Covariances, Robustness, and Variational Bayes

Tamara Broderick
ITT Career Development
Assistant Professor,
MIT

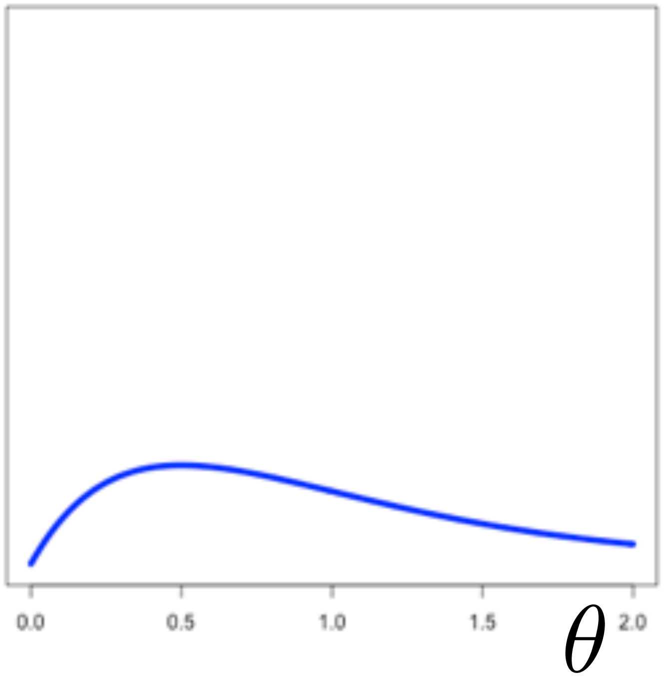
With: Ryan Giordano, Rachael Meager,
Jonathan H. Huggins, Michael I. Jordan

http://www.tamarabroderick.com/tutorials.html

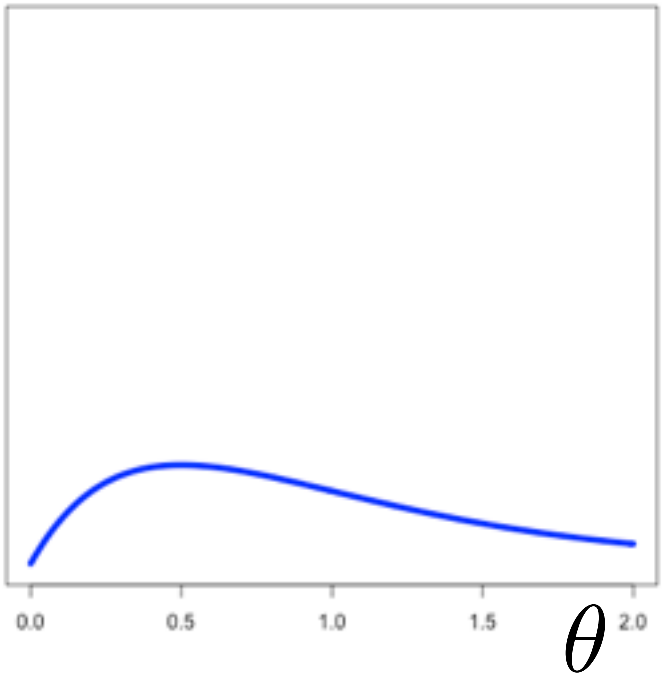
- Bayesian inference

- Bayesian inference $p(\theta)$

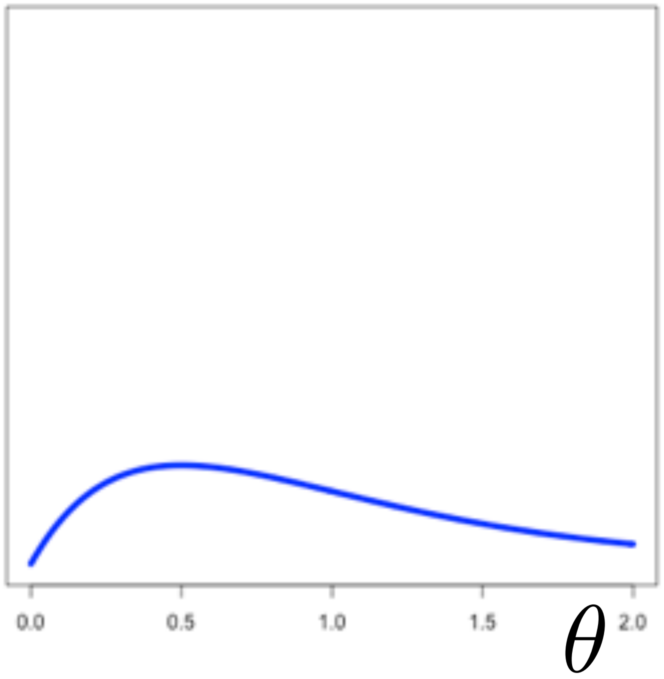
- Bayesian inference $p(\theta)$



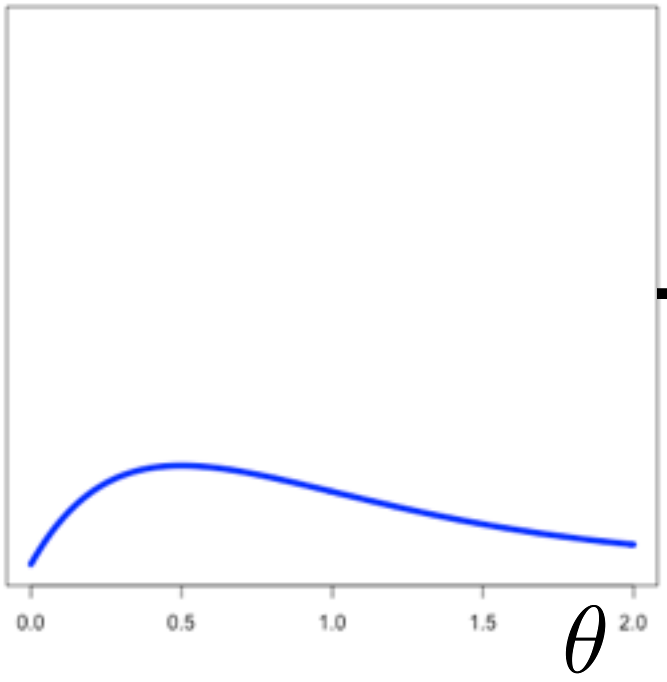
- Bayesian inference $p(y|\theta)p(\theta)$



- Bayesian inference $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$

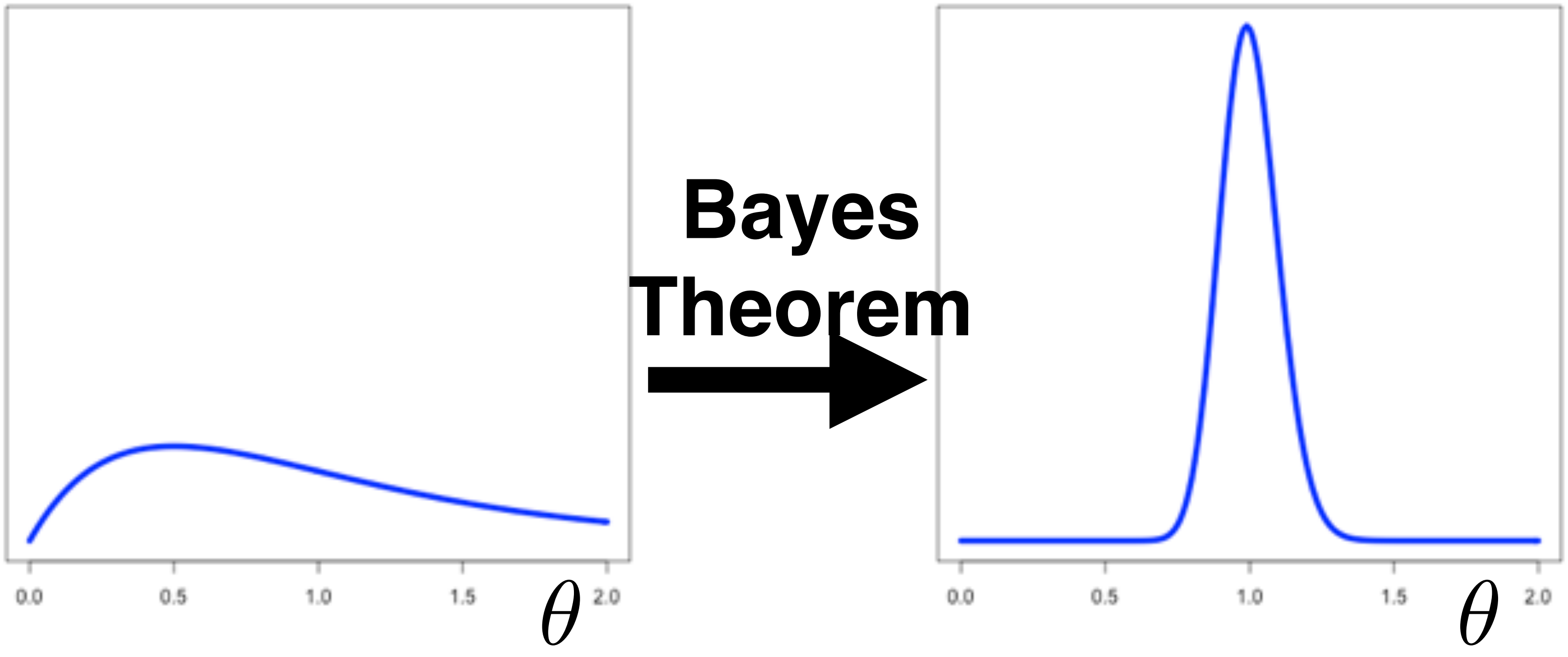


- Bayesian inference $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$

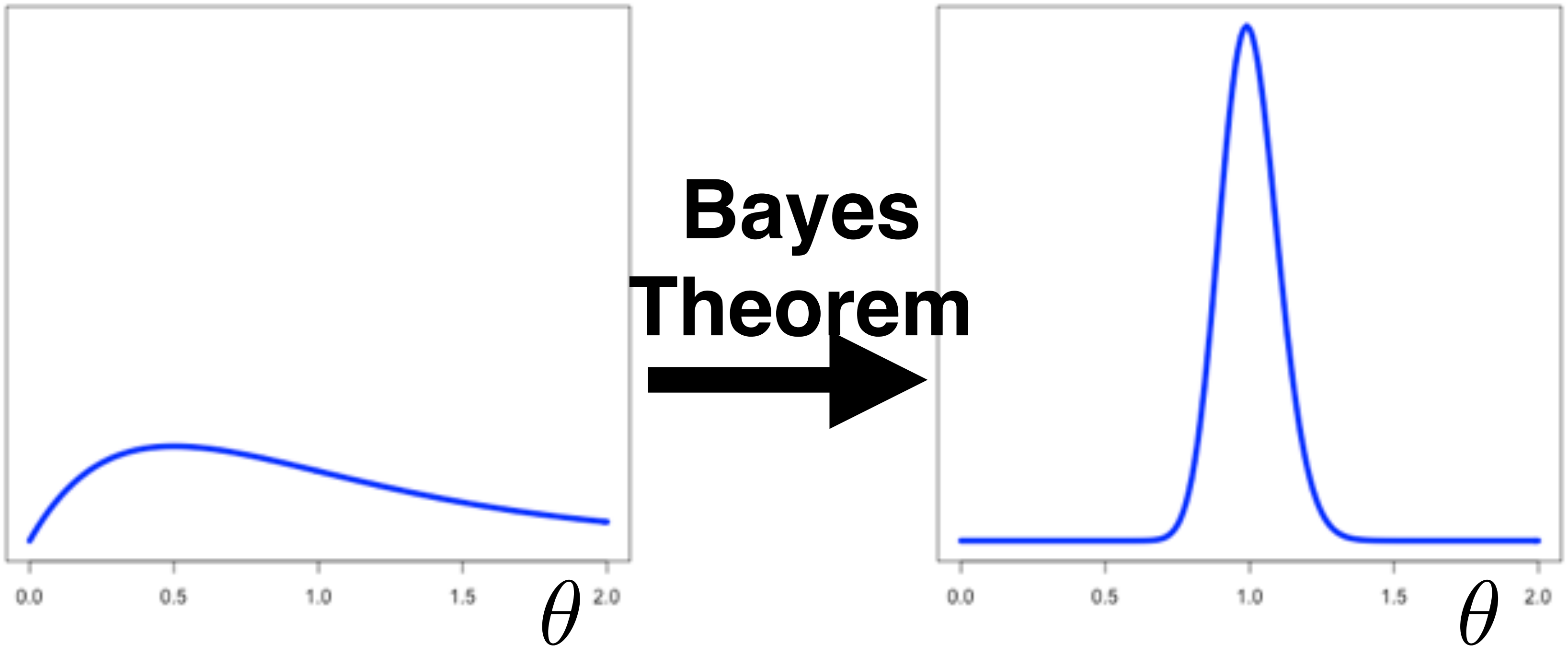


Bayes Theorem →

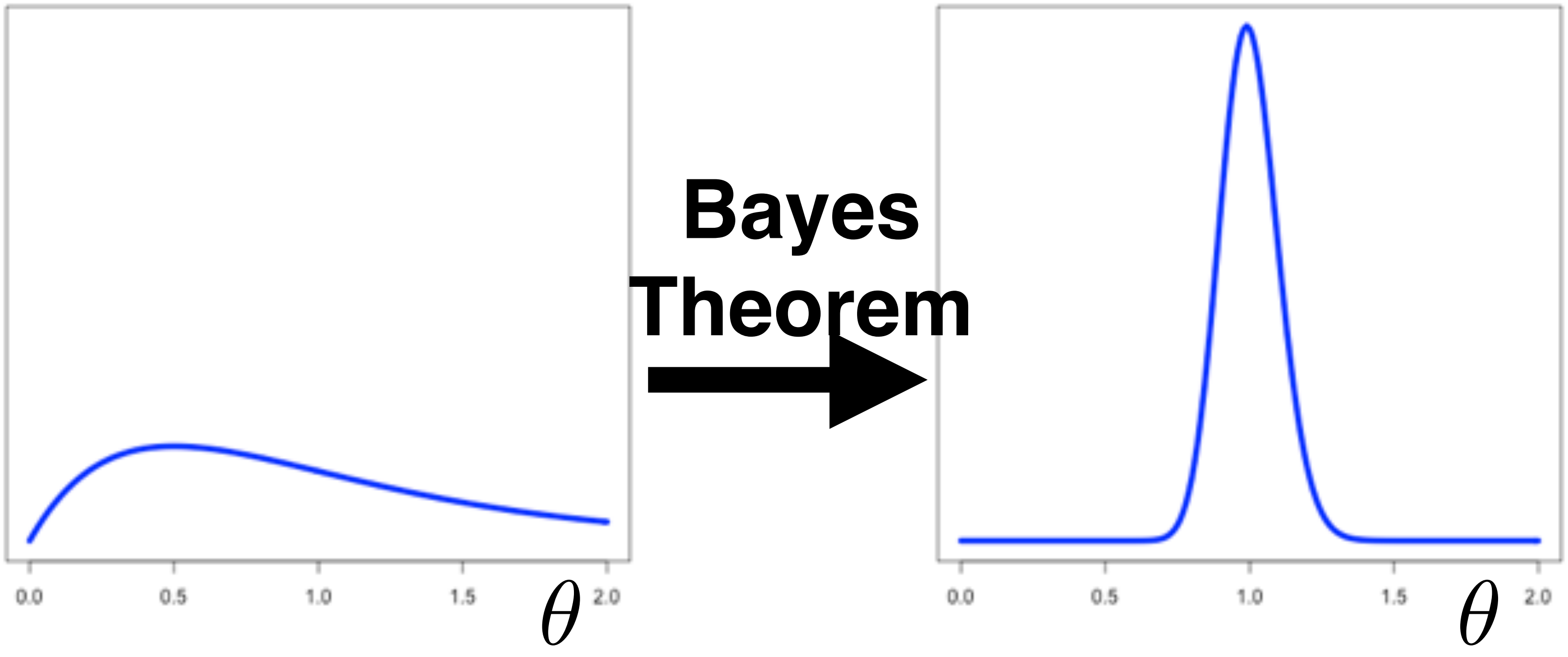
- Bayesian inference $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$



- Bayesian inference $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Challenge: Express knowledge in a distribution (prior, likelihood)

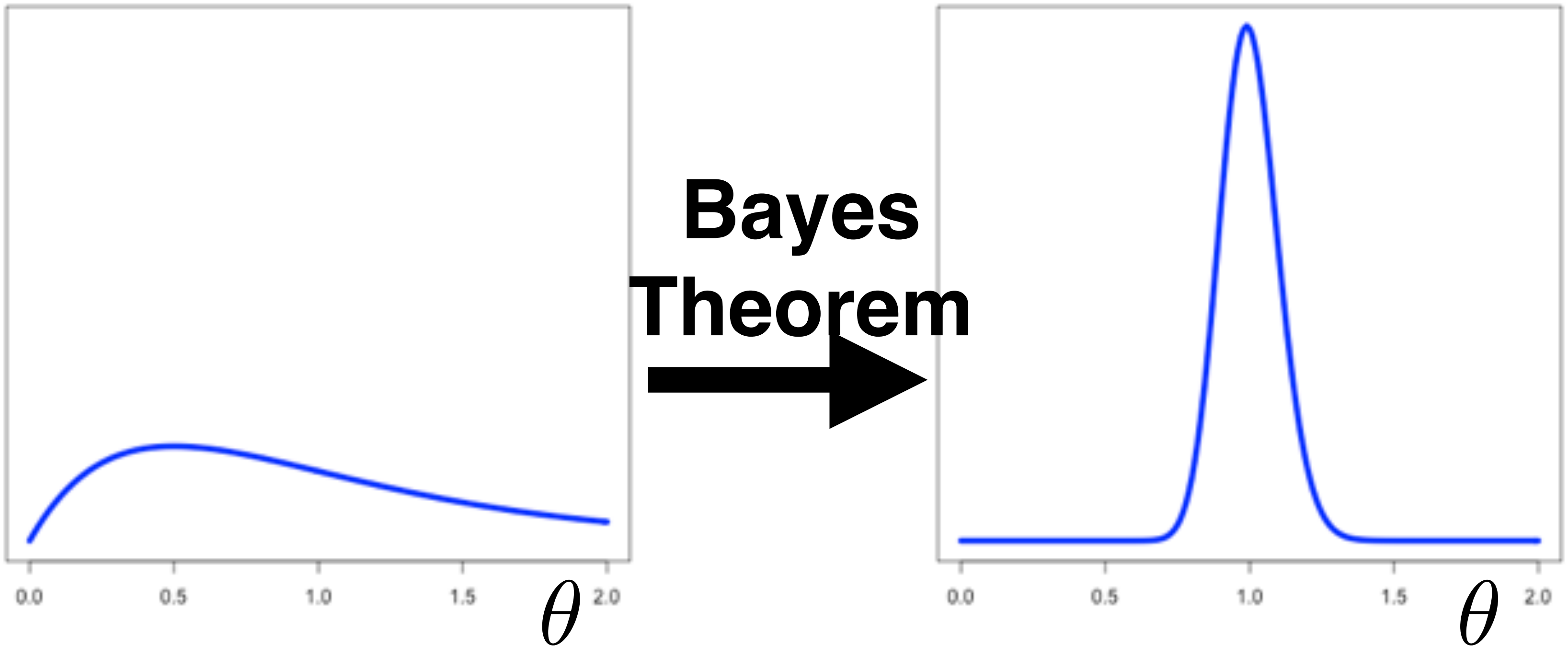


- Bayesian inference $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Challenge: Express knowledge in a distribution (prior, likelihood)
    - Time-consuming



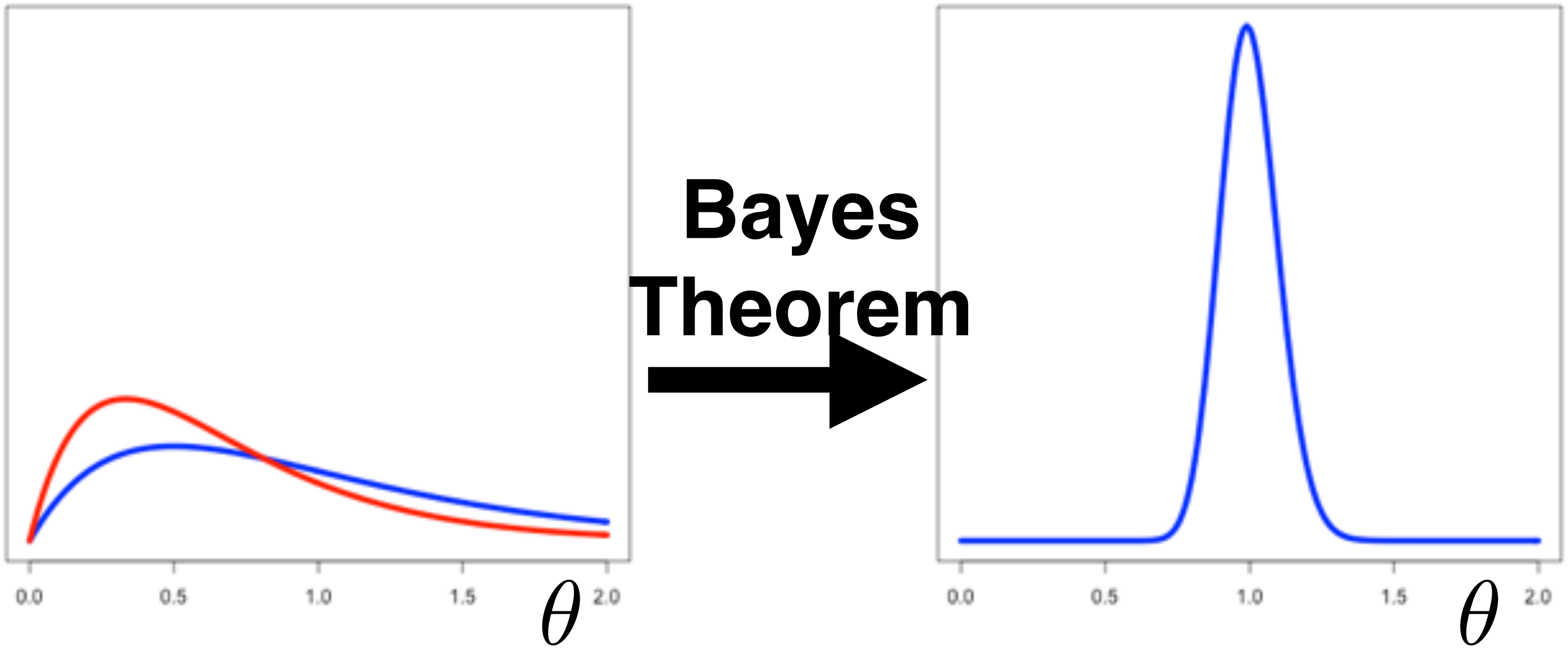
Bayes Theorem

- Bayesian inference $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Challenge: Express knowledge in a distribution (prior, likelihood)
  - Time-consuming; subjective



- Bayesian inference $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Challenge: Express knowledge in a distribution (prior, likelihood)
    - Time-consuming; subjective

Some reasonable priors

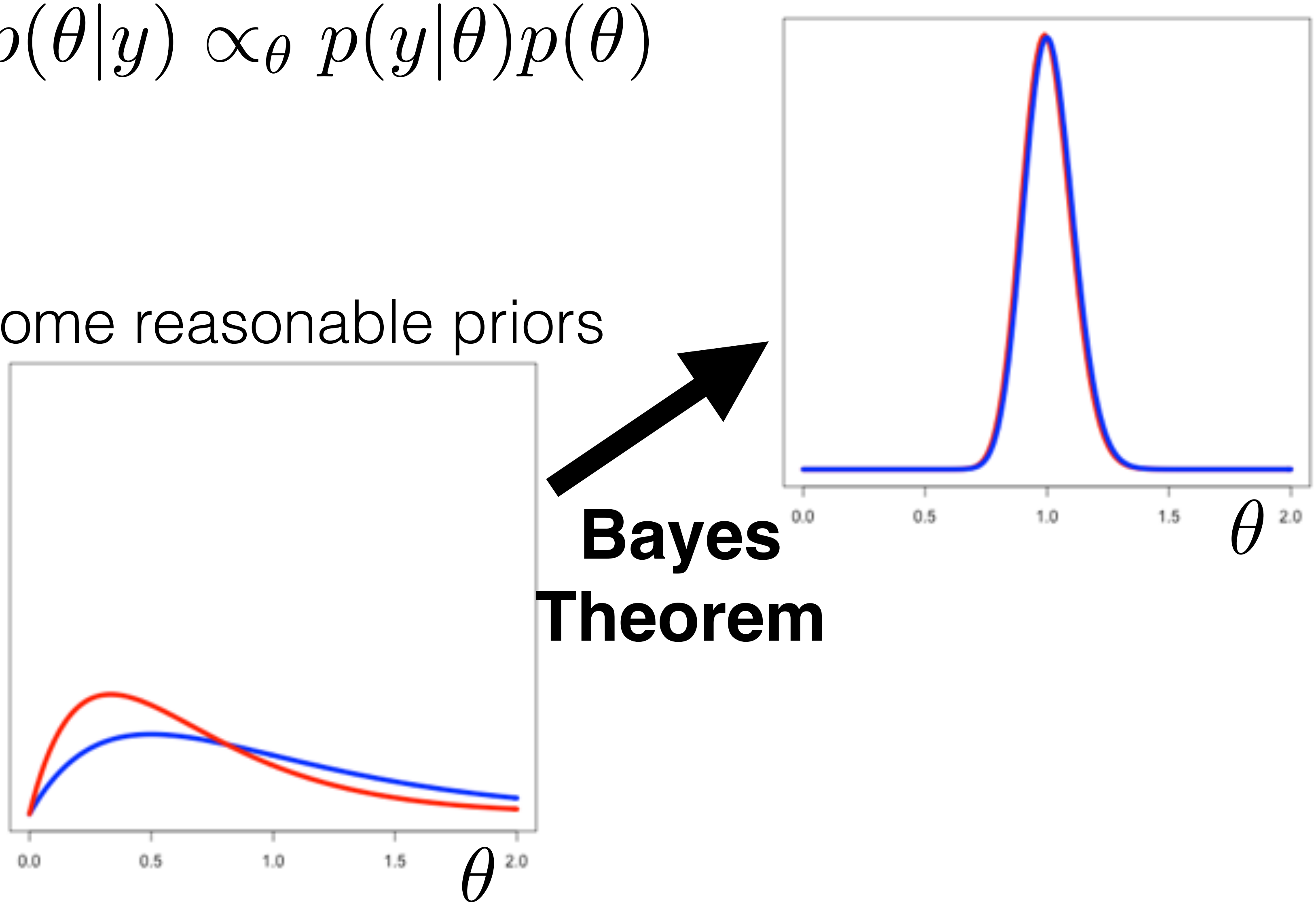


Bayes Theorem

- Bayesian inference

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

- Challenge: Express knowledge in a distribution (prior, likelihood)
  - Time-consuming; subjective

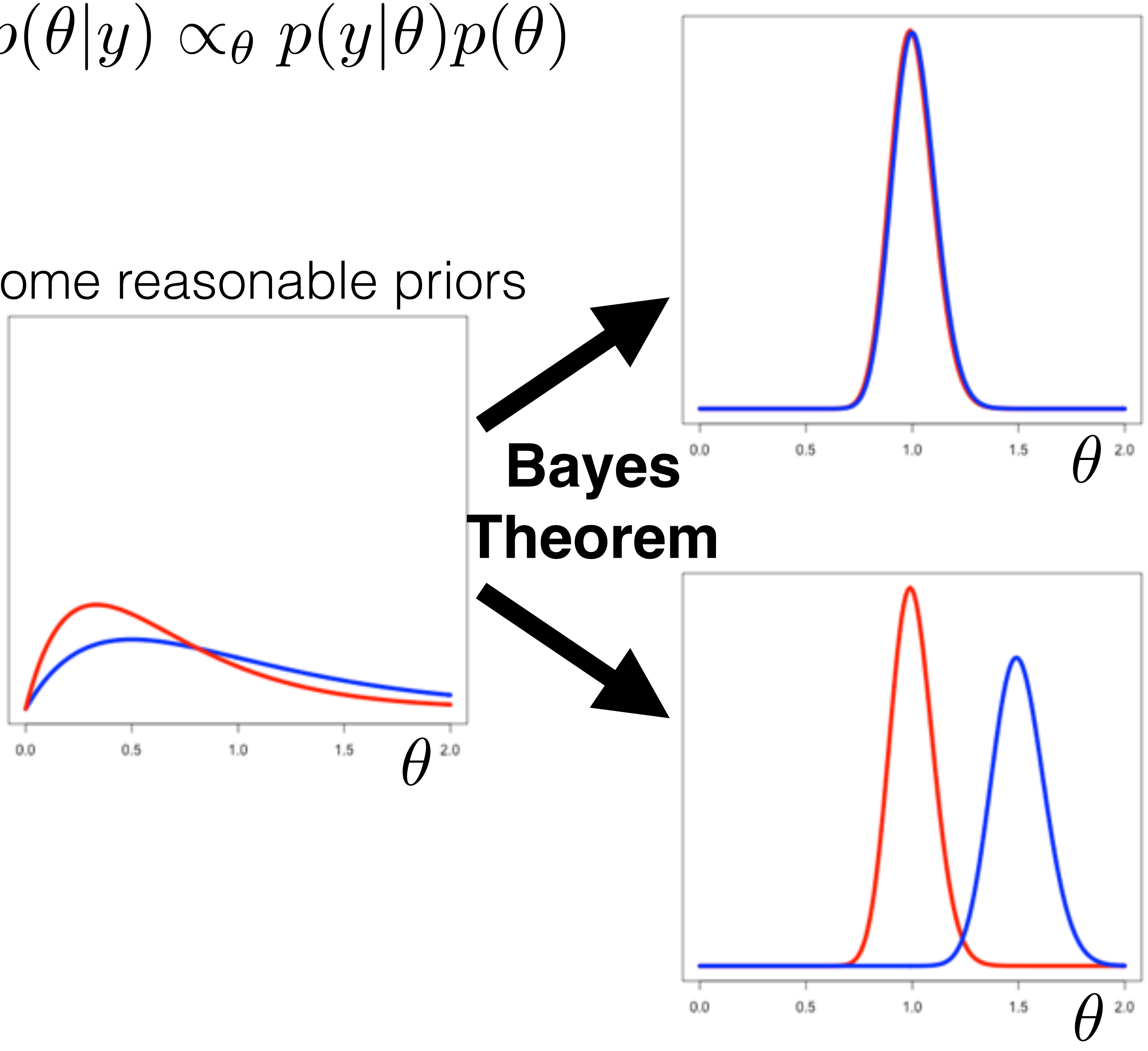


Some reasonable priors

Bayes Theorem

- Bayesian inference
- Challenge: Express knowledge in a distribution (prior, likelihood)
    - Time-consuming; subjective

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



Some reasonable priors

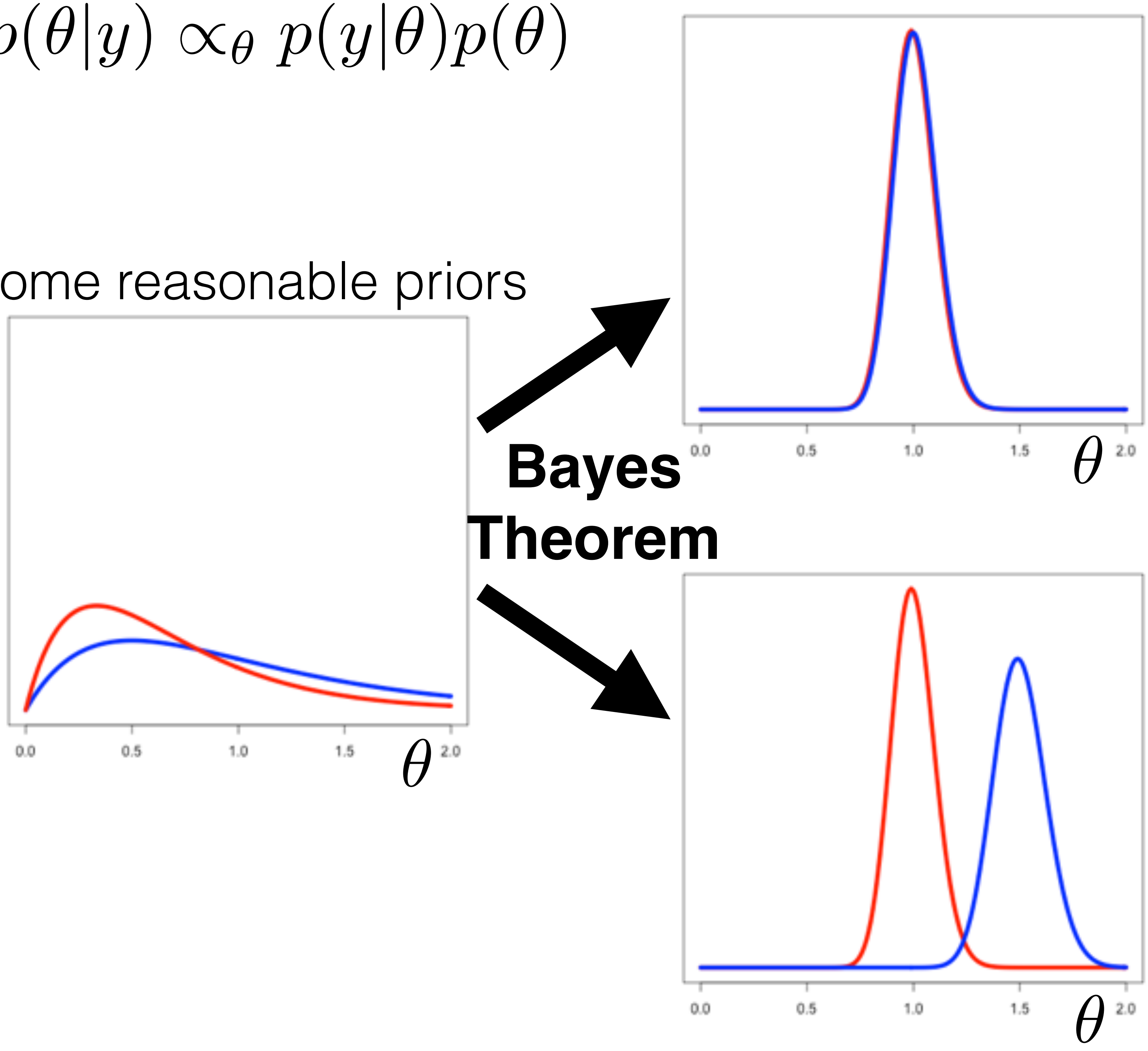
Bayes Theorem

- Bayesian inference

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

- Challenge: Express knowledge in a distribution (prior, likelihood)
    - Time-consuming; subjective; complex models

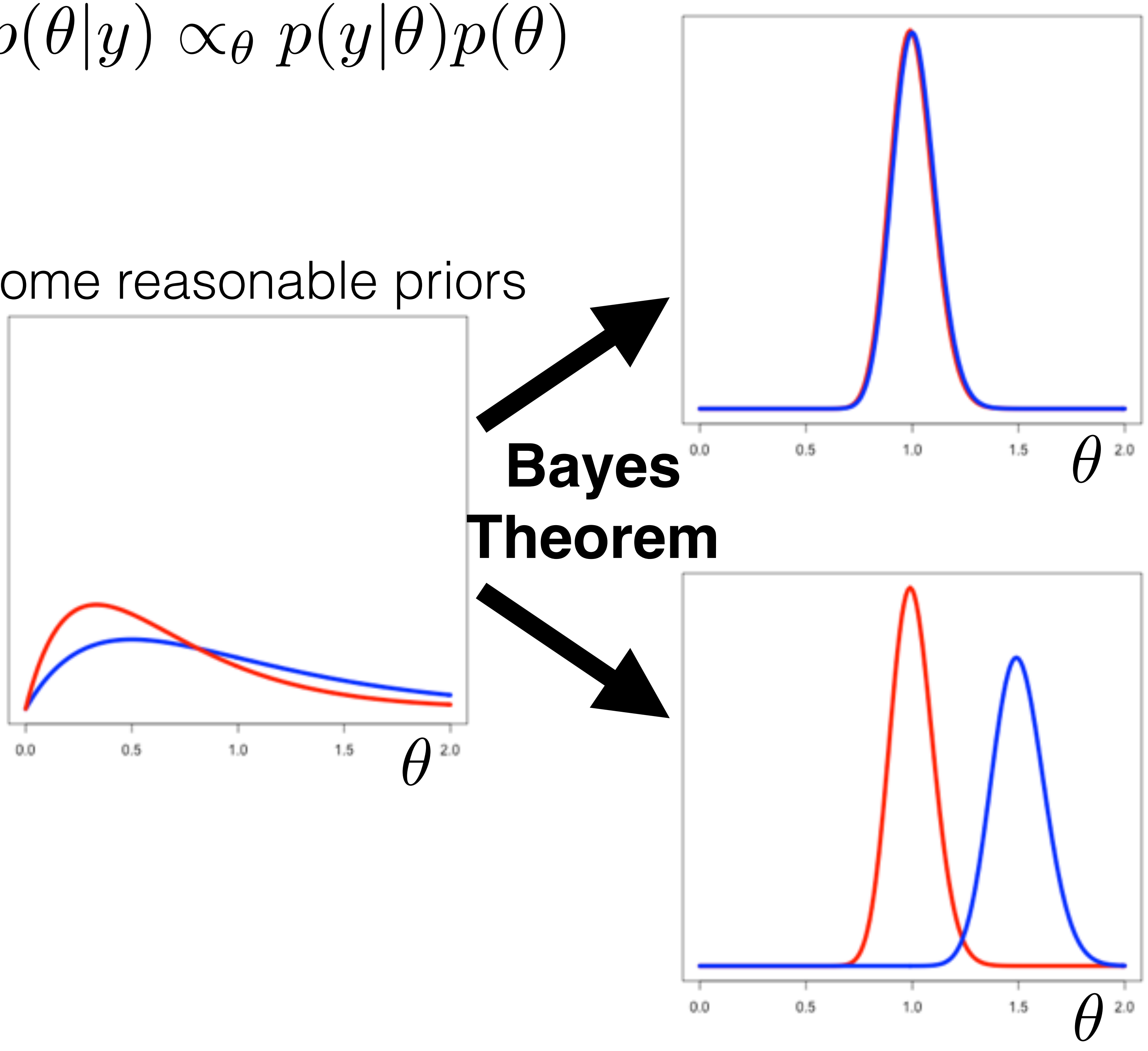
Some reasonable priors



Bayes Theorem

robustness quantification

- Bayesian inference $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Challenge: Express knowledge in a distribution (prior, likelihood)
  - Time-consuming; subjective; complex models

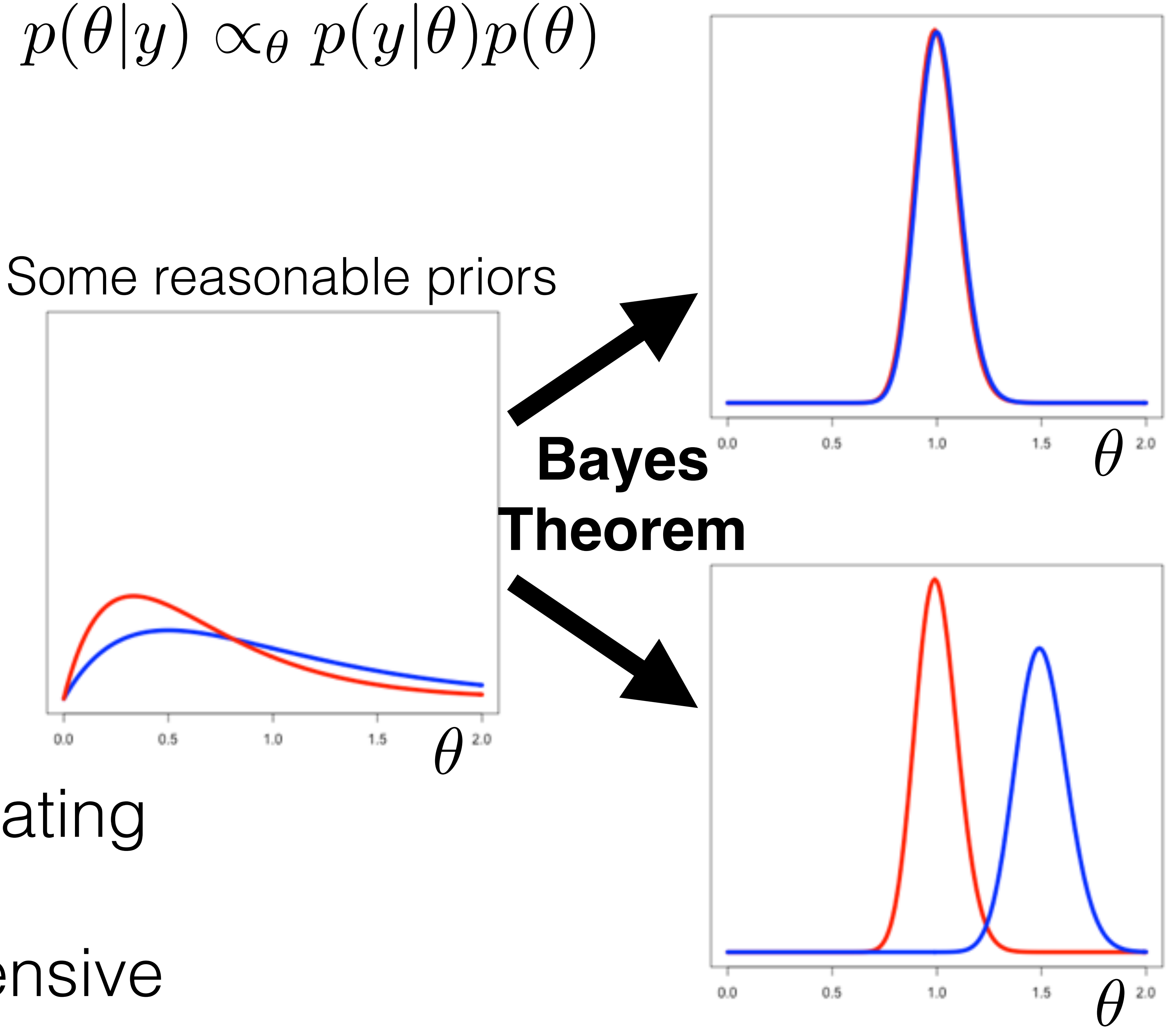


Some reasonable priors

Bayes Theorem

robustness quantification

- Bayesian inference $p(\theta|y) \propto_{\theta} p(y|\theta)p(\theta)$
- Challenge: Express knowledge in a distribution (prior, likelihood)
  - Time-consuming; subjective; complex models
- Challenge: Approximating the posterior can be computationally expensive

