Fast Robustness Quantification with Variational Bayes

Tamara Broderick
ITT Career Development
Assistant Professor,
MIT

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

- Bayesian inference

- Bayesian inference
    - Complex, modular models

- Bayesian inference
  - Complex, modular models; posterior distribution

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

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

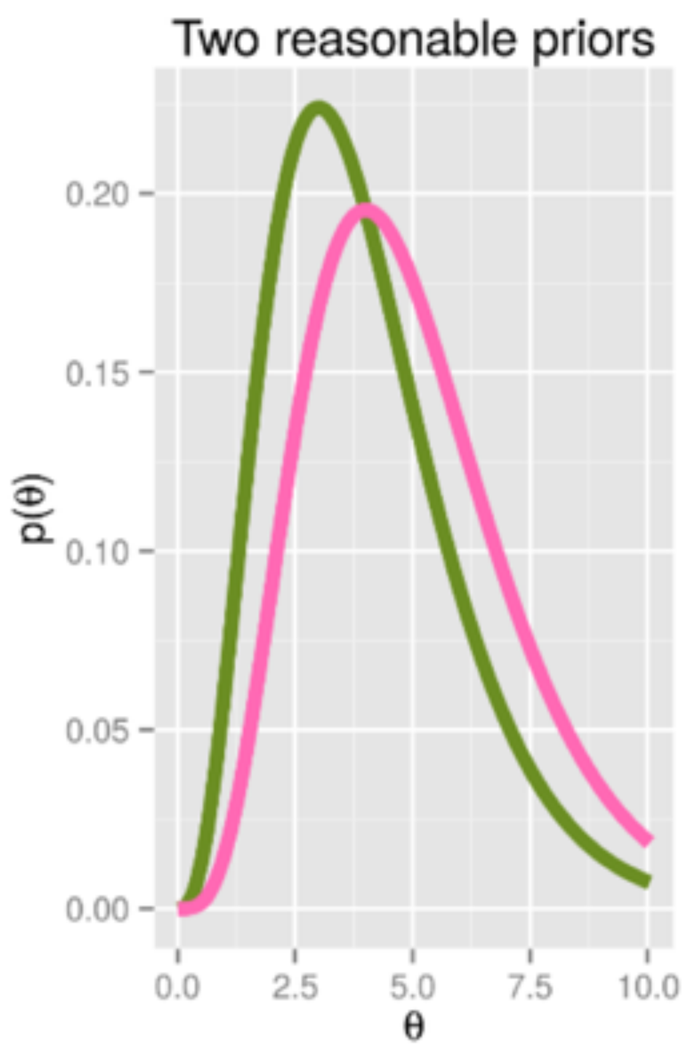
- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
    - Complex, modular models; posterior distribution

- Bayesian inference $p(\theta|x) \propto_{\theta} p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges

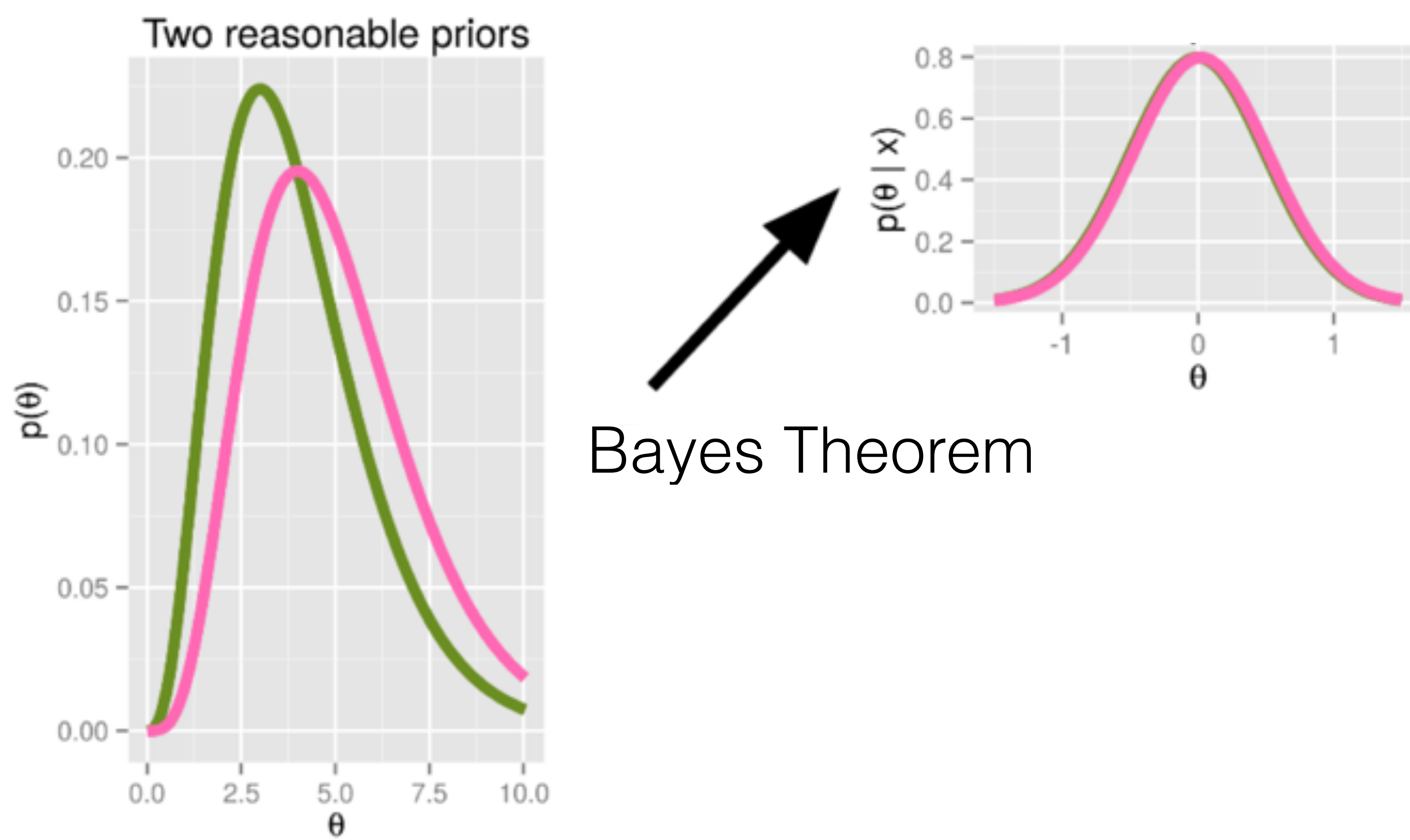
- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming; subjective

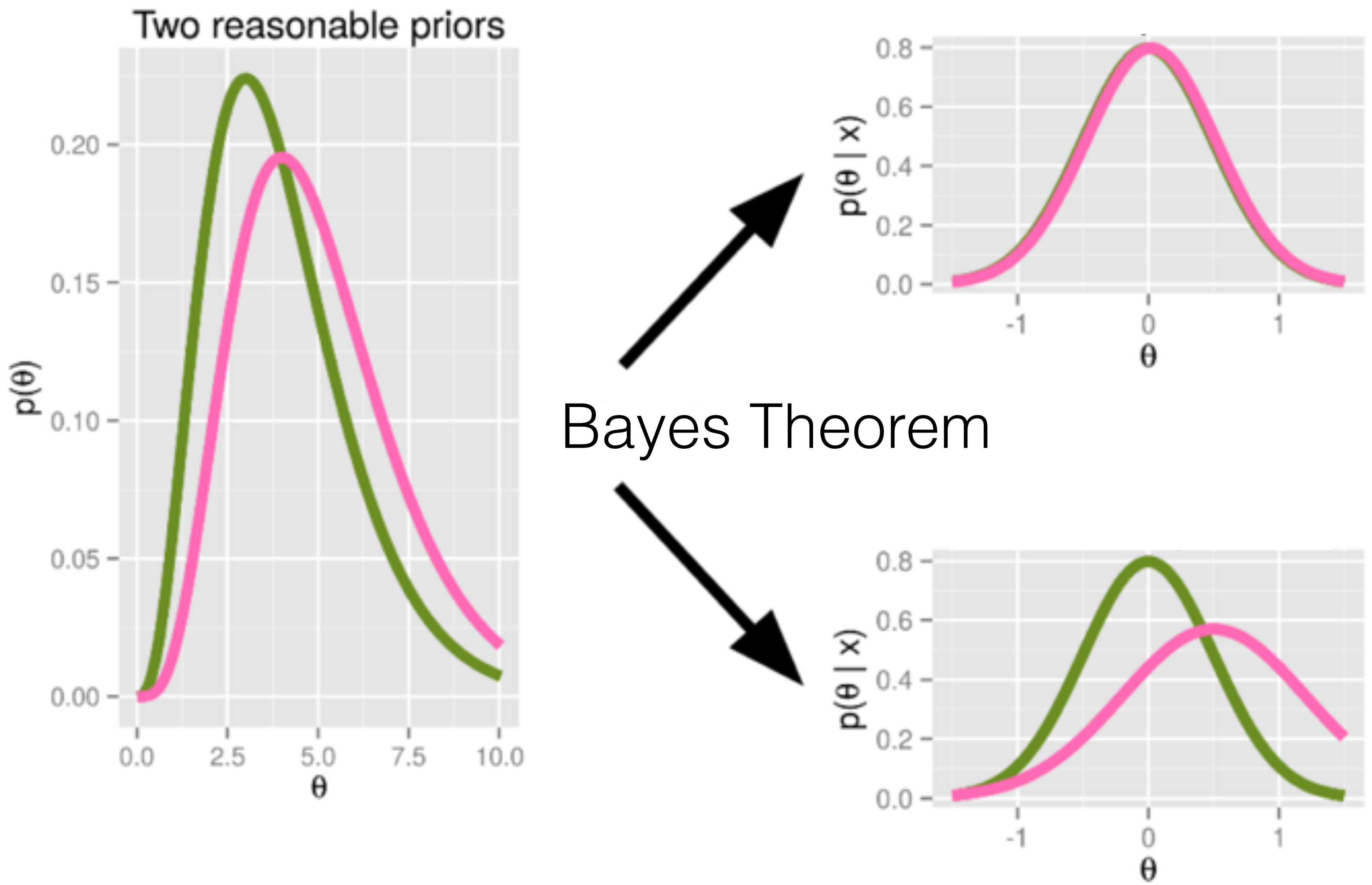
- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming; subjective



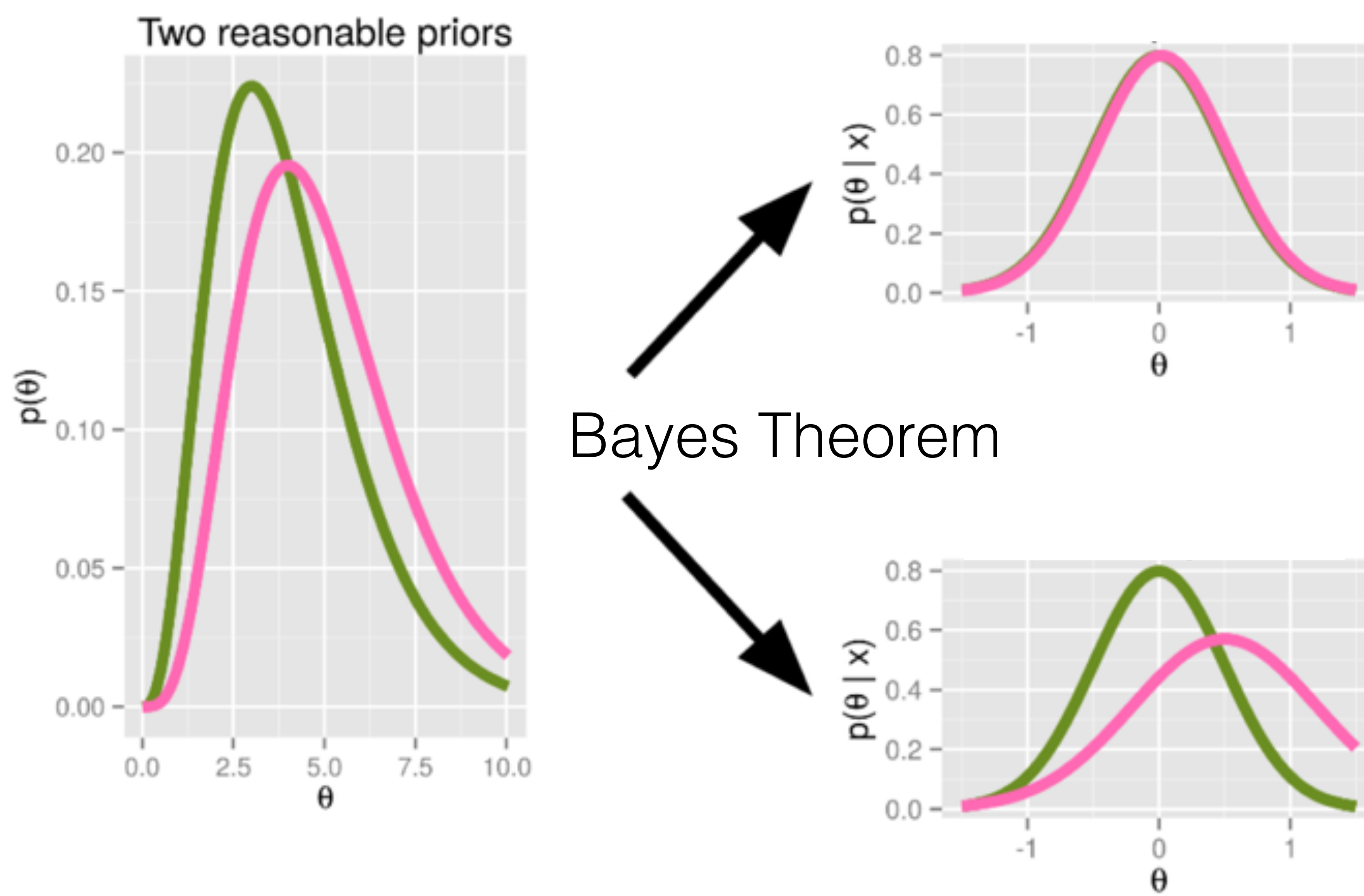
- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming; subjective



- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming; subjective

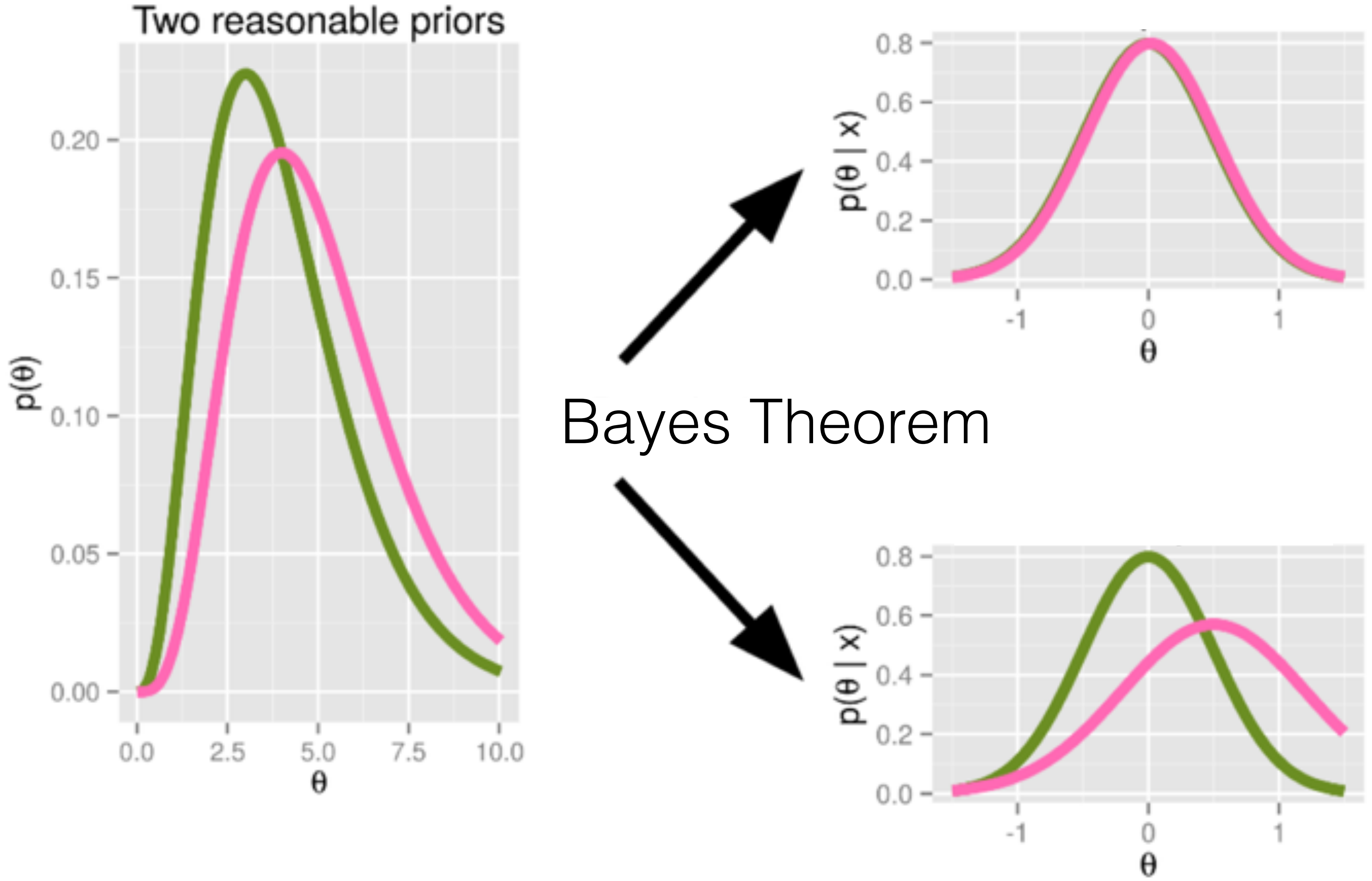


- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming; subjective; complex models



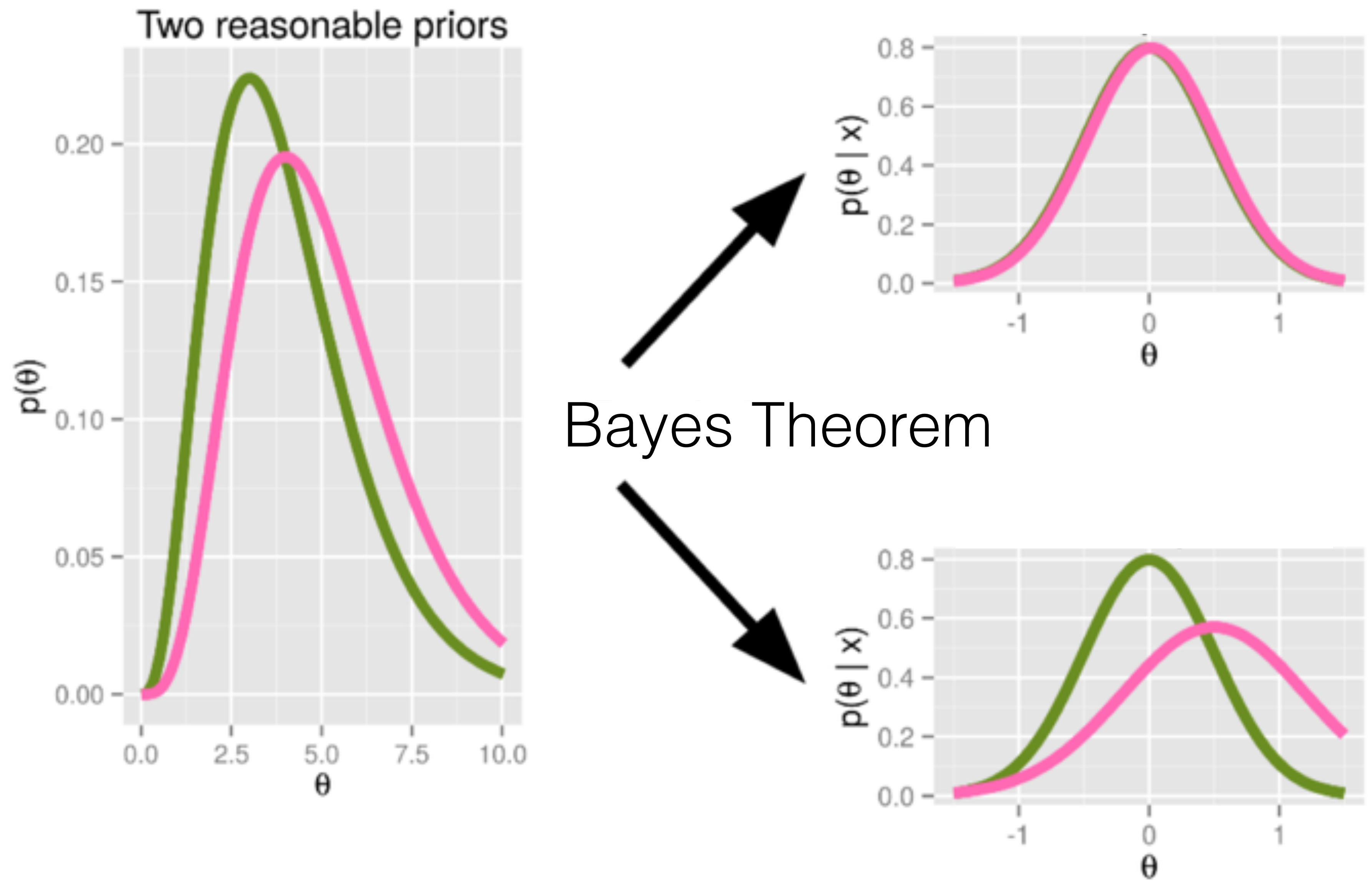
Robustness quantification

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming; subjective; complex models



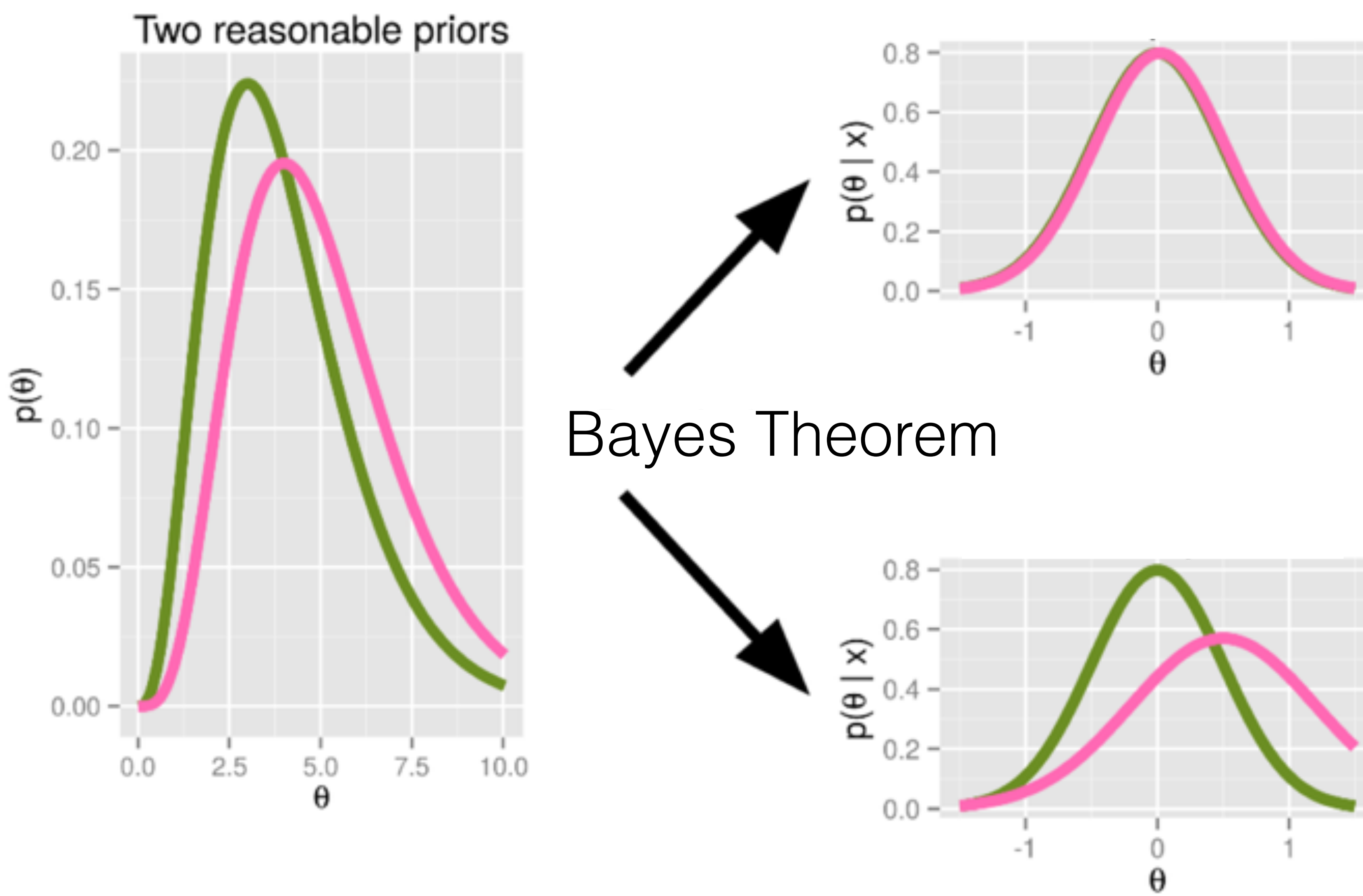
Robustness quantification

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming; subjective; complex models
- Robustness



