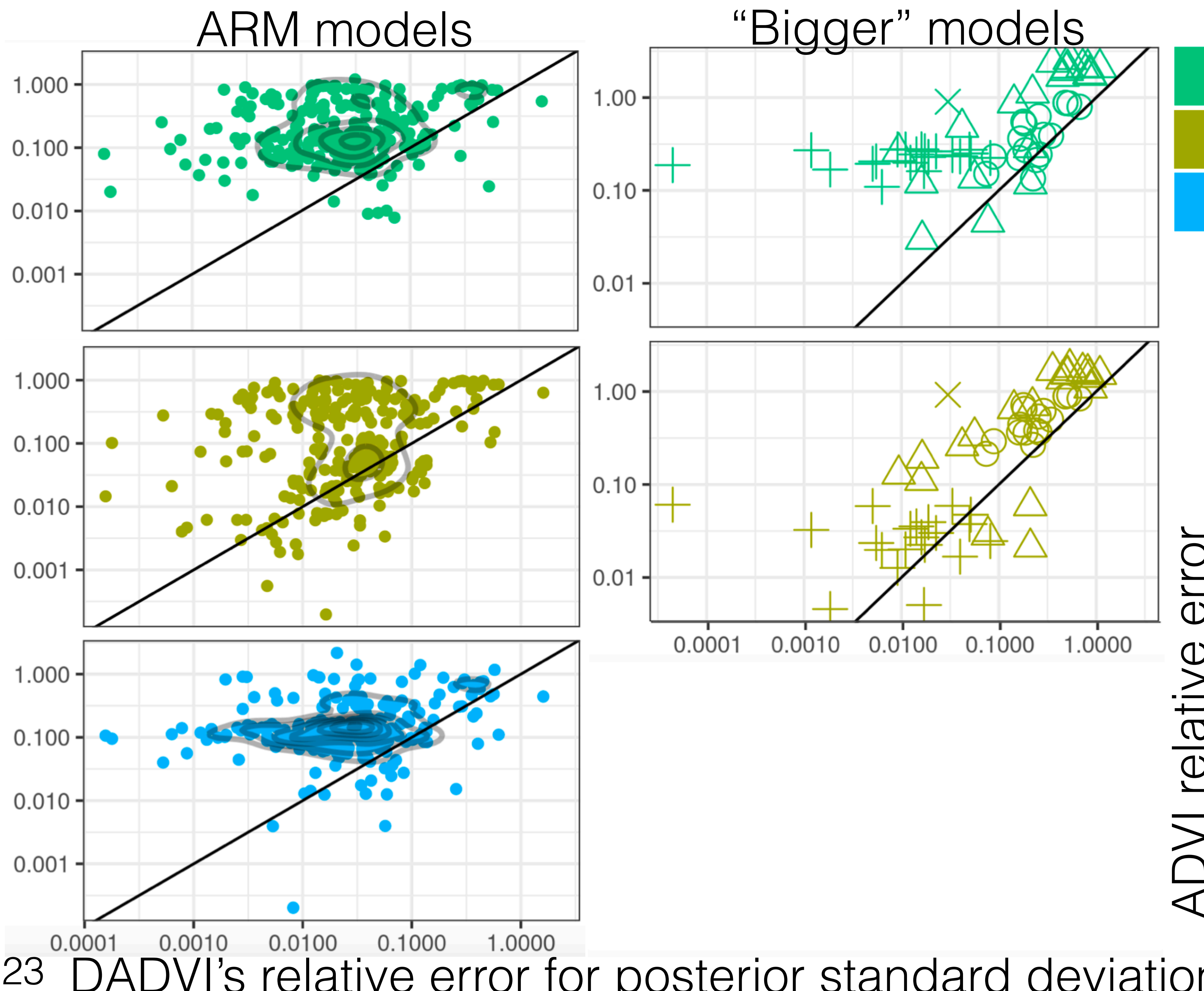


- Point for each parameter (can be a vector)
- If a point is above the diagonal line, using DADVI + linear response is better