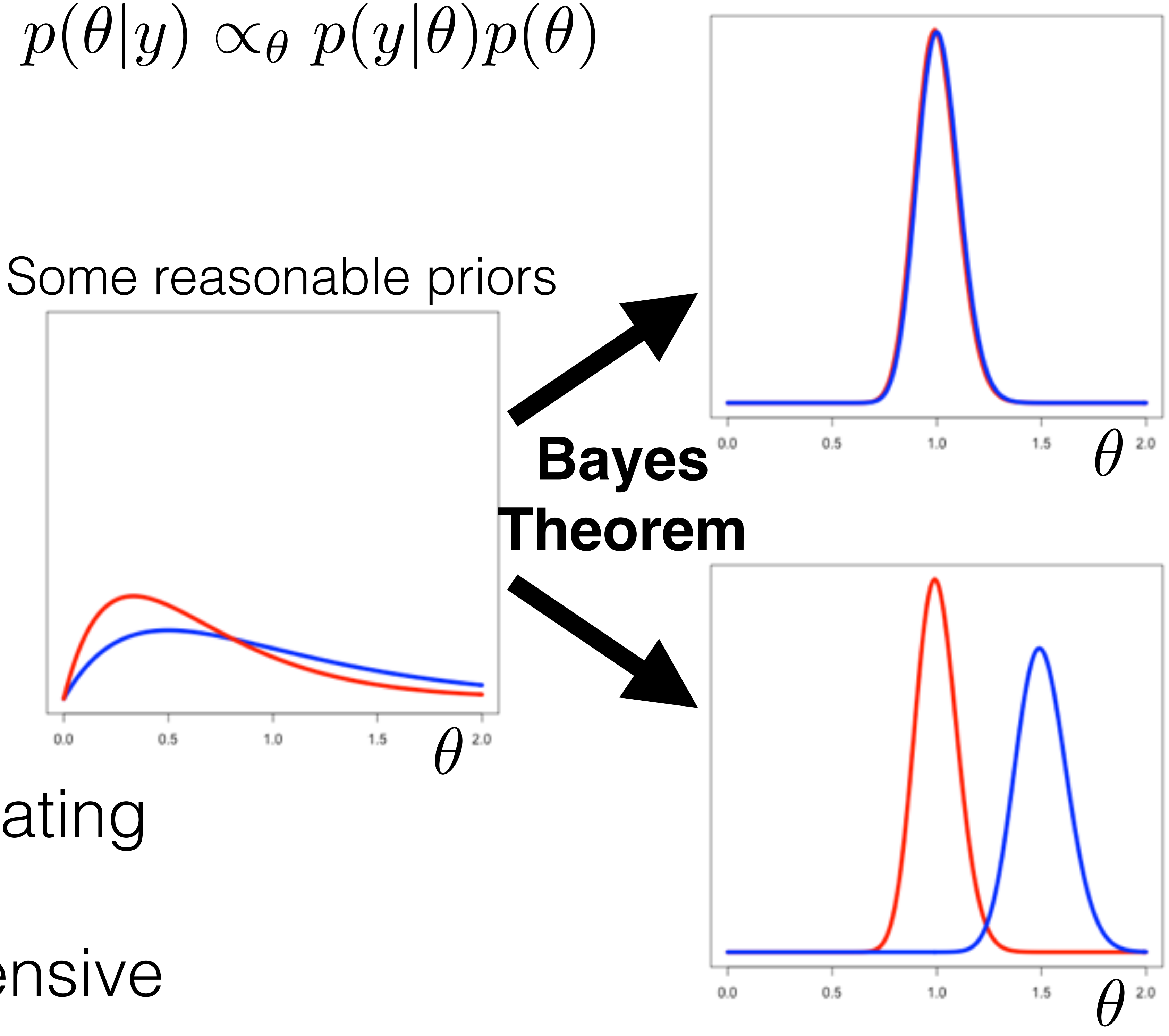


Some reasonable priors

Bayes Theorem

robustness quantification

- Bayesian inference $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Challenge: Express knowledge in a distribution (prior, likelihood)
  - Time-consuming; subjective; complex models



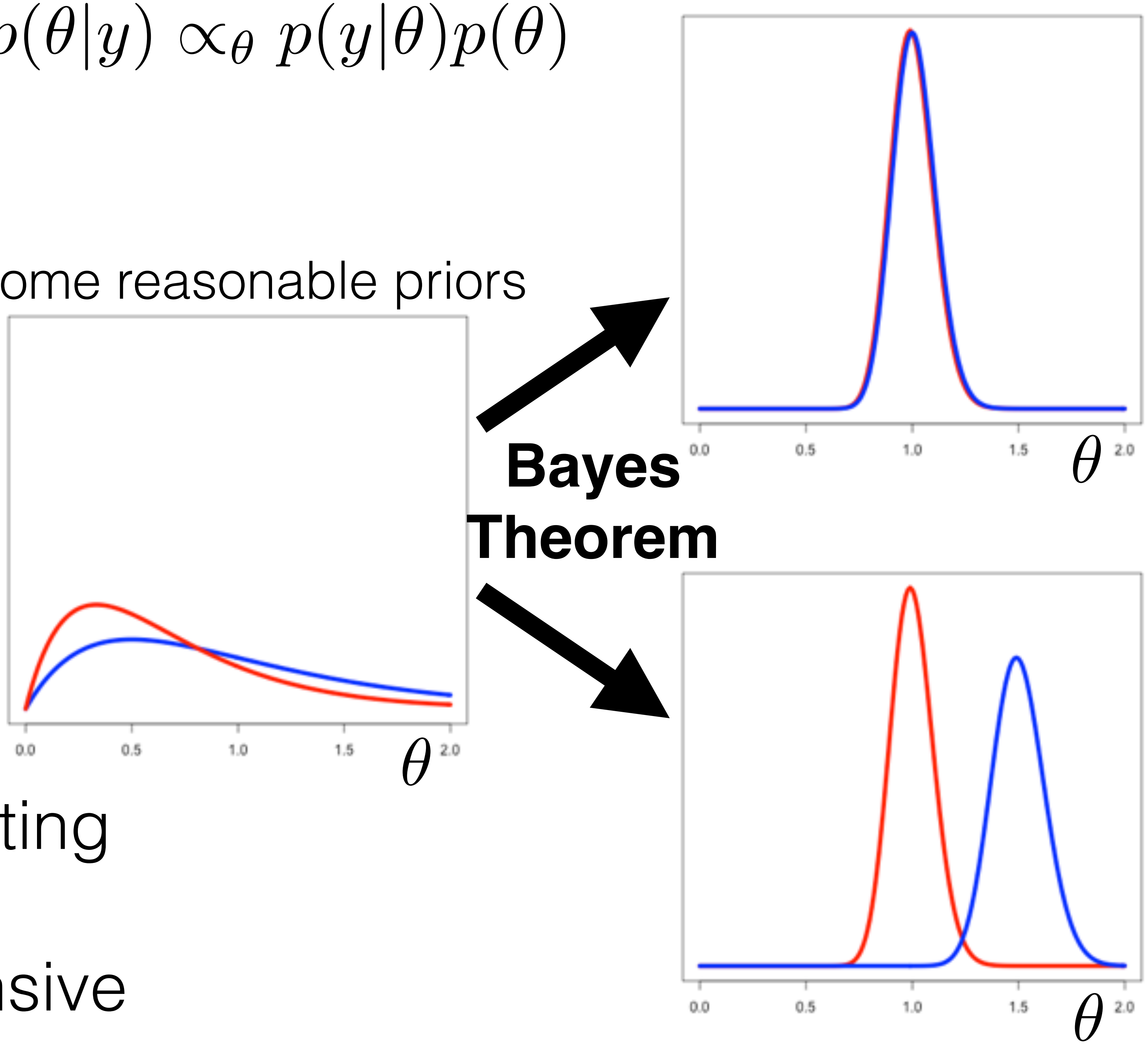
Some reasonable priors

Bayes Theorem

- Challenge: Approximating the posterior can be computationally expensive
  - Markov Chain Monte Carlo

robustness quantification

- Bayesian inference $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Challenge: Express knowledge in a distribution (prior, likelihood)
  - Time-consuming; subjective; complex models

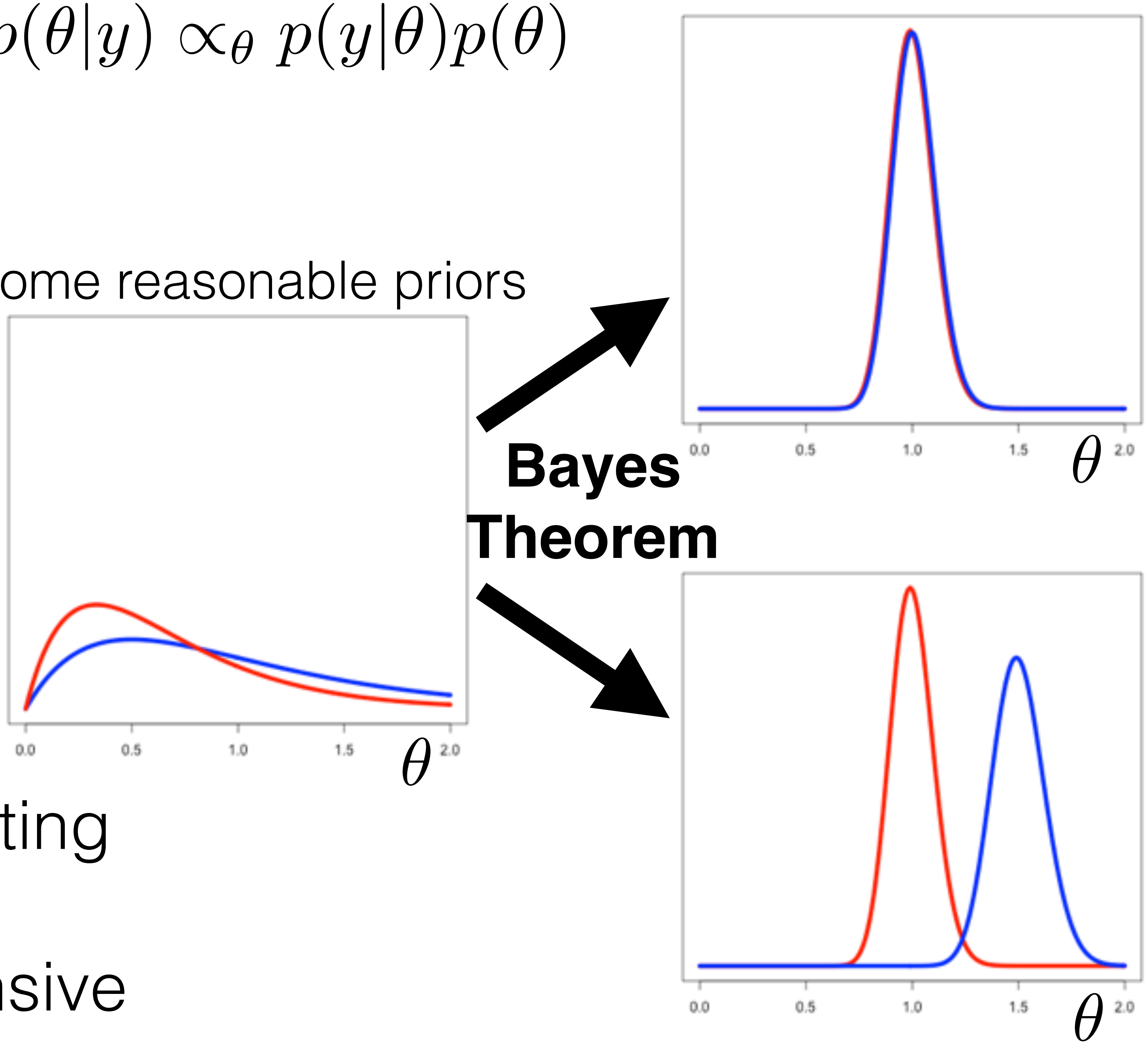


- Challenge: Approximating the posterior can be computationally expensive
  - Markov Chain Monte Carlo

variational Bayes

Uncertainty & robustness quantification

- Bayesian inference $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Challenge: Express knowledge in a distribution (prior, likelihood)
  - Time-consuming; subjective; complex models
- Challenge: Approximating the posterior can be computationally expensive
  - Markov Chain Monte Carlo



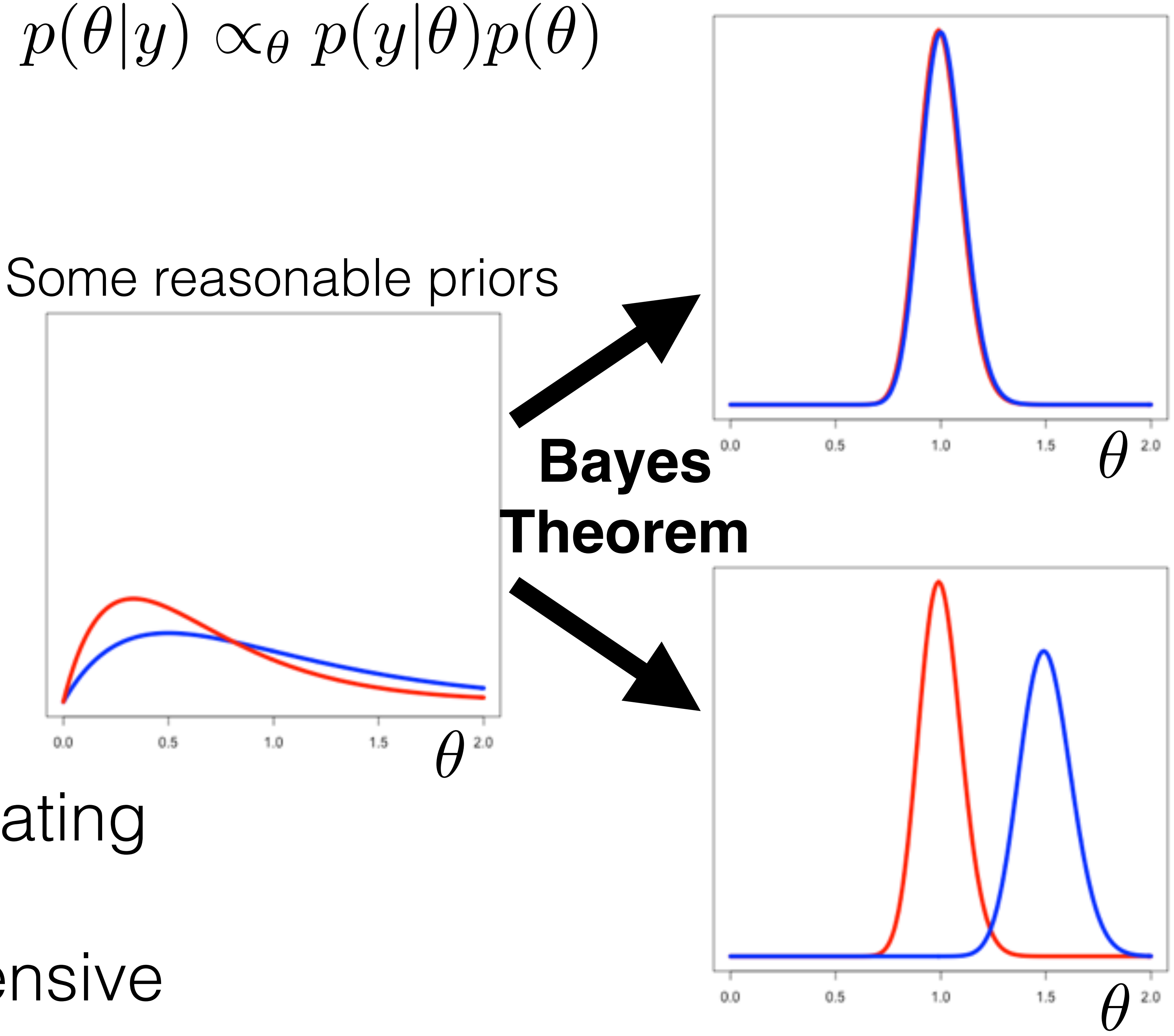
Some reasonable priors

Bayes Theorem

variational Bayes

Uncertainty & robustness quantification

- Bayesian inference $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Challenge: Express knowledge in a distribution (prior, likelihood)
  - Time-consuming; subjective; complex models



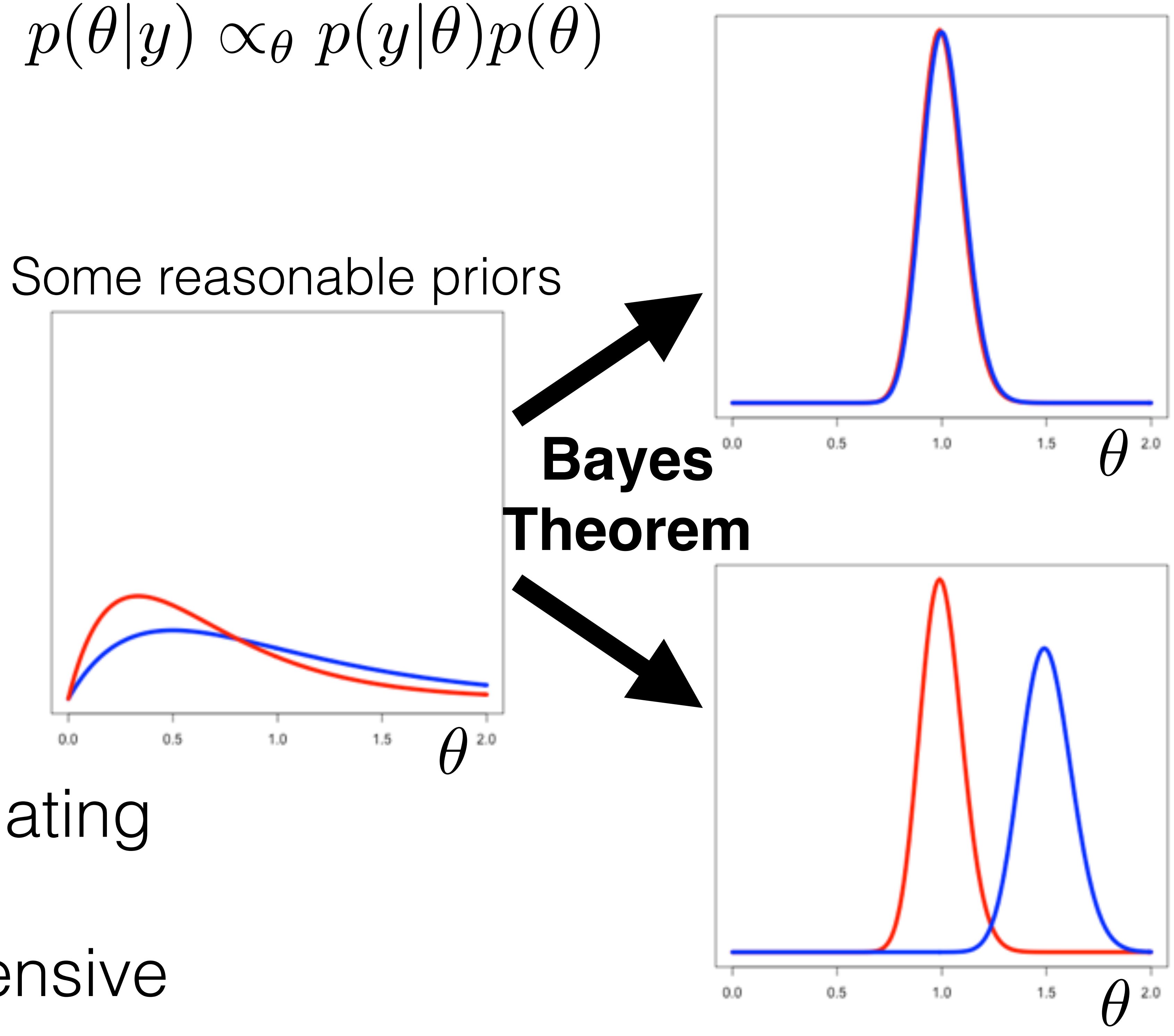
Some reasonable priors

Bayes Theorem

- Challenge: Approximating the posterior can be computationally expensive
  - Markov Chain Monte Carlo
- We propose: *linear response variational Bayes*

Uncertainty & robustness quantification

- Bayesian inference $p(\theta|y) \propto_{\theta} p(y|\theta)p(\theta)$
- Challenge: Express knowledge in a distribution (prior, likelihood)
  - Time-consuming; subjective; complex models



Some reasonable priors

Bayes Theorem

- Challenge: Approximating the posterior can be computationally expensive
  - Markov Chain Monte Carlo
- We propose: *linear response variational Bayes*

[see also Opper, Winther 2003]

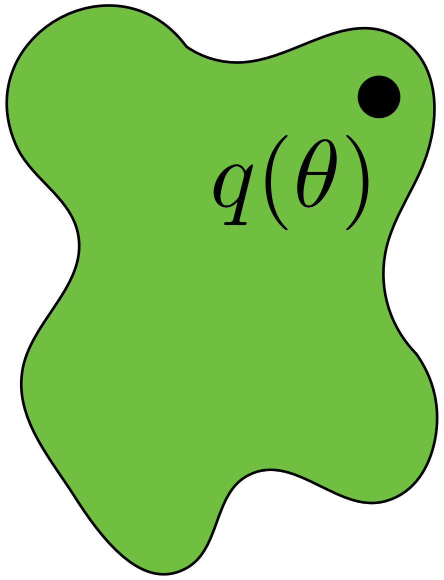
Roadmap

- Challenges of VB
- Accurate uncertainties from VB
- Accurate robustness quantification from VB
  - Big idea: derivatives/perturbations are relatively easy in VB

Roadmap

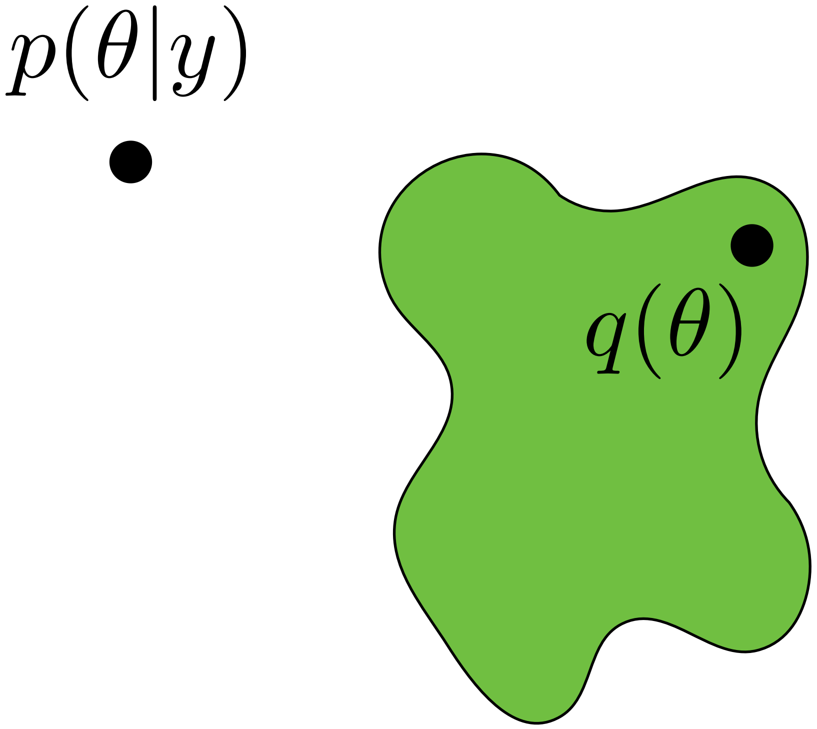
- Challenges of VB
- Accurate uncertainties from VB
- Accurate robustness quantification from VB
  - Big idea: derivatives/perturbations are relatively easy in VB

What about uncertainty?



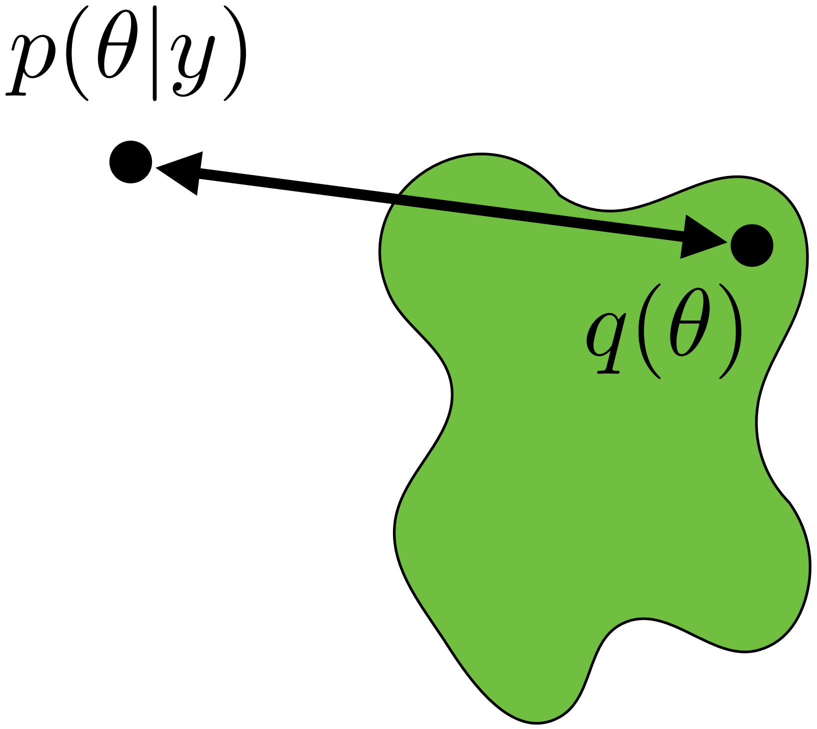
[Bishop 2006]

What about uncertainty?



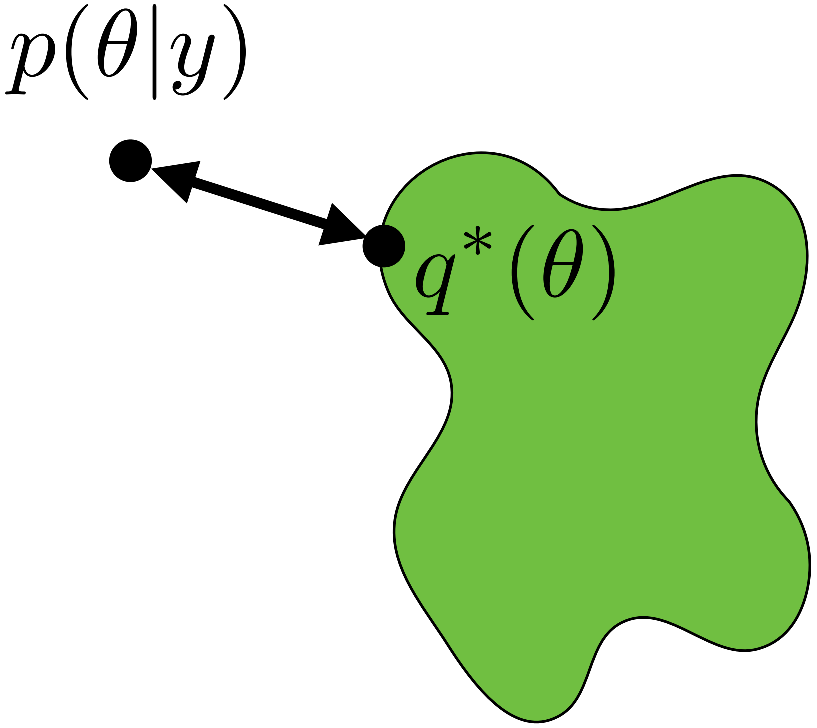
[Bishop 2006]

What about uncertainty?



[Bishop 2006]

What about uncertainty?

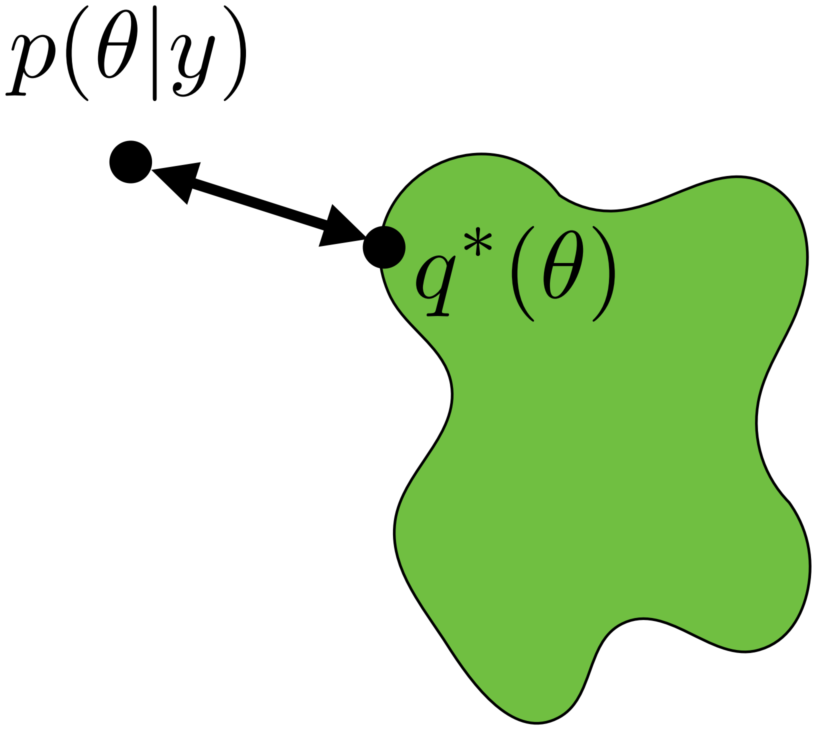


[Bishop 2006]

What about uncertainty?

- Variational Bayes

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

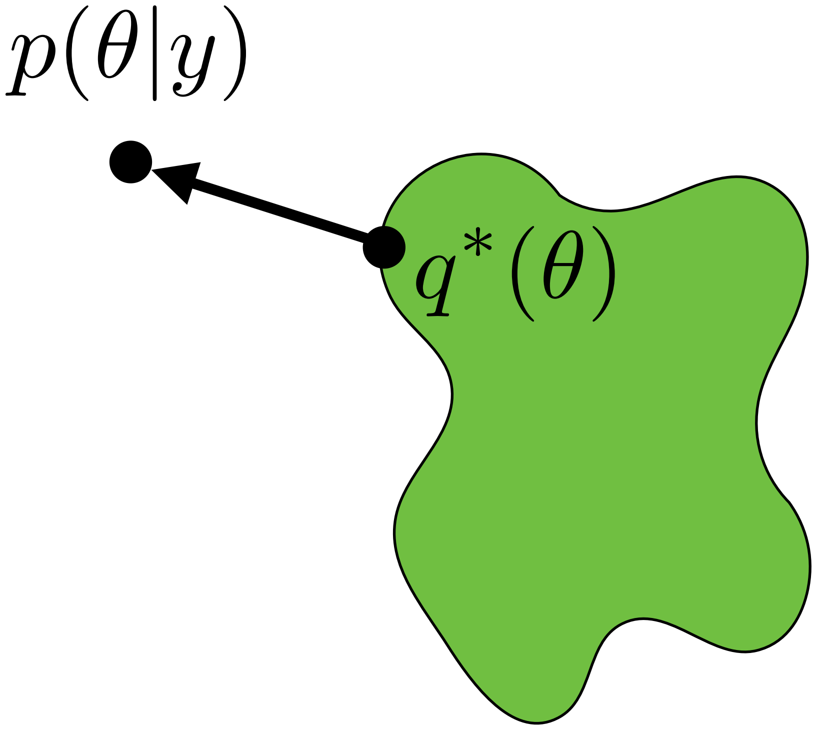


[Bishop 2006]

What about uncertainty?

- Variational Bayes

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



[Bishop 2006]

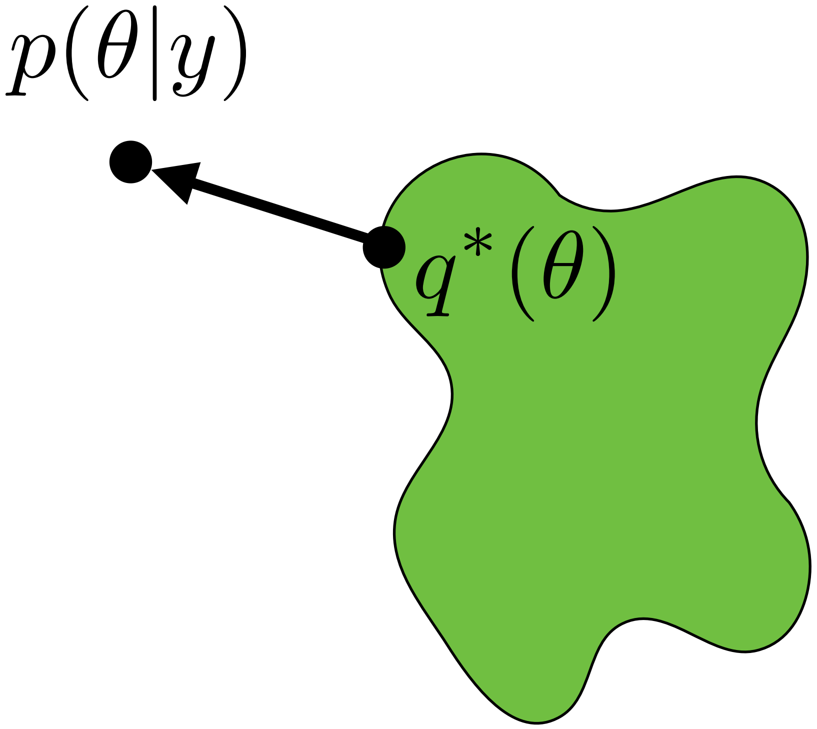
What about uncertainty?

- Variational Bayes

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

- Mean-field variational Bayes (MFVB)

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



[Bishop 2006]

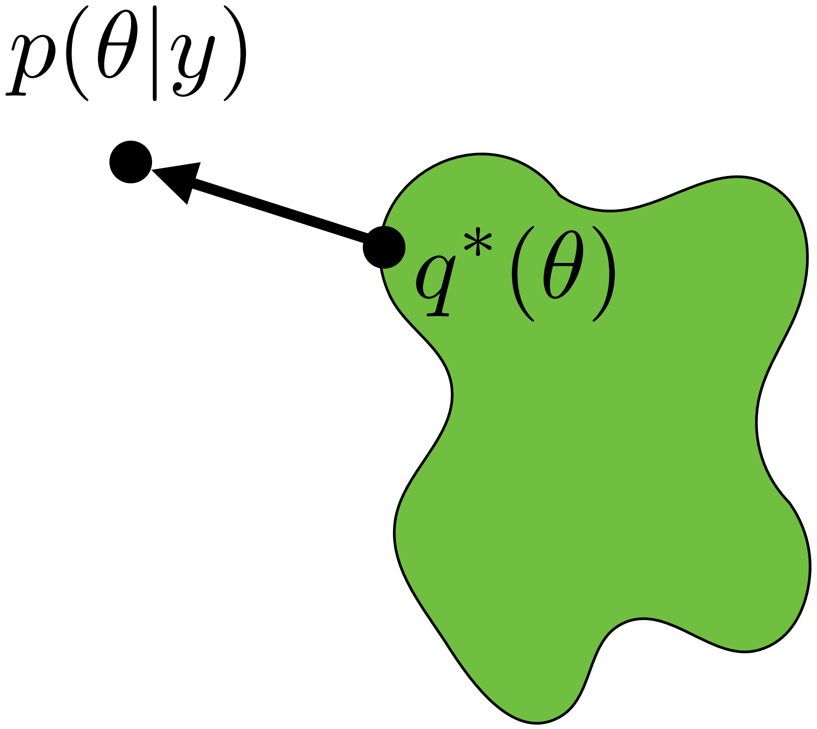
What about uncertainty?

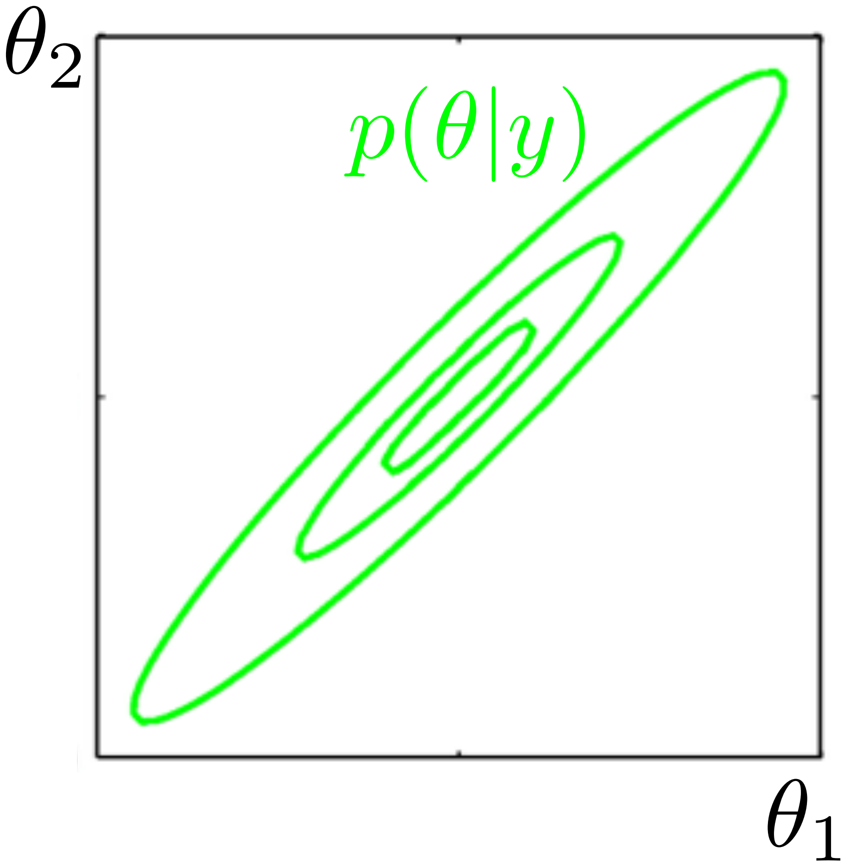
- Variational Bayes

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

- Mean-field variational Bayes (MFVB)

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





[Bishop 2006]

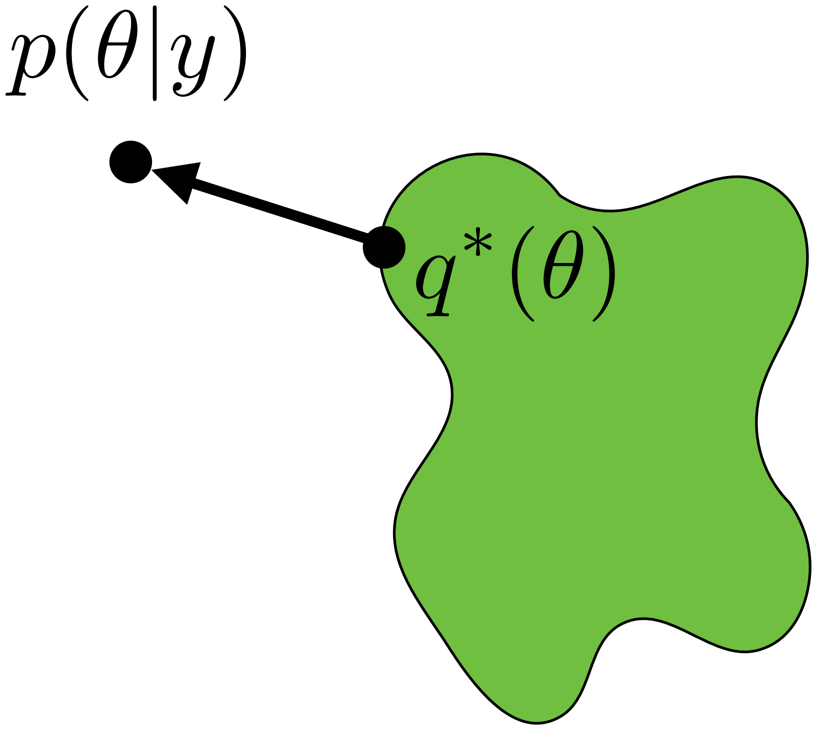
What about uncertainty?

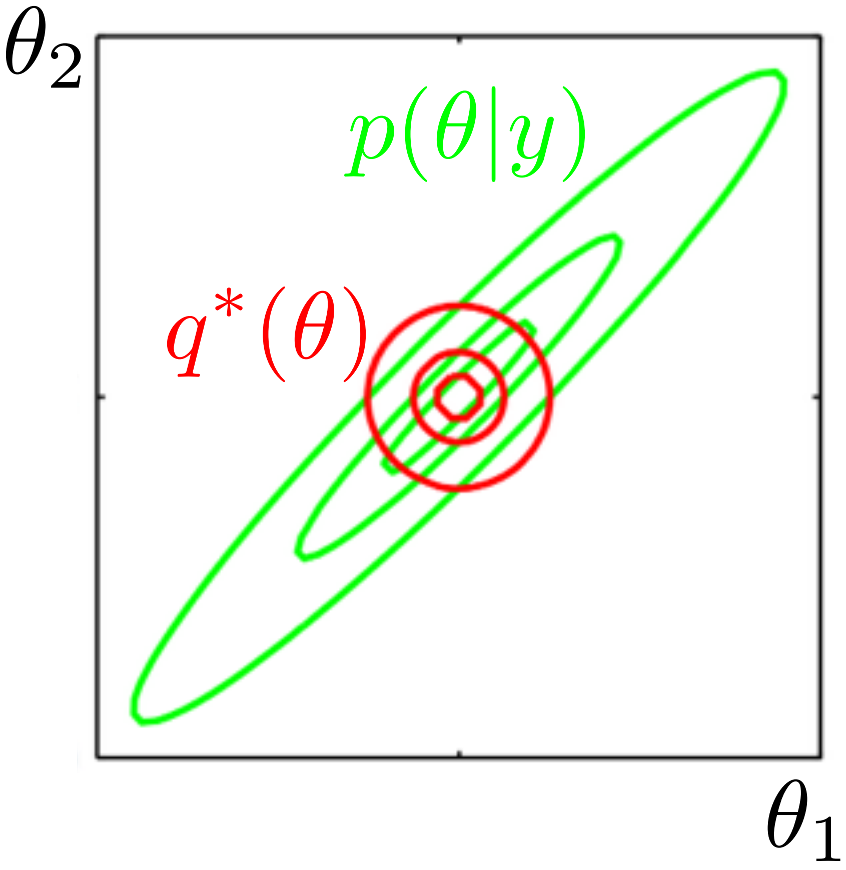
- Variational Bayes

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

- Mean-field variational Bayes (MFVB)

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





[Bishop 2006]

What about uncertainty?