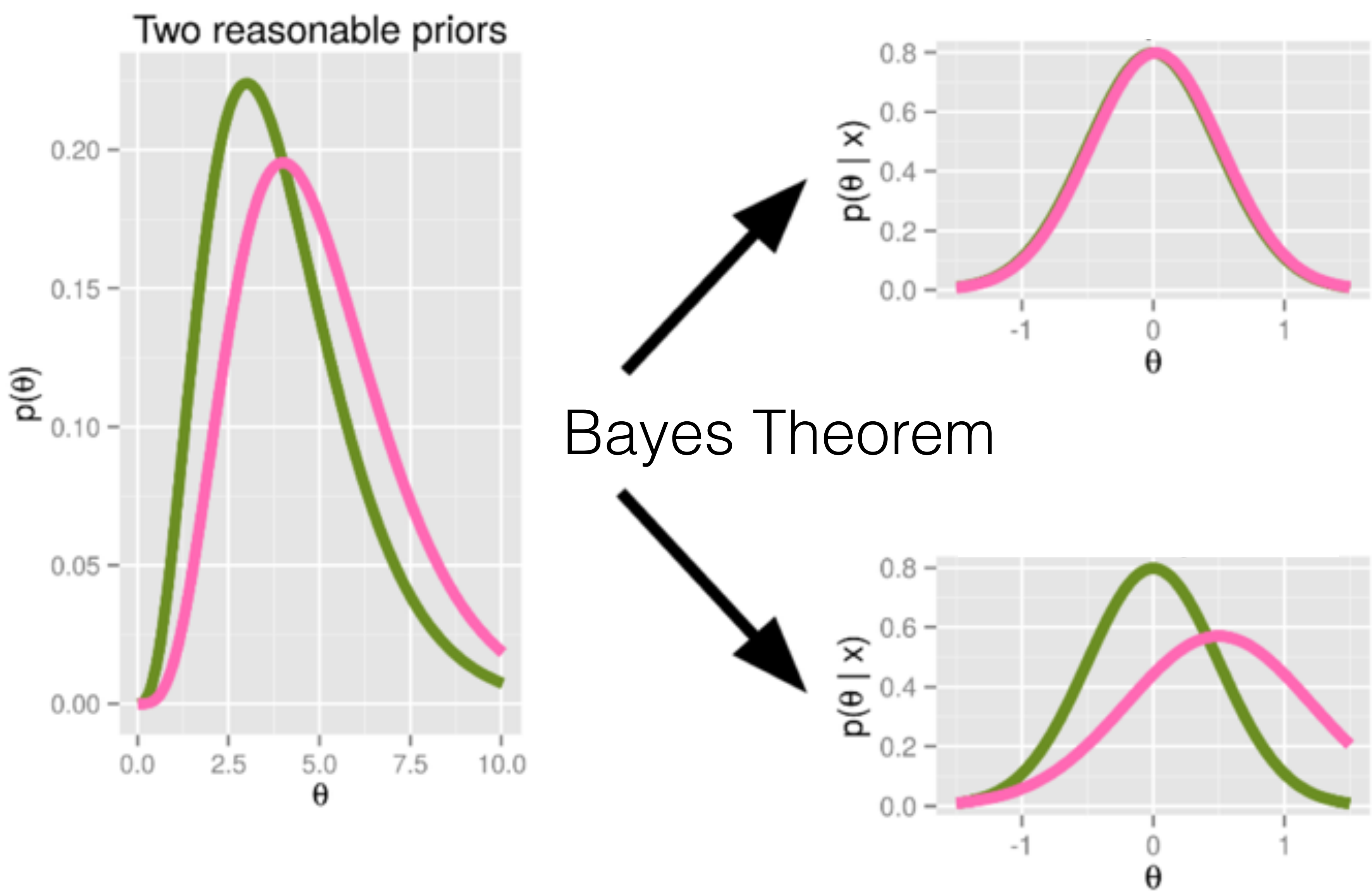
Robustness quantification

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming; subjective; complex models
- Robustness
  - Global & local



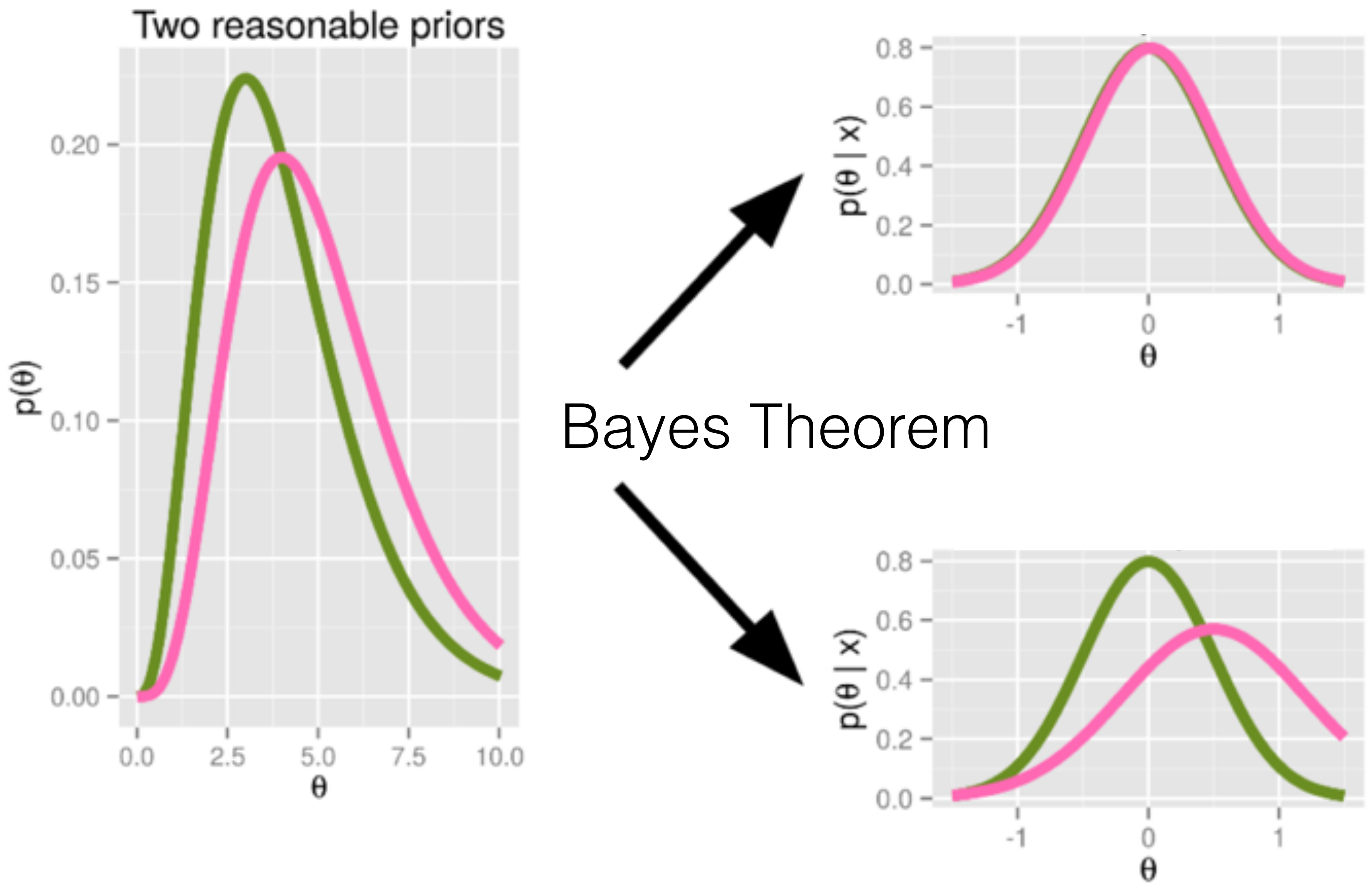
Robustness quantification

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming; subjective; complex models
- Robustness
  - Global & local
  - Rarely used



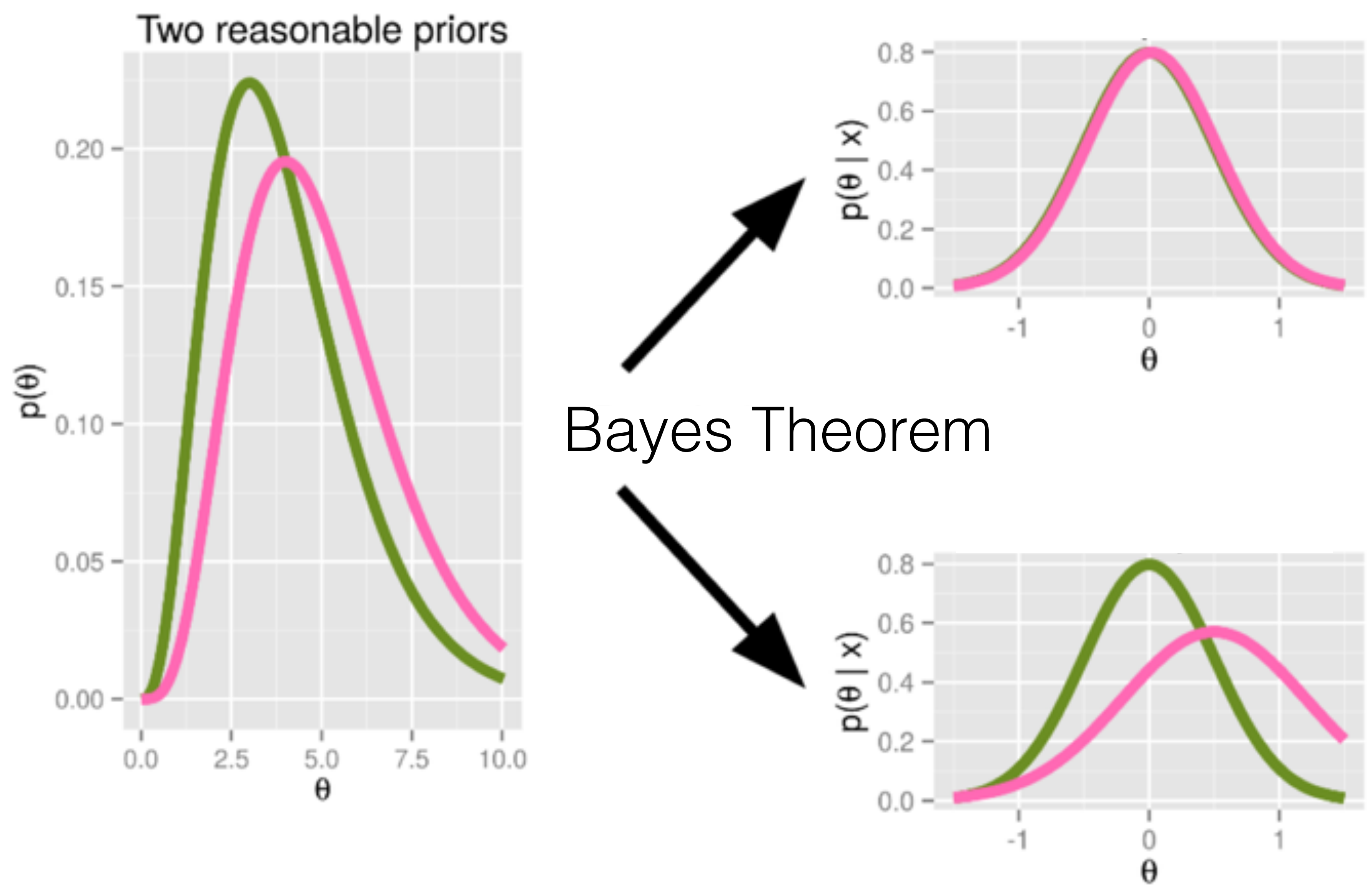
Robustness quantification

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming; subjective; complex models
- Robustness
  - Global & local
  - Rarely used
  - Approximation