 DADVI’s relative error for posterior standard deviation estimates

Quality of posterior variance estimates

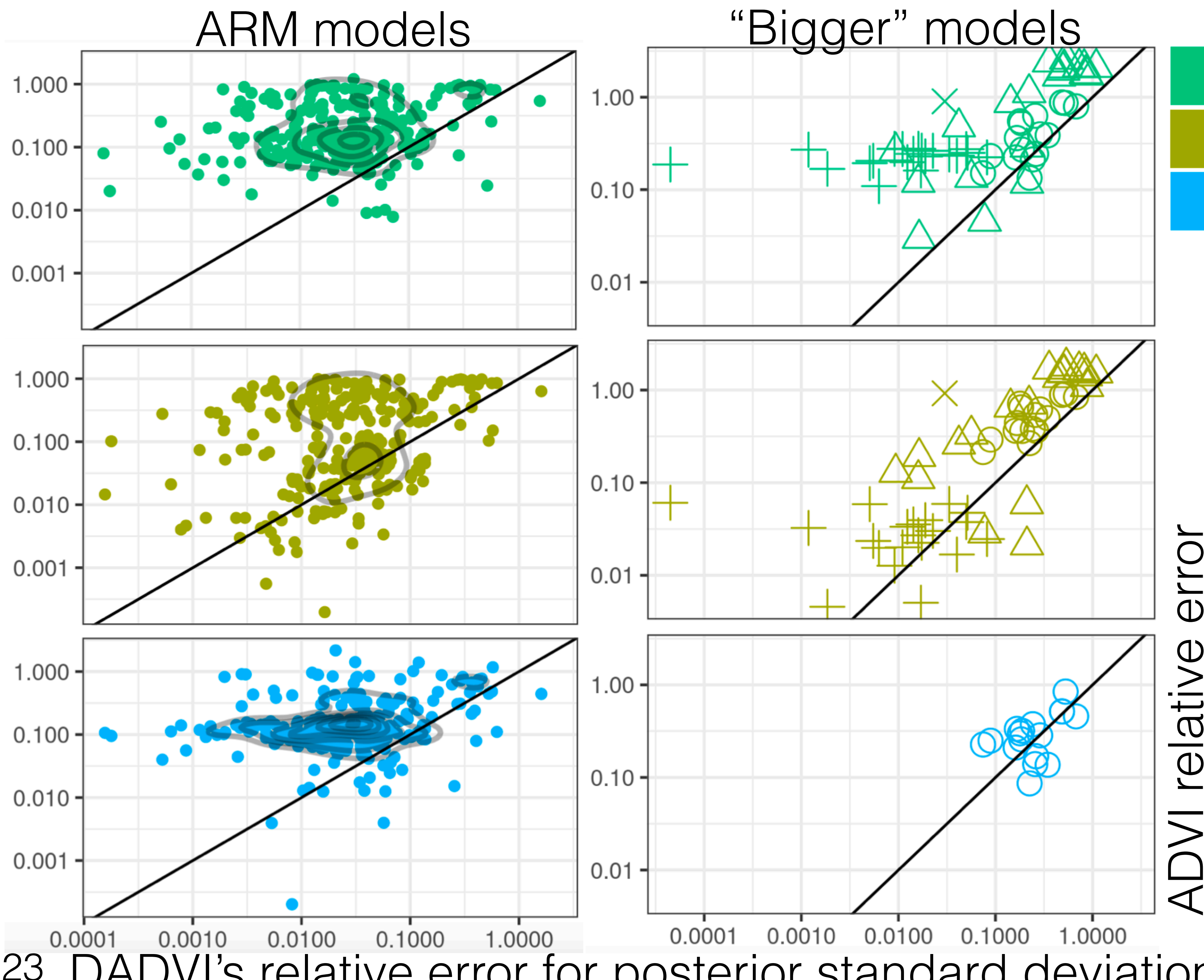
- DADVI allows a linear response (LR) correction