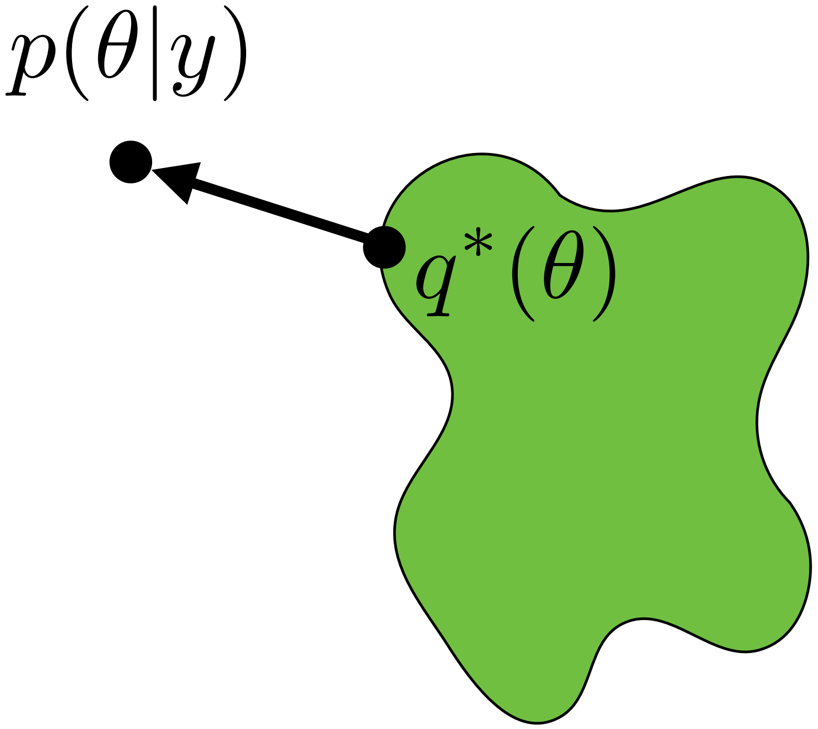
- Variational Bayes

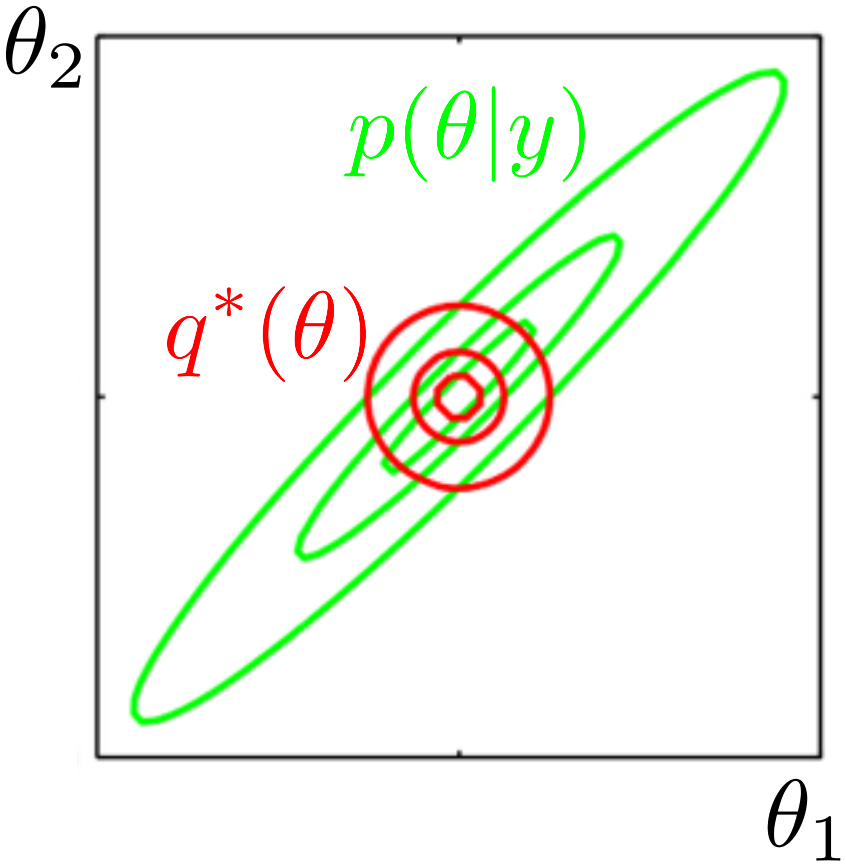
$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)





[Bishop 2006]

What about uncertainty?

- Variational Bayes

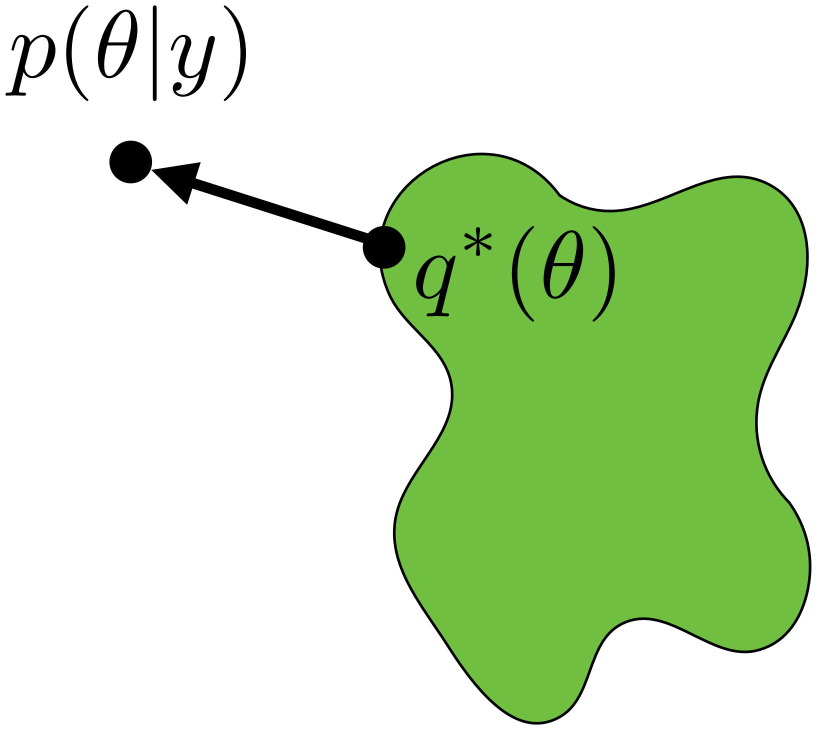
$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

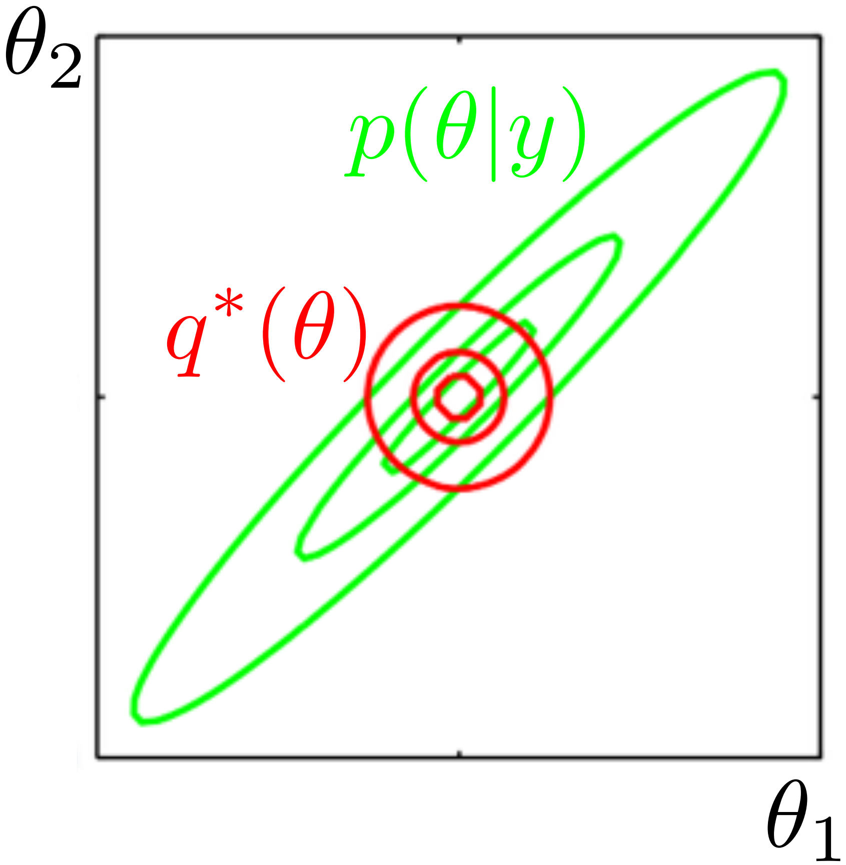
- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)

- No covariance estimates





[Bishop 2006]

What about uncertainty?

- Variational Bayes

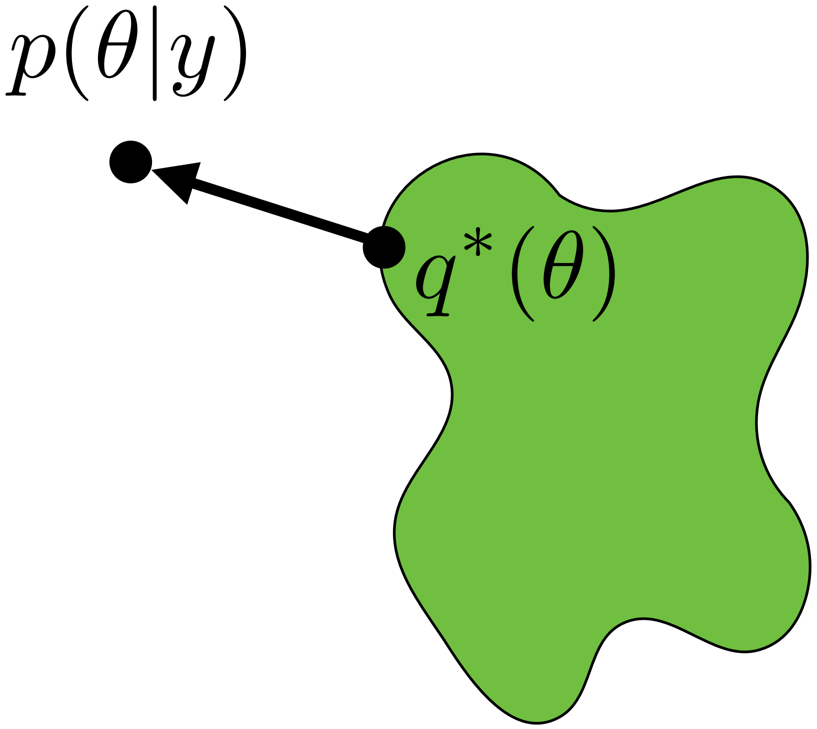
$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

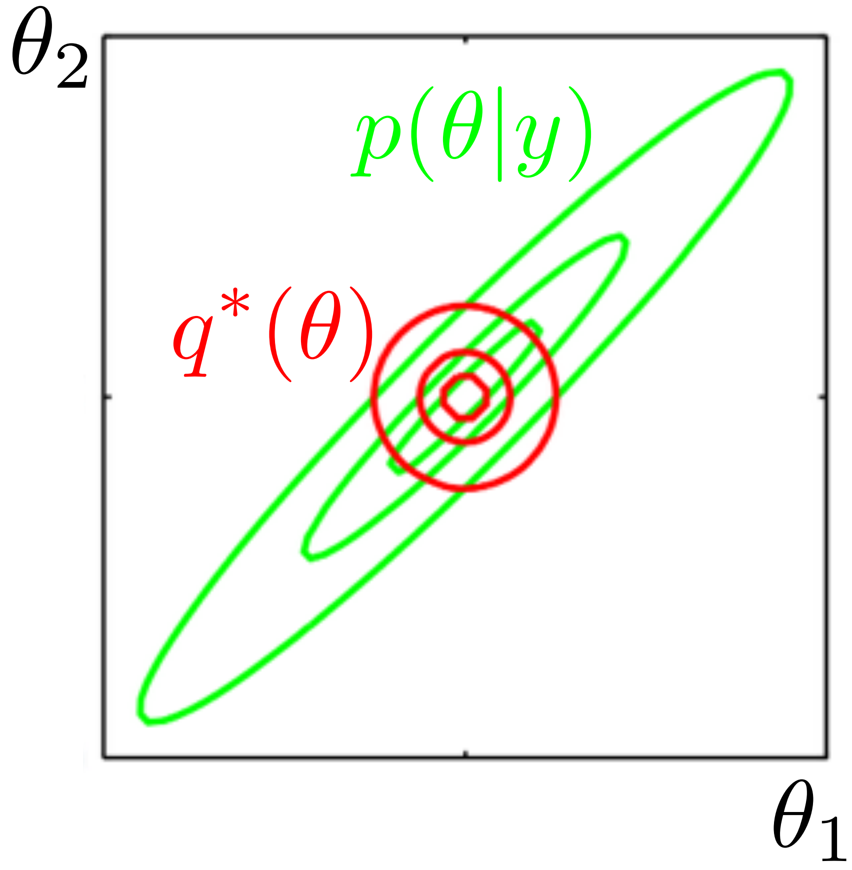
- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)

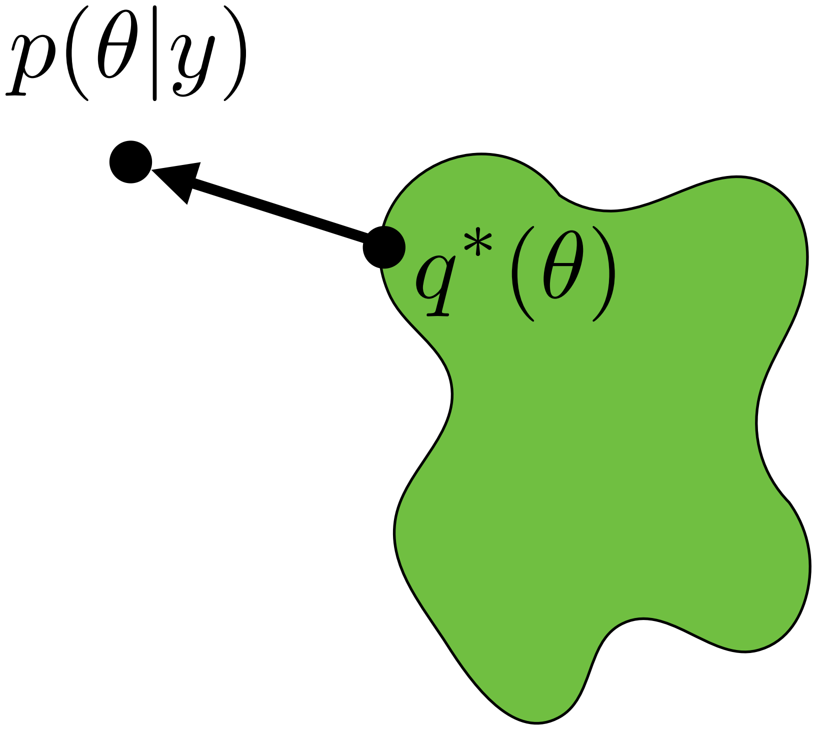
- No covariance estimates





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

What about uncertainty?



- Variational Bayes

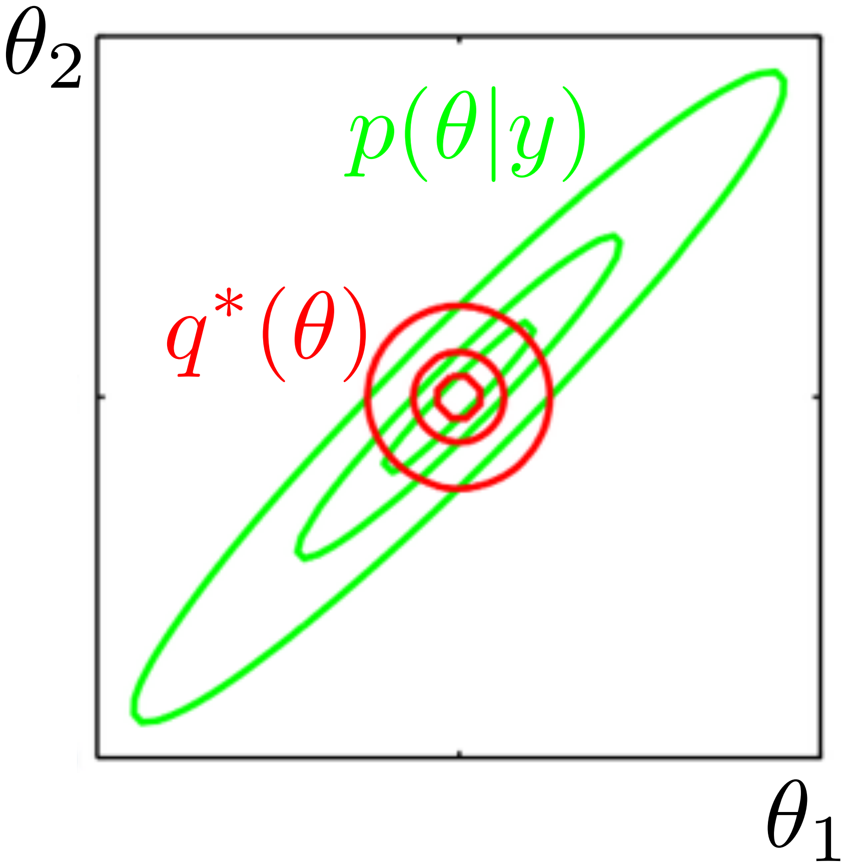
$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)

- No covariance estimates



[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011]
[Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2017]

Linear response

Linear response

- Cumulant-generating function

Linear response

- Cumulant-generating function

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

Linear response

- Cumulant-generating function

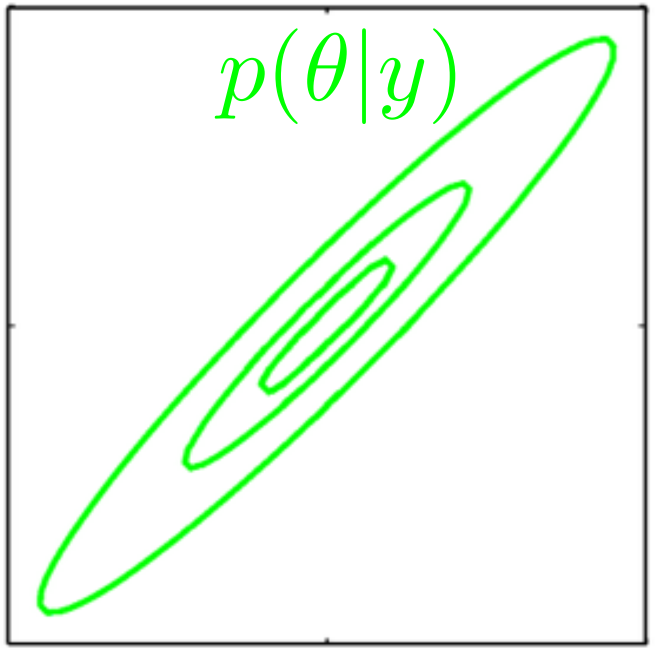
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

Linear response

- Cumulant-generating function

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

- Exact posterior covariance



[adapted from Bishop 2006]

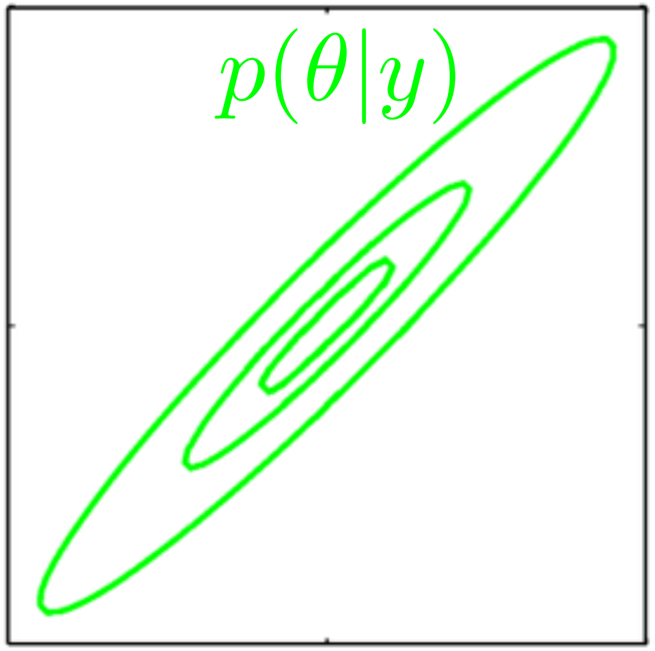
Linear response

- Cumulant-generating function

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

- Exact posterior covariance

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



[adapted from Bishop 2006]

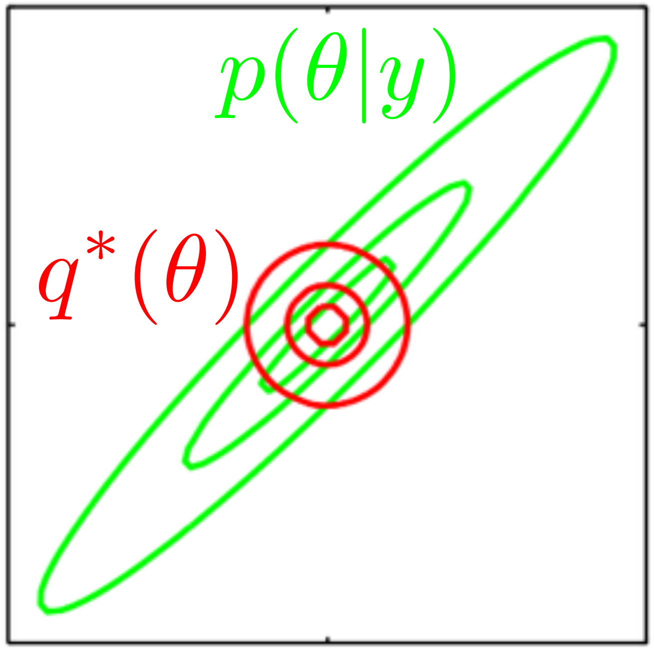
Linear response

- Cumulant-generating function

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

- Exact posterior covariance vs MFVB covariance

$$\Sigma := \frac{d^2}{dt^T dt} C_{p(\cdot|y)}(t) \bigg|_{t=0}$$



[adapted from Bishop 2006]

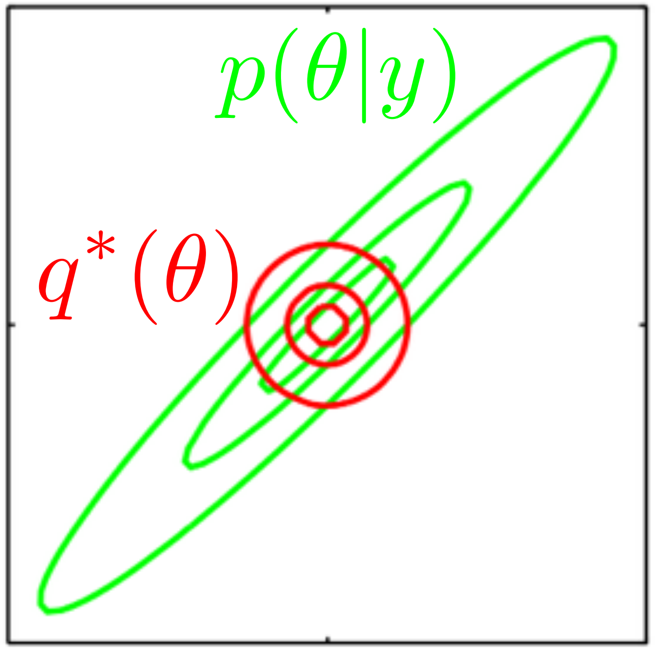
Linear response

- Cumulant-generating function

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

- Exact posterior covariance vs MFVB covariance

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



[adapted from Bishop 2006]

Linear response

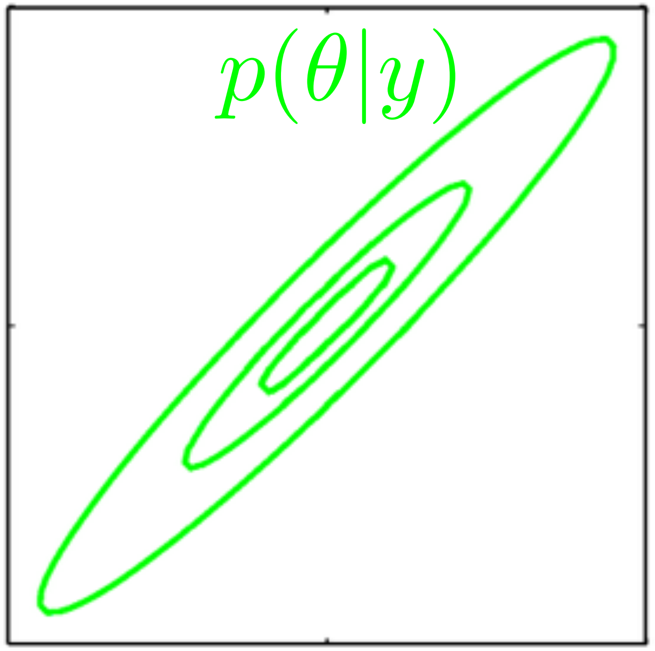
- Cumulant-generating function

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

- Exact posterior covariance vs MFVB covariance

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

- “Linear response”



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

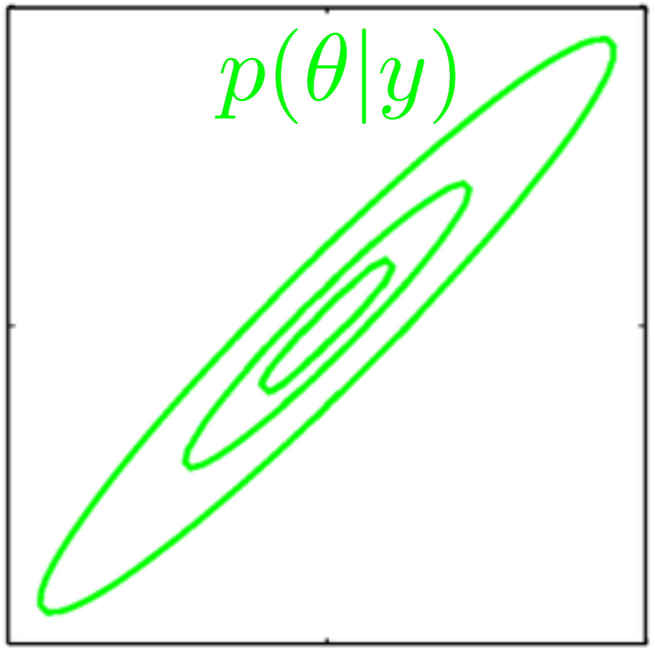
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- Exact posterior covariance vs MFVB covariance

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

- "Linear response"

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



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

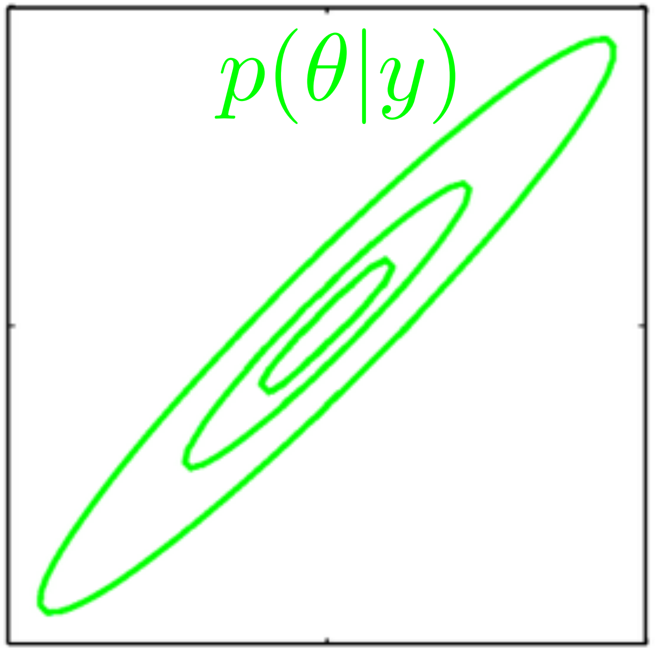
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- Exact posterior covariance vs MFVB covariance

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

- "Linear response"

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



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

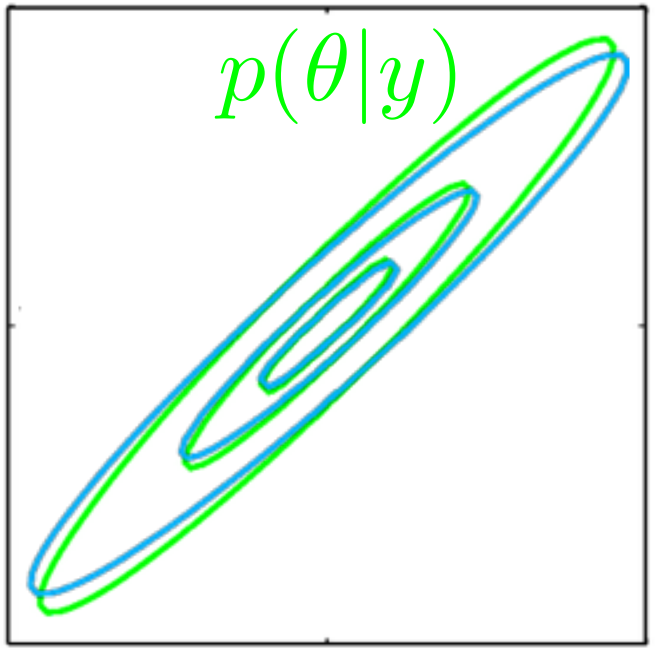
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- Exact posterior covariance vs MFVB covariance

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

- "Linear response"

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



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

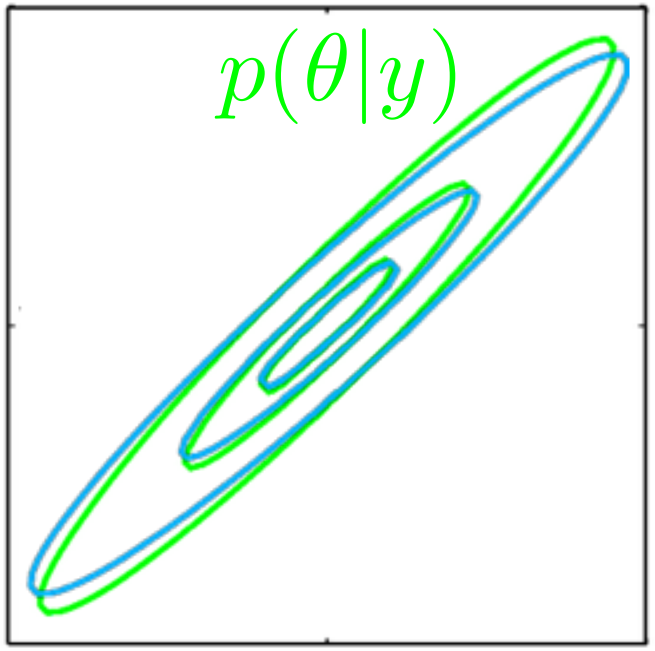
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- Exact posterior covariance vs MFVB covariance

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

- "Linear response"

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



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

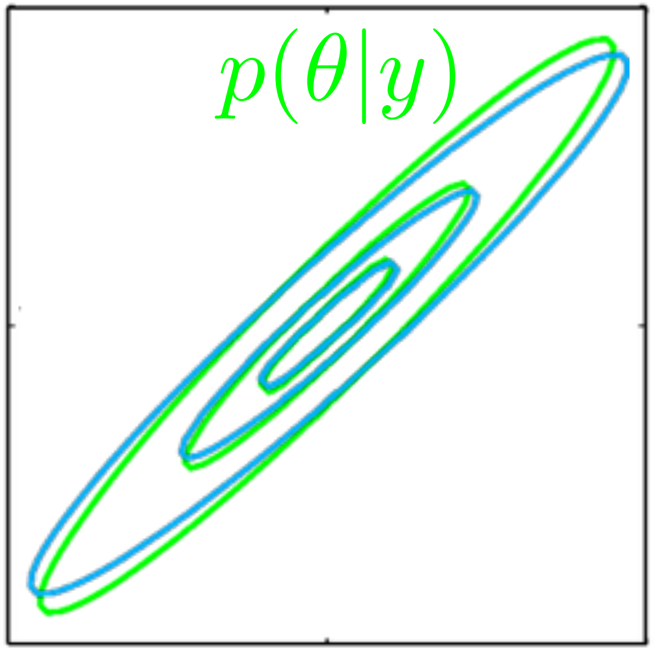
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- Exact posterior covariance vs MFVB covariance

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

- “Linear response”

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



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

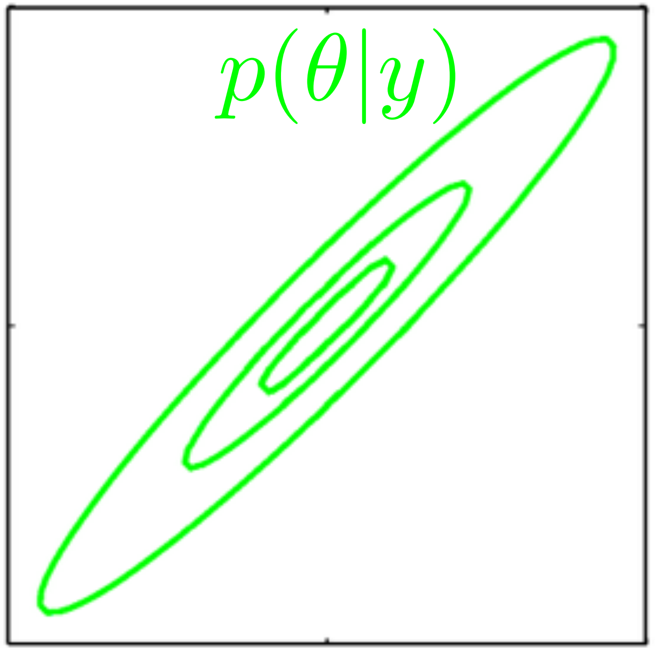
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- Exact posterior covariance vs MFVB covariance

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

- "Linear response"

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



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

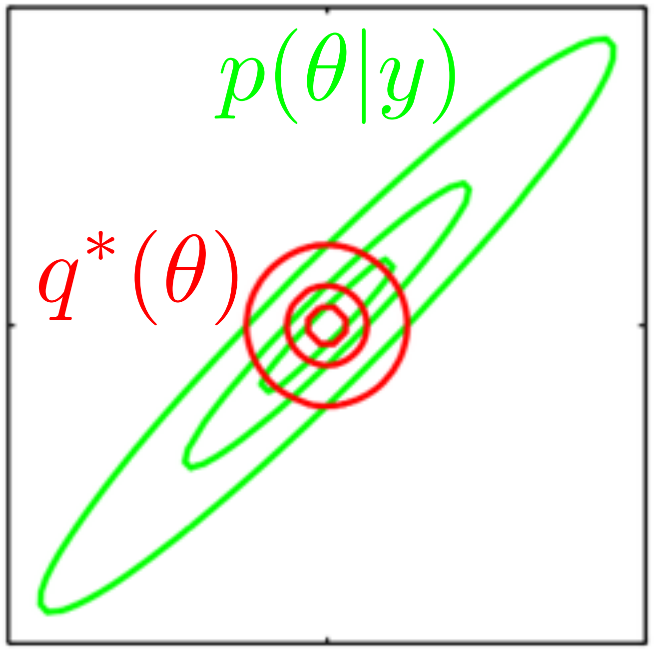
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- Exact posterior covariance vs MFVB covariance

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

- "Linear response"

$\log p_t(\theta) := \log p(\theta|y) + t^T \theta - C(t)$, MFVB $q_{\eta^*(t)}$



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

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

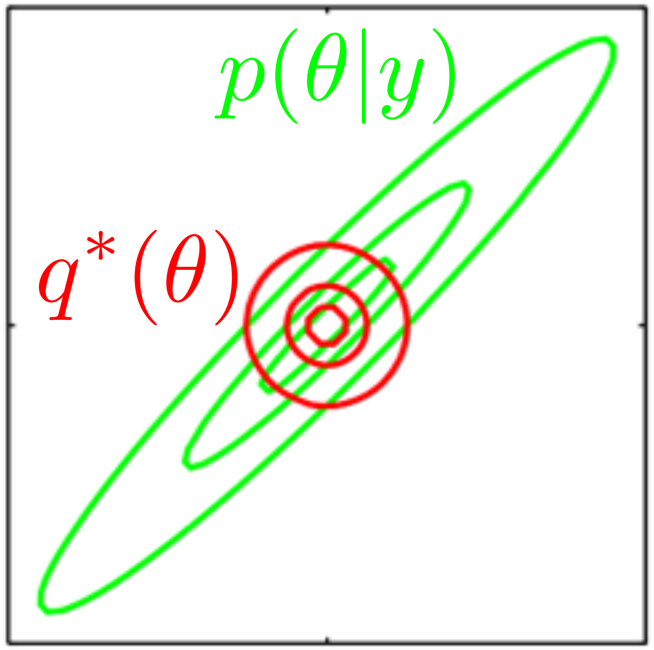
- Exact posterior covariance vs MFVB covariance

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

- "Linear response"

$\log p_t(\theta) := \log p(\theta|y) + t^T \theta - C(t)$, MFVB $q_{\eta^*(t)}$

- The LRVB approximation



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

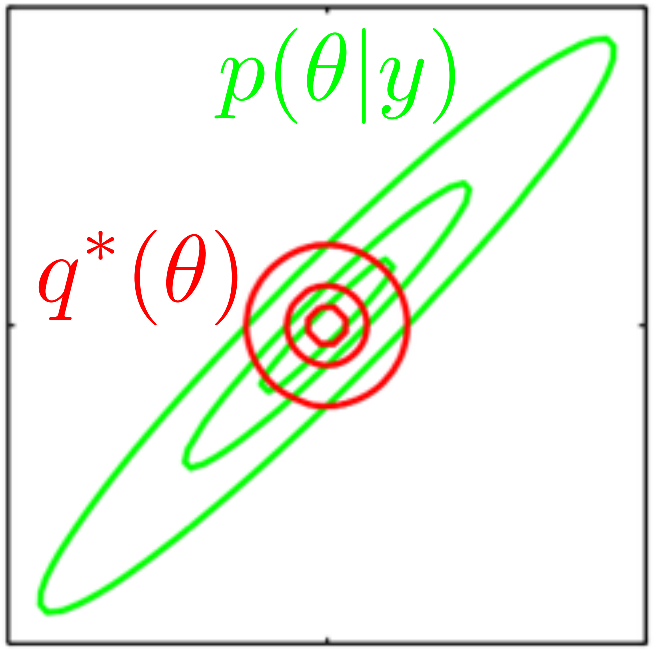
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- Exact posterior covariance vs MFVB covariance

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

- "Linear response"

$\log p_t(\theta) := \log p(\theta|y) + t^T \theta - C(t)$, MFVB $q_{\eta^*(t)}$



[adapted from Bishop 2006]

- The LRVB approximation

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

Linear response

- Cumulant-generating function

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

- Exact posterior covariance vs MFVB covariance

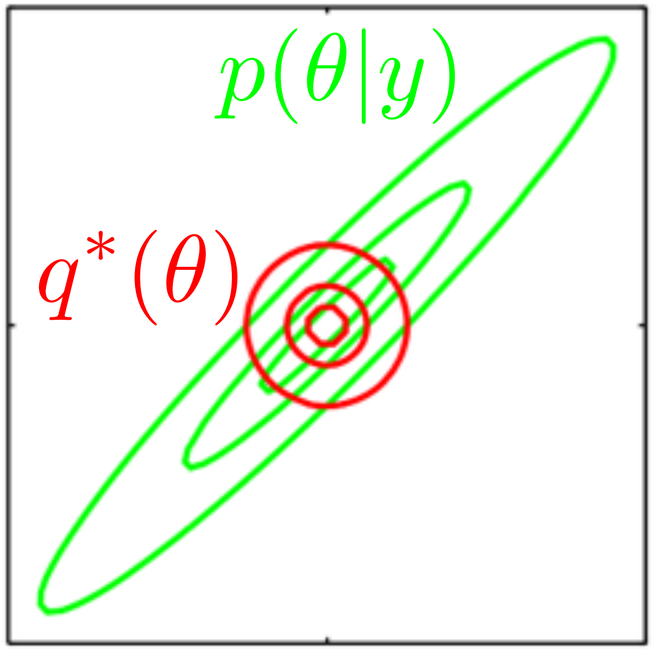
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|y)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q_{\eta^*}}(t) \right|_{t=0}$$

- “Linear response”