Robustness quantification

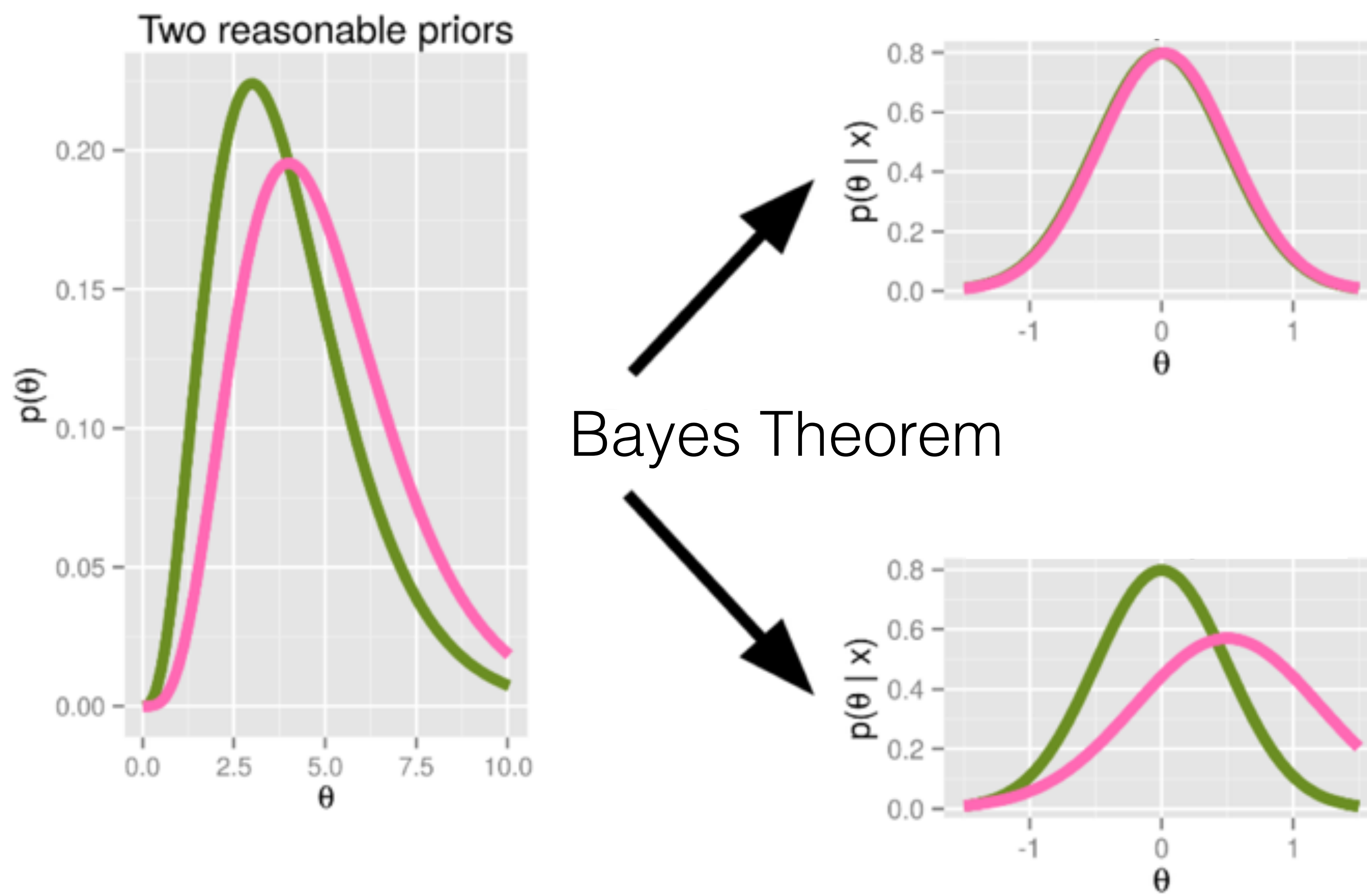
- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
    - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
    - Time-consuming; subjective; complex models
- Robustness
    - Global & local
    - Rarely used
    - Approximation, MCMC



Robustness quantification

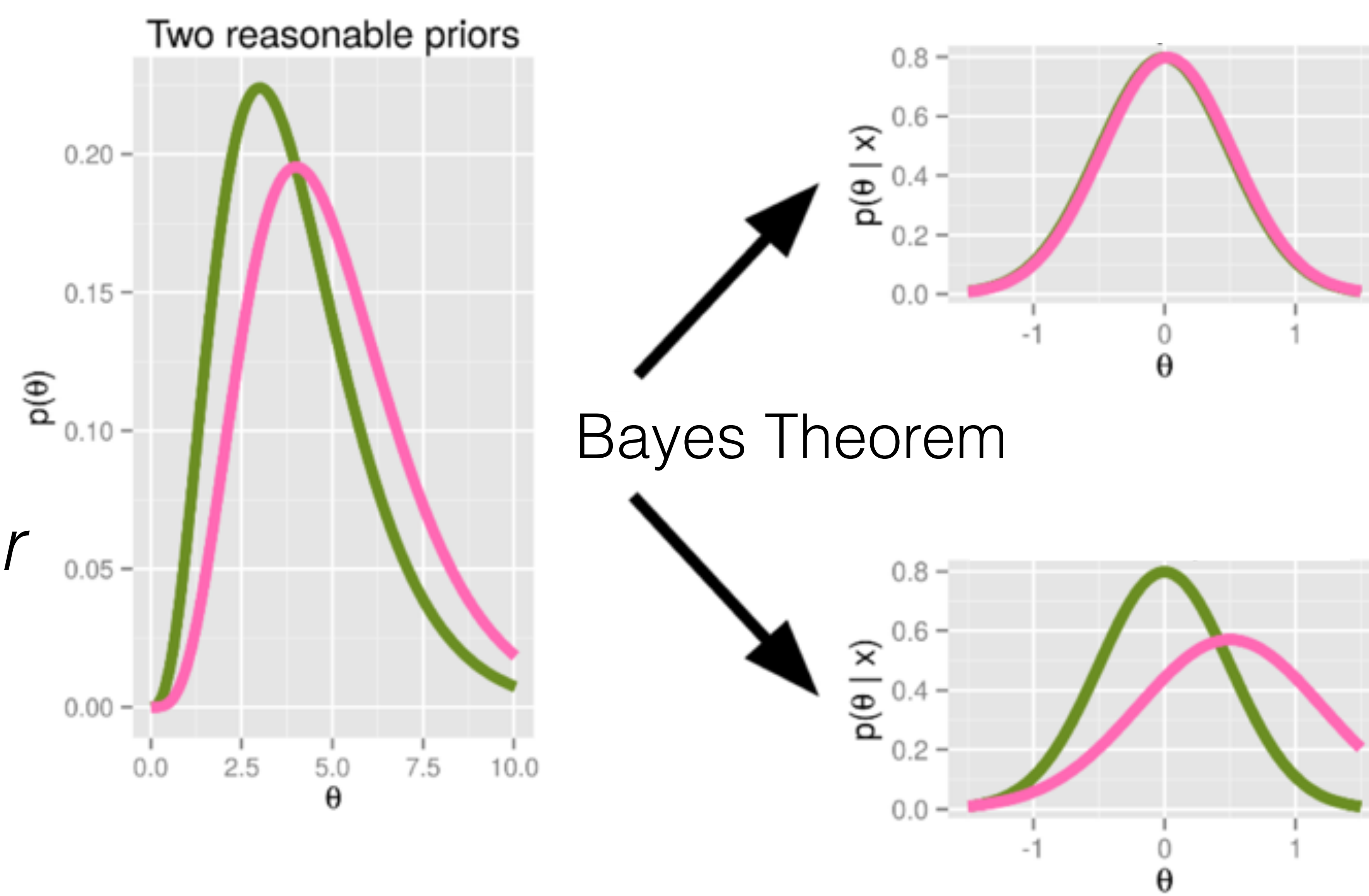
- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming; subjective; complex models
- Robustness
  - Global & local
  - Rarely used
  - Approximation, MCMC
- Our solution:

  variational Bayes



Robustness quantification

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Have to express prior beliefs in a distribution: challenges
  - Time-consuming; subjective; complex models
- Robustness
  - Global & local
  - Rarely used
  - Approximation, MCMC
- Our solution: *linear response variational Bayes*



Robustness quantification

Robustness quantification

- Variational Bayes as an alternative to MCMC

Robustness quantification

- Variational Bayes as an alternative to MCMC
- Challenges of VB

Robustness quantification

- Variational Bayes as an alternative to MCMC
- Challenges of VB
- Accurate uncertainties from VB

Robustness quantification

- Variational Bayes as an alternative to MCMC
- Challenges of VB
- Accurate uncertainties from VB
- Accurate robustness quantification from VB

Robustness quantification

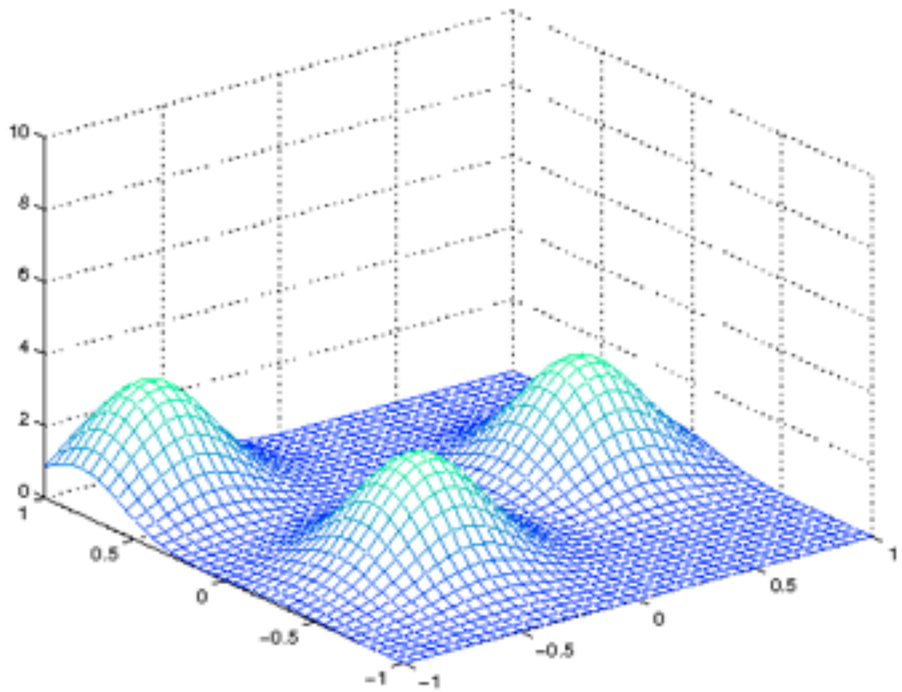
- Variational Bayes as an alternative to MCMC
- Challenges of VB
- Accurate uncertainties from VB
- Accurate robustness quantification from VB
- Big idea: derivatives/perturbations are easy in VB

Variational Bayes

Variational Bayes

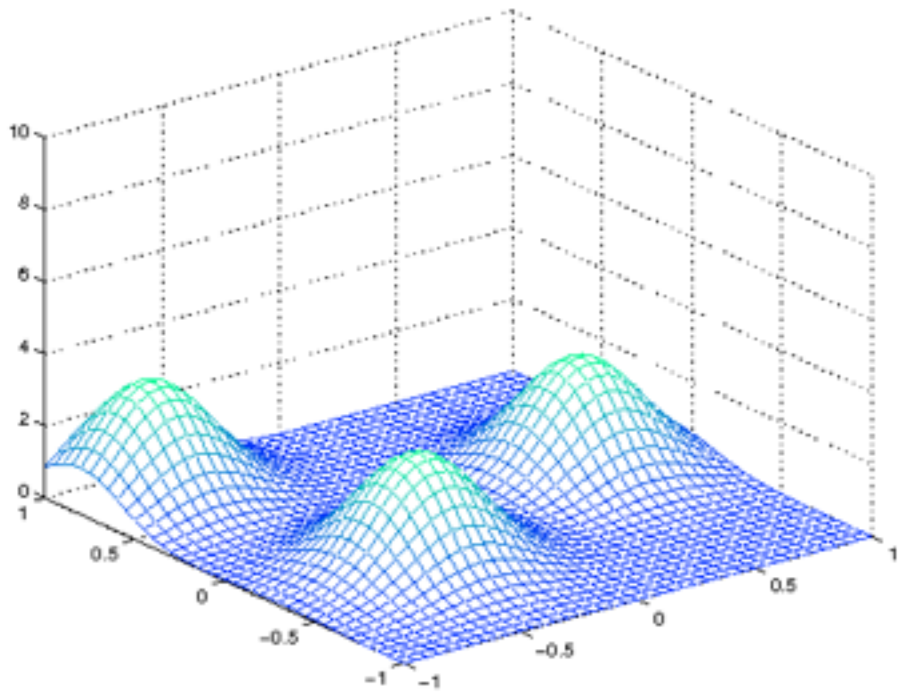
- Variational Bayes (VB)