- Point for each parameter (can be a vector)
- If a point is above the diagonal line, using DADVI + linear response is better

Quality of posterior variance estimates

- DADVI allows a linear response (LR) correction



- Point for each parameter (can be a vector)
- If a point is above the diagonal line, using DADVI + linear response is better

Some takeaways

Some takeaways

- Sometimes it takes too long to run MCMC

Some takeaways

- Sometimes it takes too long to run MCMC
- ADVI is a great starting point for using variational Bayes

Some takeaways

- Sometimes it takes too long to run MCMC
- ADVI is a great starting point for using variational Bayes
  - No need for manual derivations; in major languages

Some takeaways

- Sometimes it takes too long to run MCMC
- ADVI is a great starting point for using variational Bayes
  - No need for manual derivations; in major languages
  - Mean-field is the computationally faster option

Some takeaways

- Sometimes it takes too long to run MCMC
- ADVI is a great starting point for using variational Bayes
  - No need for manual derivations; in major languages
  - Mean-field is the computationally faster option
  - If you can run full-rank, can you run MCMC?

Some takeaways

- Sometimes it takes too long to run MCMC
- ADVI is a great starting point for using variational Bayes
  - No need for manual derivations; in major languages
  - Mean-field is the computationally faster option
  - If you can run full-rank, can you run MCMC?
- VB is known to typically struggle with estimating posterior variances. Sometimes it has trouble with posterior means

Some takeaways

- Sometimes it takes too long to run MCMC
- ADVI is a great starting point for using variational Bayes
  - No need for manual derivations; in major languages
  - Mean-field is the computationally faster option
  - If you can run full-rank, can you run MCMC?
- VB is known to typically struggle with estimating posterior variances. Sometimes it has trouble with posterior means
- DADVI can help with: (1) More reliable optimization, (2) No tuning, (3) Clear convergence. Specific to our work: (4) Posterior (co)variance corrections via linear response, (5) Estimates of sampling variability, (6) Small # of samples, (7) Sensitivity

Some takeaways

- Sometimes it takes too long to run MCMC
- ADVI is a great starting point for using variational Bayes
  - No need for manual derivations; in major languages
  - Mean-field is the computationally faster option
  - If you can run full-rank, can you run MCMC?
- VB is known to typically struggle with estimating posterior variances. Sometimes it has trouble with posterior means
- DADVI can help with: (1) More reliable optimization, (2) No tuning, (3) Clear convergence. Specific to our work: (4) Posterior (co)variance corrections via linear response, (5) Estimates of sampling variability, (6) Small # of samples, (7) Sensitivity
  - In any case, it's worth being aware of ADVI challenges

What to read/do next

Textbooks and Reviews

- Murphy. *Probabilistic Machine Learning: Advanced Topics*, Ch 10. 2023.
- Bishop. *Pattern Recognition and Machine Learning*, Ch 10. 2006.
- Blei, Kucukelbir, McAuliffe. Variational inference: A review for statisticians, *JASA* 2017.
- MacKay. *Information Theory, Inference, and Learning Algorithms*, Ch 33. 2003.
- Ormerod, Wand. Explaining variational approximations. *Amer Stat* 2010.

Do the exercises, and try it out!

- ADVI is a great place to start

Example Languages

- PyMC, Stan, Edward

Refs for Experiments Etc.