$\log p_t(\theta) := \log p(\theta|y) + t^T \theta - C(t)$, MFVB $q_{\eta^*(t)}$

- The LRVB approximation

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



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

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

- Exact posterior covariance vs MFVB covariance

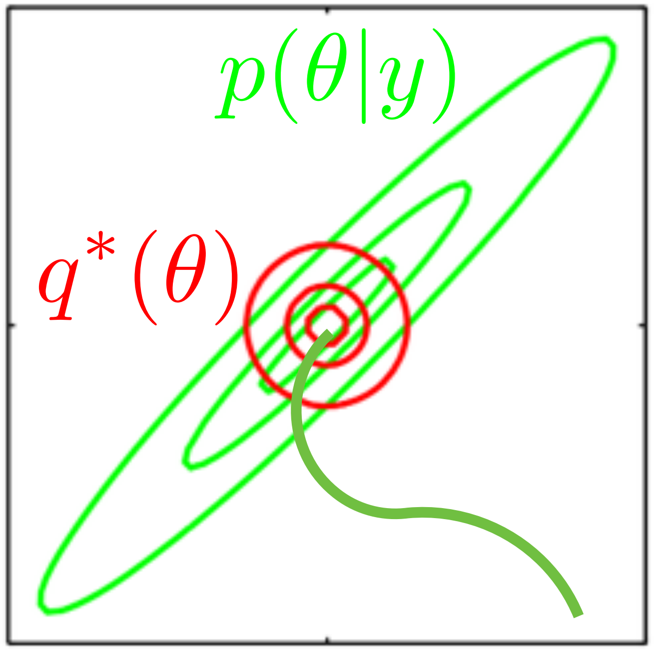
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|y)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q_{\eta^*}}(t) \right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|y) + t^T\theta - C(t)$, MFVB $q_{\eta^*(t)}$

- The LRVB approximation

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



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

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

- Exact posterior covariance vs MFVB covariance

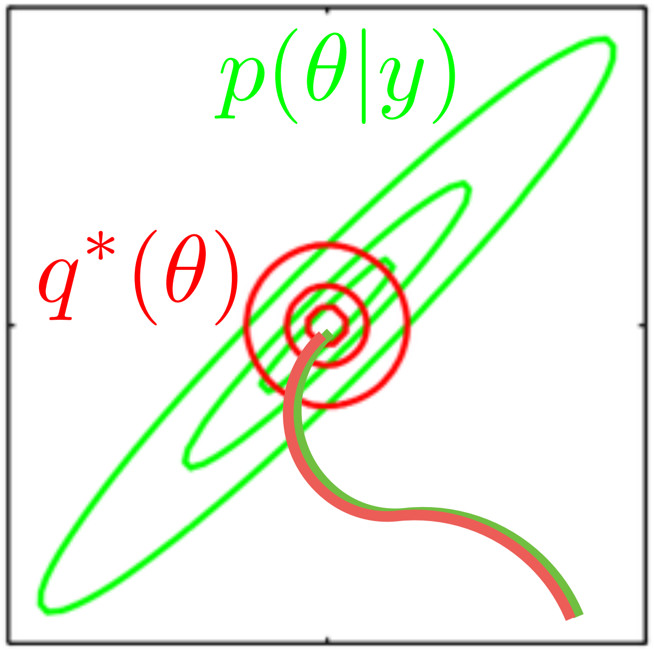
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|y)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q_{\eta^*}}(t) \right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|y) + t^T \theta - C(t)$, MFVB $q_{\eta^*(t)}$

- The LRVB approximation

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



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

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

- Exact posterior covariance vs MFVB covariance

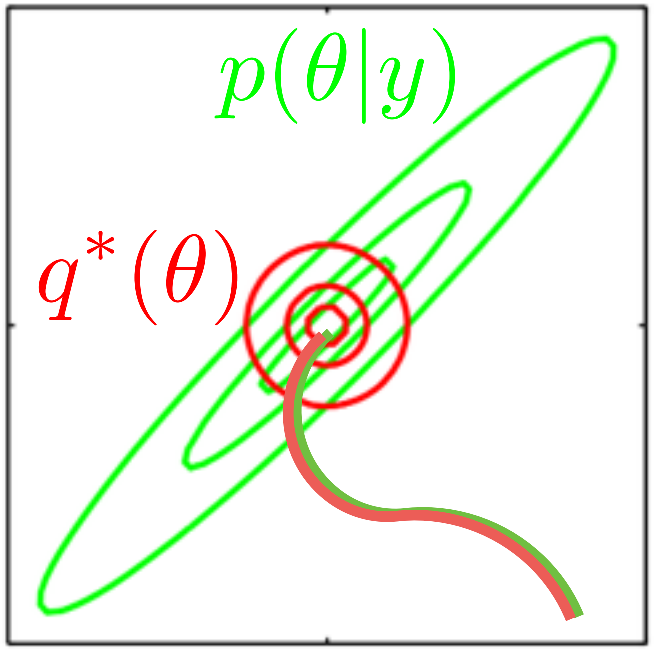
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|y)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q_{\eta^*}}(t) \right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|y) + t^T \theta - C(t)$, MFVB $q_{\eta^*(t)}$

- The LRVB approximation

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



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

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

- Exact posterior covariance vs MFVB covariance

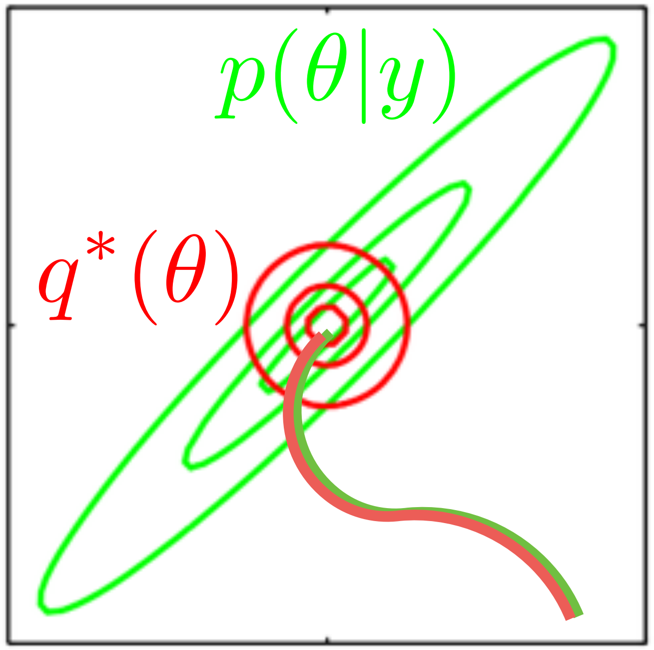
$$\Sigma := \left.\frac{d^2}{dt^T dt} C_{p(\cdot|y)}(t)\right|_{t=0} \qquad V := \left.\frac{d^2}{dt^T dt} C_{q_{\eta^*}}(t)\right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|y) + t^T\theta - C(t)$, MFVB $q_{\eta^*(t)}$

- The LRVB approximation

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



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

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

- Exact posterior covariance vs ~~M~~FVB covariance

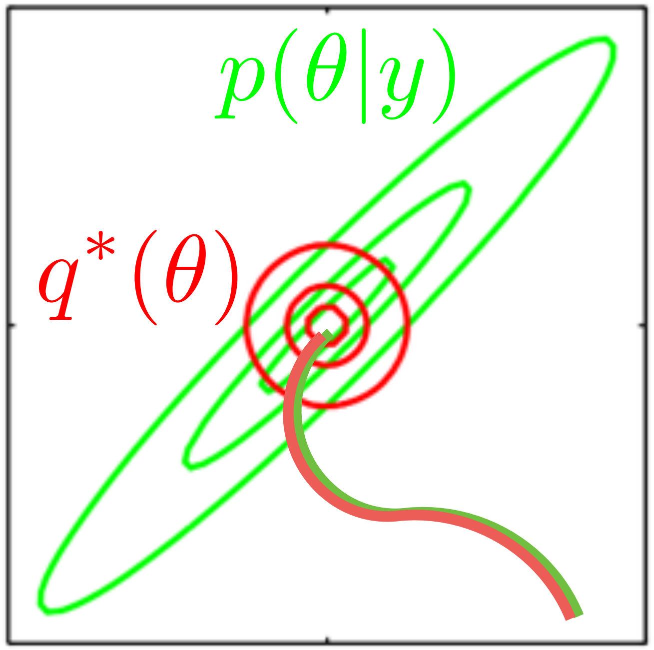
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|y)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q_{\eta^*}}(t) \right|_{t=0}$$

- "Linear response"

$\log p_t(\theta) := \log p(\theta|y) + t^T \theta - C(t)$, ~~M~~FVB $q_{\eta^*(t)}$

- The LRVB approximation

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



[adapted from Bishop 2006]

Linear response

- Cumulant-generating function

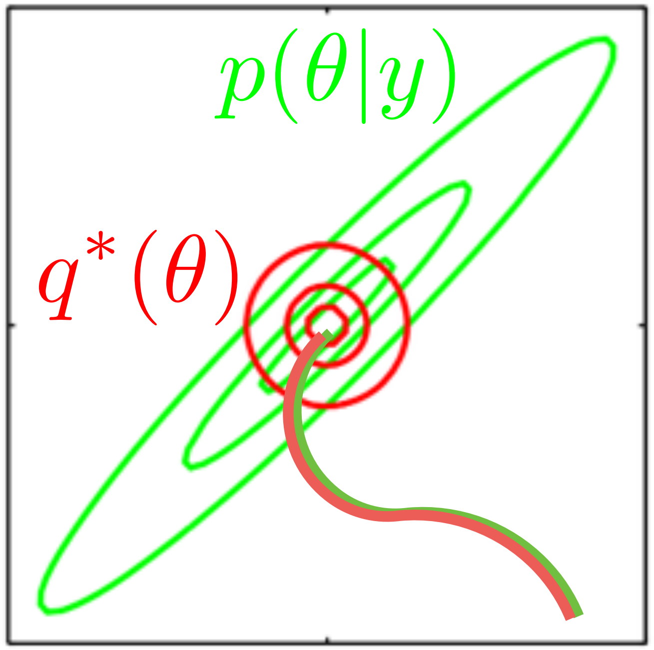
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- Exact posterior covariance vs ~~M~~FVB covariance

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

- "Linear response"

$\log p_t(\theta) := \log p(\theta|y) + t^T \theta - C(t)$, ~~M~~FVB $q_{\eta^*(t)}$



[adapted from Bishop 2006]

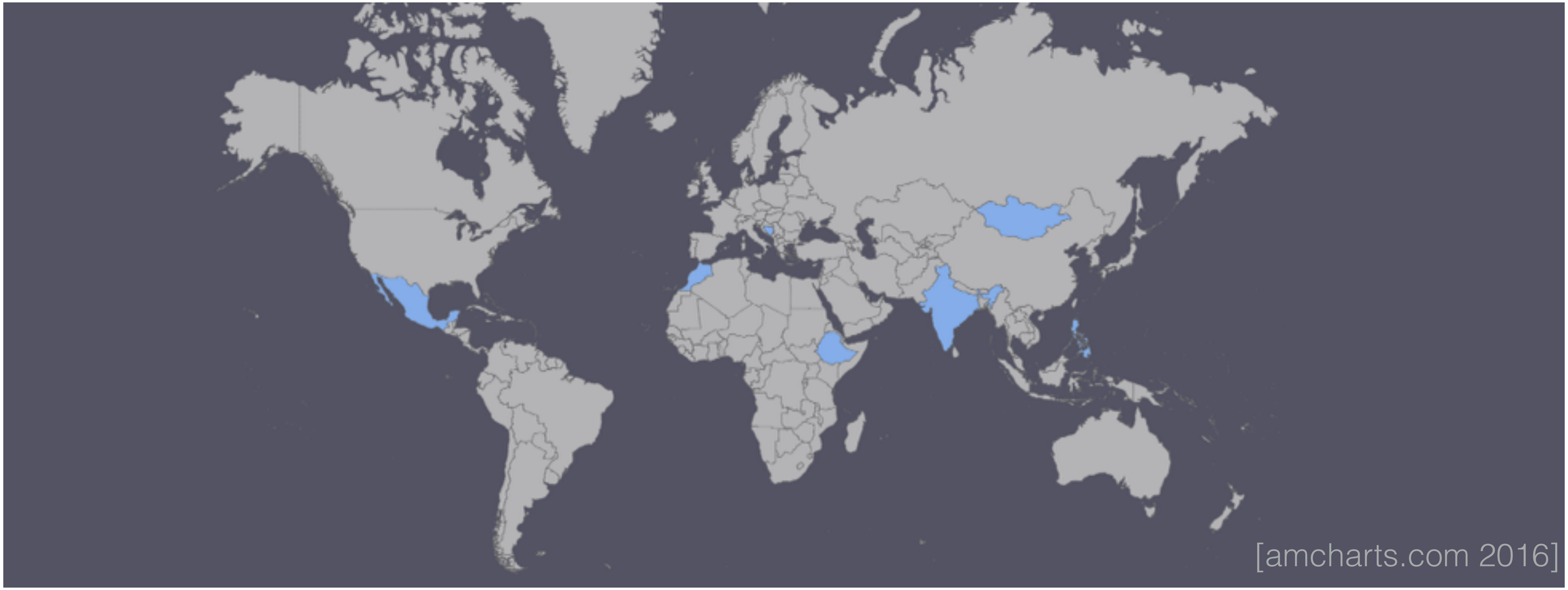
- The LRVB approximation

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

[board]

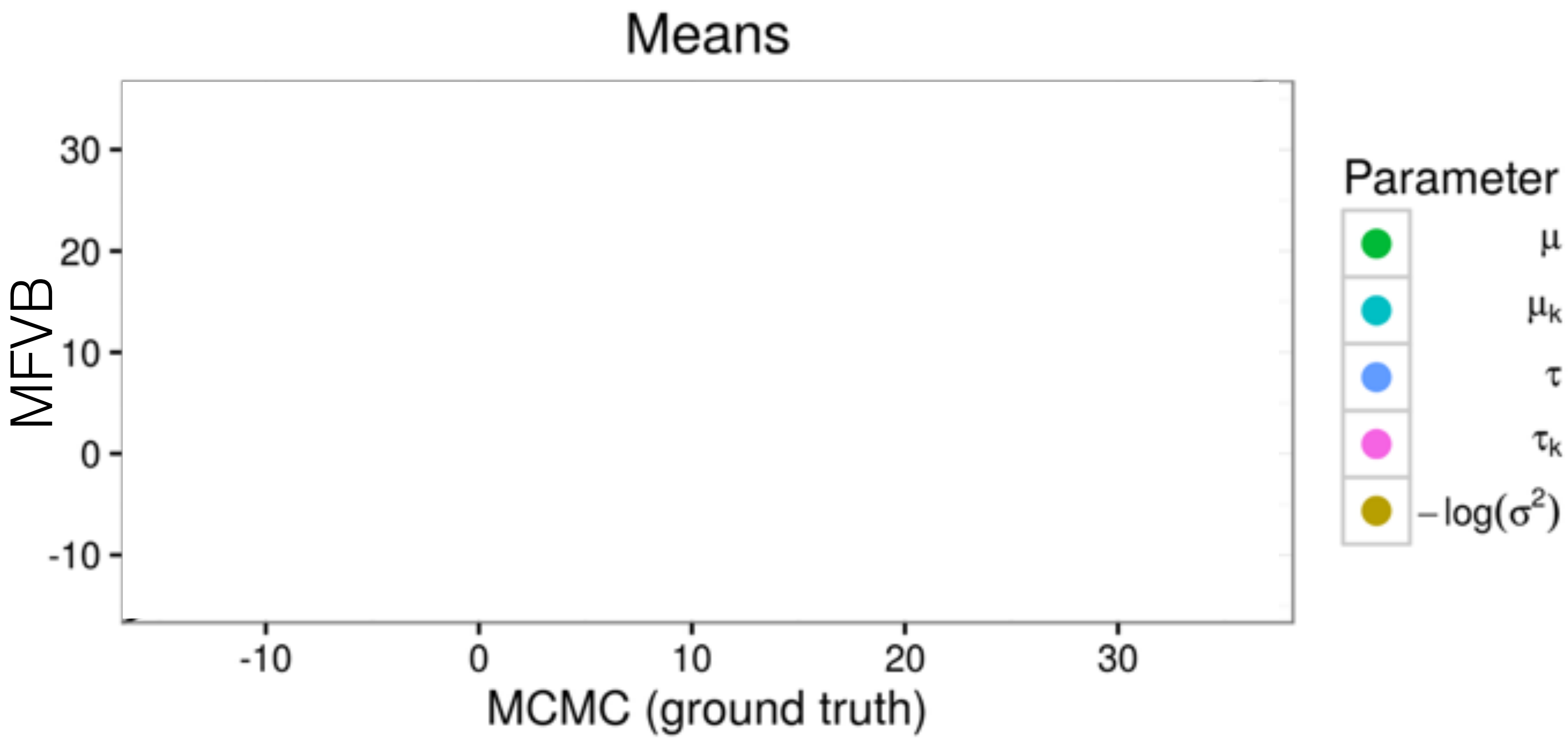
Microcredit Experiment

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



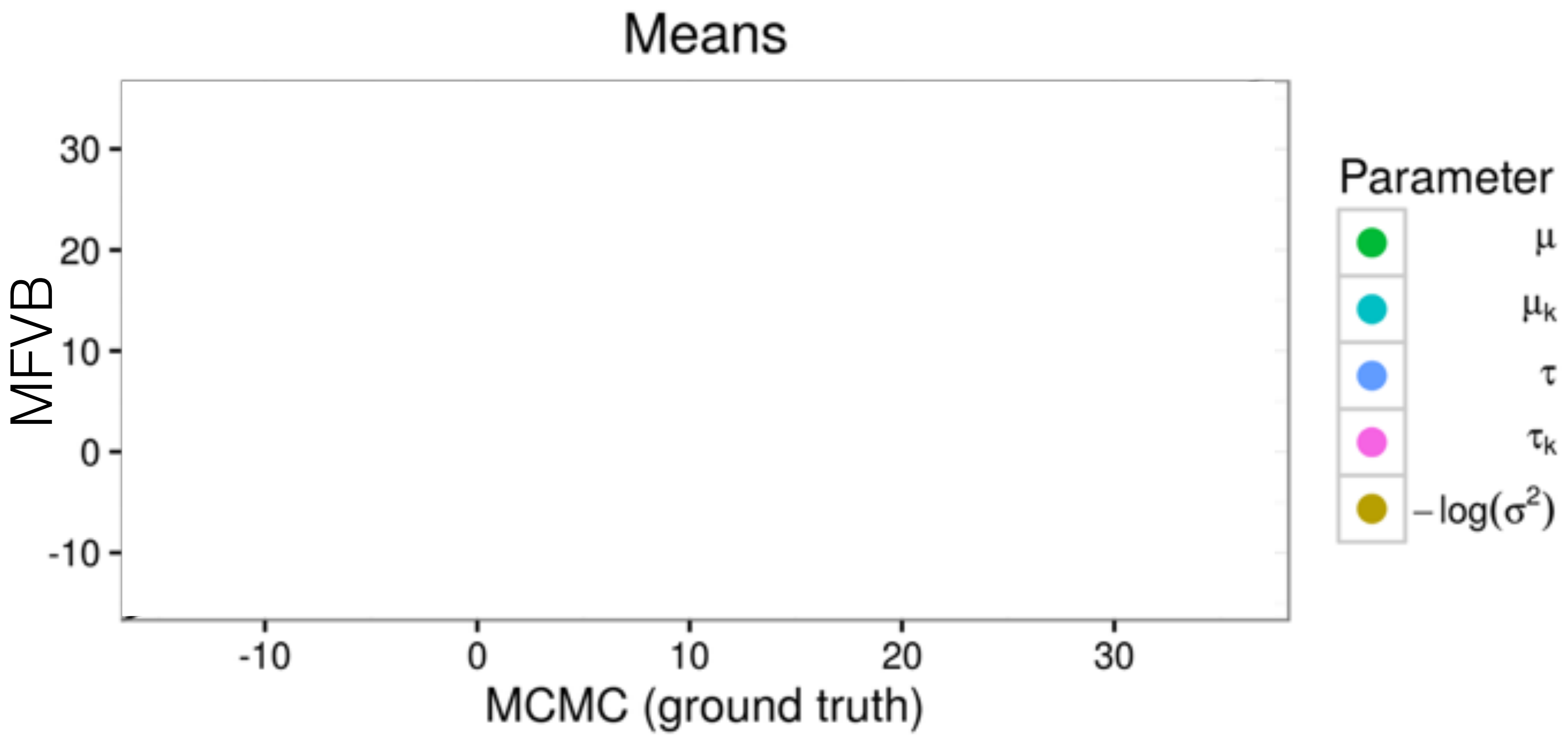
[amcharts.com 2016]

Microcredit Experiment



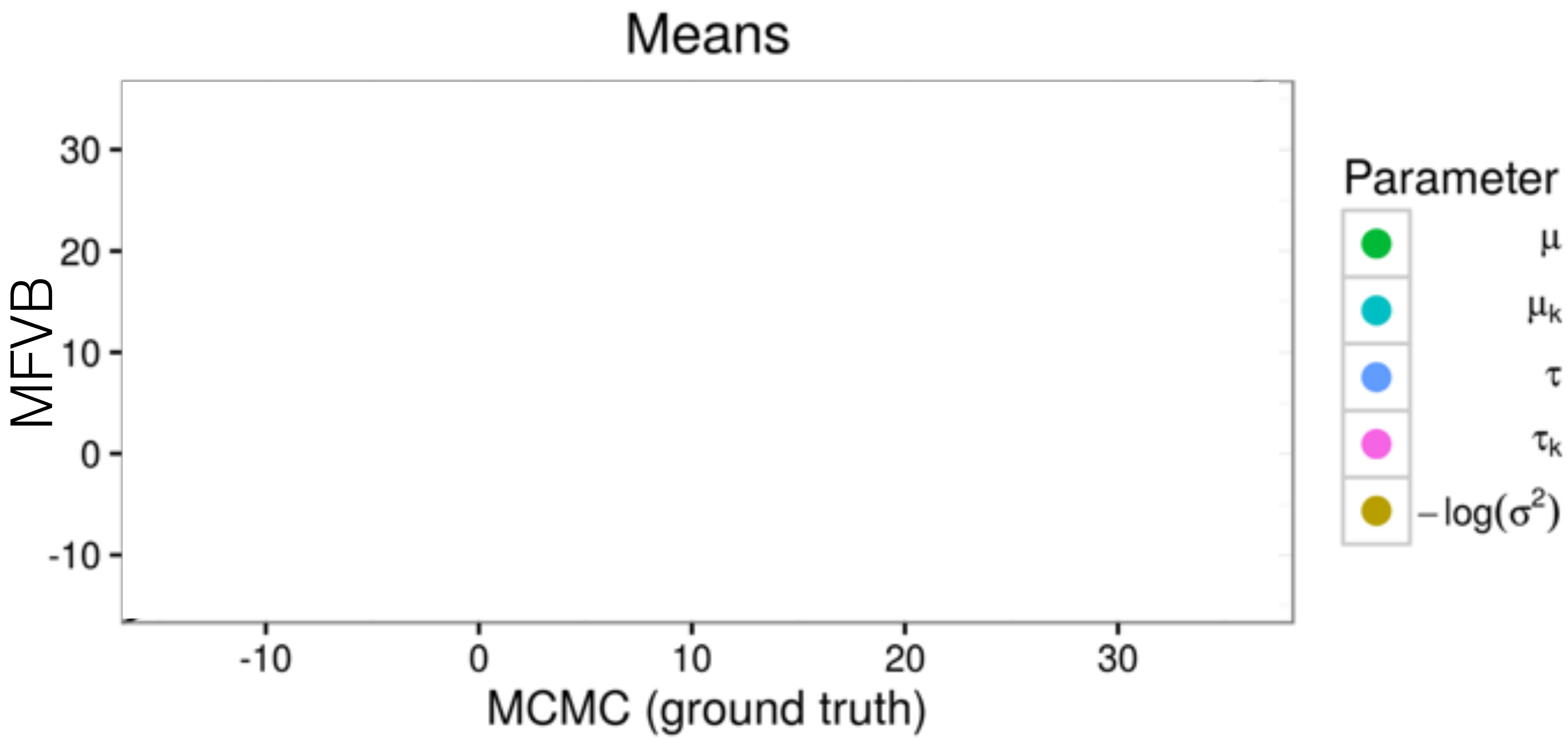
Microcredit Experiment

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



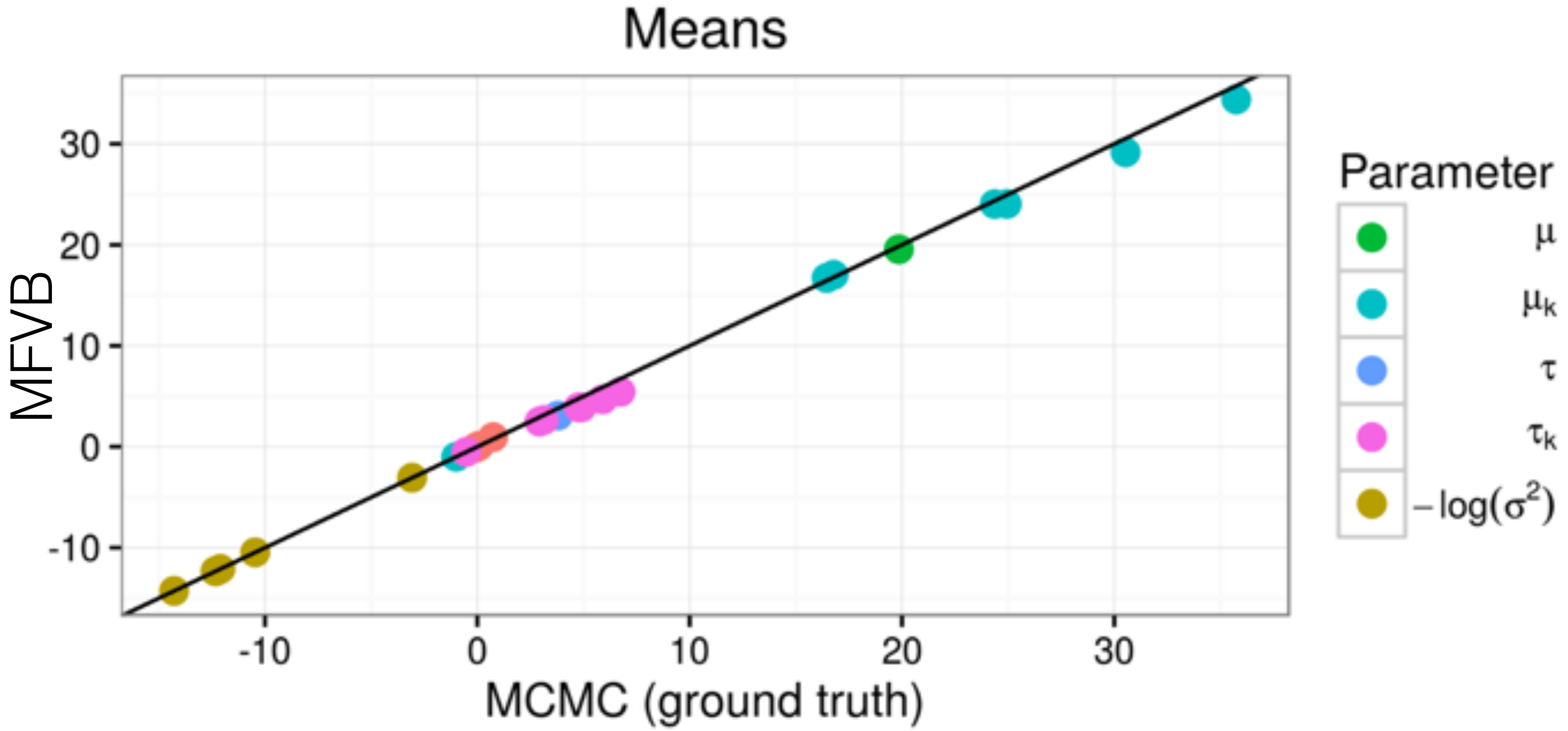
Microcredit Experiment

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



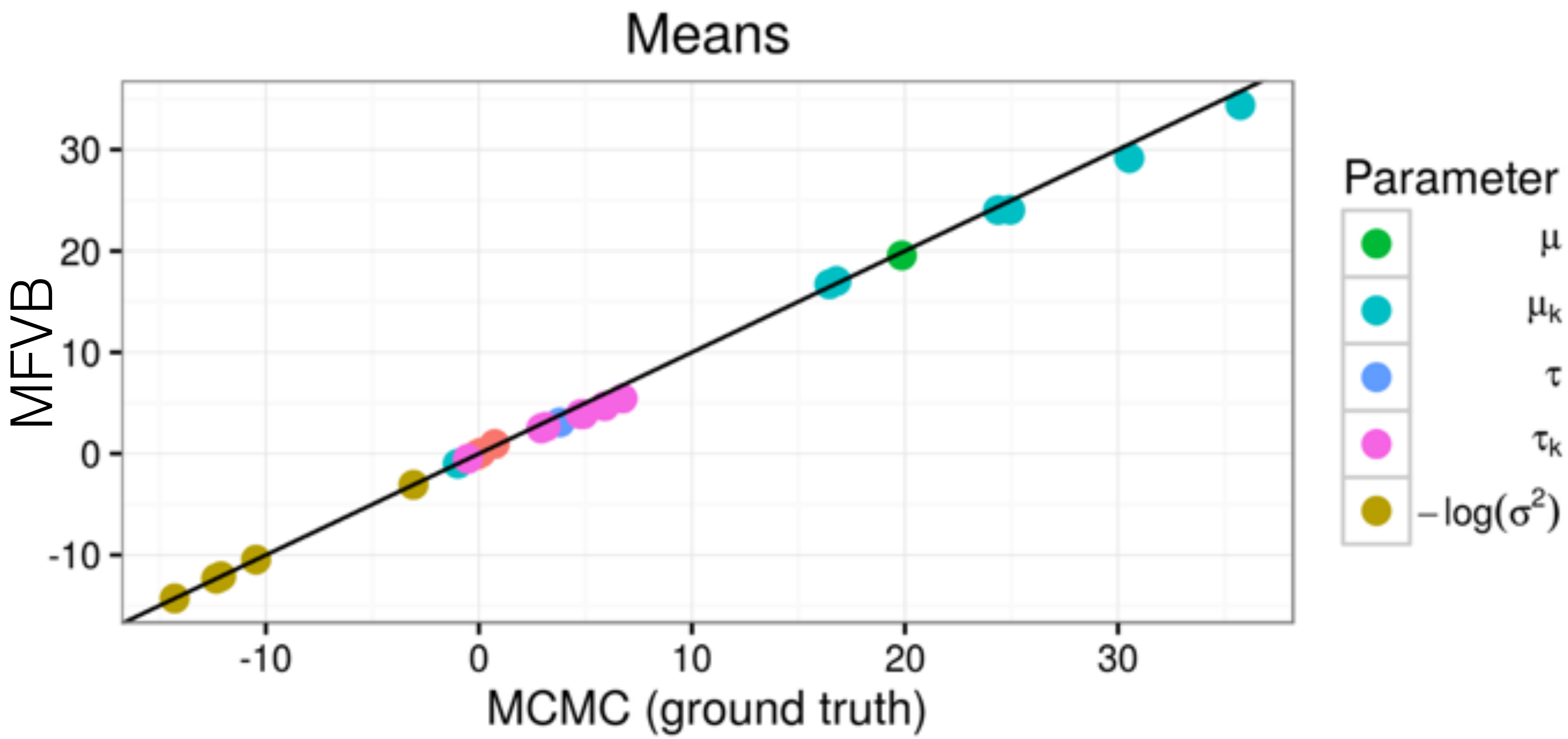
Microcredit Experiment

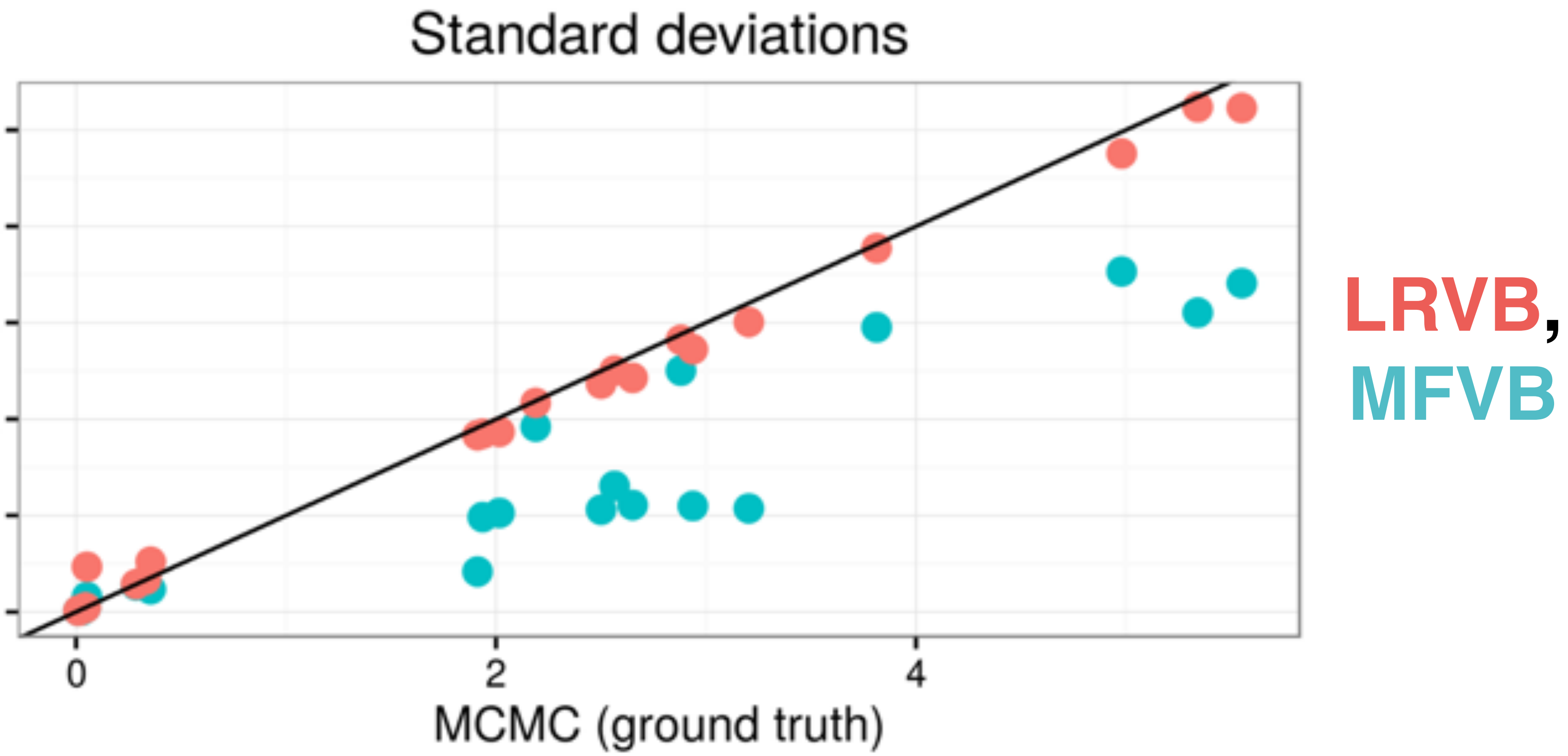
- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**



Microcredit Experiment

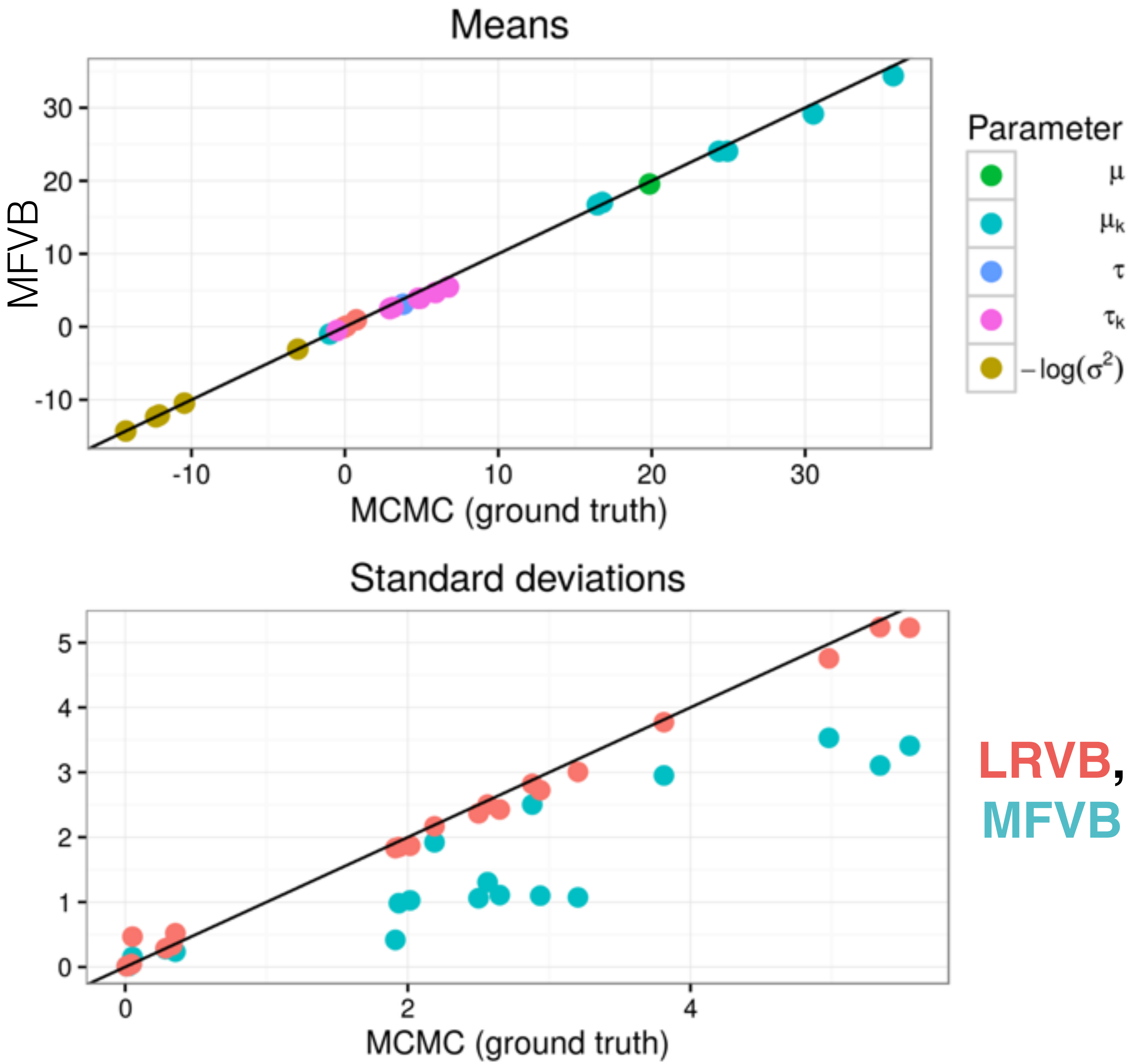
- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**