Variational Bayes



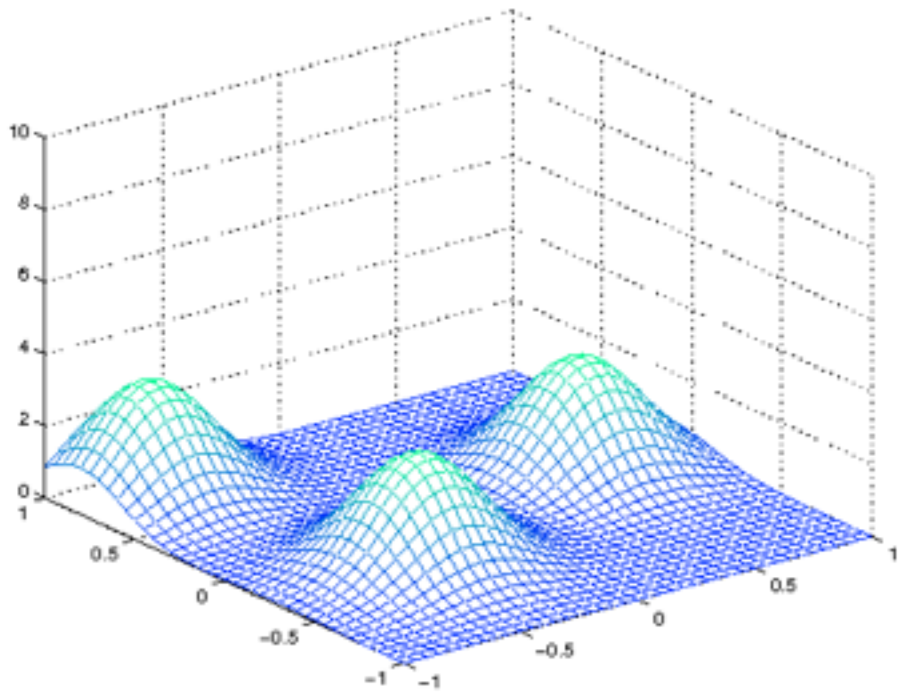
- Variational Bayes (VB)