- R Giordano, T Broderick, and MI Jordan. Covariances, robustness, and variational Bayes. *JMLR* 2018.
- R Giordano*, M Ingram*, and T Broderick. Black Box Variational Inference with a Deterministic Objective: Faster, More Accurate, and Even More Black Box. *JMLR* 2024. (ArXiv 2023. *equal contribution)
- Burroni, Domke, Sheldon. Sample Average Approximation for Black-Box VI. ArXiv 2023.
- R Giordano, T Broderick, and MI Jordan. Linear response methods for accurate covariance estimates from mean field variational Bayes. *NeurIPS* 2015.
- R Giordano, T Broderick, R Meager, JH Huggins, and MI Jordan. Fast robustness quantification with variational Bayes. *ICML Data4Good Workshop* 2016.
- J Huggins, M Kasprzak, T Campbell, T Broderick. Validated Variational Inference via Practical Posterior Error Bounds. *AISTATS* 2020.

References (1/6)

R Bardenet, A Doucet, and C Holmes. "On Markov chain Monte Carlo methods for tall data." *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." *Journal of Machine Learning Research*, 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. *NeurIPS* 2013.

J Burroni, J Domke, D Sheldon. Sample Average Approximation for Black-Box VI. ArXiv:2304.06803, 2023.

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

BK Fosdick. *Modeling Heterogeneity within and between Matrices and Arrays*. Doctoral dissertation, 2013.

R Giordano, T Broderick, and MI Jordan. "Linear response methods for accurate covariance estimates from mean field variational Bayes." *NeurIPS* 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. *Journal of Machine Learning Research*, 2018.

References (2/6)

R Giordano*, M Ingram*, and T Broderick. Black Box Variational Inference with a Deterministic Objective: Faster, More Accurate, and Even More Black Box. *Journal of Machine Learning Research*, 2024. (ArXiv:2304.05527, 2023. *equal contribution)

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.

J Huggins, T Campbell, M Kasprzak, T Broderick. Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach, 2018. ArXiv:1809.09505.

J Huggins, M Kasprzak, T Campbell, T Broderick. Validated Variational Inference via Practical Posterior Error Bounds. *AISTATS* 2020.

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

A Kucukelbir, D Tran, R Ranganath, A Gelman, and DM Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 2017.

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

KP Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

References (3/6)

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

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

B Wang and M Titterington. Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Workshop on Artificial Intelligence and Statistics*, 2004.

M Welandawe, MR Andersen, A Vehtari, and JH Huggins. "Robust, automated, and accurate black-box variational inference." ArXiv:2203.15945 (2022).

Y Yao, A Vehtari, D Simpson, and A Gelman. Yes, but Did It Work?: Evaluating Variational Inference. *ICML* 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.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation." *Journal of Machine Learning Research*. 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.

Event Horizon Telescope Collaboration et al, *The Astrophysical Journal Letters* 930 L12, 2022.

Fletcher, S. The First Milky Way Black Hole Image Lets Scientists Test Physics. *Scientific American*, Sept 2022.

Fred Hutch News Service Staff. “Science papers you should be reading about the coronavirus.” March 2020. https://www.fredhutch.org/en/news/center-news/2020/03/coronavirus-latest-scientific-research.html. Accessed 2020 March 21.

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)

Heidemanns, Merlin, Andrew Gelman, and G. Elliott Morris. "An updated dynamic Bayesian forecasting model for the US presidential election." *Harvard Data Science Review* 2.4 (2020): 10-1162.

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.

McMahan, Peter, and Daniel A. McFarland. "Creative destruction: the structural consequences of scientific curation." *American Sociological Review* 86.2 (2021): 341-376.

Meager, Rachael. “Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of 7 randomized experiments.” *AEJ: Applied*, 2019.

Meager, Rachael. “Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature.” *American Economic Review*, 2022.

Poorter, Lourens, et al. "Multidimensional tropical forest recovery." *Science* 374.6573 (2021): 1370-1376.

Spertus, John A., et al. "Health-status outcomes with invasive or conservative care in coronary disease." *New England Journal of Medicine* 382.15 (2020): 1408-1419.

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.

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)

A. Kongrut. 23 Jan 2020. Obtained from: https://www.bangkokpost.com/opinion/opinion/1841569/bungling-govt-is-losing-the-pm2-5-war