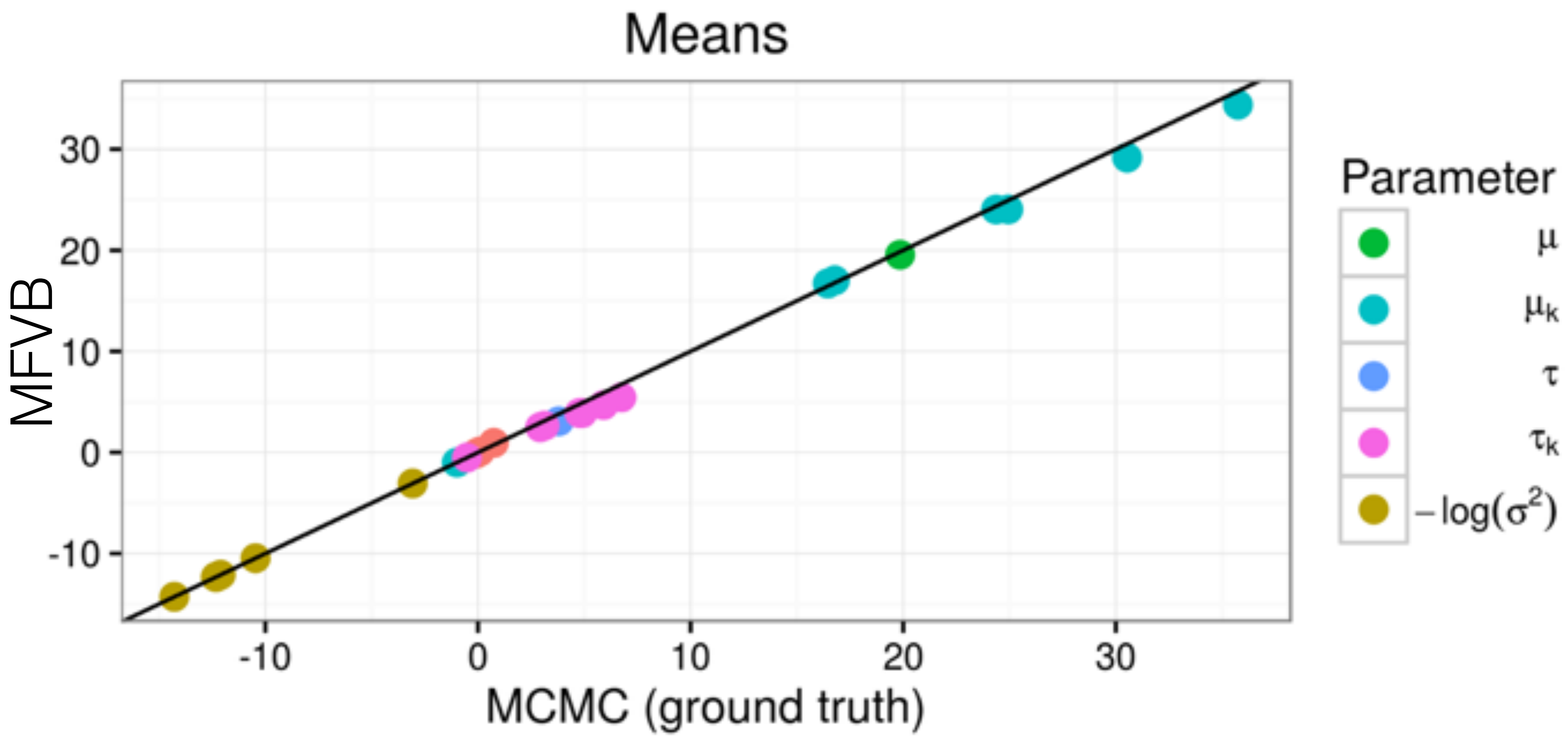
Microcredit Experiment

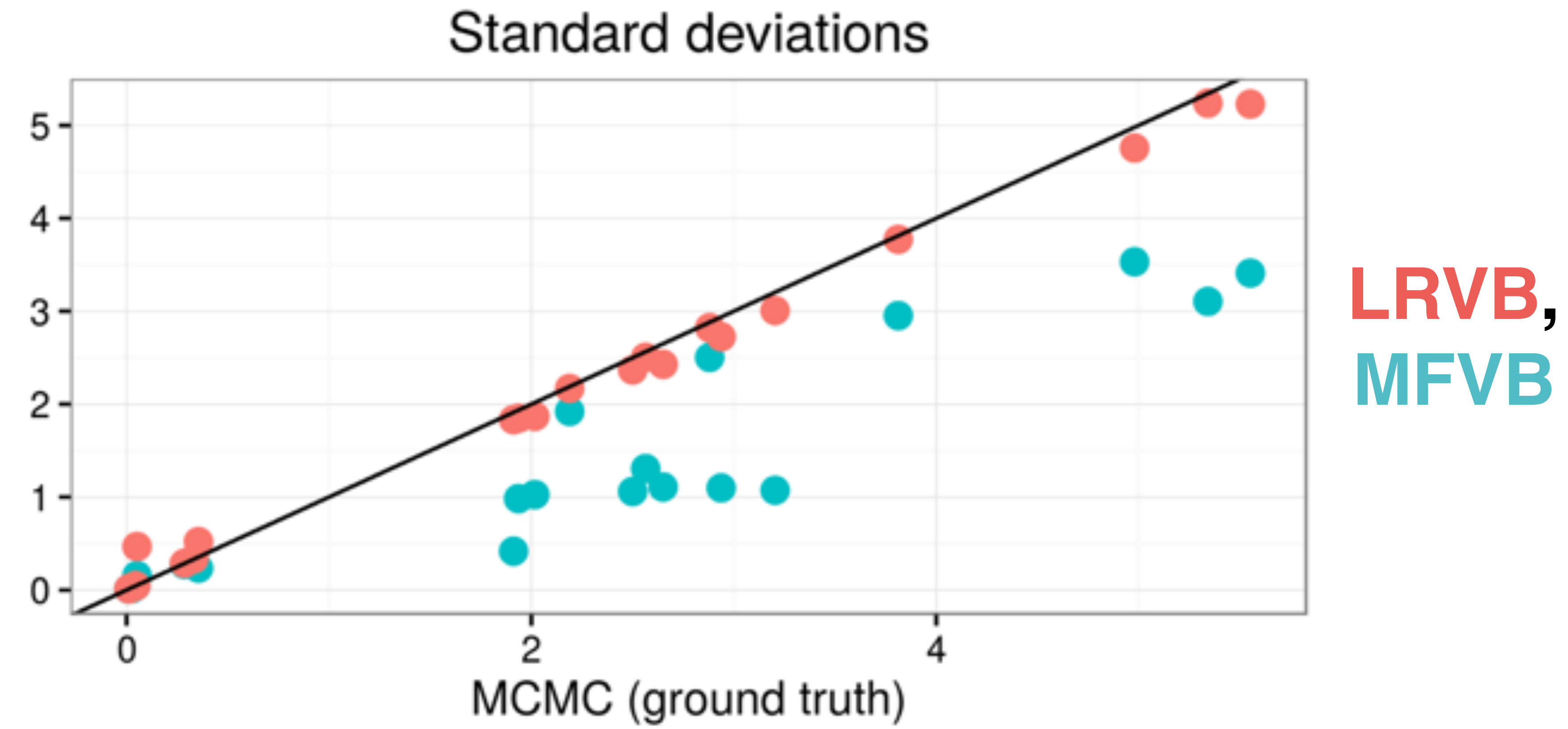
- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**
- *τ* mean (MFVB): 3.08 USD PPP



Microcredit Experiment

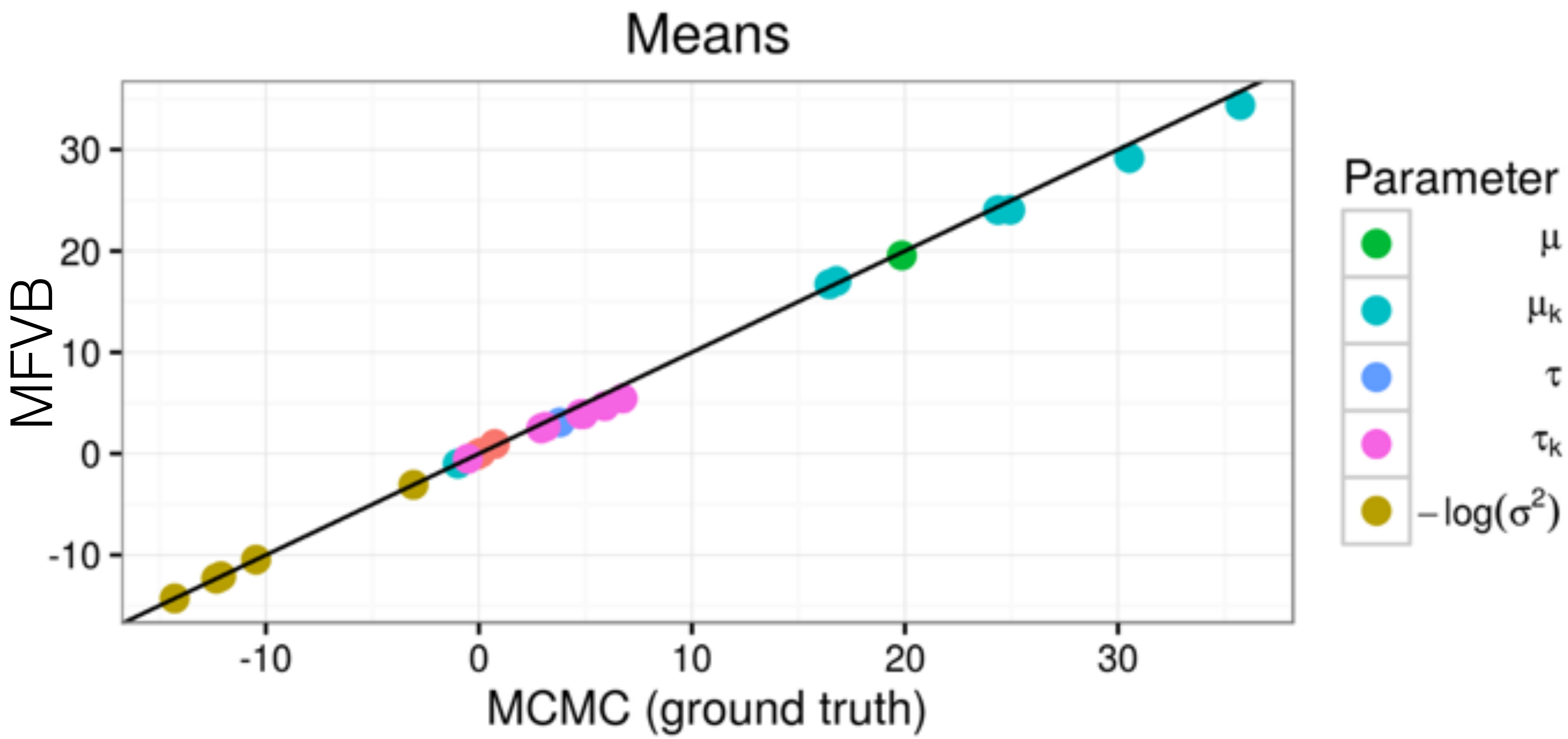
- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**
- τ mean (MFVB): 3.08 USD PPP
- τ std dev (LRVB): 1.83 USD PPP

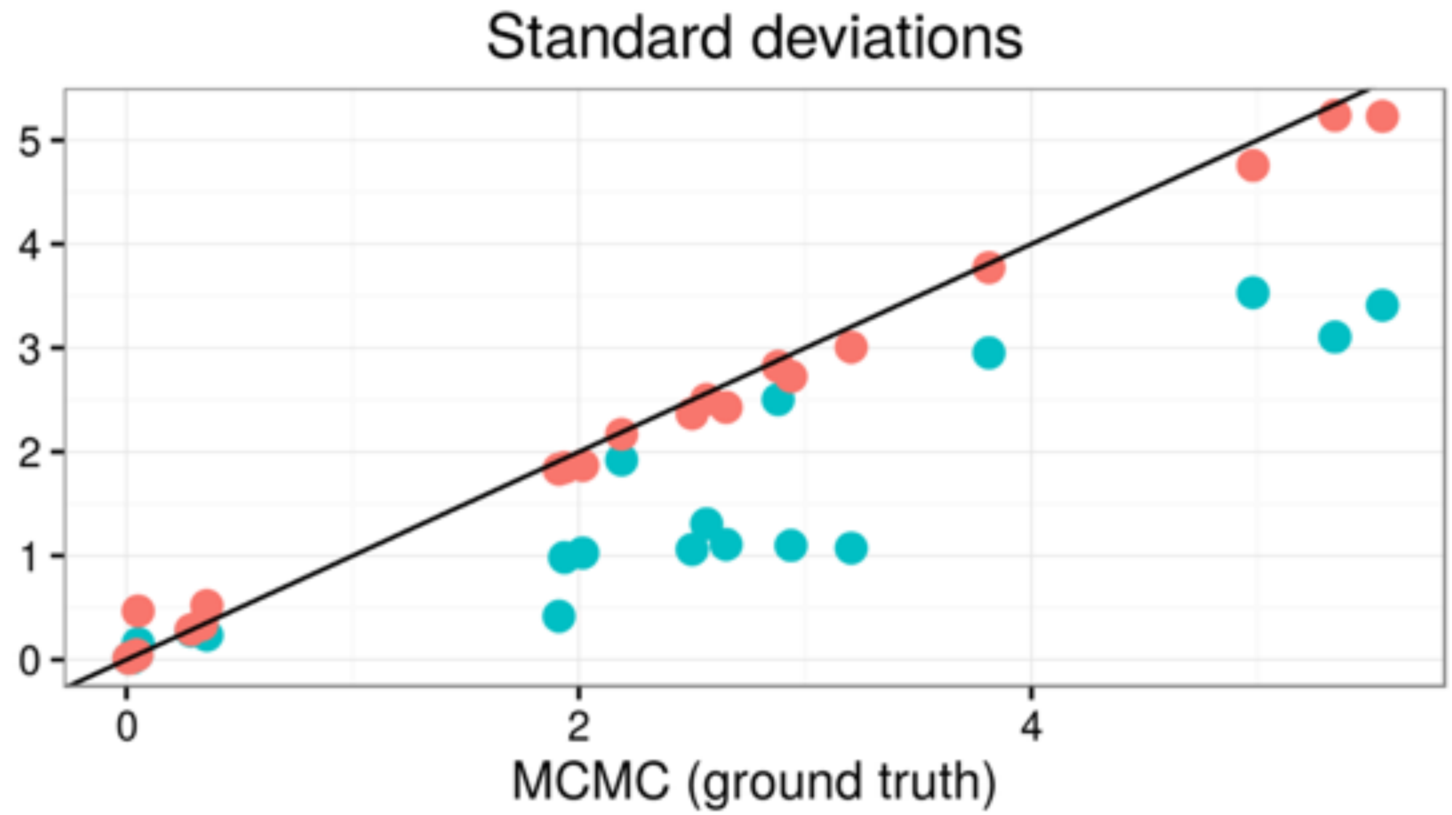




Microcredit Experiment

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





LRVB, **MFVB**

Roadmap

- Challenges of VB
- Accurate uncertainties from VB
- Accurate robustness quantification from VB

Roadmap

- Challenges of VB
- Accurate uncertainties from VB
- Accurate robustness quantification from VB

Robustness quantification

- Bayes Theorem

$$p(\theta|y)$$

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

Robustness quantification

- Bayes Theorem

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

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

Robustness quantification

- Bayes Theorem

$$\begin{aligned} p_\alpha(\theta) &:= p(\theta|y,\alpha) \\ &\propto_\theta p(y|\theta)p(\theta|\alpha) \end{aligned}$$

Robustness quantification

- Bayes Theorem

$$
\begin{aligned}
p_\alpha(\theta) := p(\theta|y,\alpha) \\
\propto_\theta p(y|\theta)p(\theta|\alpha)
\end{aligned}
$$

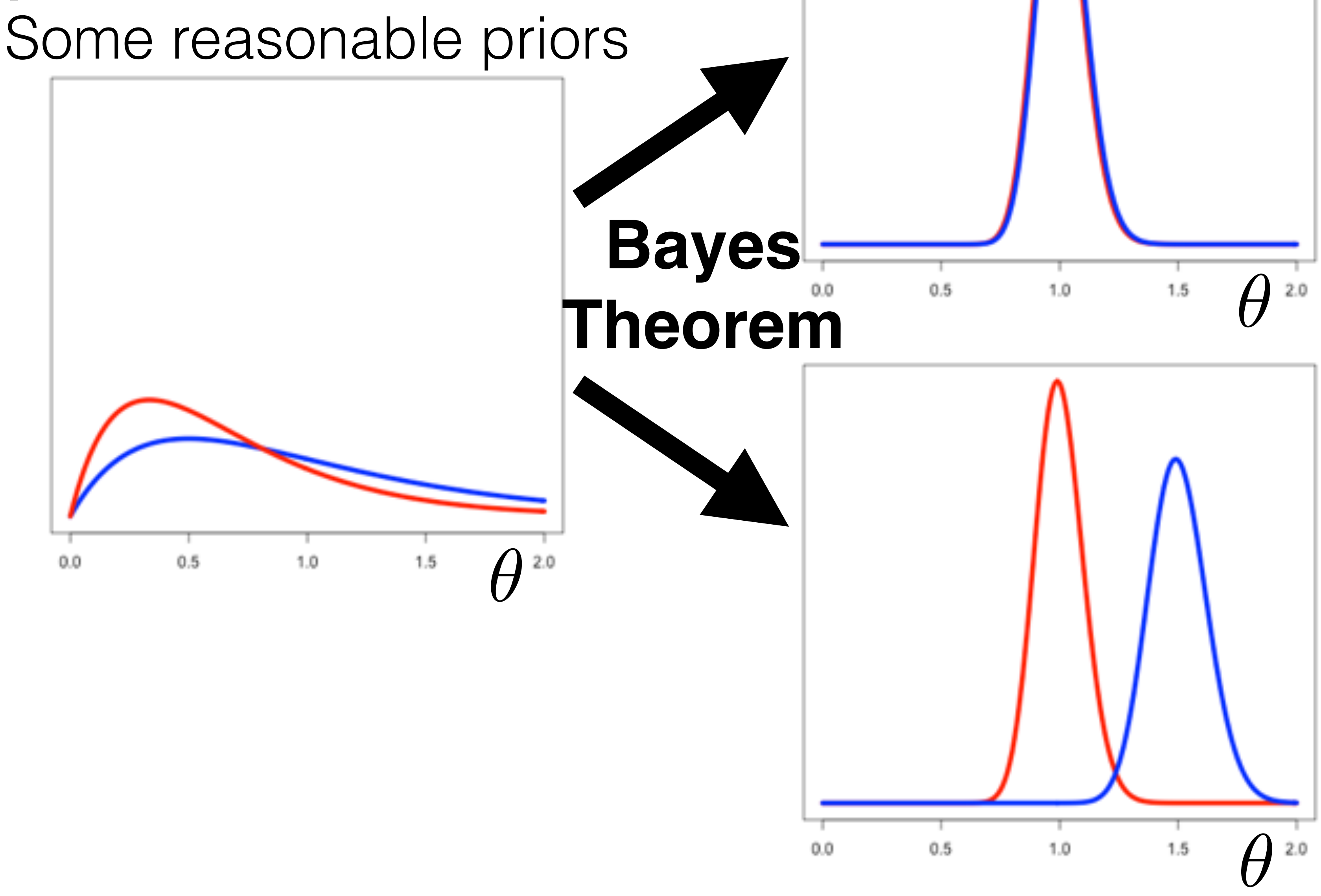
- Sensitivity

Robustness quantification

- Bayes Theorem

$$\begin{aligned} p_\alpha(\theta) &:= p(\theta|y,\alpha) \\ &\propto_\theta p(y|\theta)p(\theta|\alpha) \end{aligned}$$

- Sensitivity


Some reasonable priors

Bayes Theorem

Robustness quantification

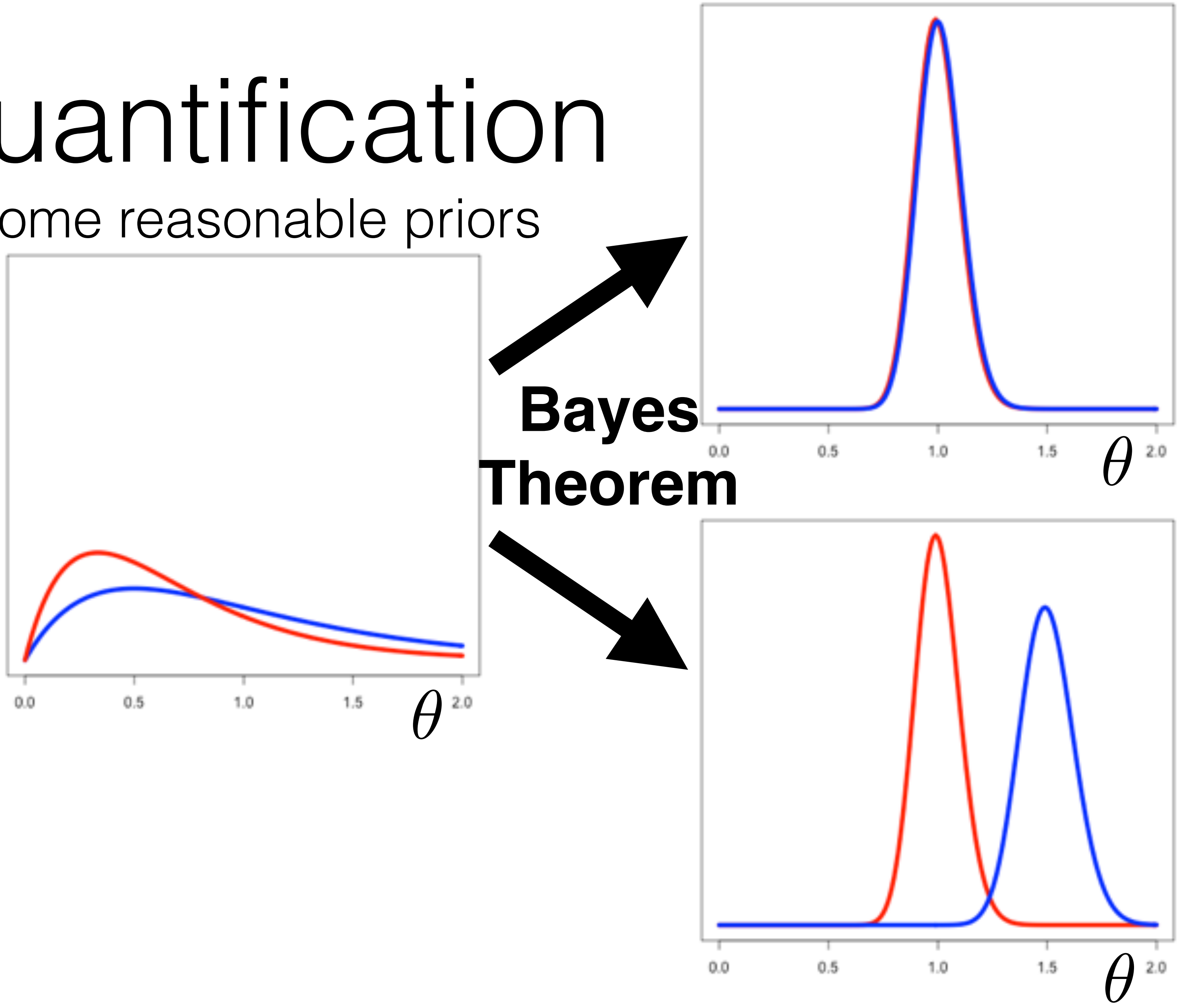
- Bayes Theorem

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

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

- Sensitivity

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


Some reasonable priors

Bayes Theorem

Robustness quantification

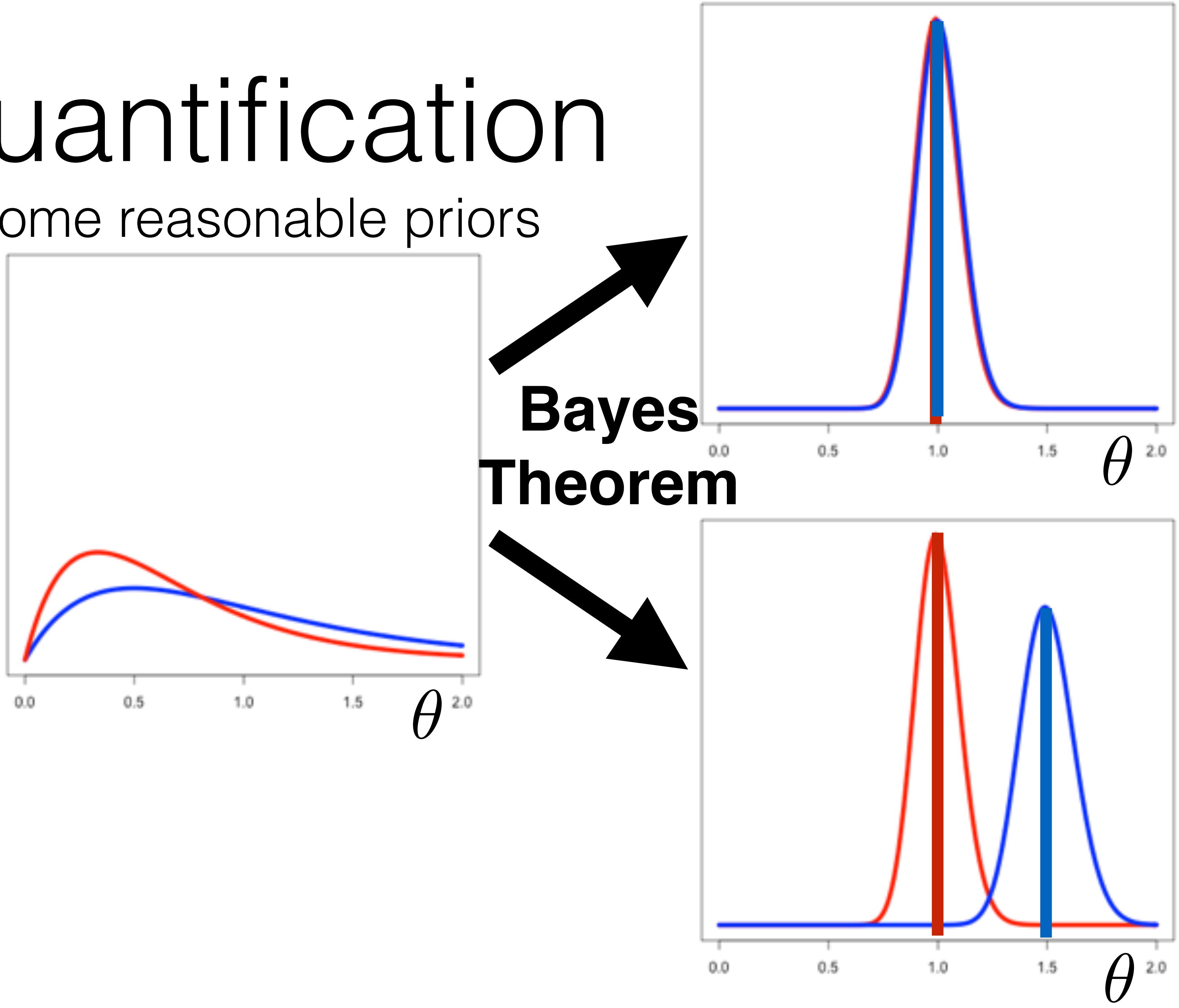
- Bayes Theorem

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

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

- Sensitivity

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



Some reasonable priors

Bayes Theorem

Robustness quantification

- Bayes Theorem

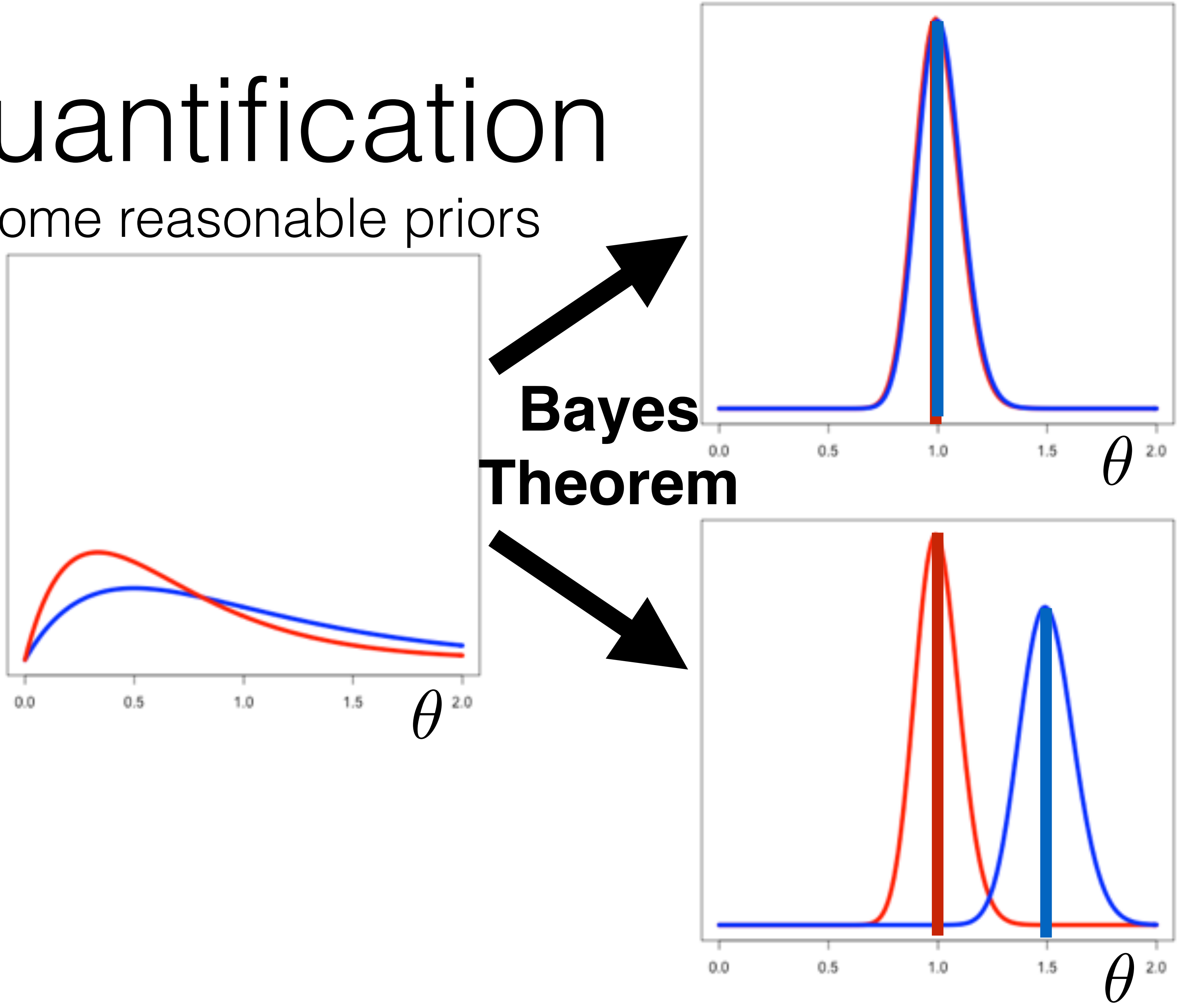
$$p_\alpha(\theta) := p(\theta|y, \alpha)$$

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

- Sensitivity (local)

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

Some reasonable priors



Bayes Theorem

Robustness quantification

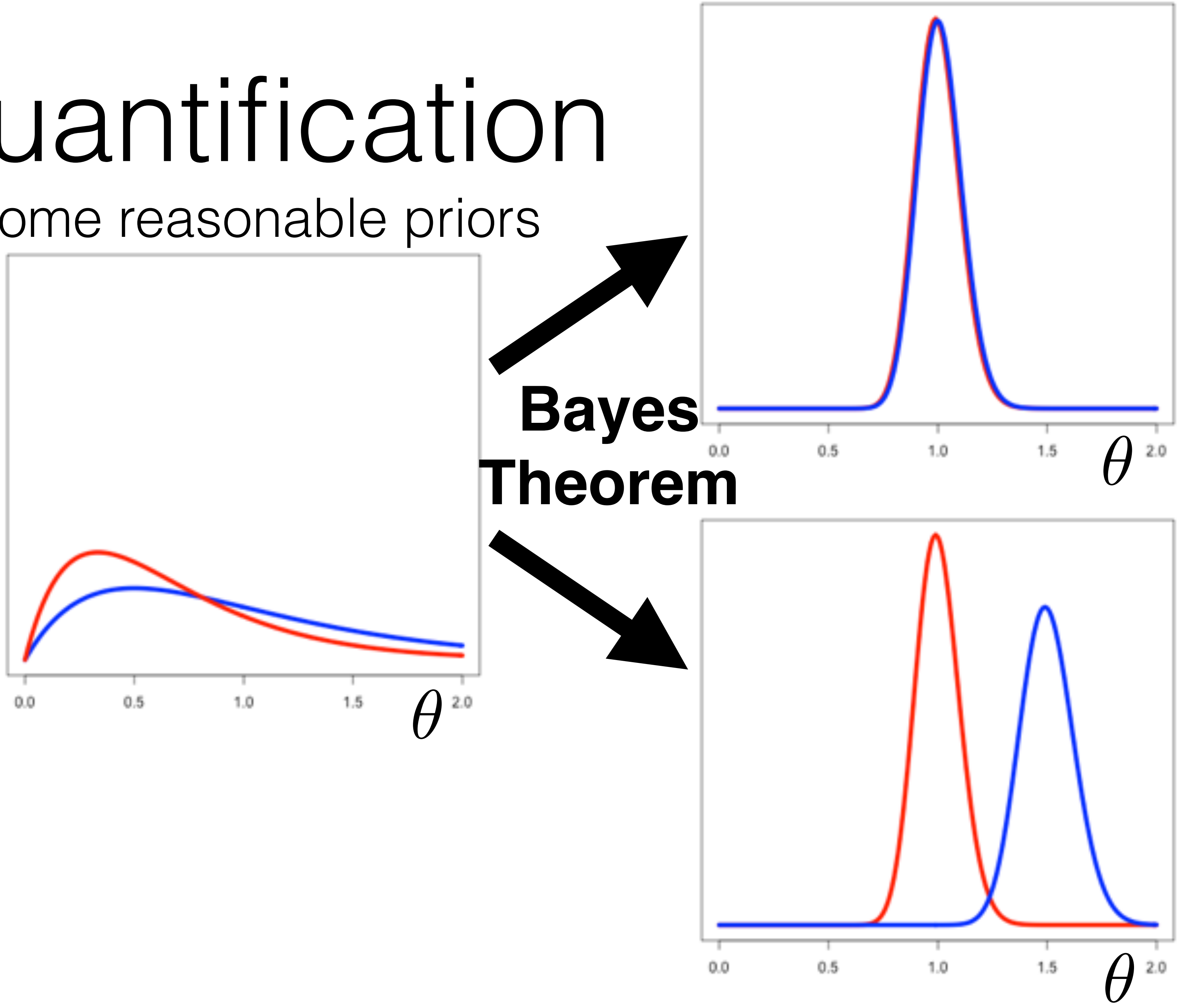
- Bayes Theorem

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

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

- Sensitivity (local)

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



Some reasonable priors

Bayes Theorem

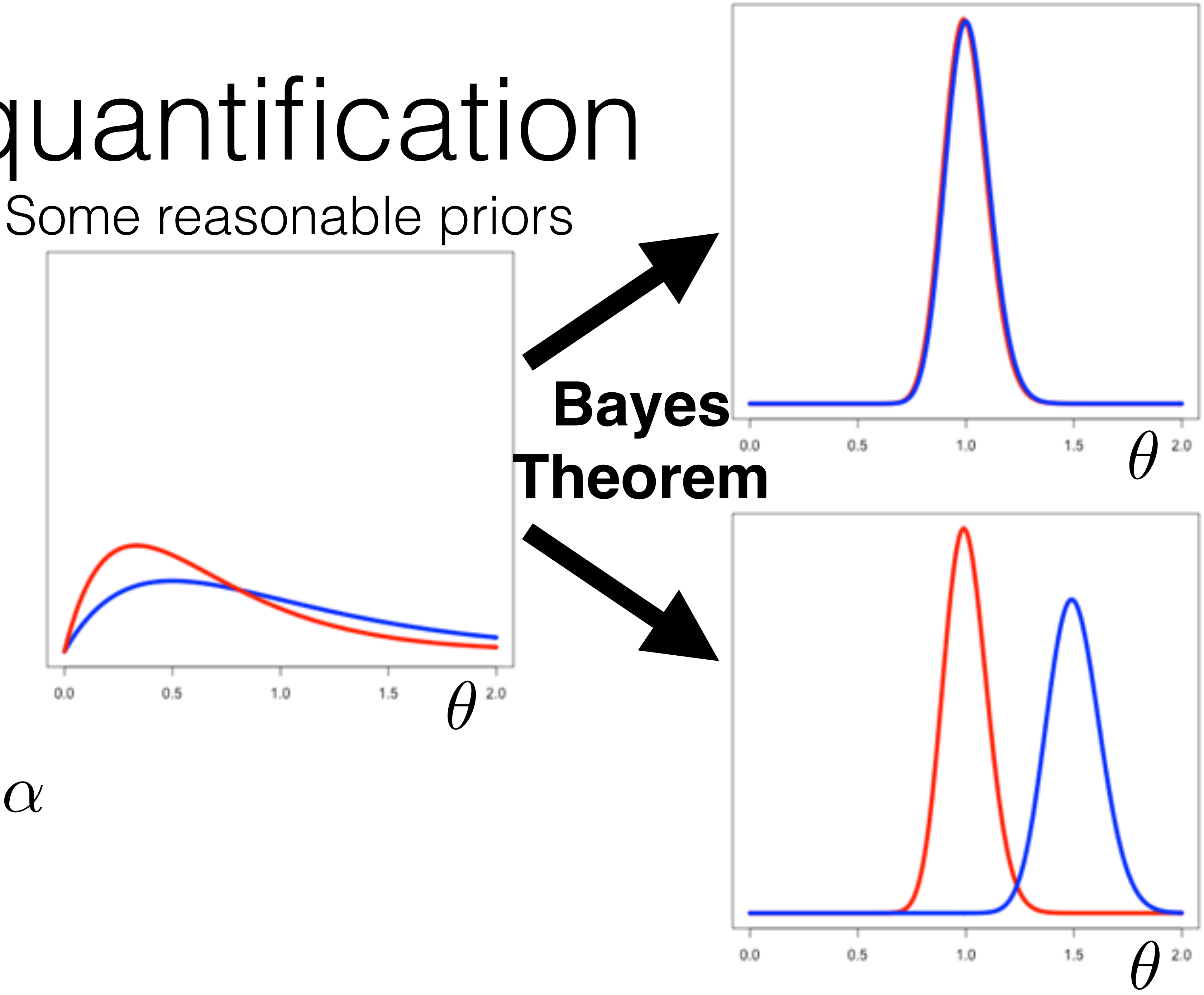
Robustness quantification

- Bayes Theorem

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

- Sensitivity (local)

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



Some reasonable priors

Bayes Theorem

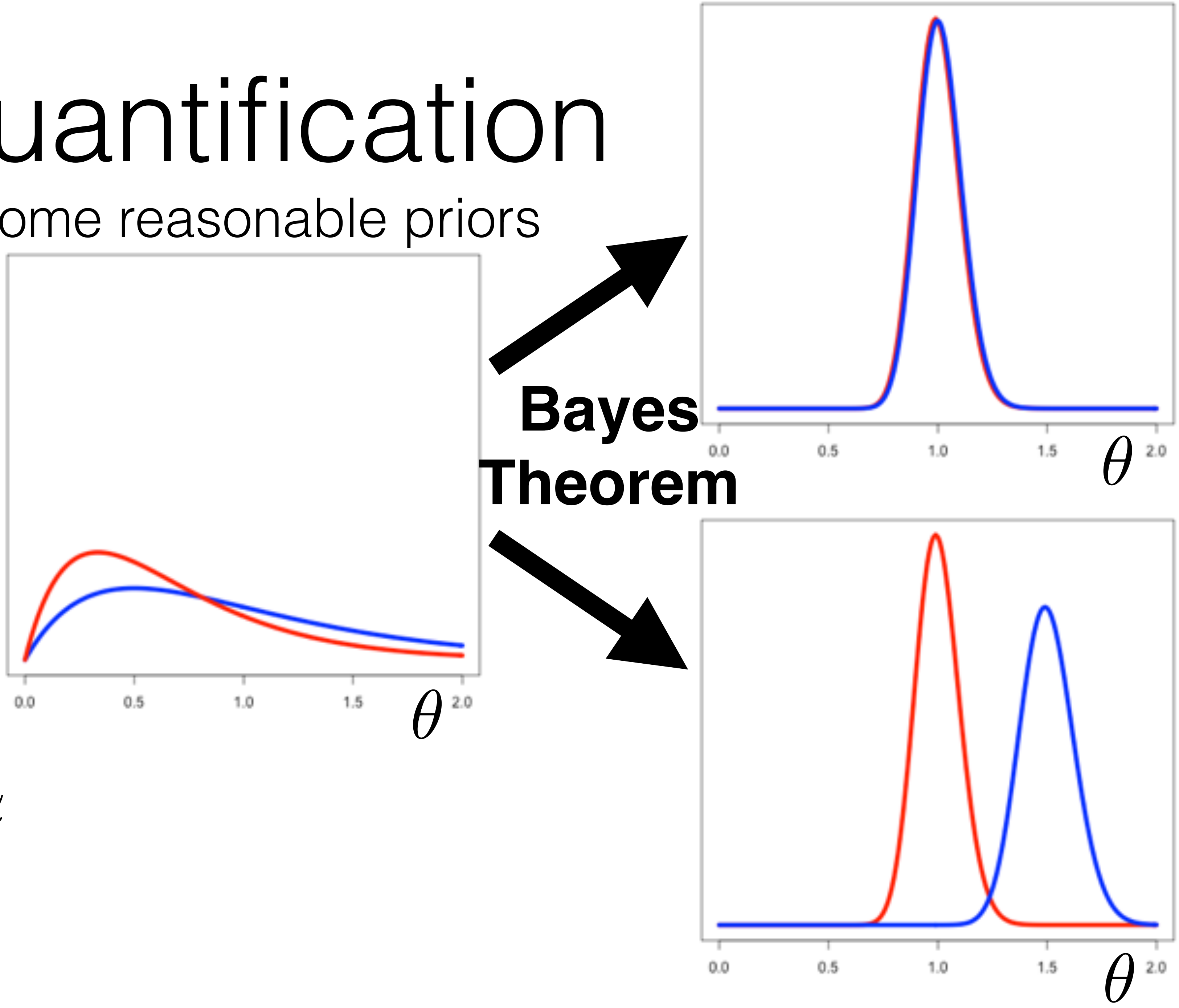
Robustness quantification

- Bayes Theorem

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

- Sensitivity (local)

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


Some reasonable priors

Bayes Theorem

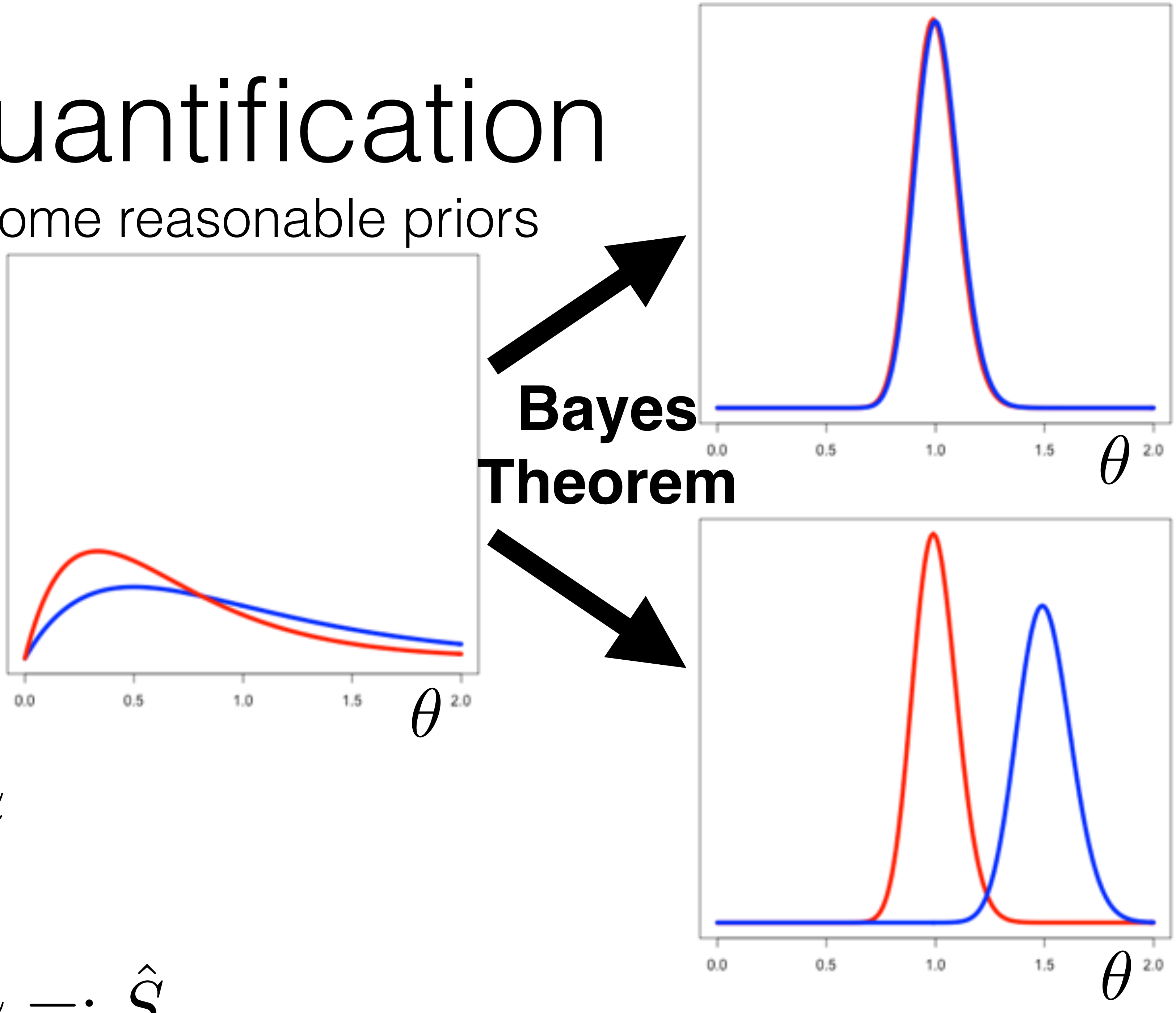
Robustness quantification

- Bayes Theorem

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

- Sensitivity (local)