Variational Bayes

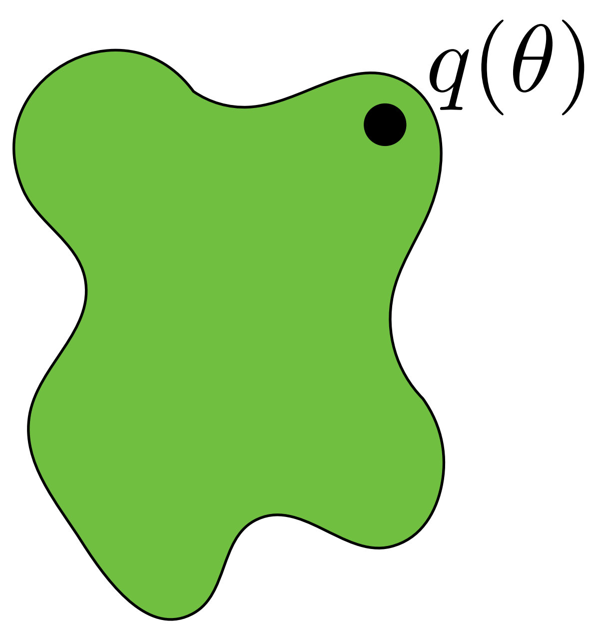


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$

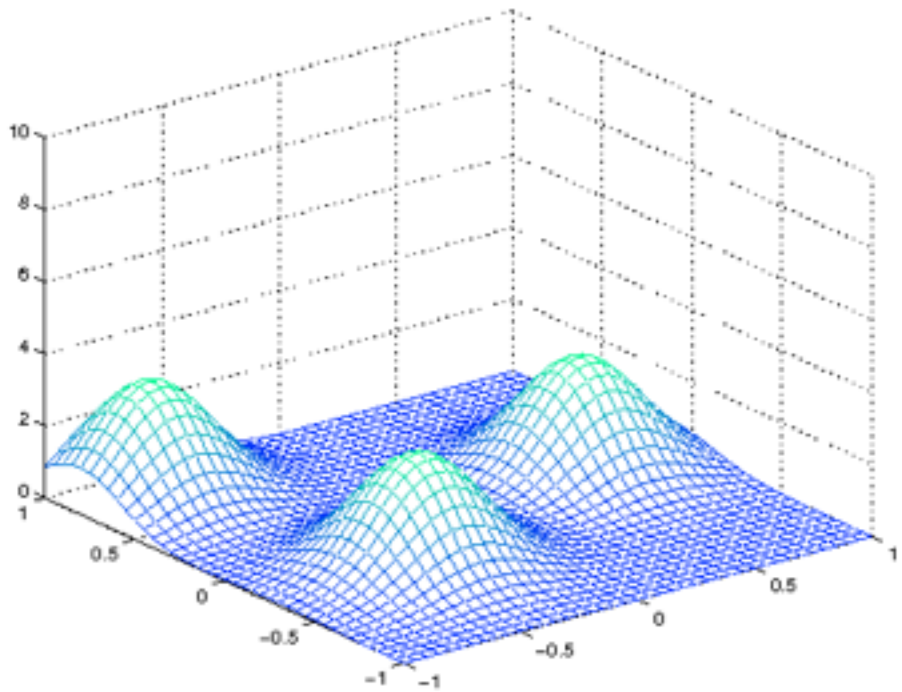
Variational Bayes

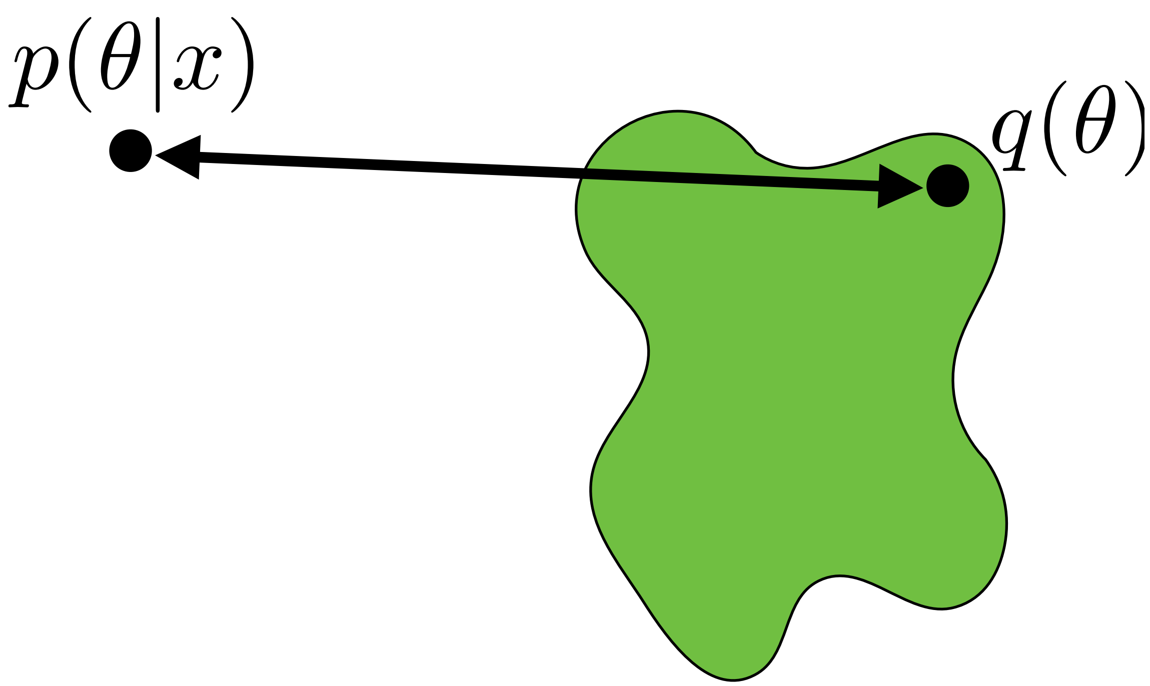


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$



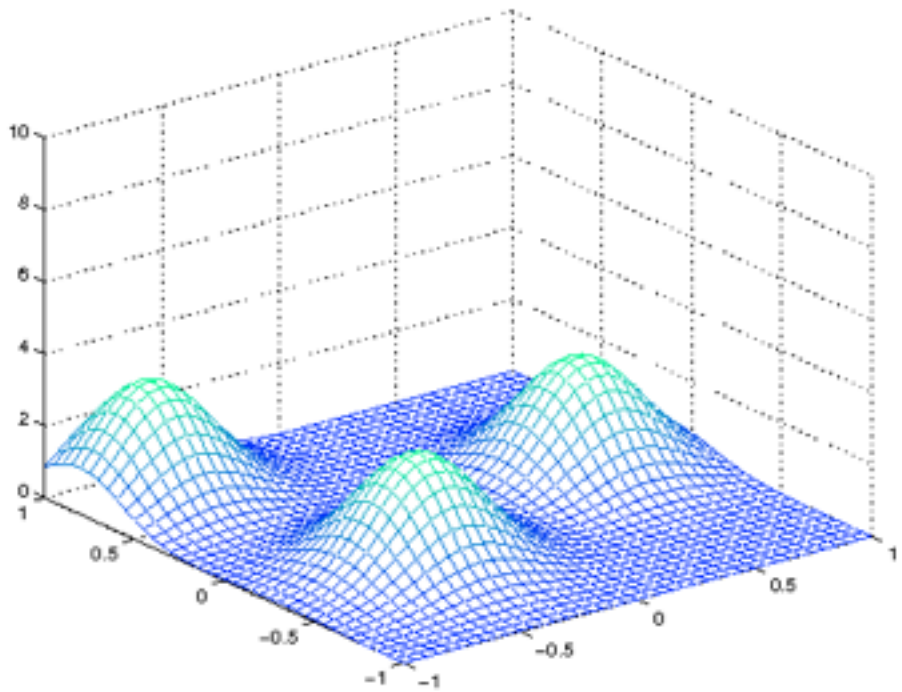
Variational Bayes



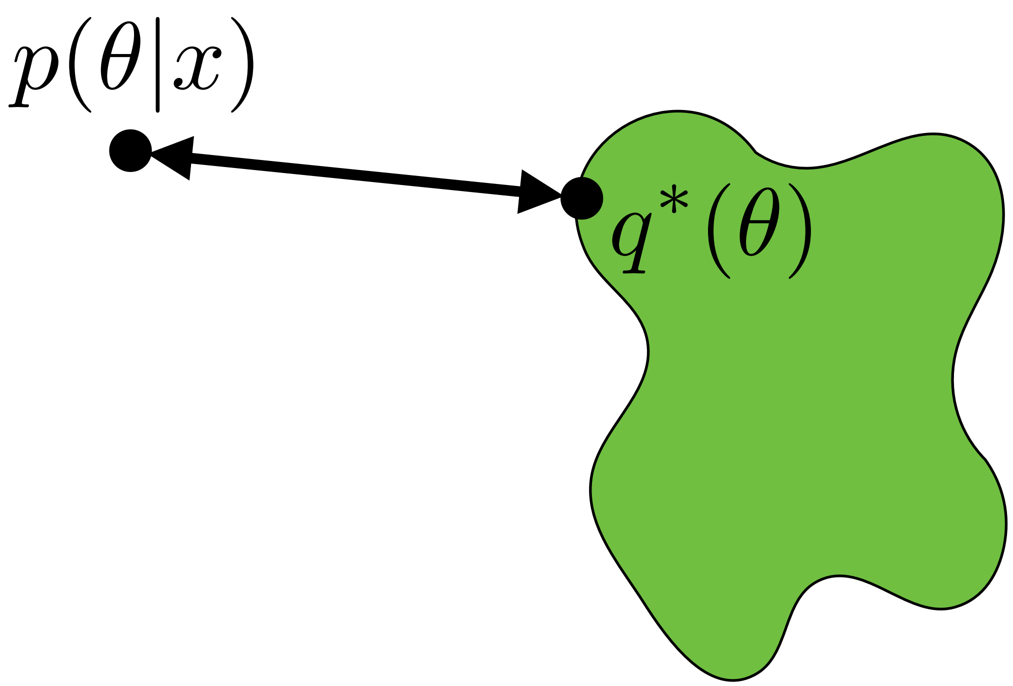


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$

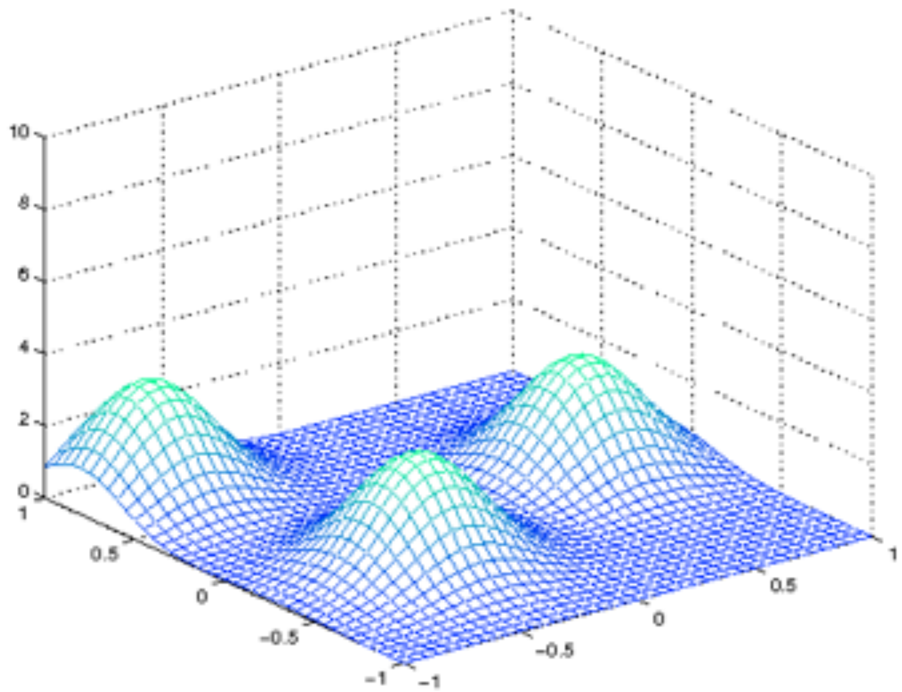
Variational Bayes

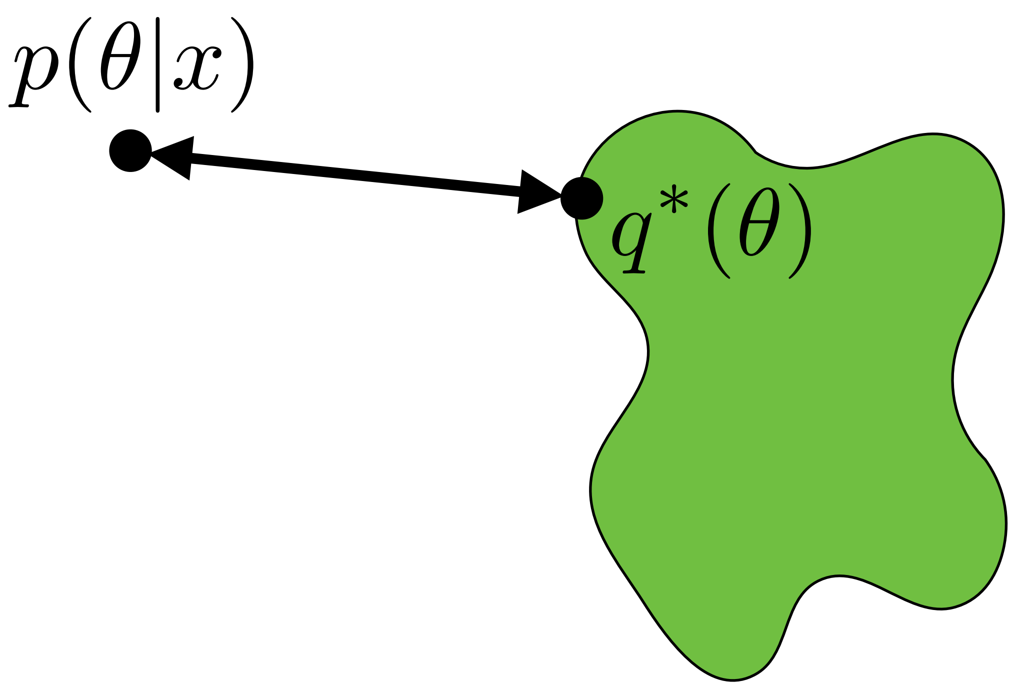


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$



Variational Bayes

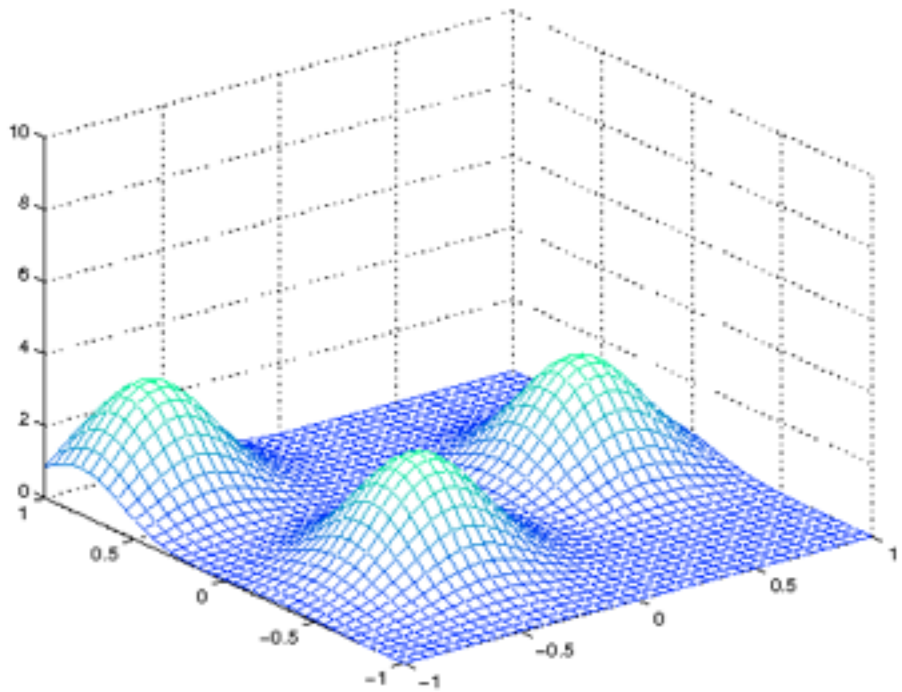




- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

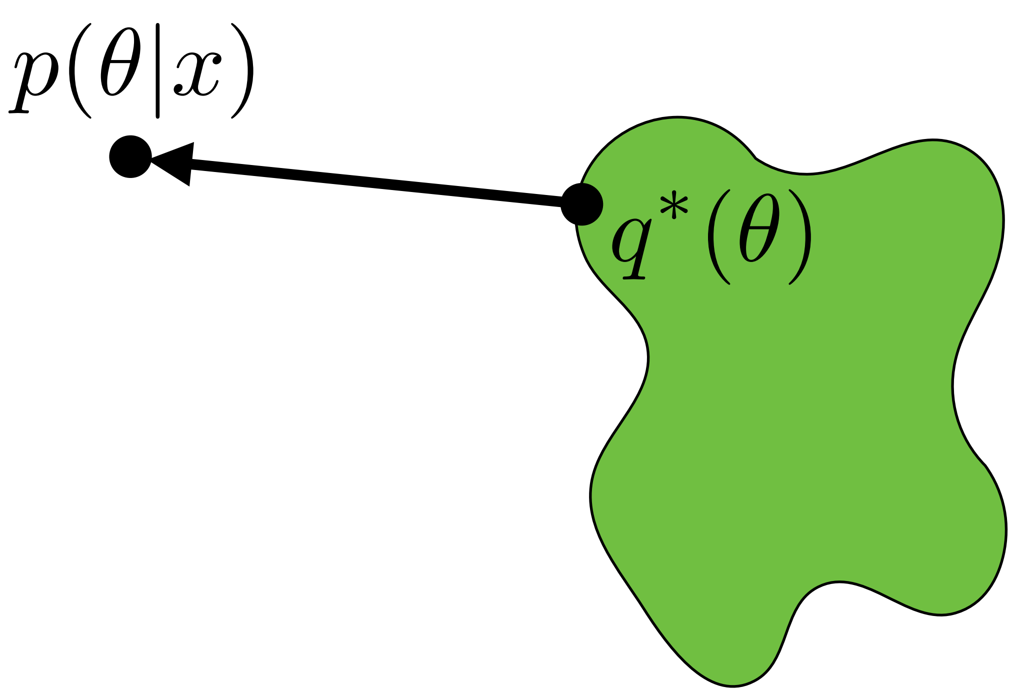
$$KL(q||p(\cdot|x))$$

Variational Bayes

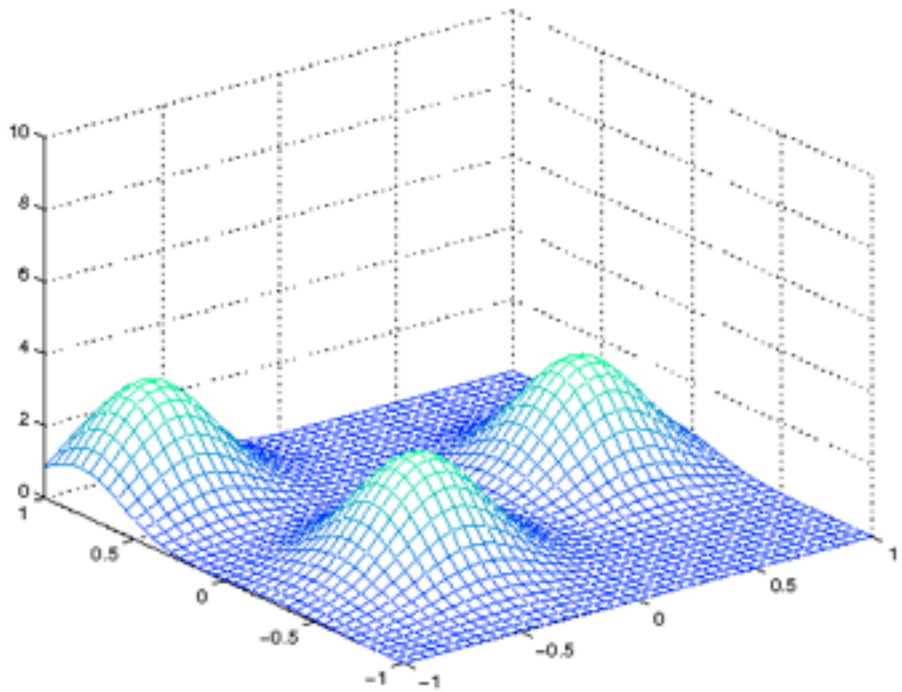


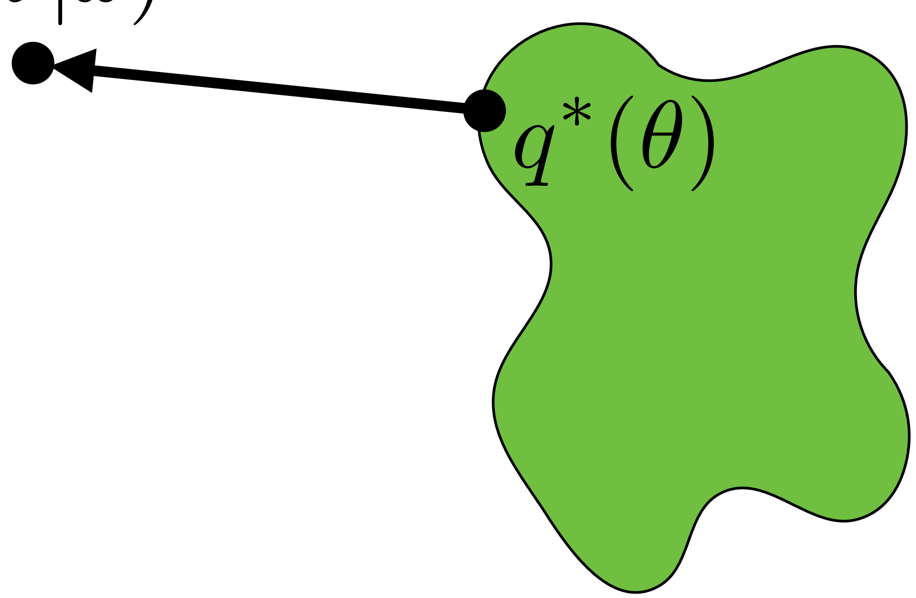
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