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


Some reasonable priors

Bayes Theorem

Robustness quantification

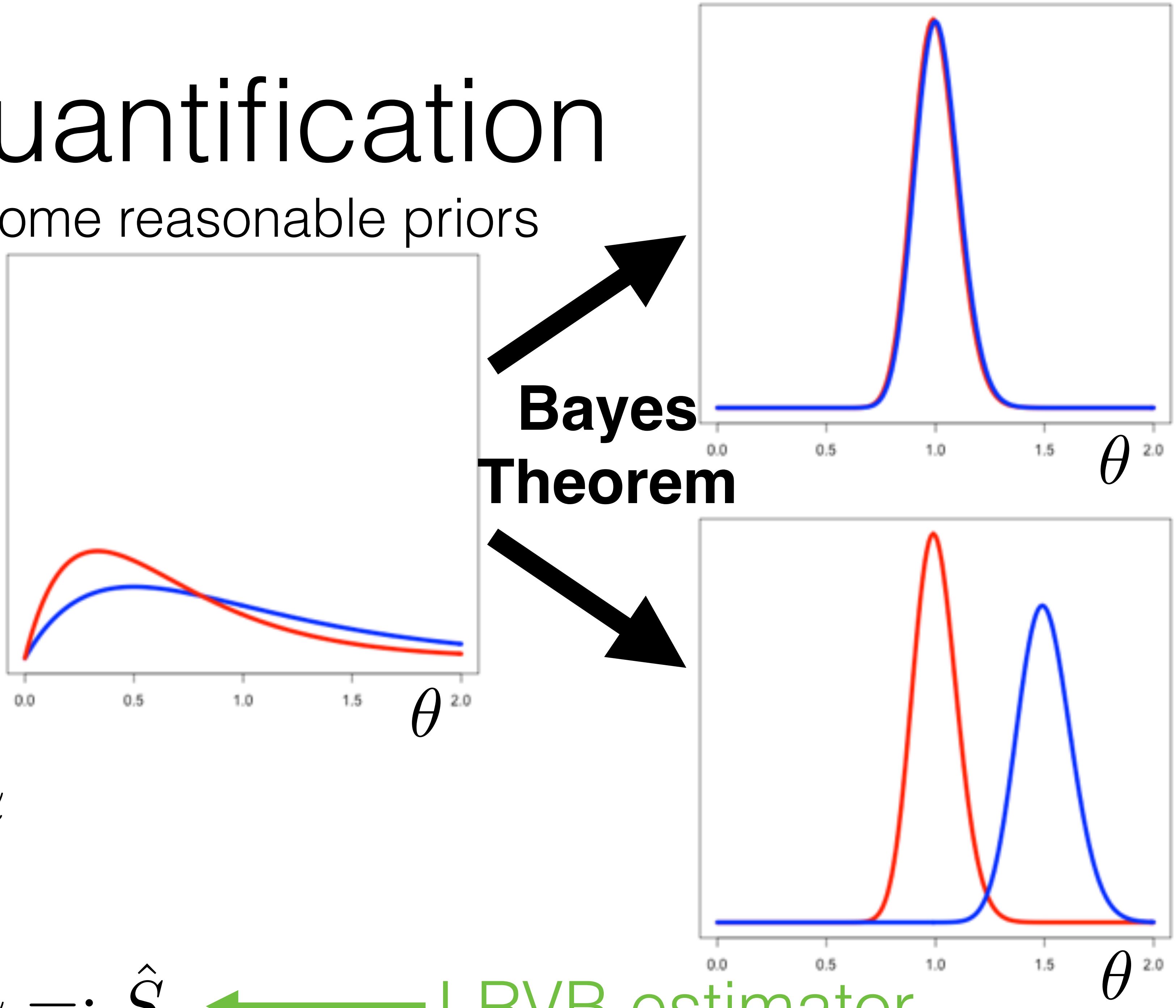
- Bayes Theorem

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

- Sensitivity (local)

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

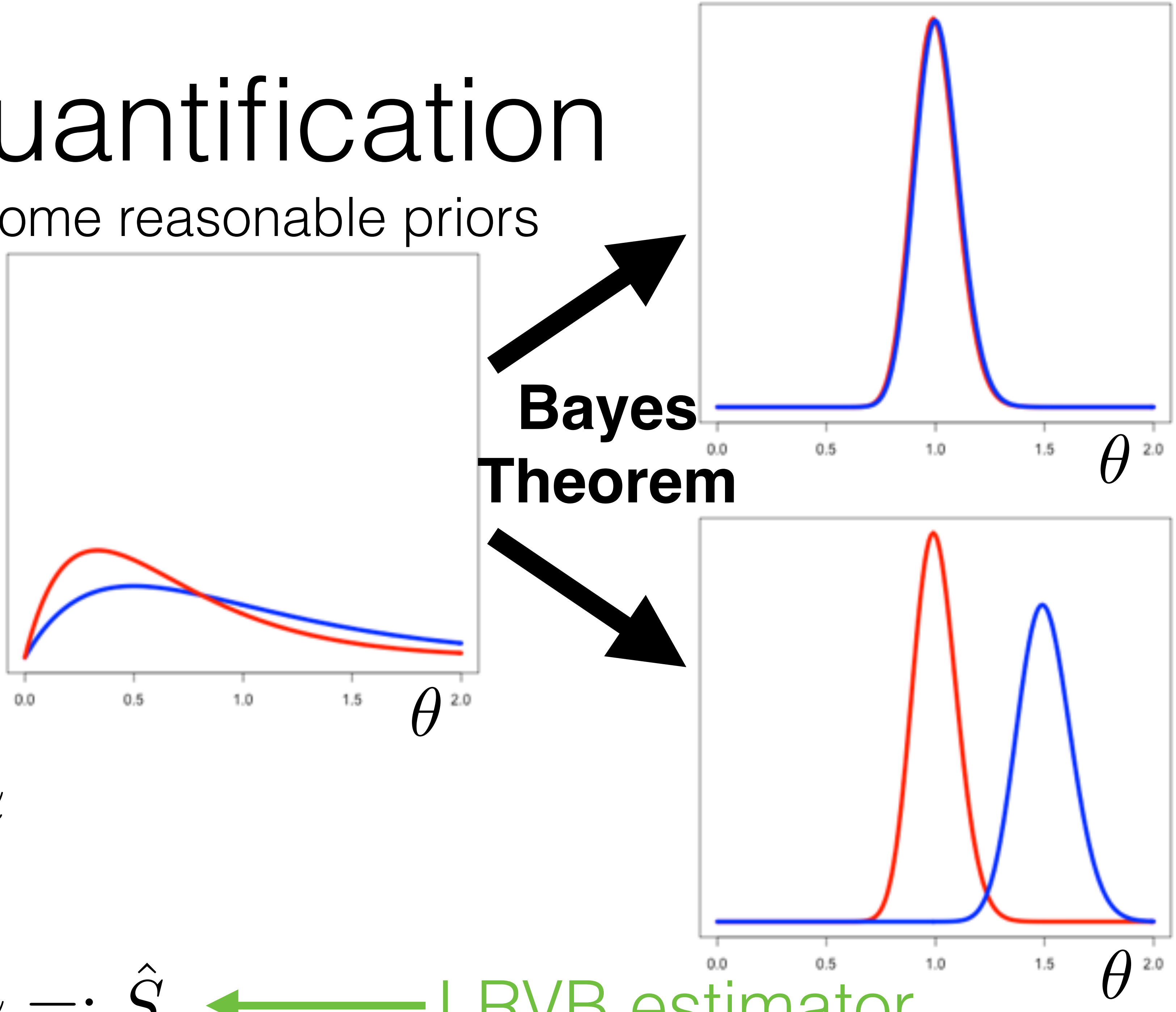
Some reasonable priors



Bayes Theorem

LRVB estimator

Robustness quantification



- Bayes Theorem

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

- Sensitivity (local)

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

LRVB estimator

- Recall: our general LRVB formula applies for:

$$\log p_t(\theta) = \log p(\theta|y) + f(\theta,t) - \mathrm{Const}(t)$$

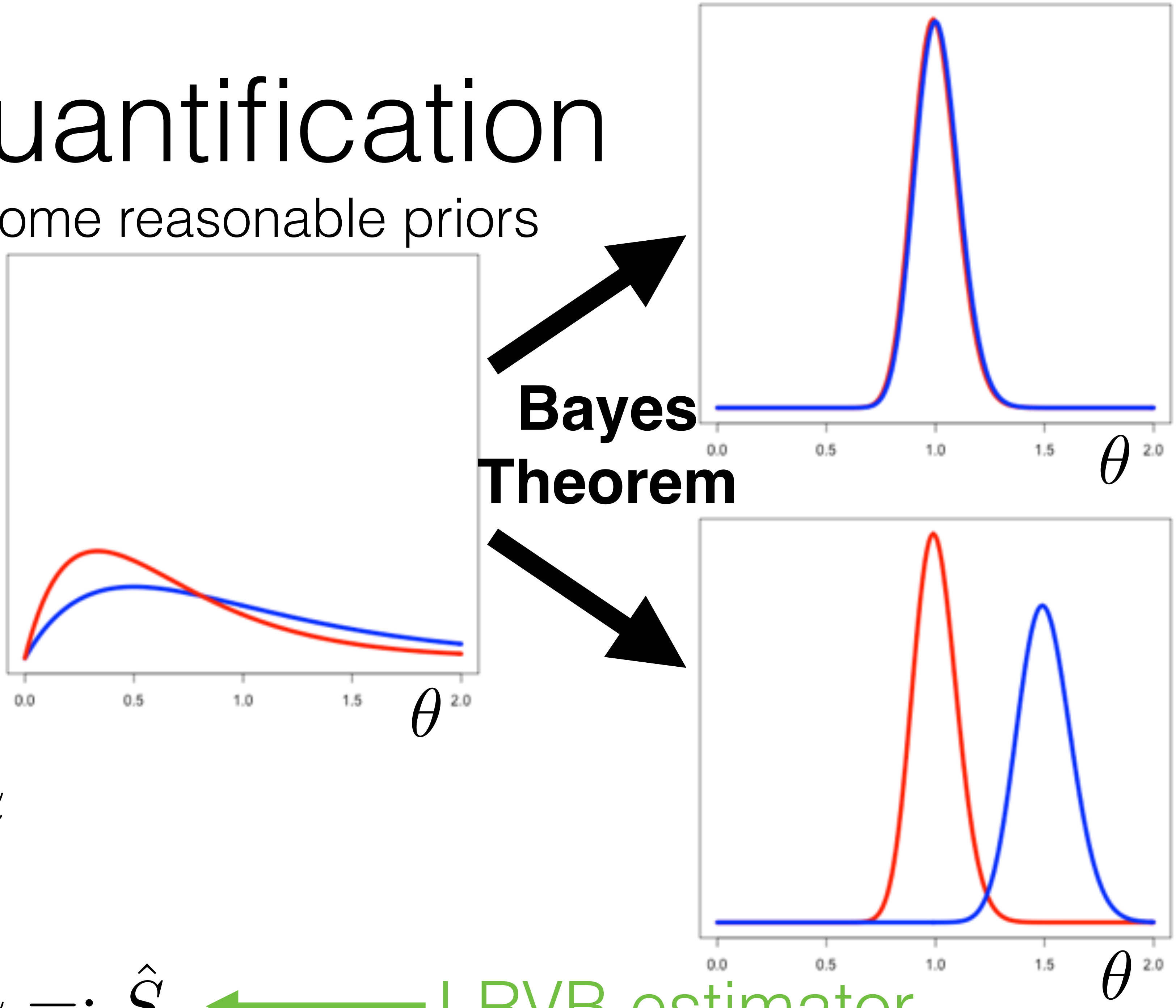
Robustness quantification

- Bayes Theorem

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

- Sensitivity (local)

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



Some reasonable priors

Bayes Theorem

LRVB estimator

- Recall: our general LRVB formula applies for:

$$\log p_t(\theta) = \log p(\theta | y) + f(\theta, t) - \text{Const}(t)$$

[board]

Microcredit Experiment

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

1 if microcredit

profit

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

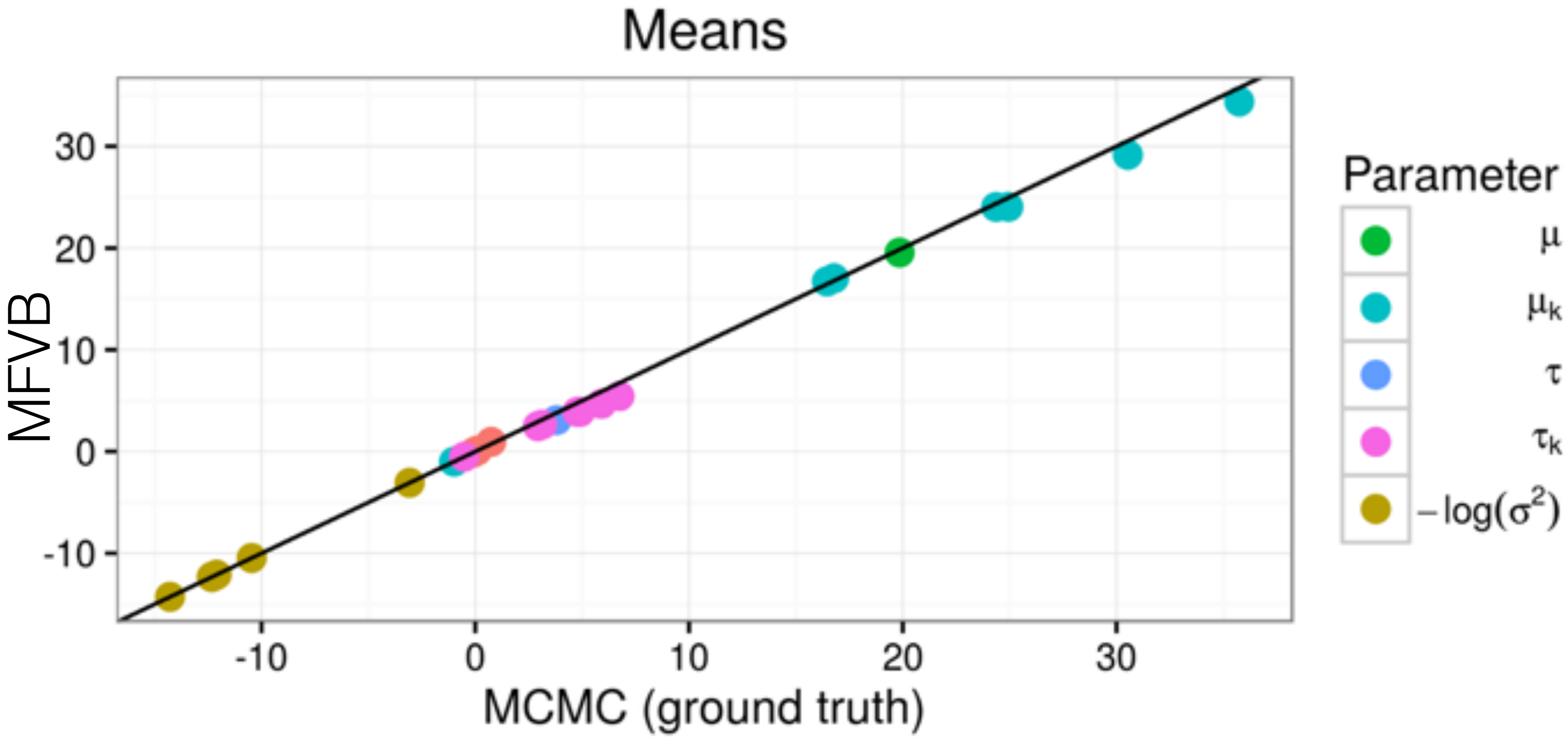
- Priors and hyperpriors:

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

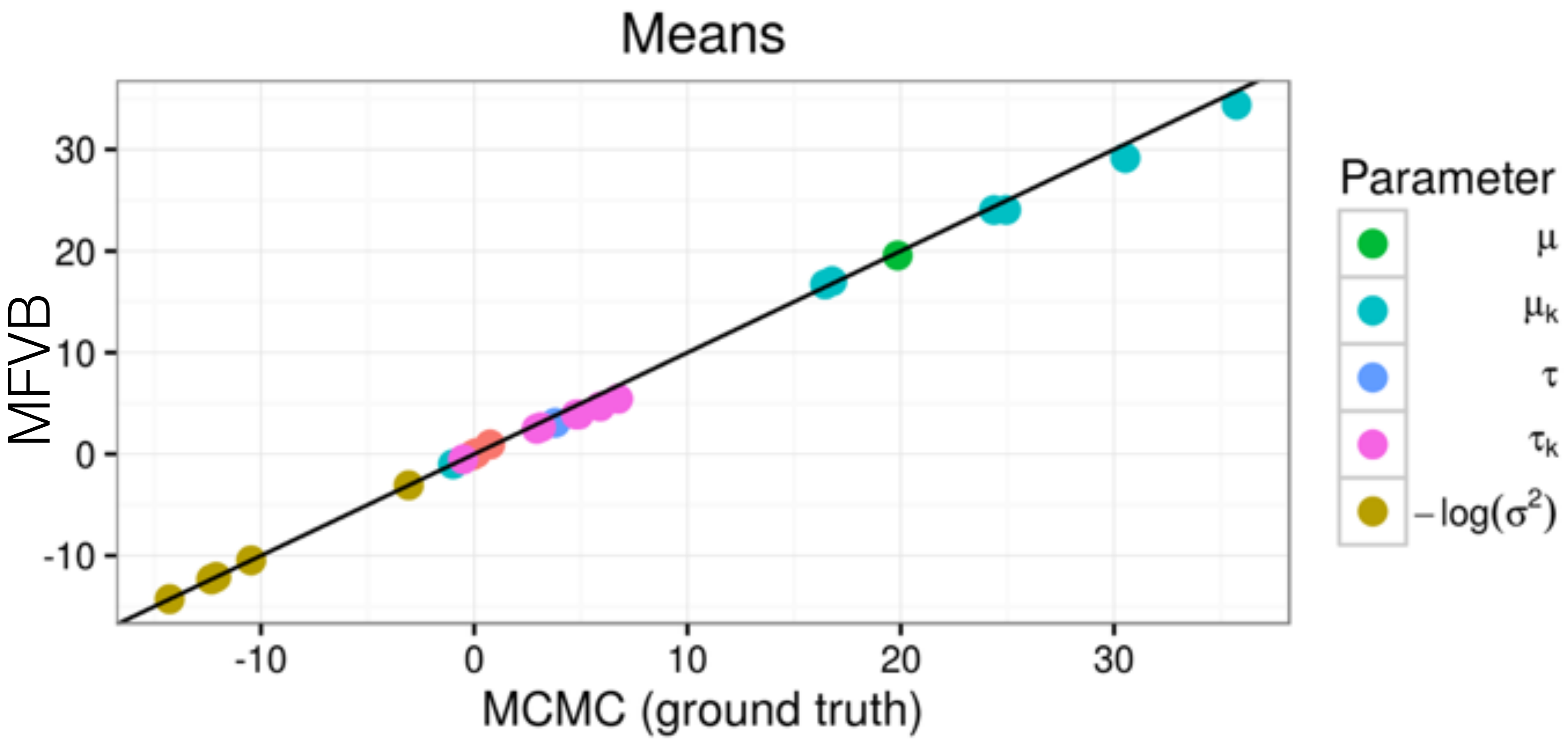
$$\sigma_k^{-2} \overset{iid}{\sim} \Gamma(a, b) \qquad C \sim \mathrm{Sep\&LKJ}(\eta, c, d)$$

Microcredit Experiment

Microcredit Experiment



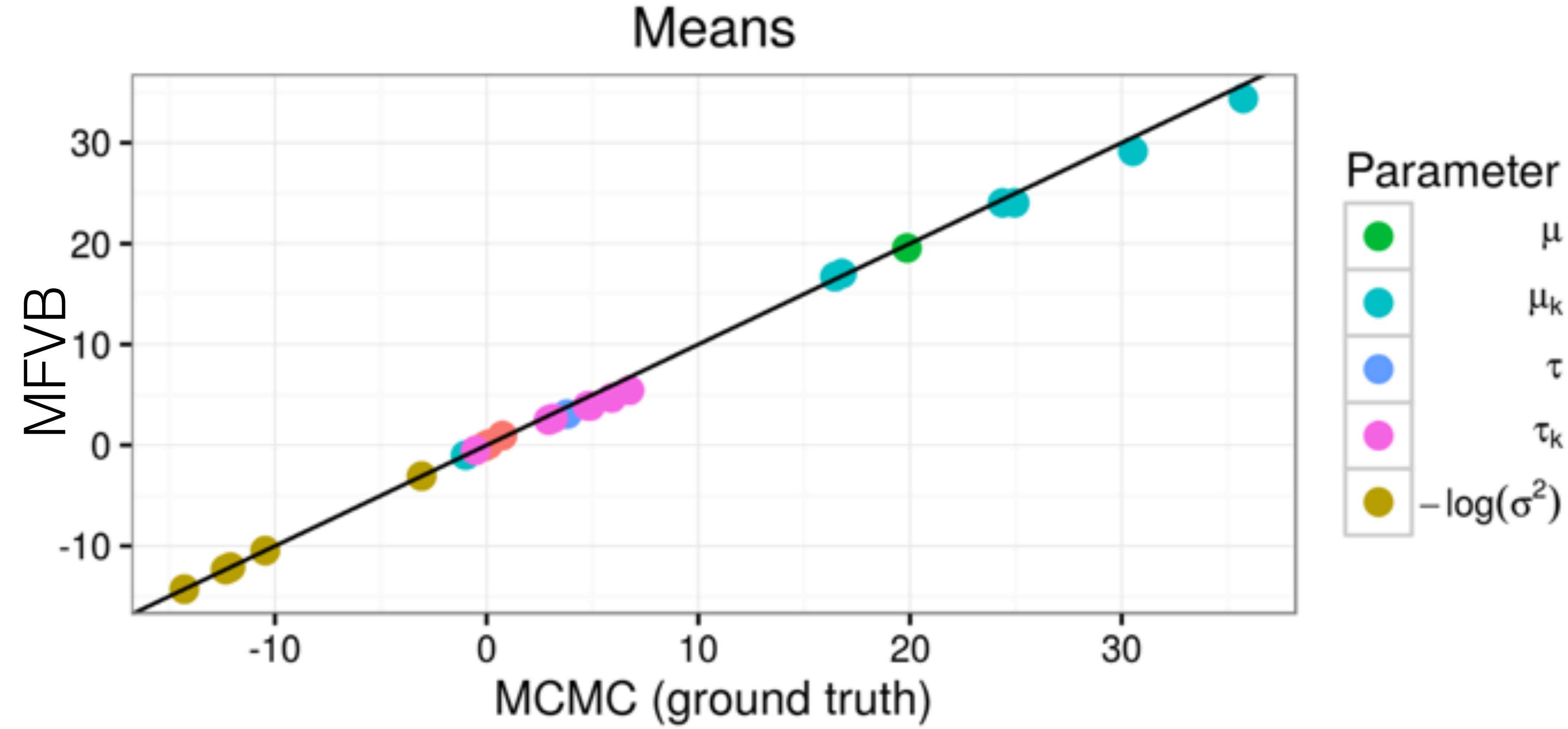
Microcredit Experiment

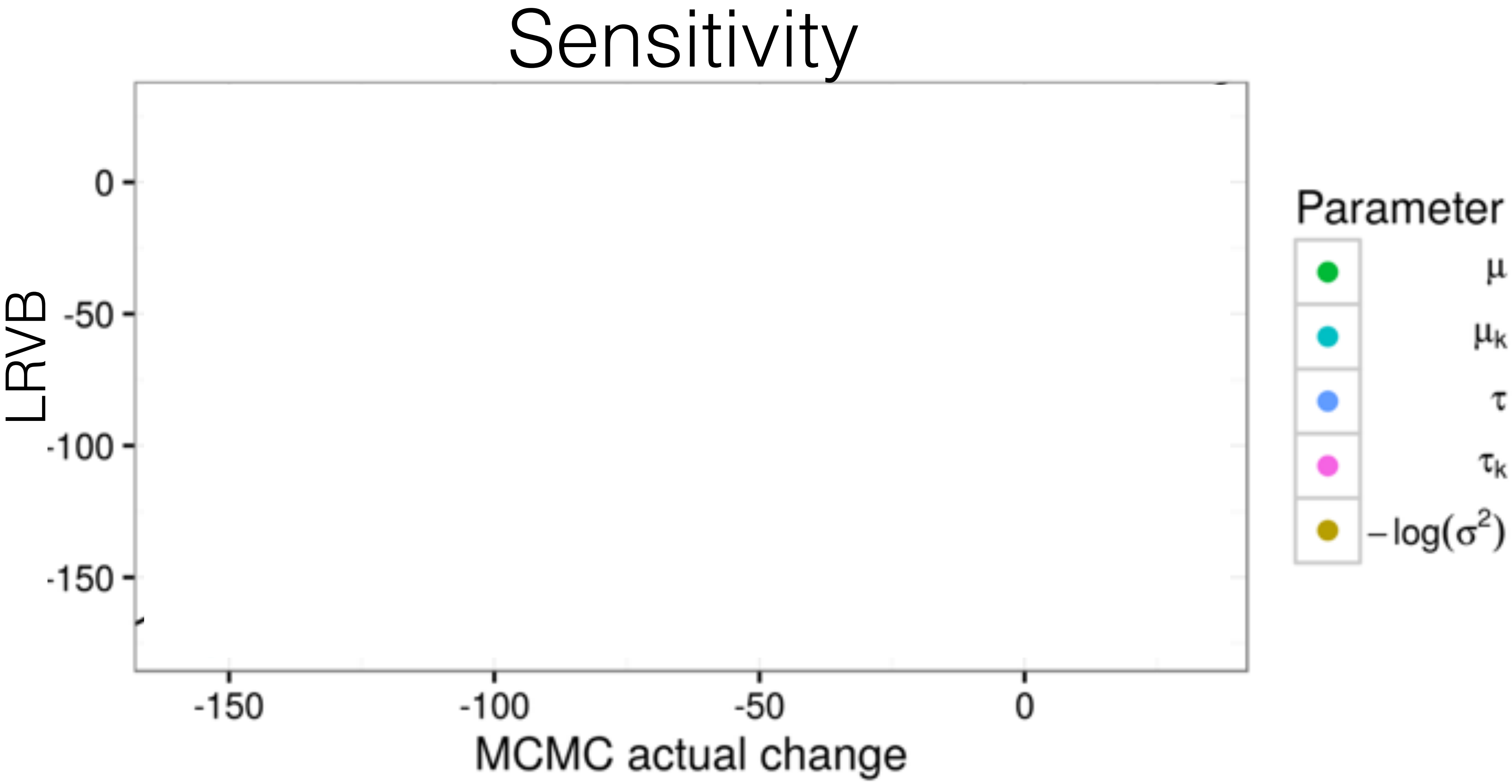


- Perturb Λ_{11}: 0.03 ➔ 0.04

Microcredit Experiment

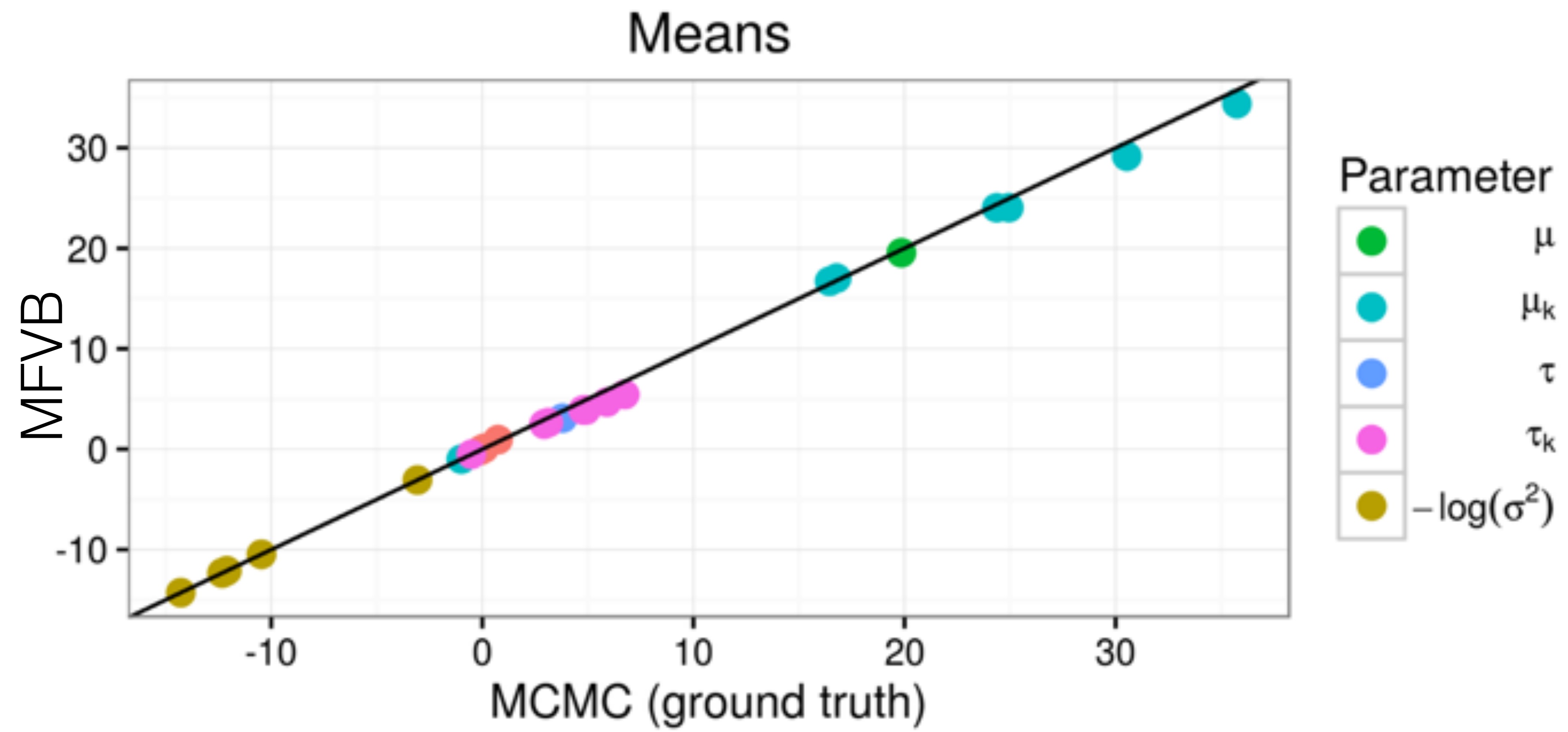
- Perturb Λ_{11}: 0.03 ➔ 0.04



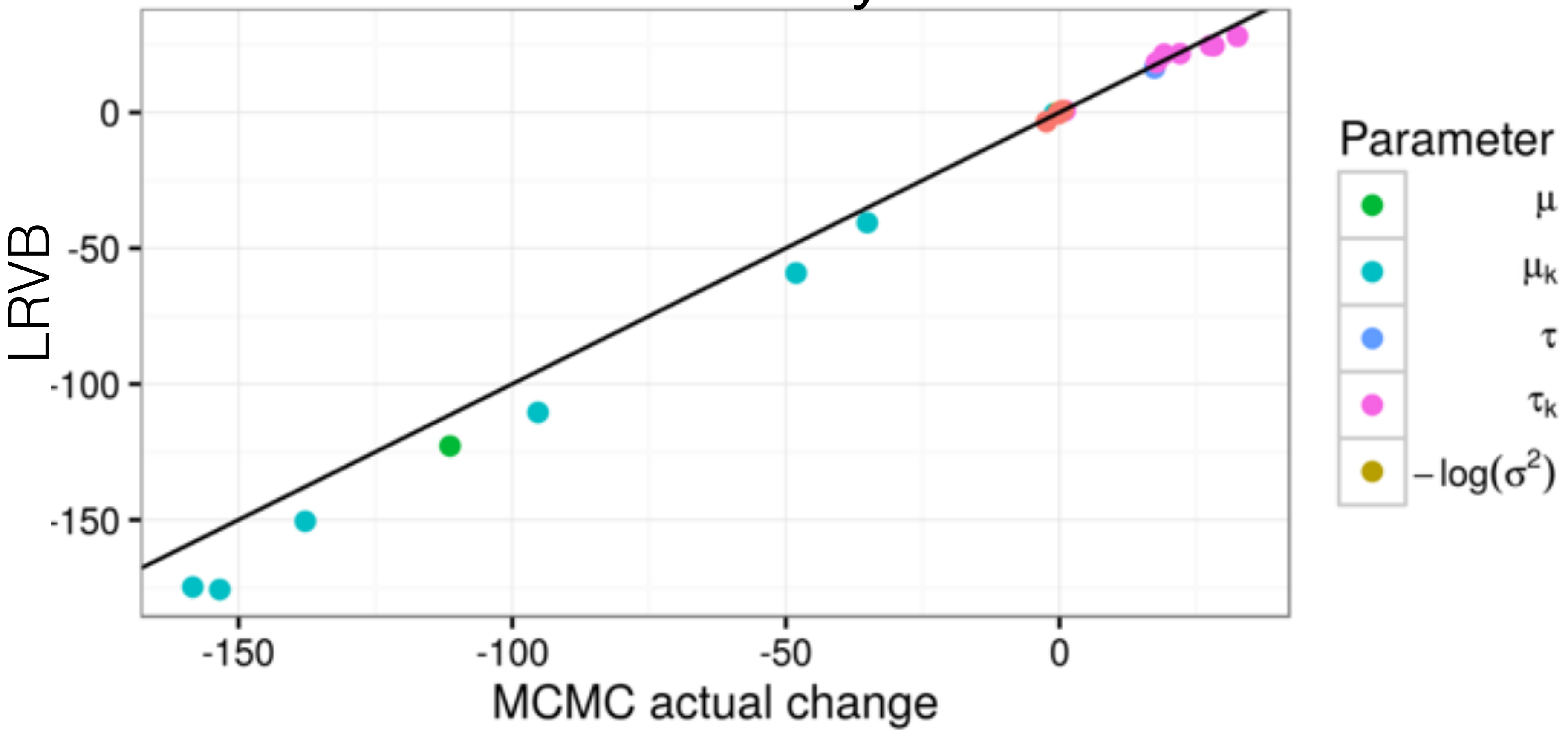


Microcredit Experiment

- Perturb Λ_{11}: 0.03 ➔ 0.04



Sensitivity



Microcredit Experiment

Microcredit Experiment

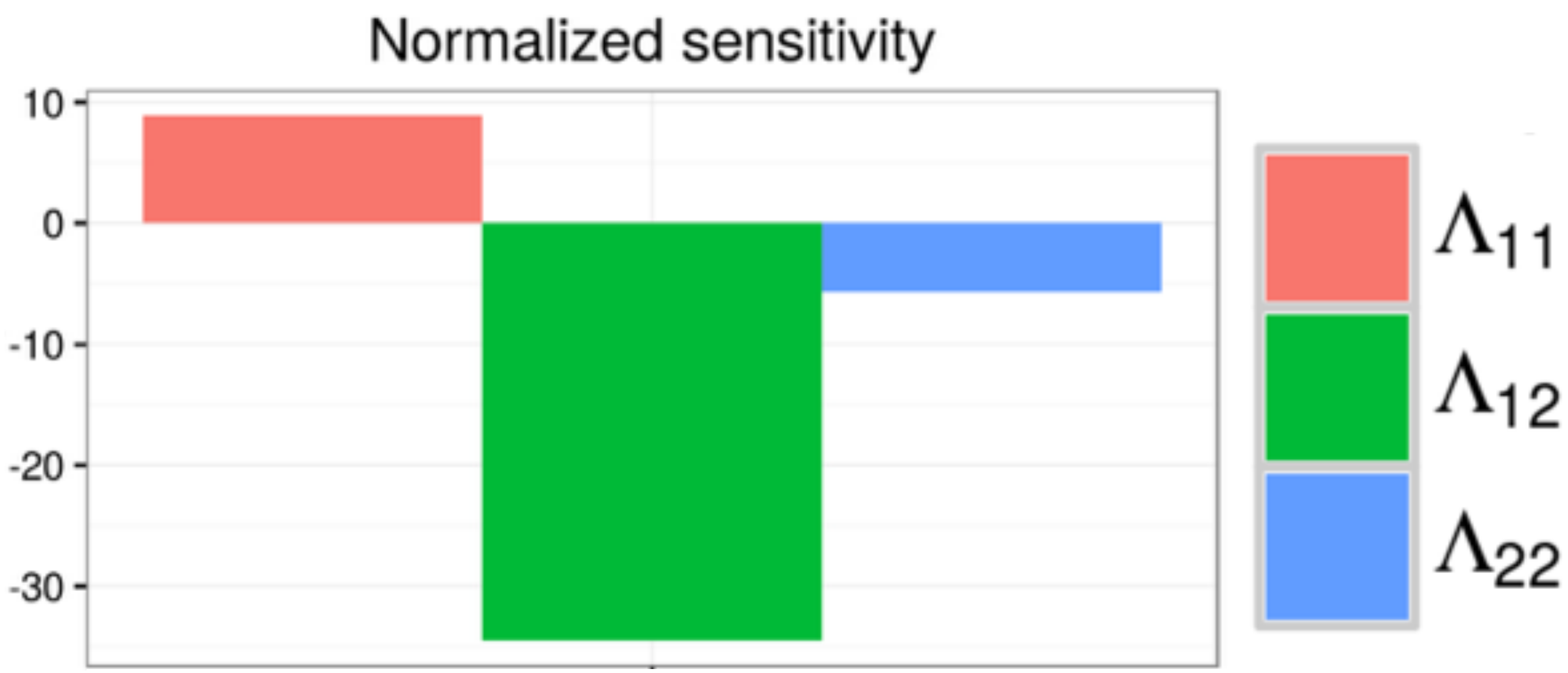
- Sensitivity of the expected microcredit effect (τ)

Microcredit Experiment

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

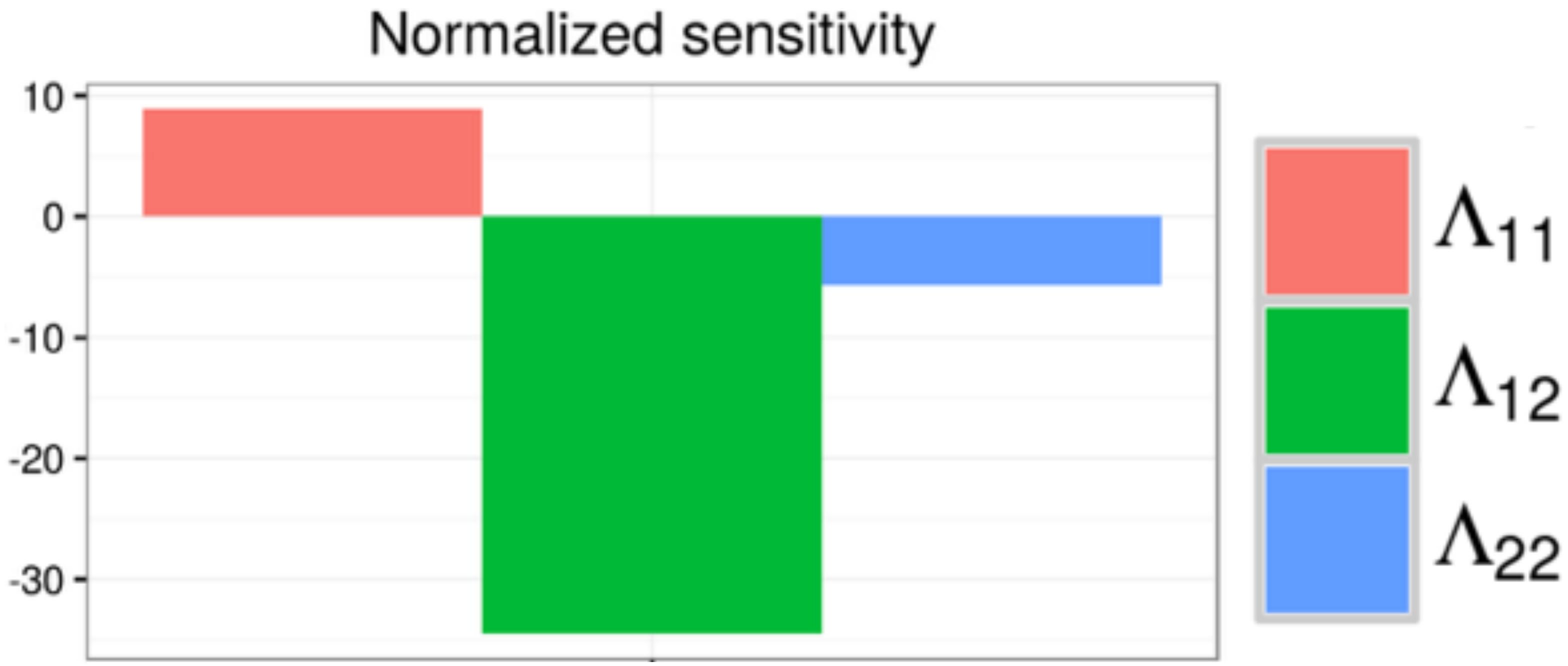
Microcredit Experiment

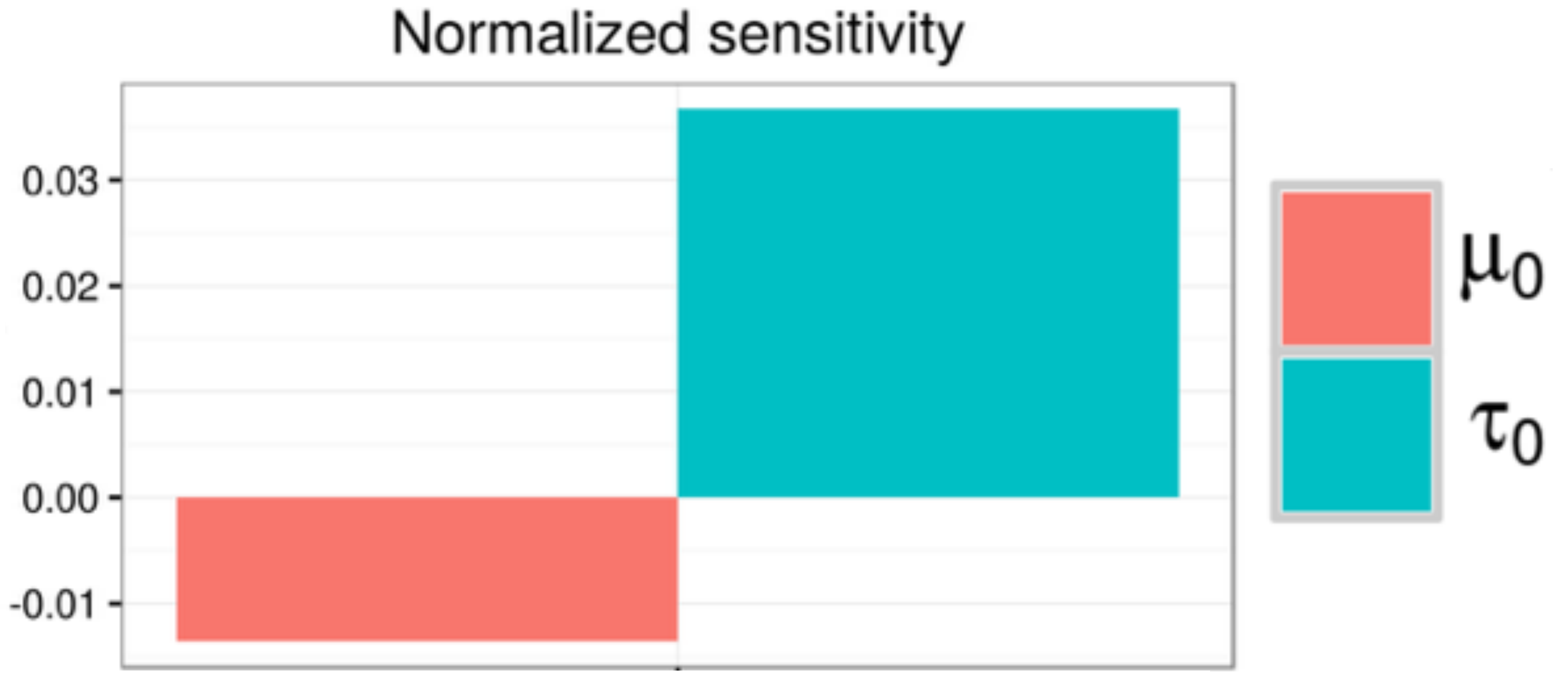
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of τ std devs



Microcredit Experiment

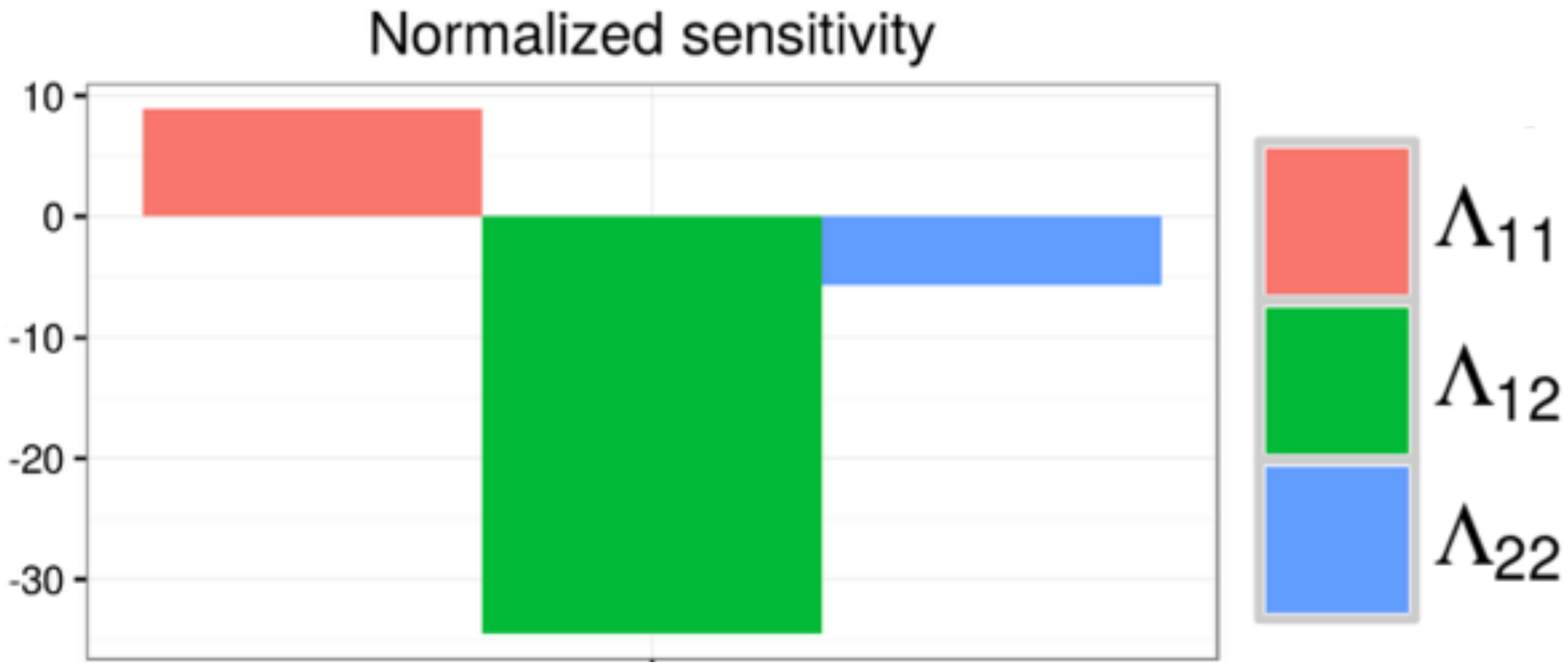
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of τ std devs

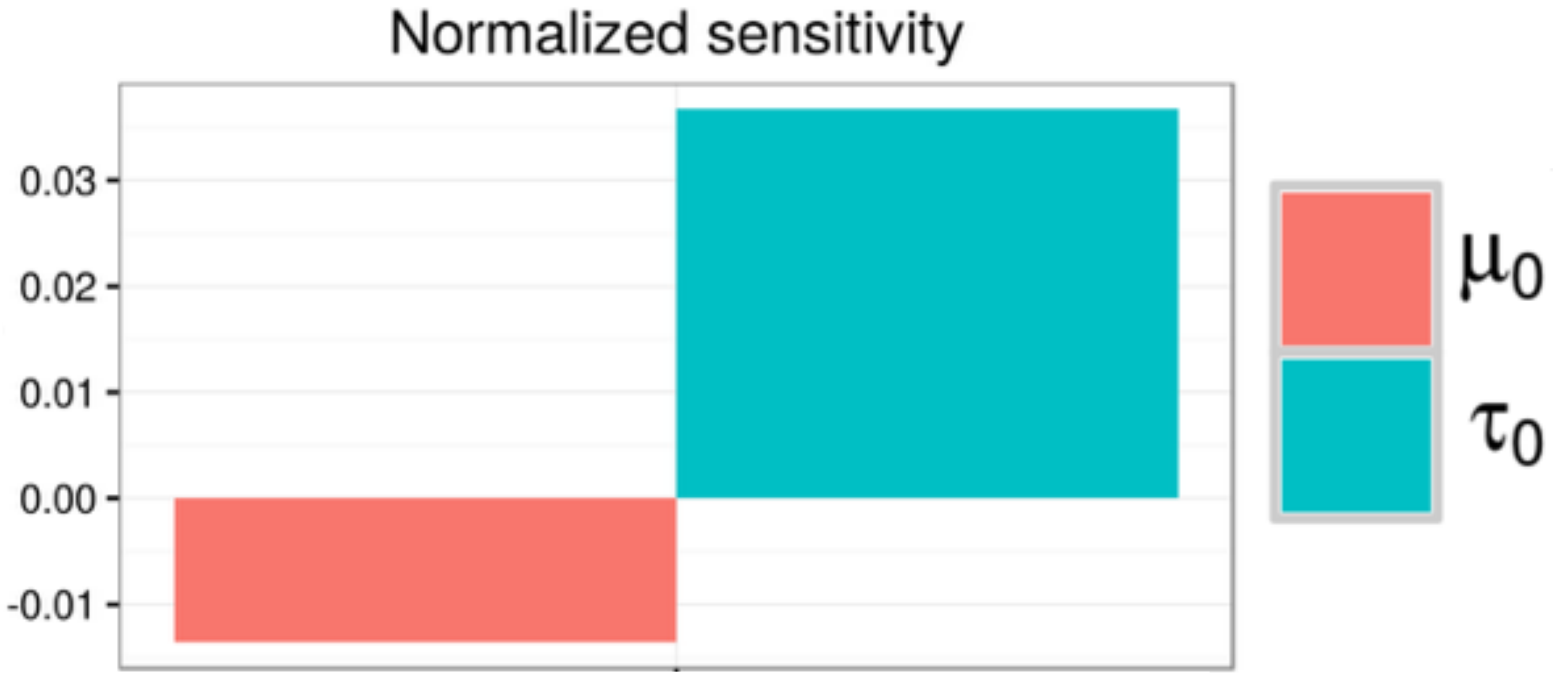




Microcredit Experiment

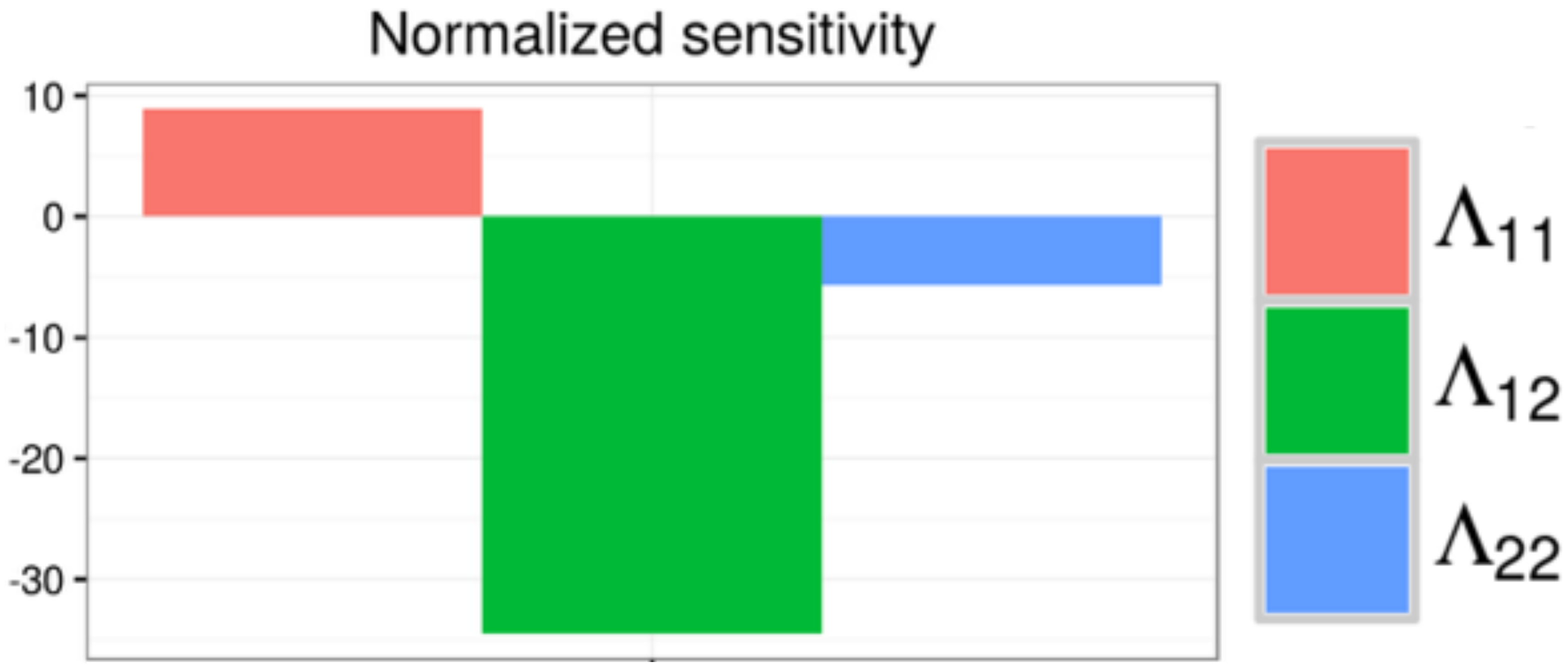
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of τ std devs
- τ mean (MFVB): 3.08 USD PPP
- τ std dev (LRVB): 1.83 USD PPP
- Mean is 1.68 std dev from 0

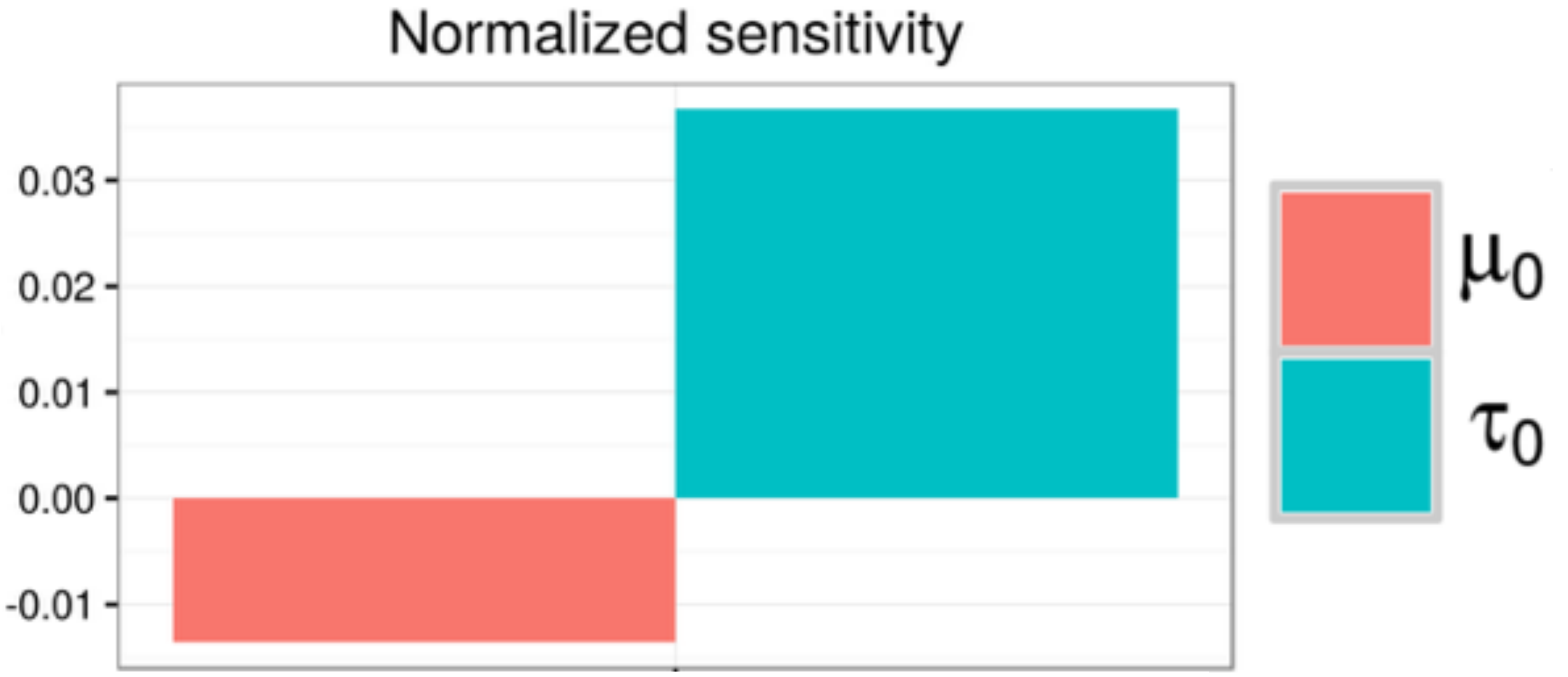




Microcredit Experiment

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

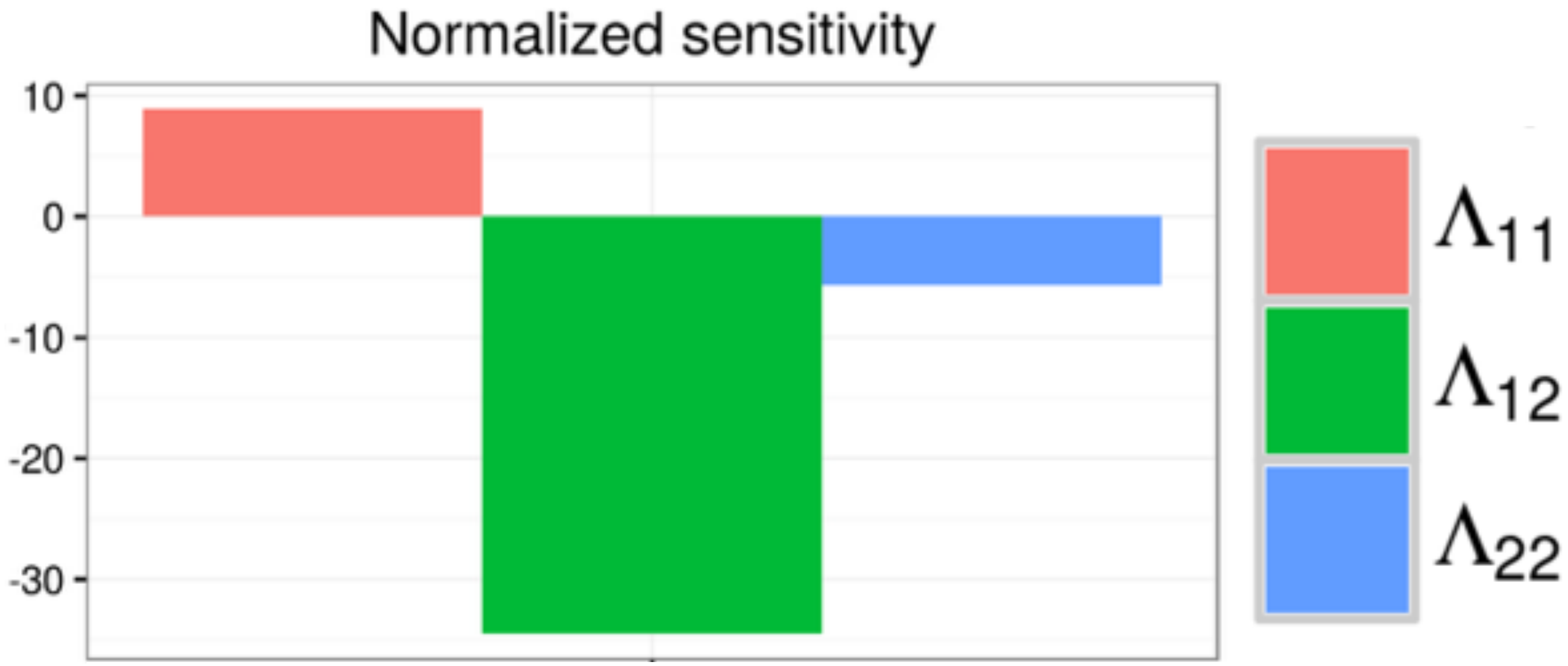


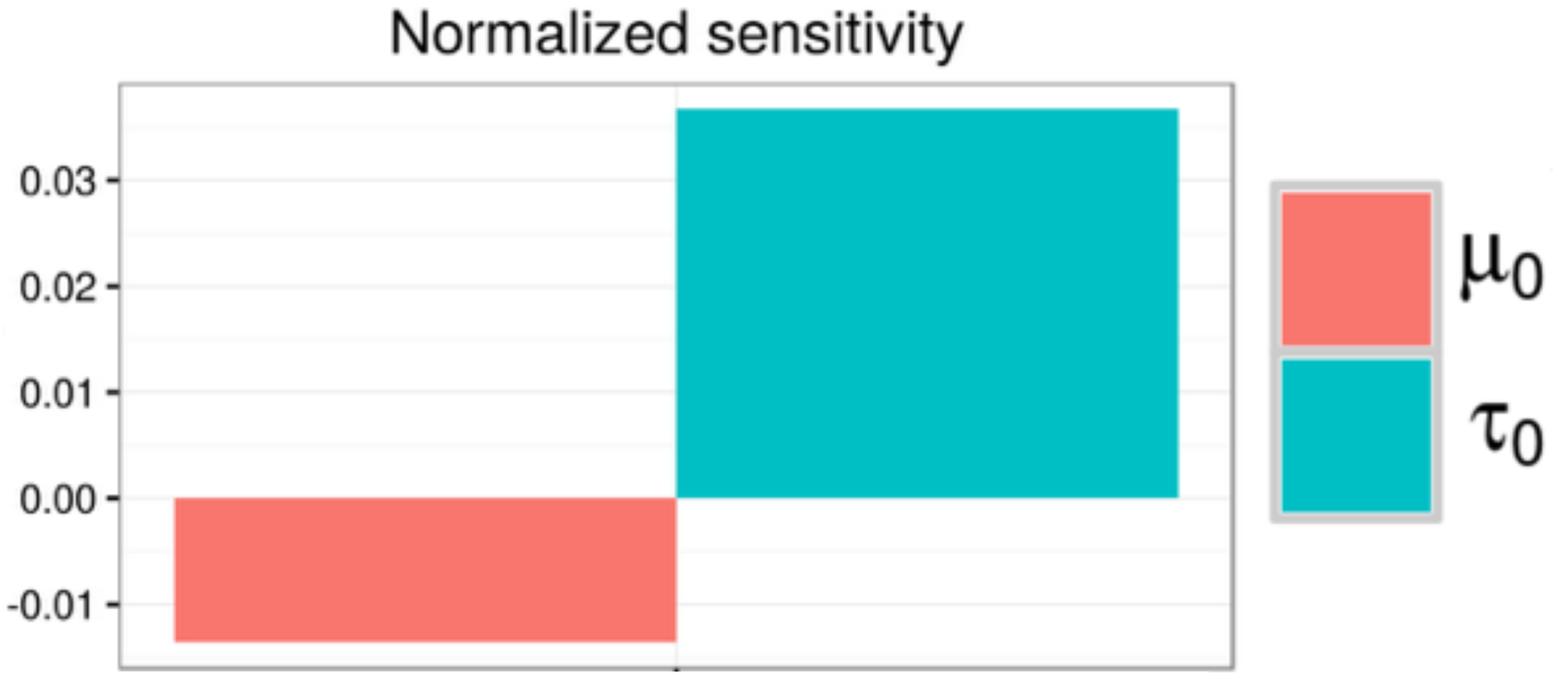


Microcredit Experiment

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

⇒ Mean > 2 std dev





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
- Model:

$$y_{kn} \sim \mathrm{Bernoulli}(p_{kn}) \qquad p_{kn} = \frac{\exp(\rho_{kn})}{1 + \exp(\rho_{kn})}$$

$$\rho_{kn} = x_{kn}^T \beta + u_k$$

  - Priors and hyperpriors:

$$u_k \sim \mathcal{N}(\mu, \sigma^2) \qquad \beta \sim \mathcal{N}(\beta_0, \mathrm{diag}(\gamma))$$

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

$$(\sigma^2)^{-1} \sim \mathrm{Gamma}(a, b)$$

Criteo Online Ads Experiment

Criteo Online Ads Experiment

- VB: 57 sec

Criteo Online Ads Experiment

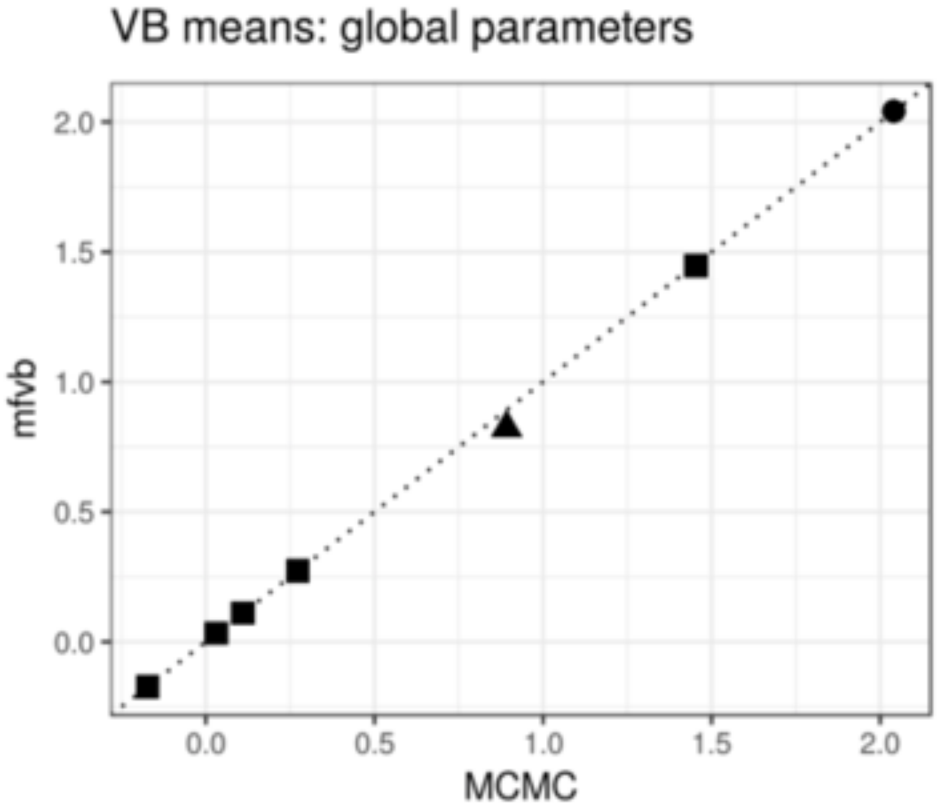
- VB: 57 sec;
  VB + LRVB:
  553 sec
  (**9.2 min**)

Criteo Online Ads Experiment

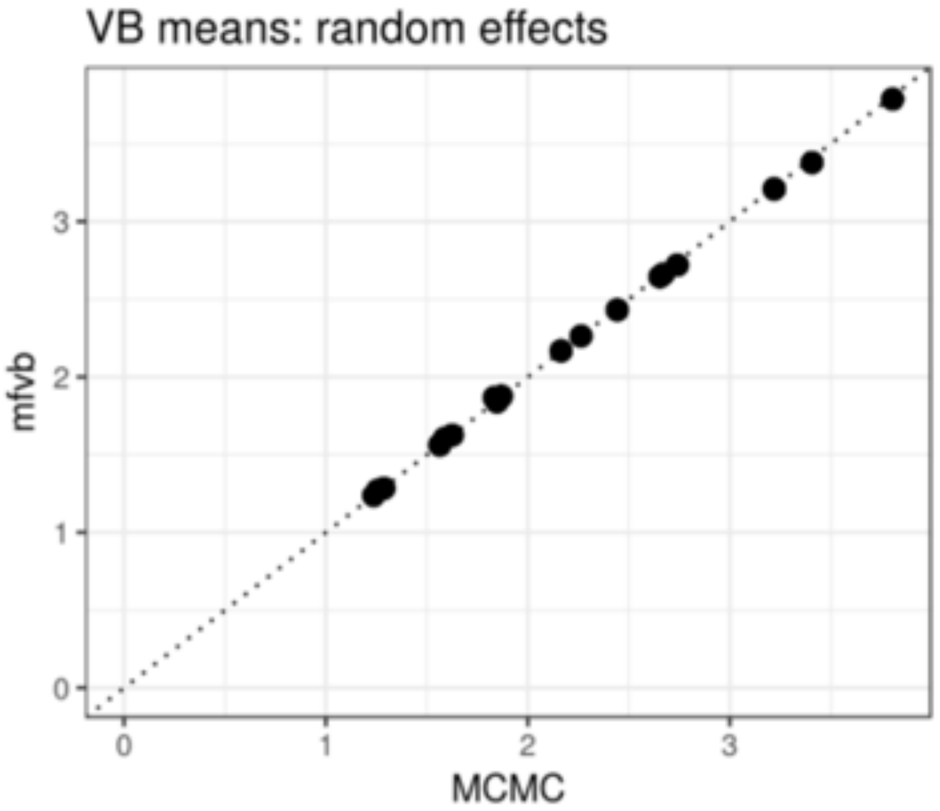
- VB: 57 sec; VB + LRVB: 553 sec (**9.2 min**)
- MCMC (5k samples): 21,066 sec (**5.85 h**)

Criteo Online Ads Experiment

- VB: 57 sec; VB + LRVB: 553 sec (**9.2 min**)
- MCMC (5k samples): 21,066 sec (**5.85 h**)



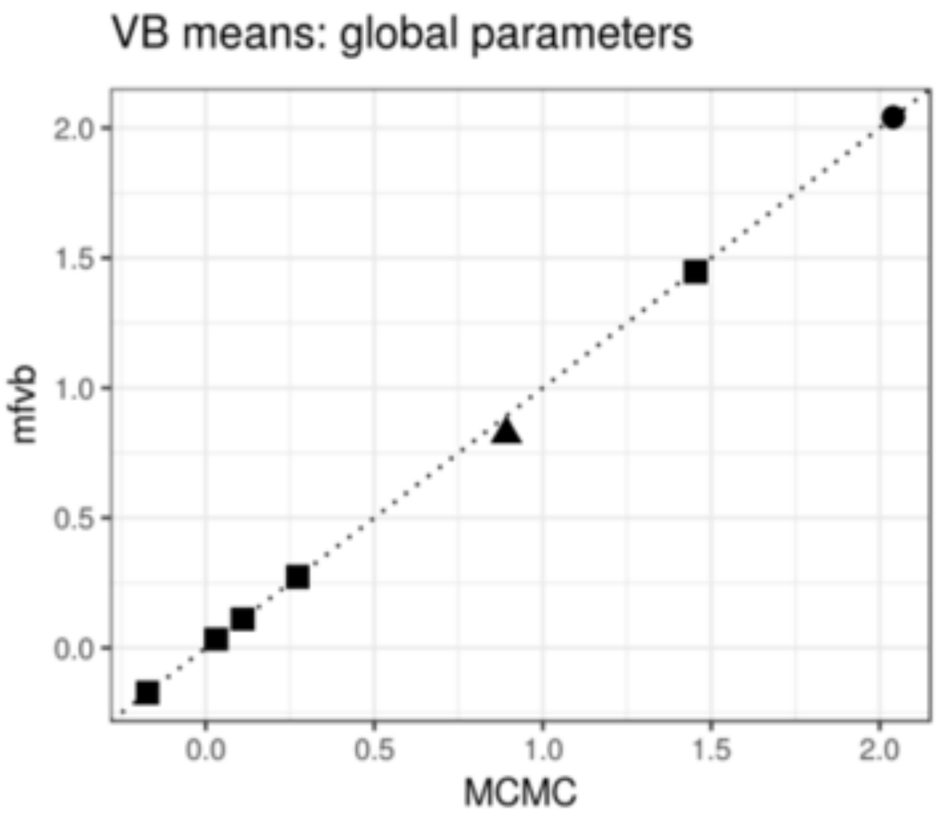
VB means: global parameters

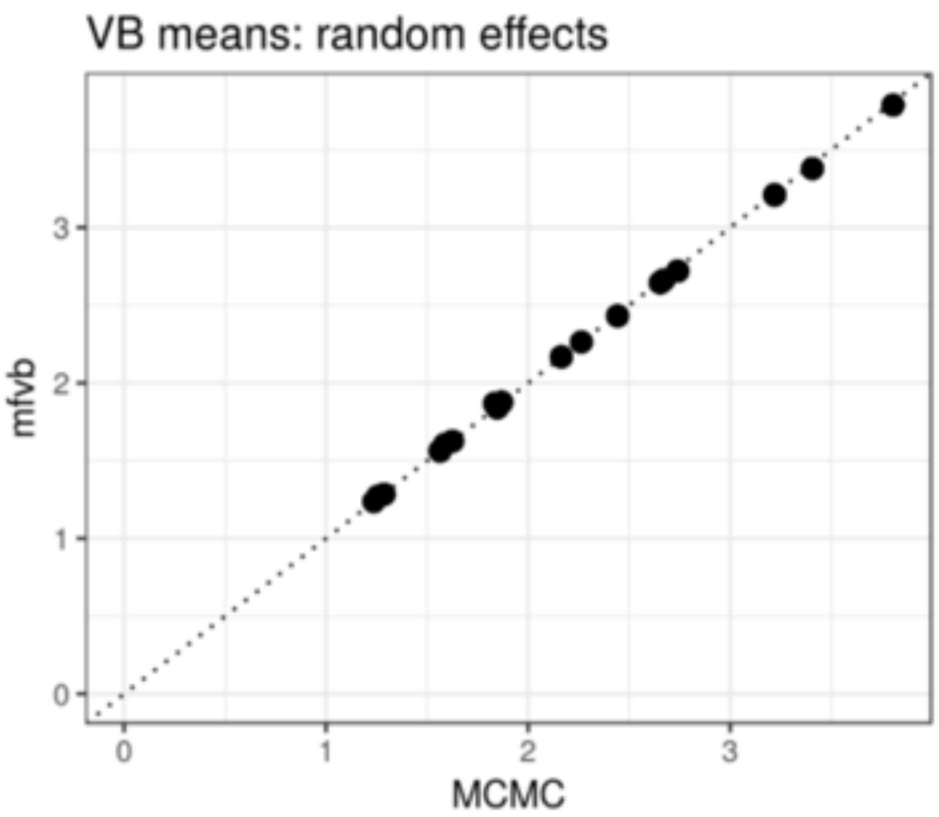


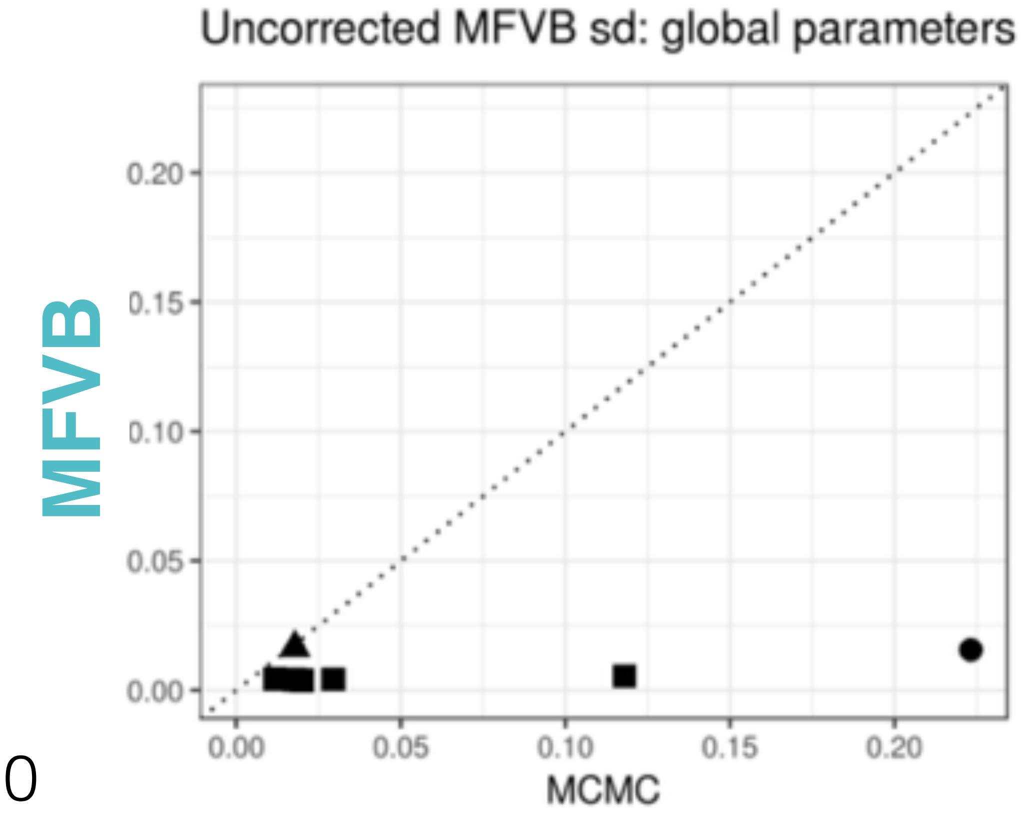
VB means: random effects

Criteo Online Ads Experiment

- VB: 57 sec; VB + LRVB: 553 sec (**9.2 min**)
- MCMC (5k samples): 21,066 sec (**5.85 h**)