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



Variational Bayes



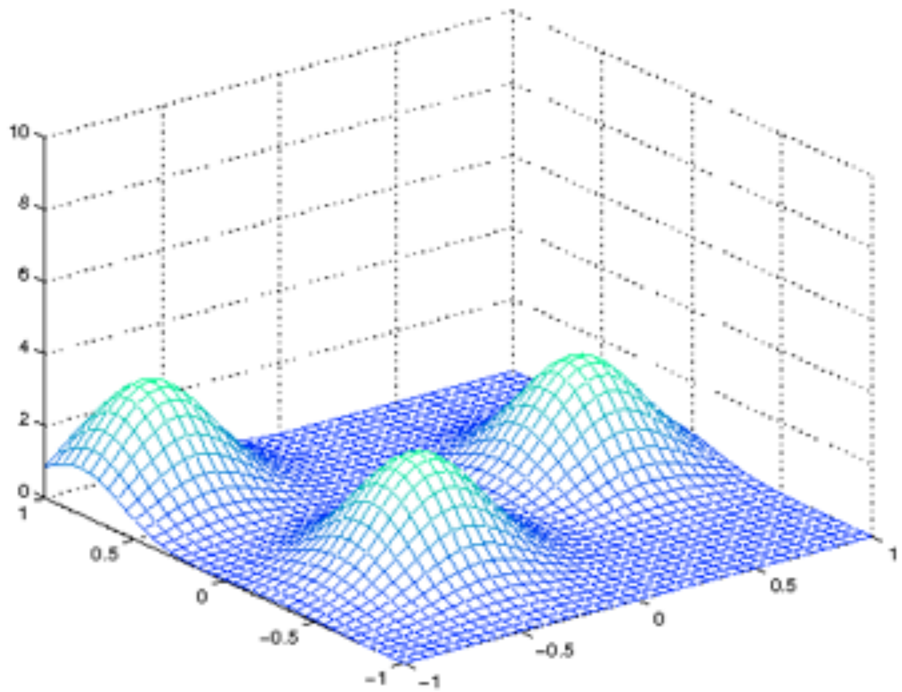


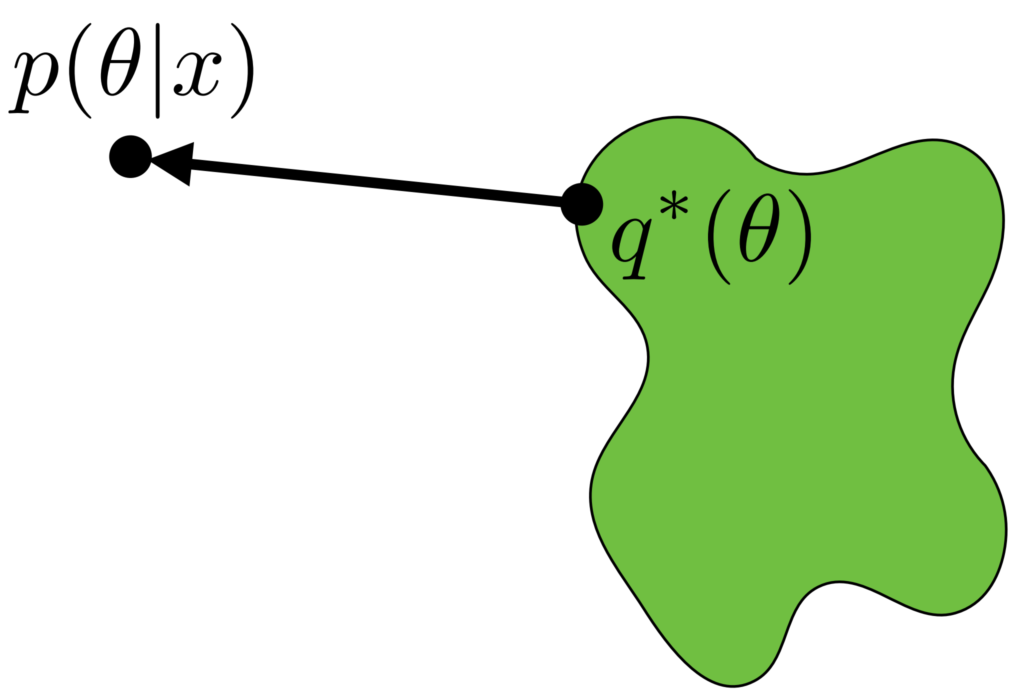
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

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

- VB practical success

Variational Bayes



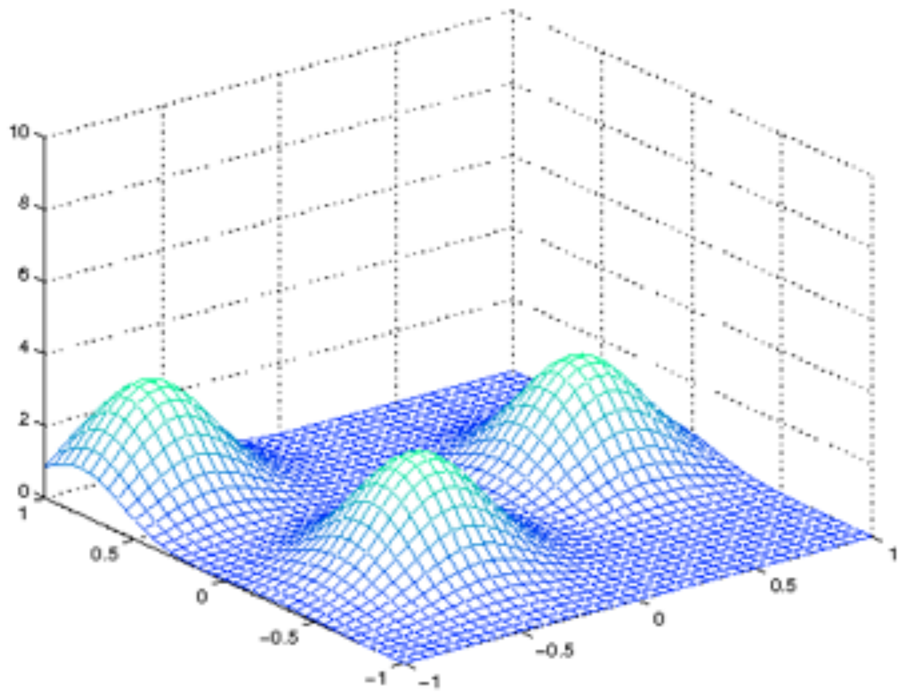


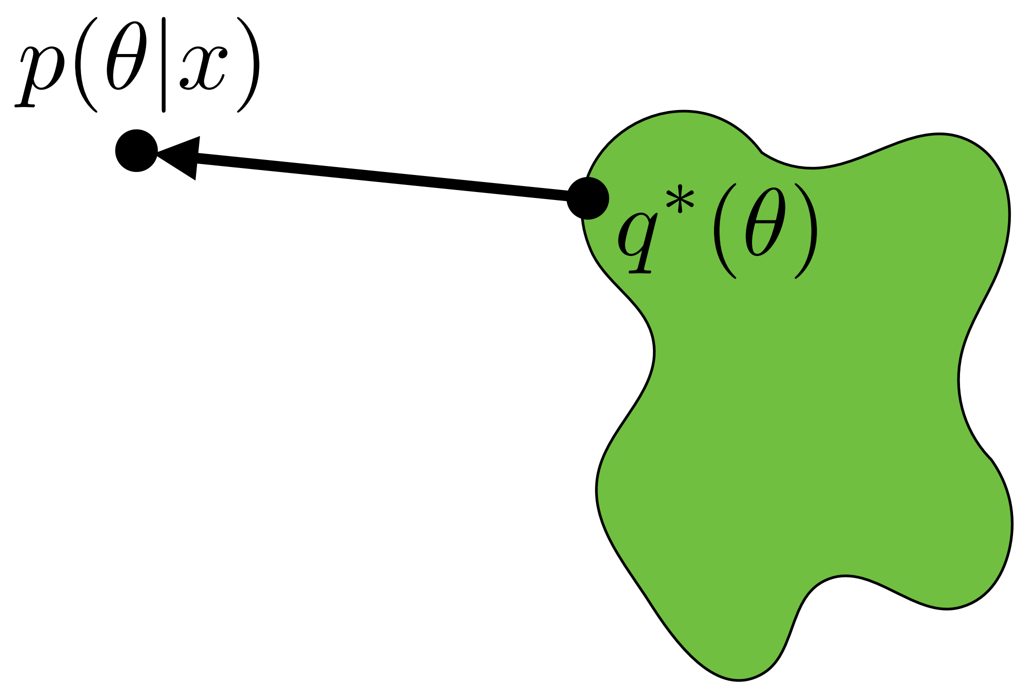
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

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

- VB practical success
  - point estimates and prediction

Variational Bayes





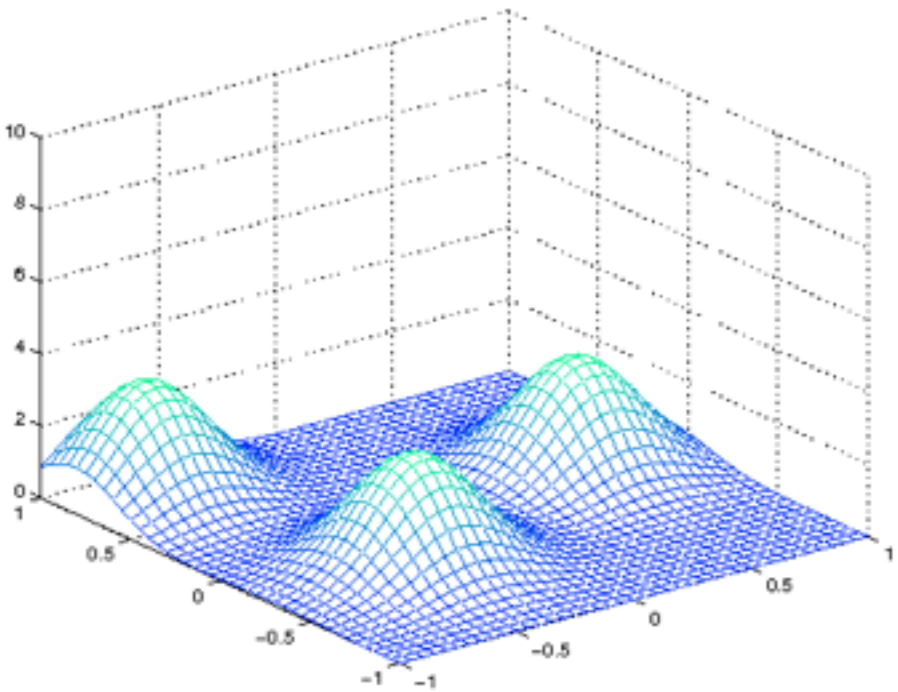
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