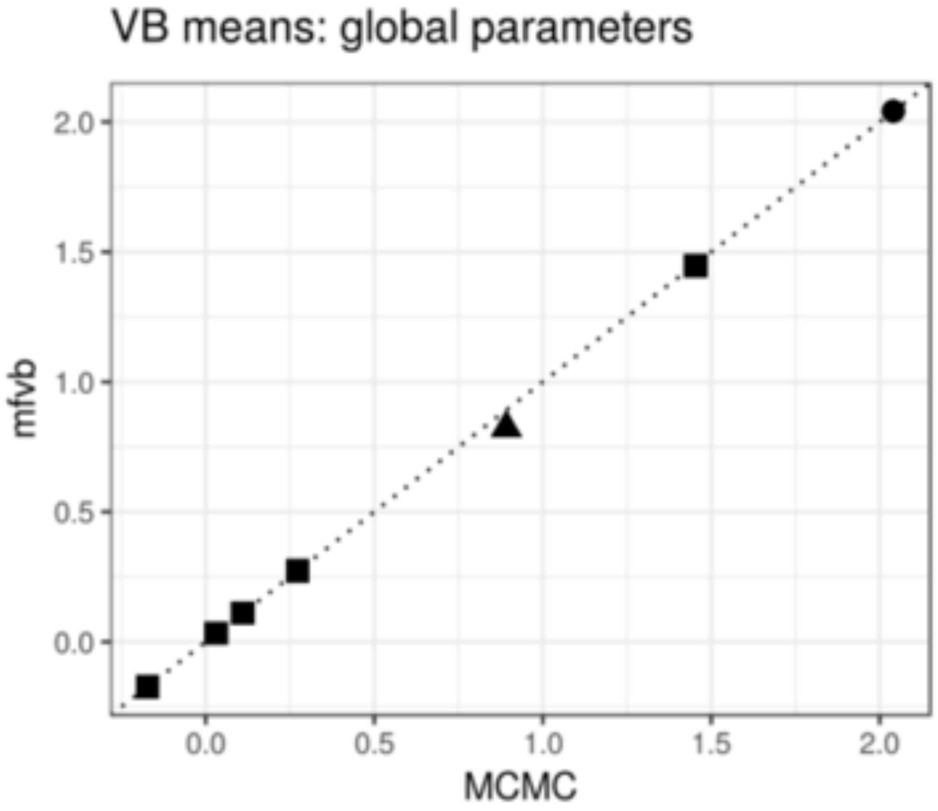


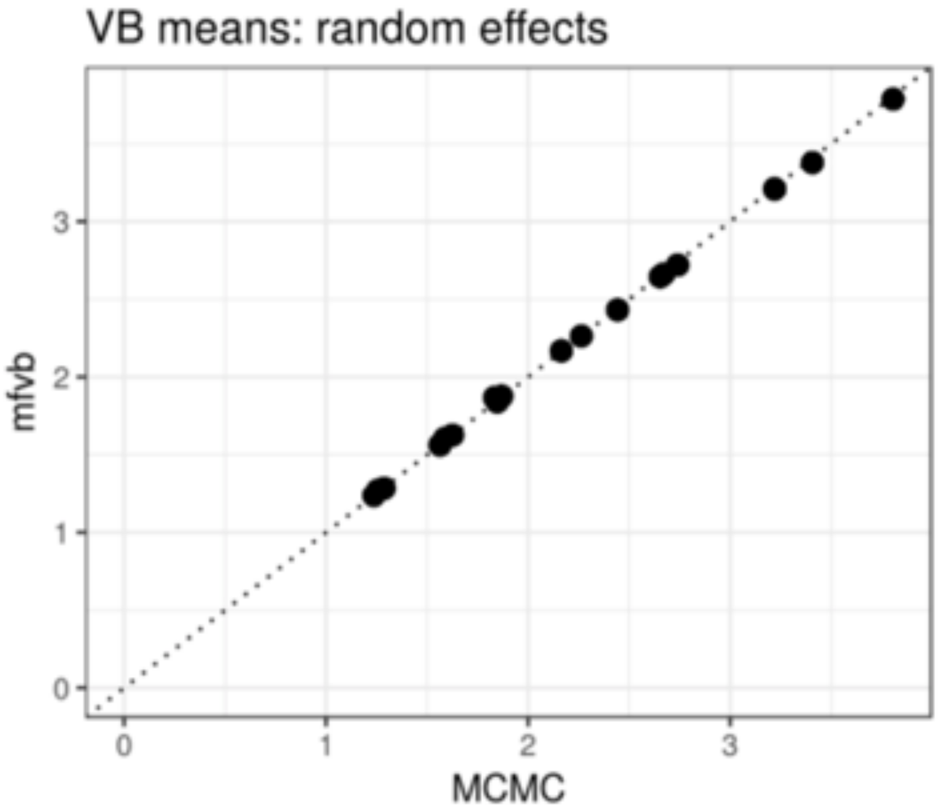


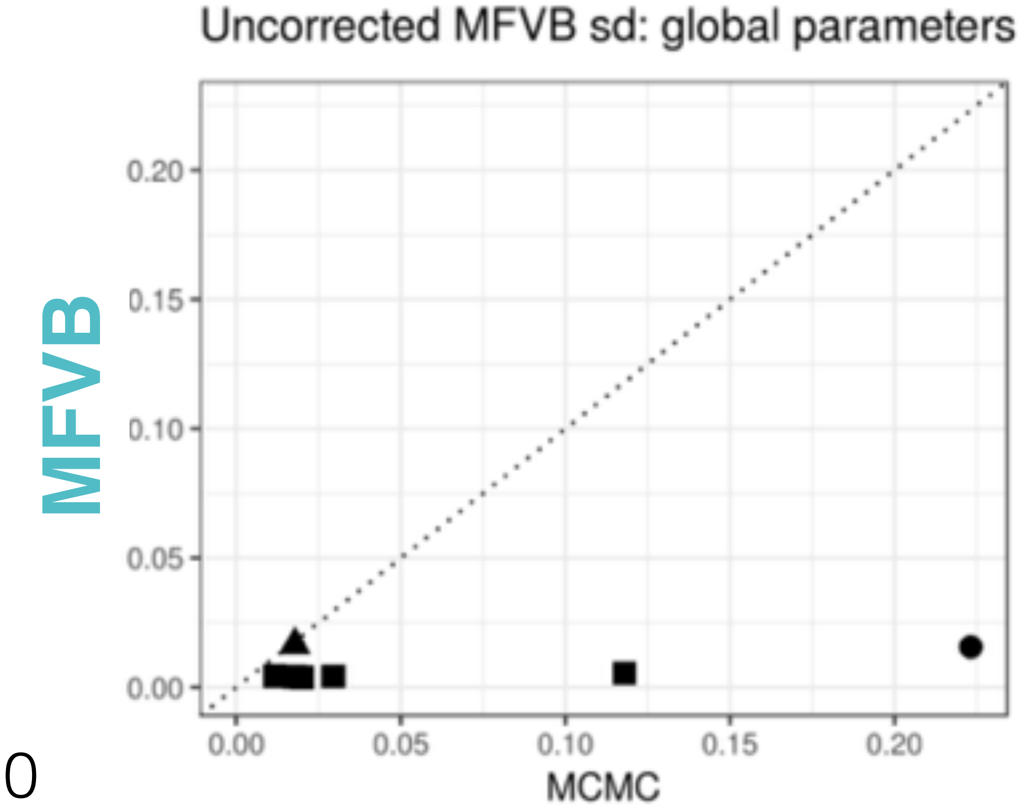


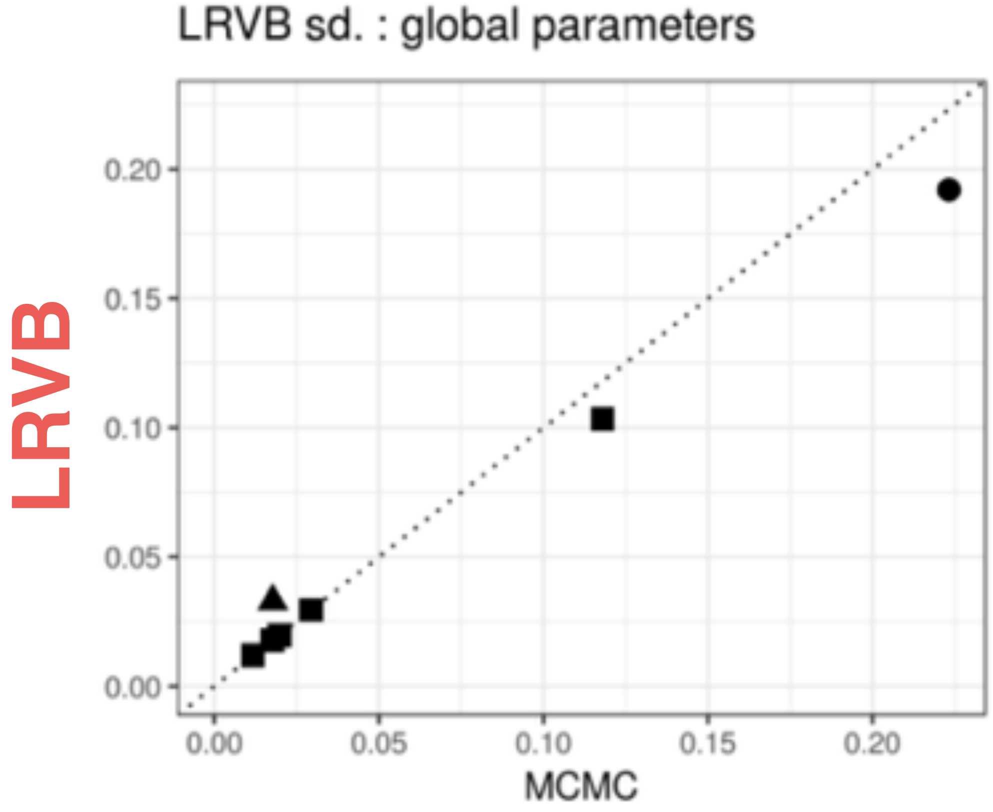
Criteo Online Ads Experiment

- VB: 57 sec; VB + LRVB: 553 sec (**9.2 min**)
- MCMC (5k samples): 21,066 sec (**5.85 h**)


VB means: global parameters

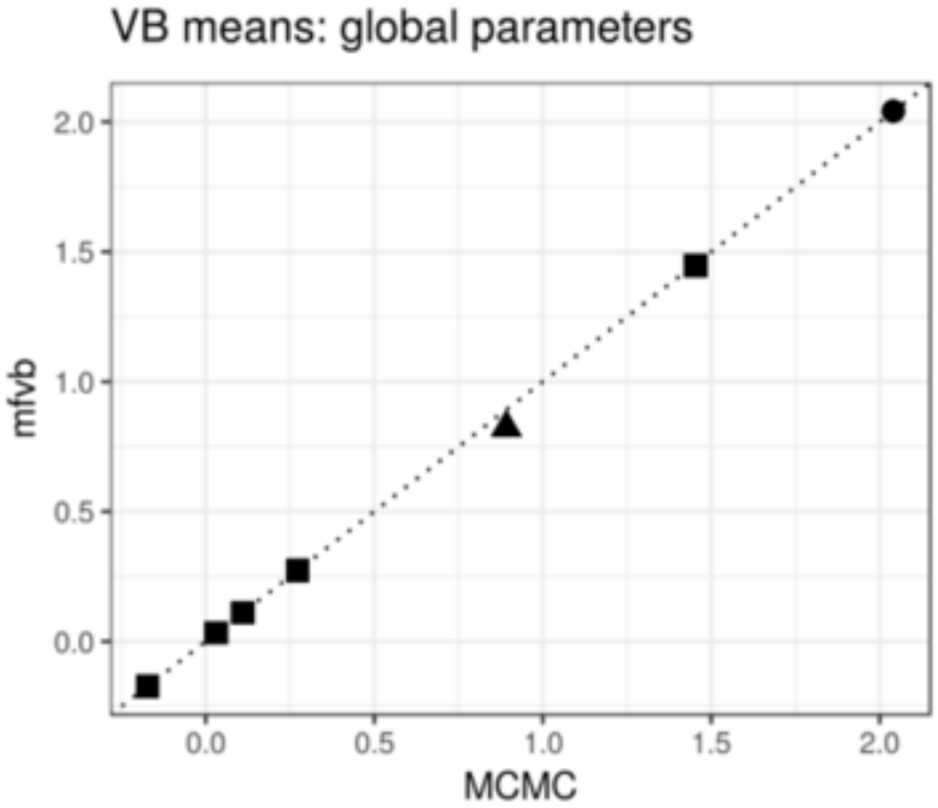

VB means: random effects

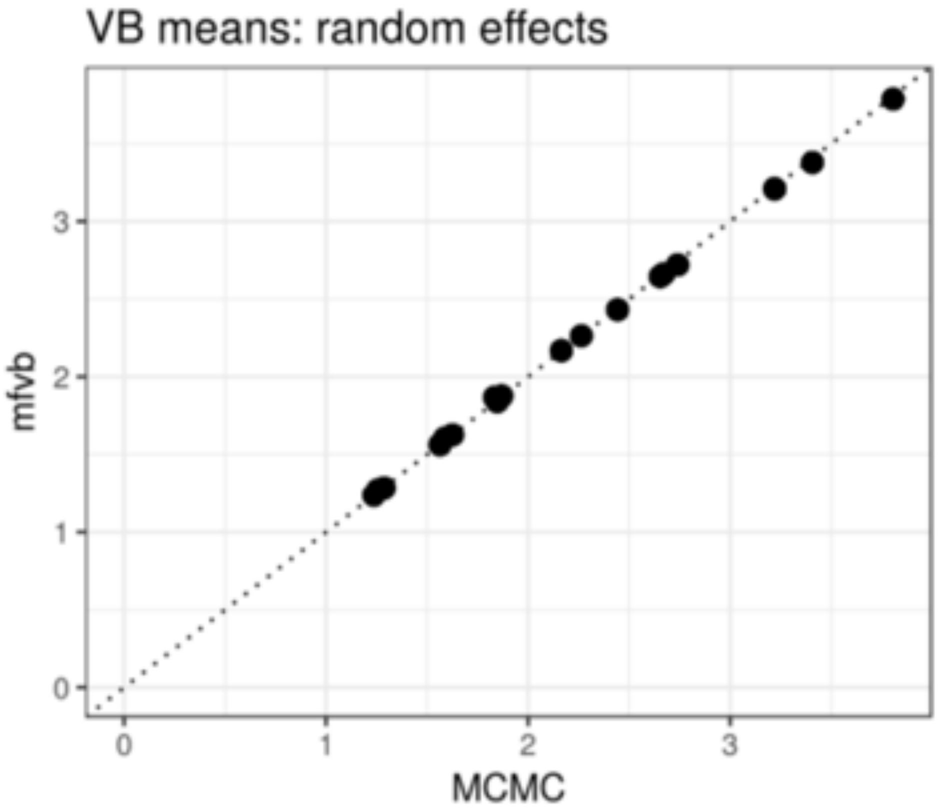

Uncorrected MFVB sd: global parameters

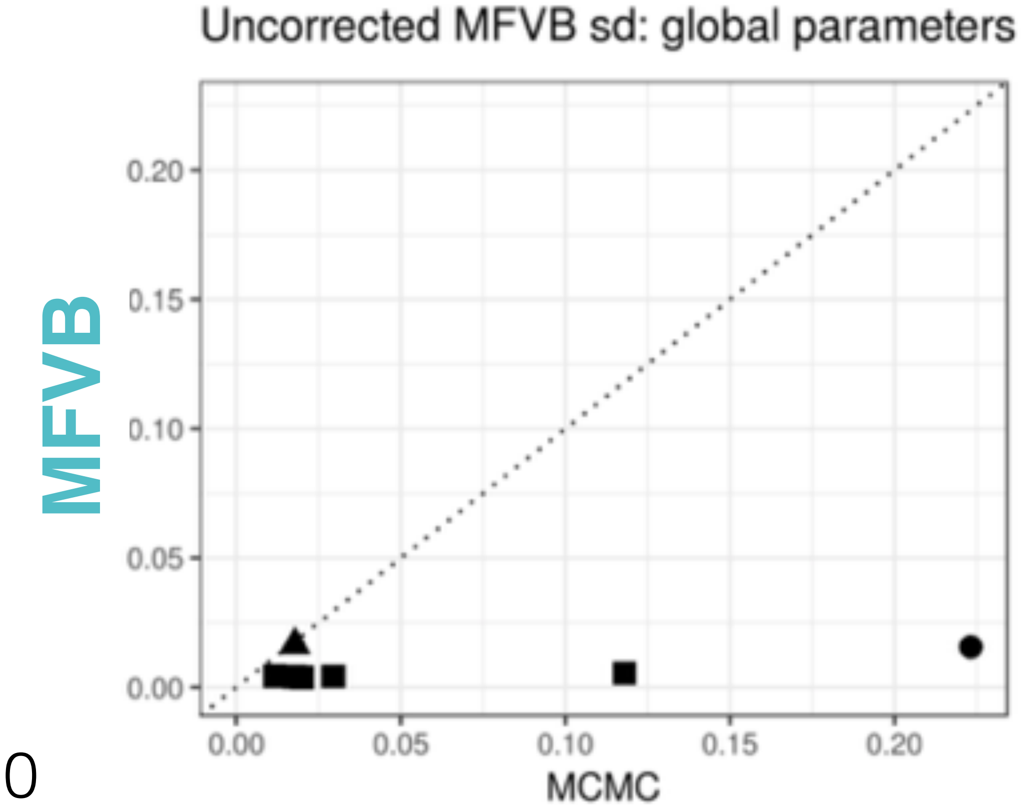

LRVB sd. : global parameters

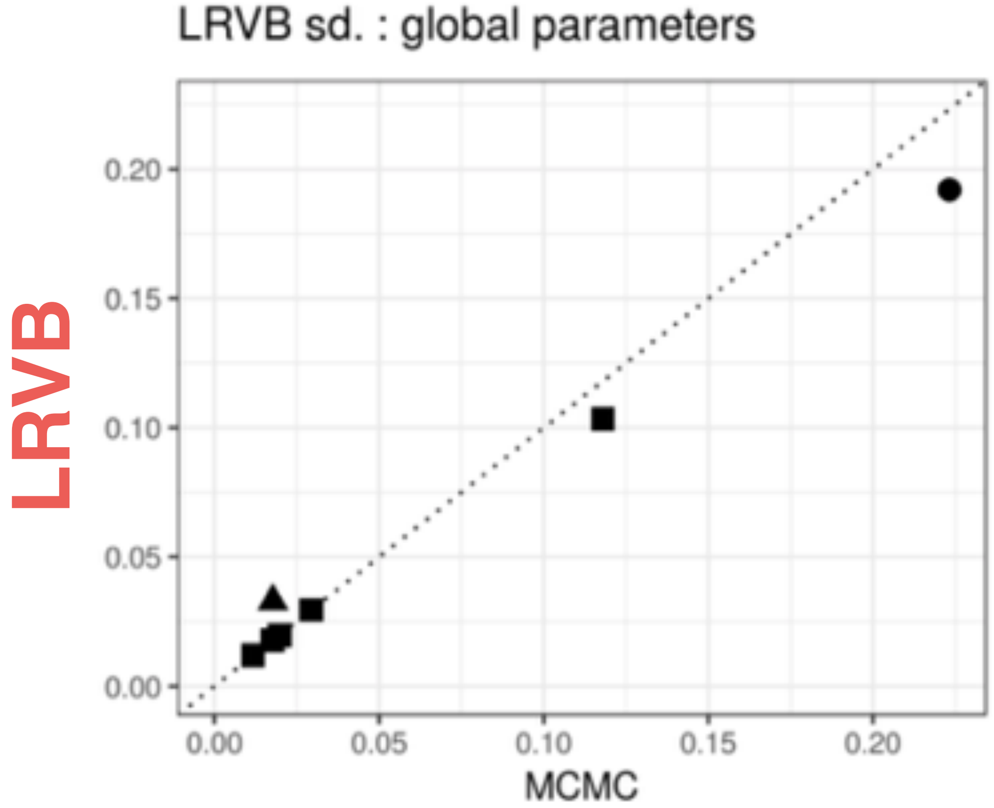
Criteo Online Ads Experiment

- VB: 57 sec; VB + LRVB: 553 sec (**9.2 min**)
- MCMC (5k samples): 21,066 sec (**5.85 h**)


VB means: global parameters


VB means: random effects

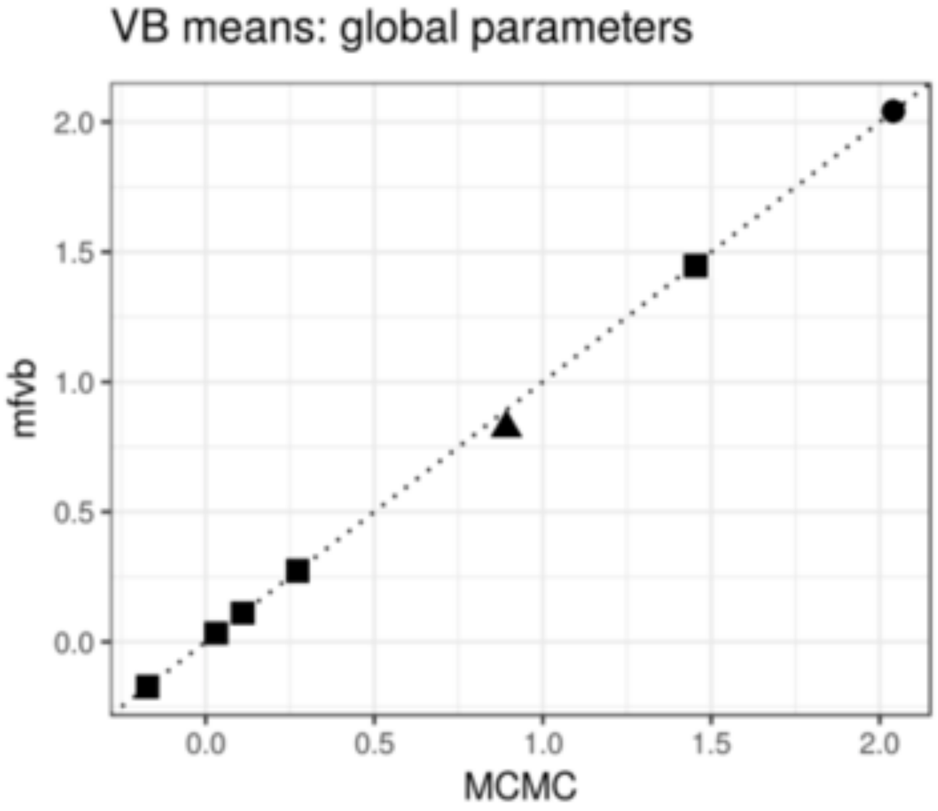

Uncorrected MFVB sd: global parameters

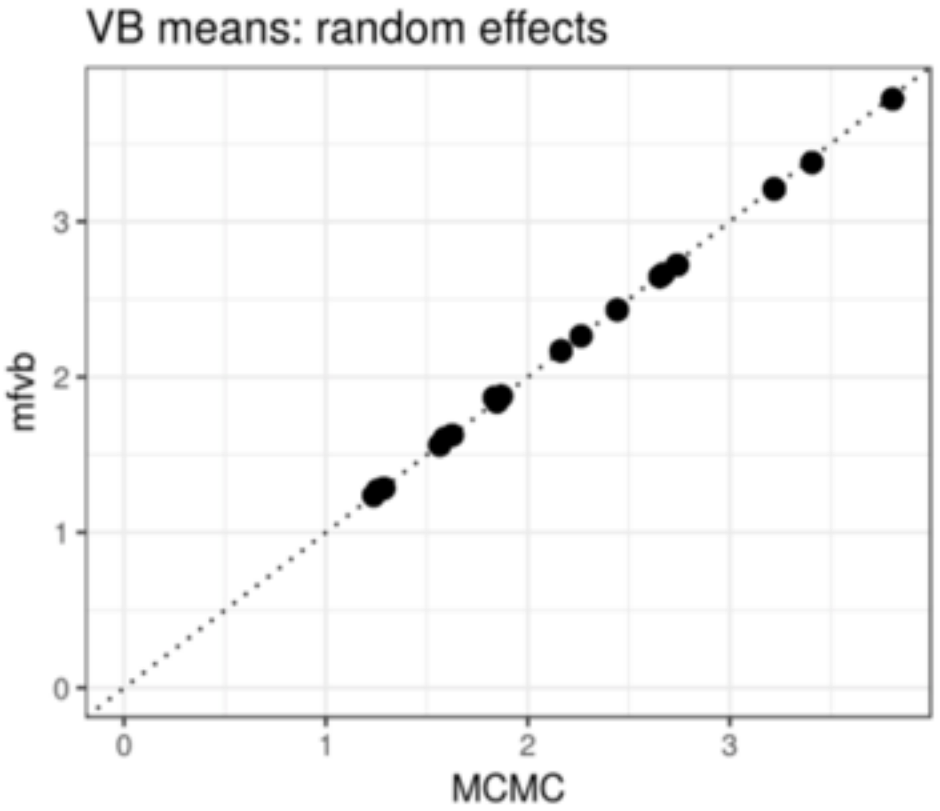

LRVB sd. : global parameters

Also good random effects sd and covariances

Criteo Online Ads Experiment

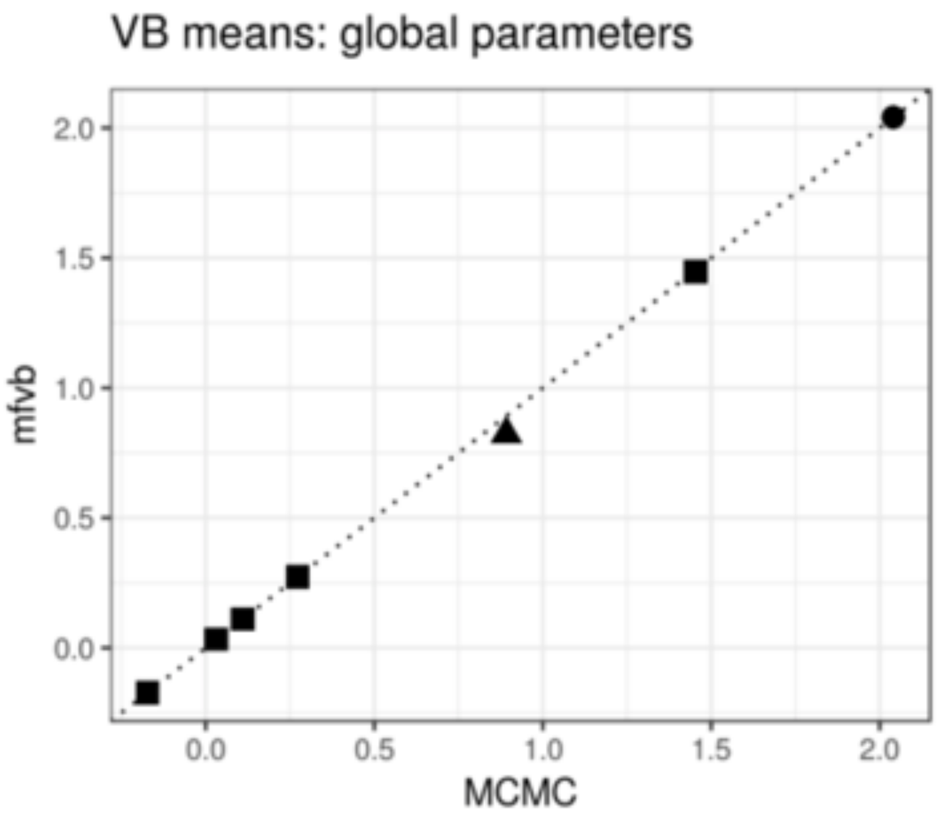
- VB: 57 sec; VB + LRVB: 553 sec (**9.2 min**)
- MCMC (5k samples): 21,066 sec (**5.85 h**)

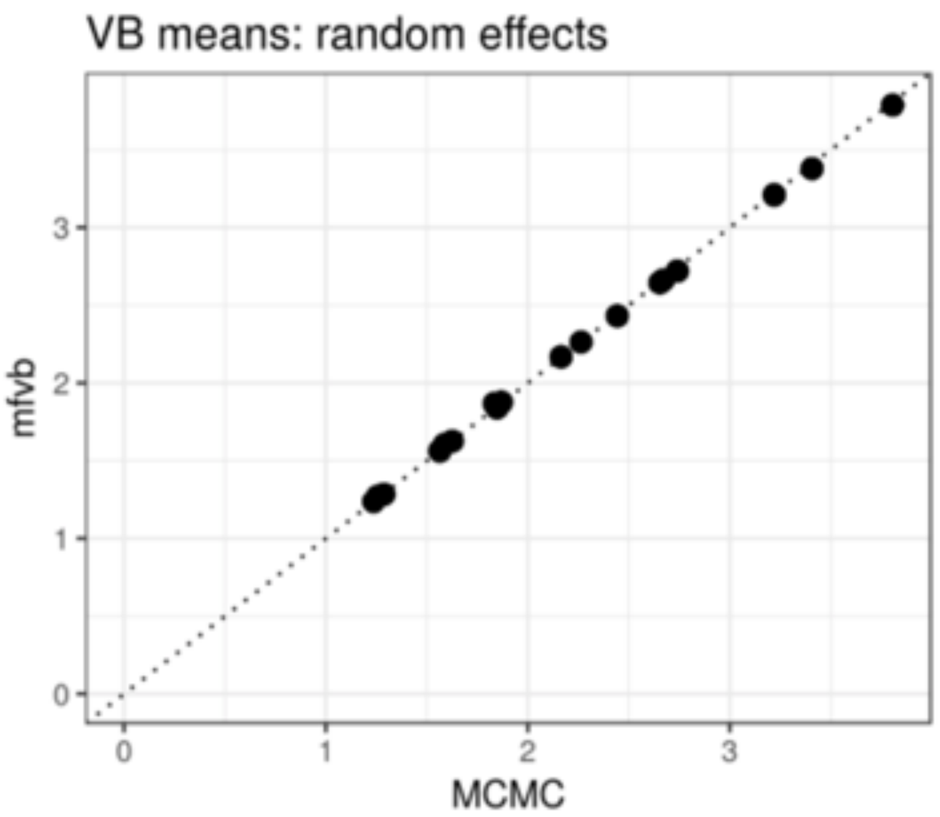

VB means: global parameters

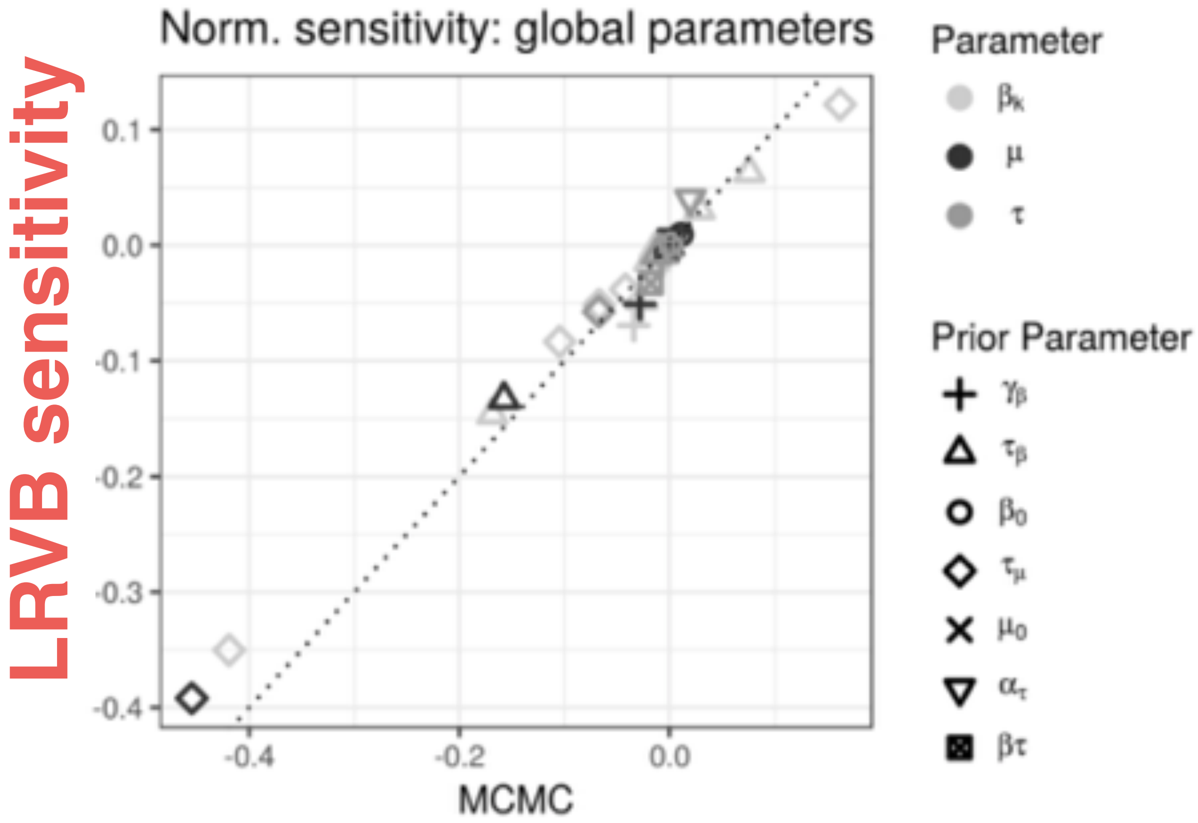

VB means: random effects

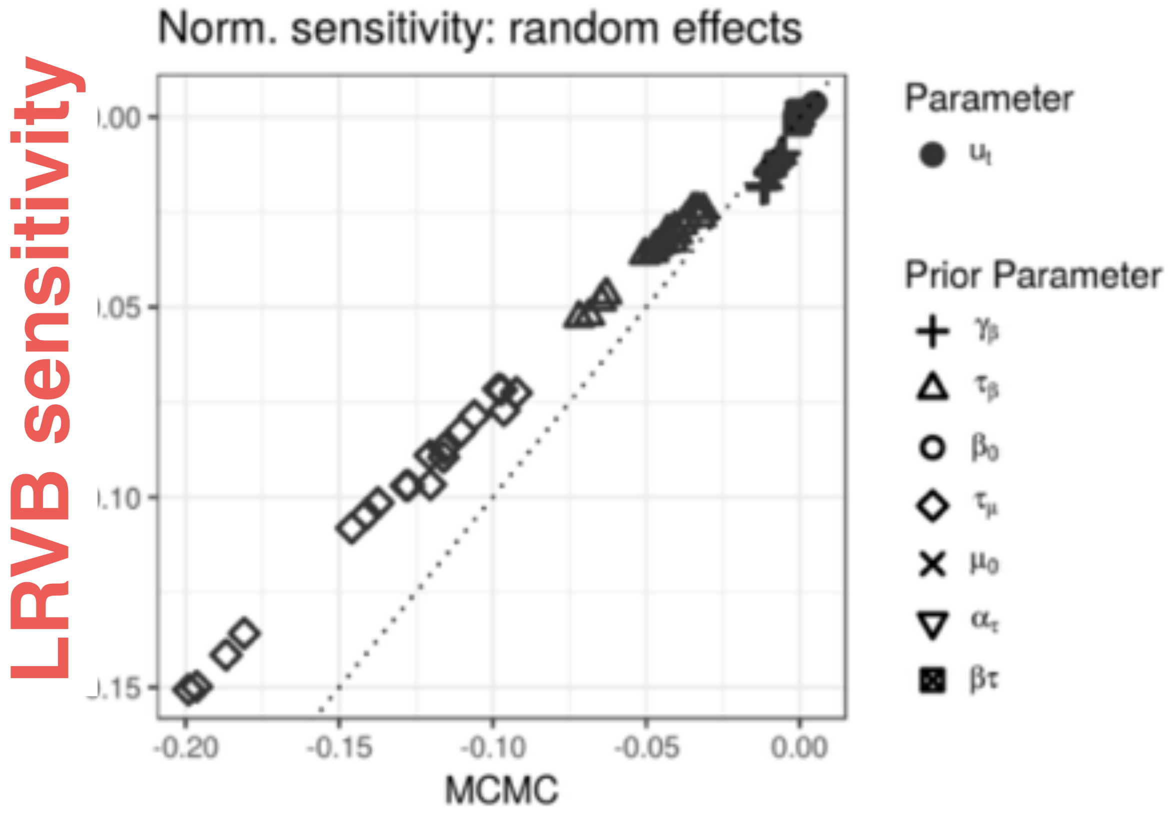
Criteo Online Ads Experiment

- VB: 57 sec; VB + LRVB: 553 sec (**9.2 min**)
- MCMC (5k samples): 21,066 sec (**5.85 h**)





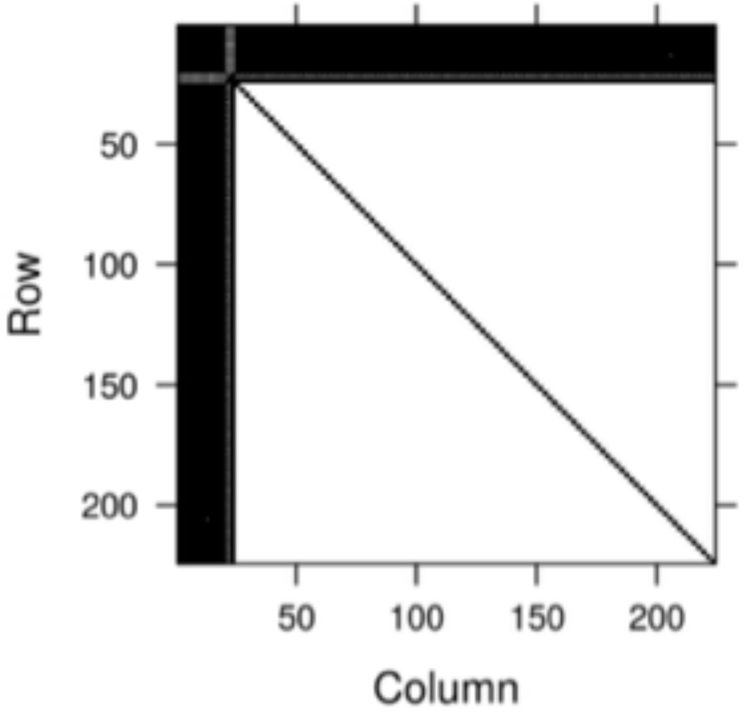




Computational complexity

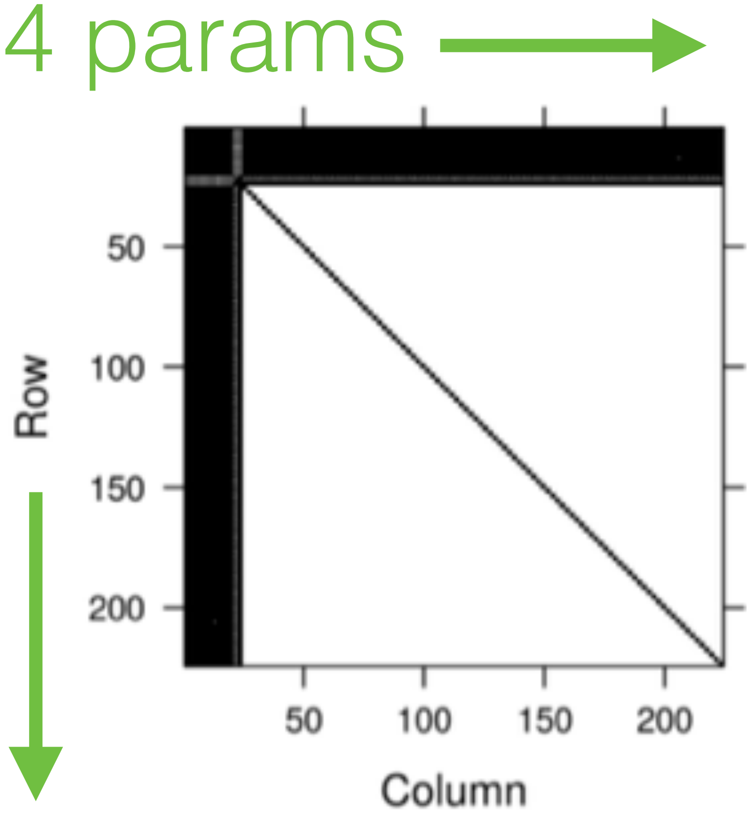
Computational complexity

- Top left submatrix for Criteo analysis



Computational complexity

- Top left submatrix for Criteo analysis



10,014 params

Posterior means: revisited

- Want to predict college GPA y_n

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

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

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

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

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

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

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

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

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

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

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

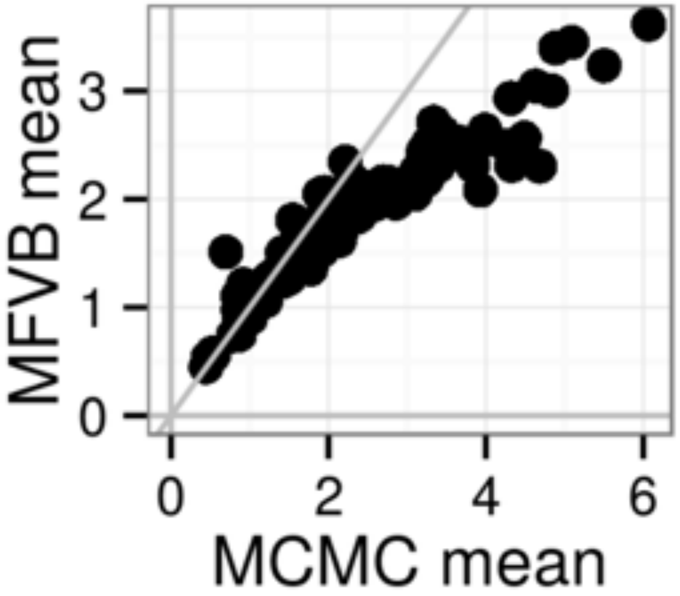
- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model: $y_n|\beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$

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

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

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

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



[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

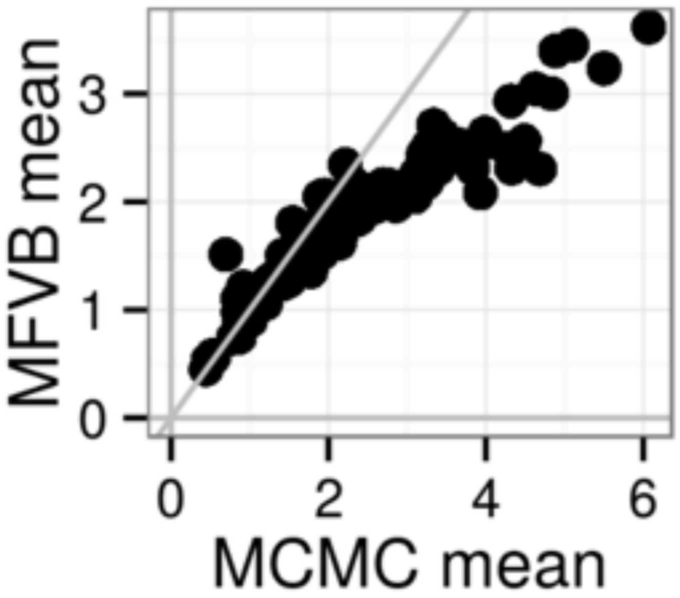
- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model: $y_n|\beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$

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

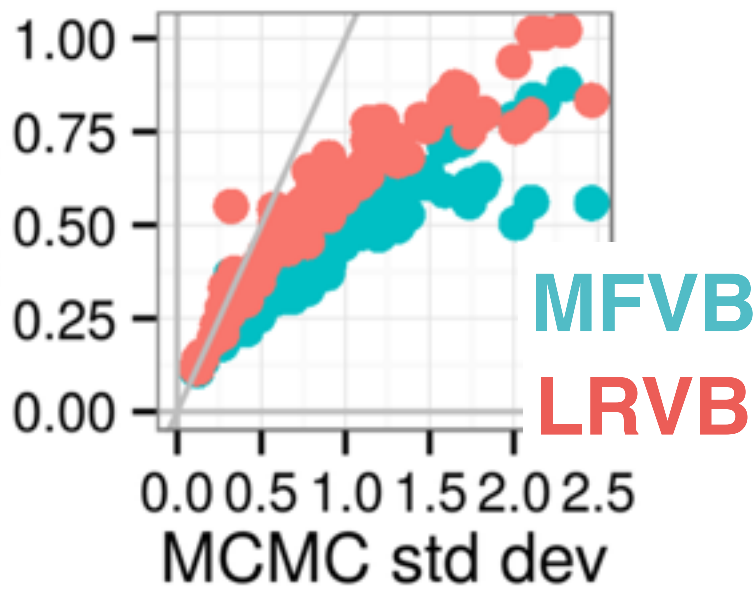
$$\beta \sim \mathcal{N}(0, \Sigma) \qquad (\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$$

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

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



$(\rho^2)^{-1}$ std dev



[Giordano, Broderick, Jordan 2015]

Conclusions

- *Linear response variational Bayes (LRVB)*: supplements VB for:
  - Fast **covariance** estimates
  - Fast **robustness** quantification
    - Priors, likelihood, data

[Giordano, Broderick, Jordan 2015, 2017]

Conclusions

- *Linear response variational Bayes (LRVB)*: supplements VB for:
  - Fast **covariance** estimates
  - Fast **robustness** quantification
    - Priors, likelihood, data
- When can we trust LRVB?

[Giordano, Broderick, Jordan 2015, 2017]

Conclusions

- *Linear response variational Bayes (LRVB)*: supplements VB for:
    - Fast **covariance** estimates
    - Fast **robustness** quantification
        - Priors, likelihood, data
- When can we trust LRVB?
- Data summarization for scalability (Next part)

[Giordano, Broderick, Jordan 2015, 2017]

References (1/2)

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

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

R Giordano, T Broderick, and MI Jordan. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, to appear. ArXiv:1709.02536.

References (1/2)

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

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

R Giordano, T Broderick, and MI Jordan. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, to appear. ArXiv:1709.02536.

R Giordano, W Stephenson, R Liu, MI Jordan, and T Broderick. Return of the infinitesimal jackknife. Under review. ArXiv:1806.00550.

References (2/2)

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

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

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

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

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

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

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

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

M Opper and O Winther. Variational linear response. *NIPS* 2003.

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

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