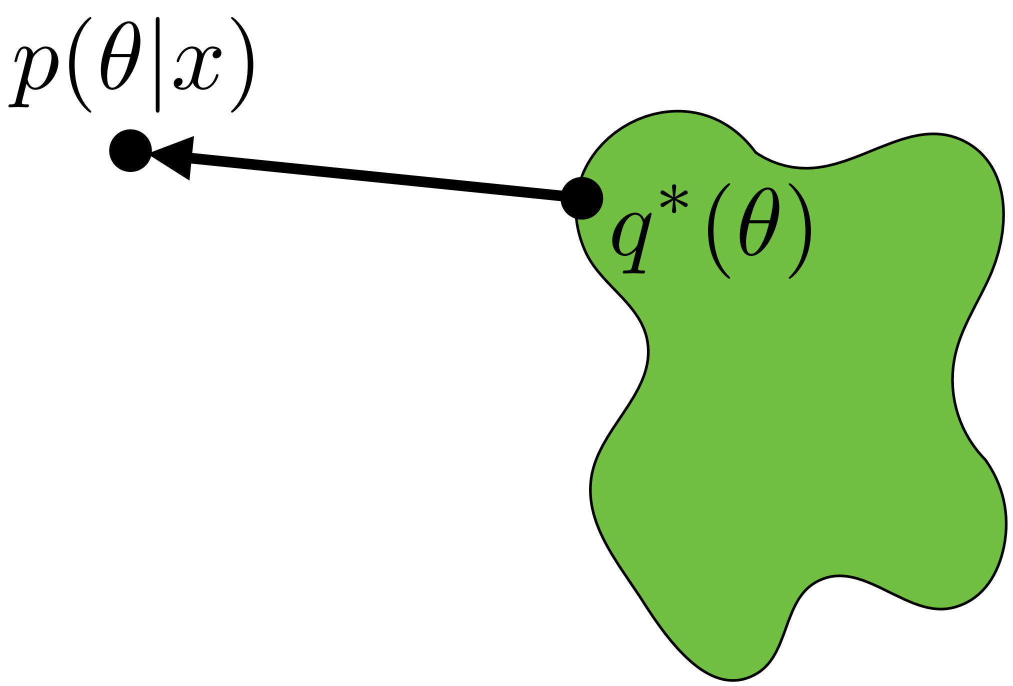
$$KL(q||p(\cdot|x))$$

- VB practical success
  - point estimates and prediction
  - fast

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

Variational Bayes





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

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

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

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

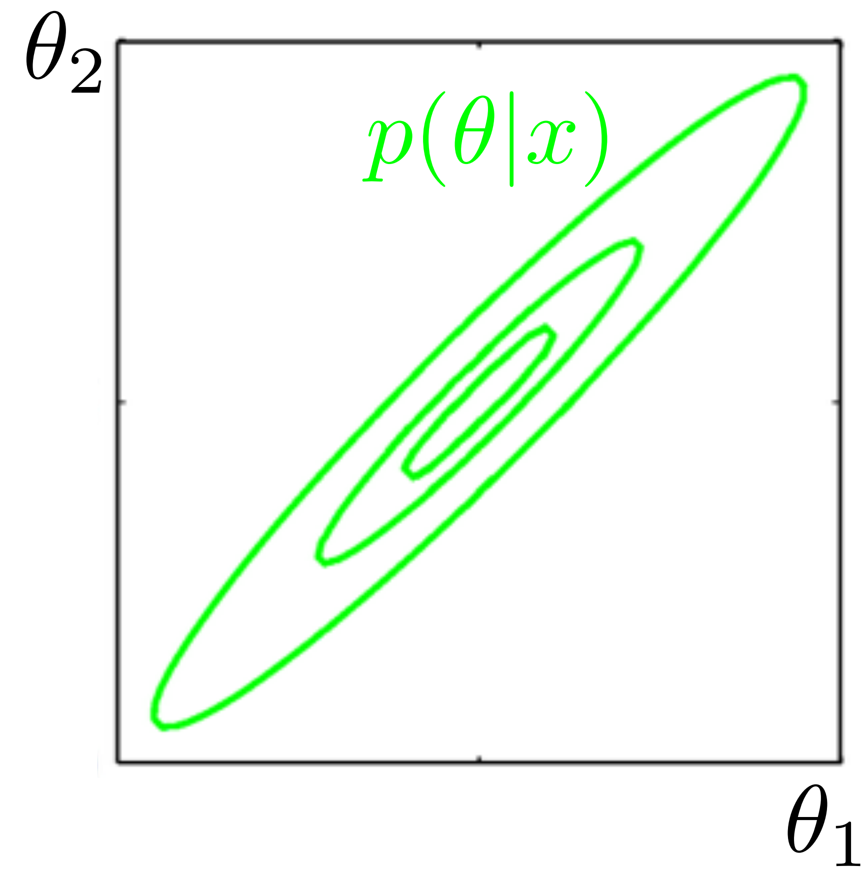
What about uncertainty?

What about uncertainty?

- Variational Bayes

What about uncertainty?

- Variational Bayes

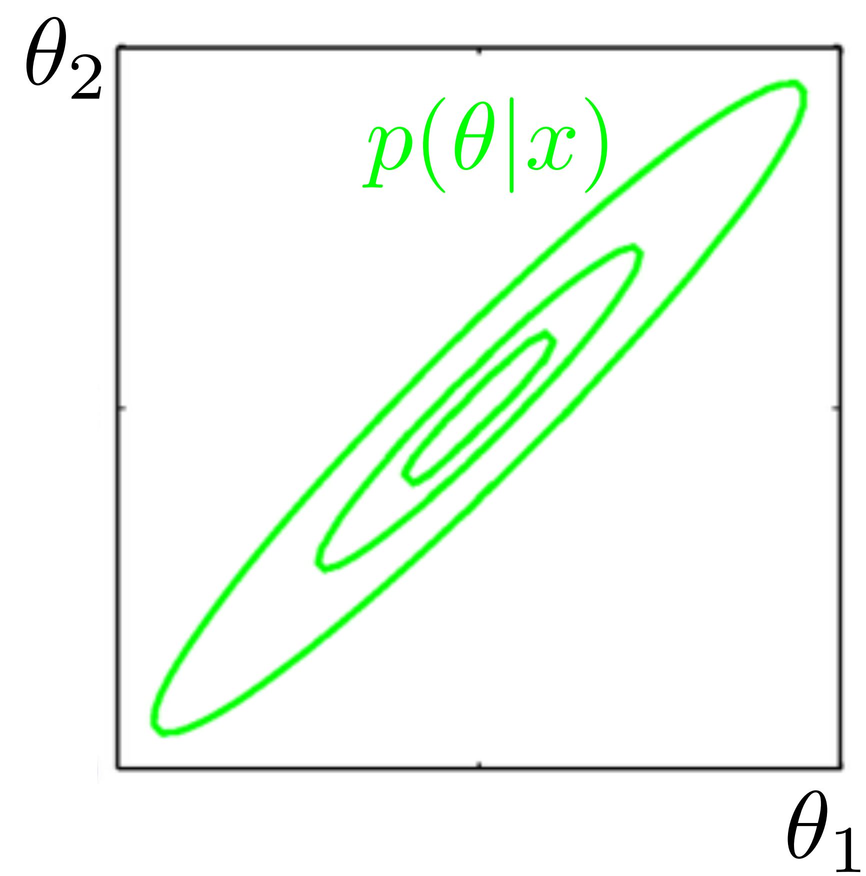


[Bishop 2006]

What about uncertainty?

- Variational Bayes

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



[Bishop 2006]

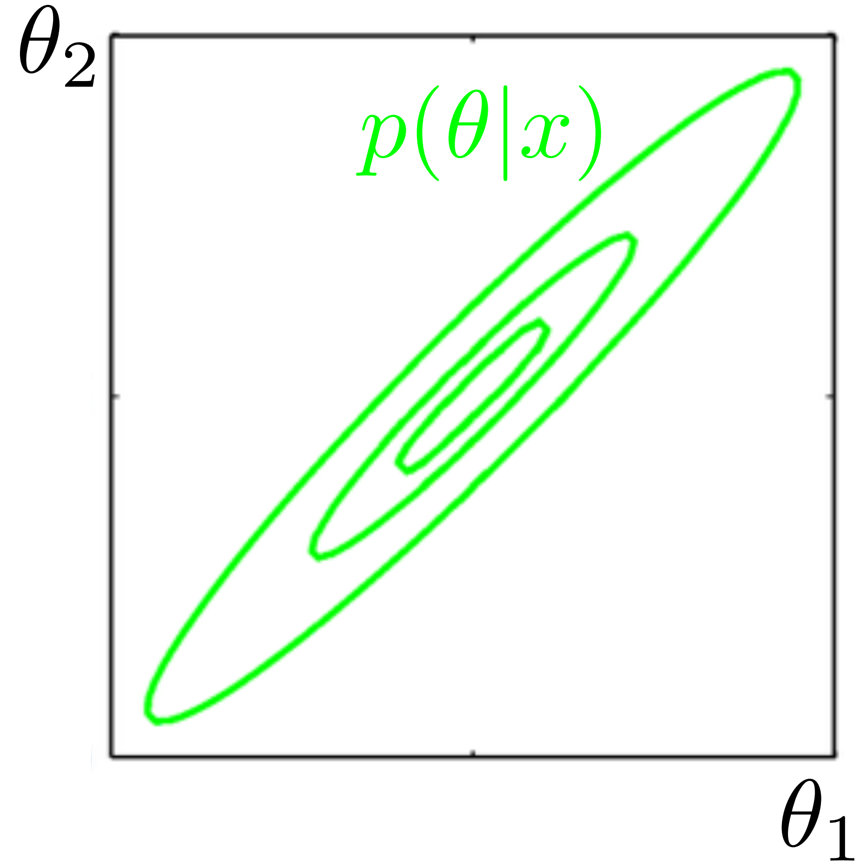
What about uncertainty?

- Variational Bayes

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

- Mean-field variational Bayes (MFVB)

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



[Bishop 2006]

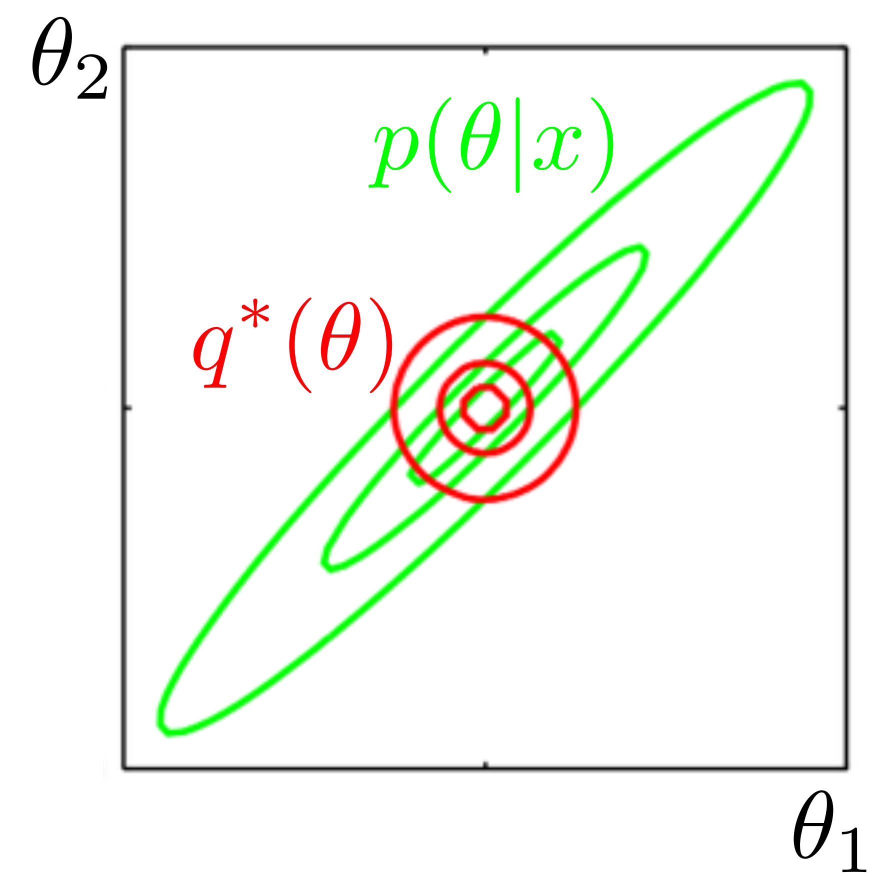
What about uncertainty?

- Variational Bayes

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

- Mean-field variational Bayes (MFVB)

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



[Bishop 2006]

What about uncertainty?

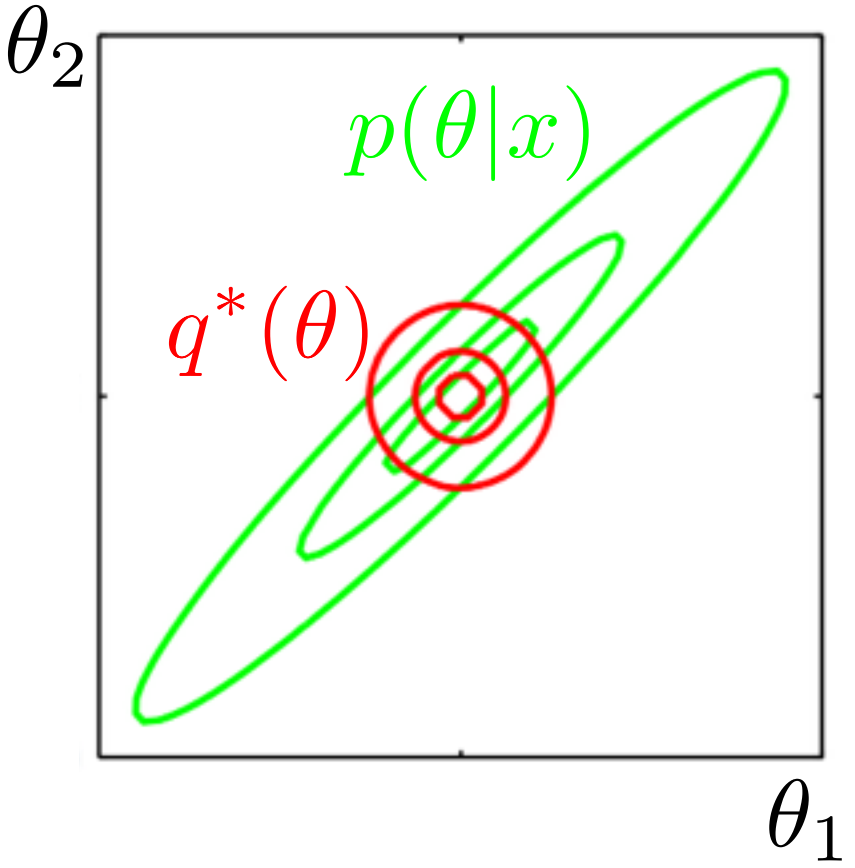
- Variational Bayes

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

- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)



[Bishop 2006]

What about uncertainty?

- Variational Bayes

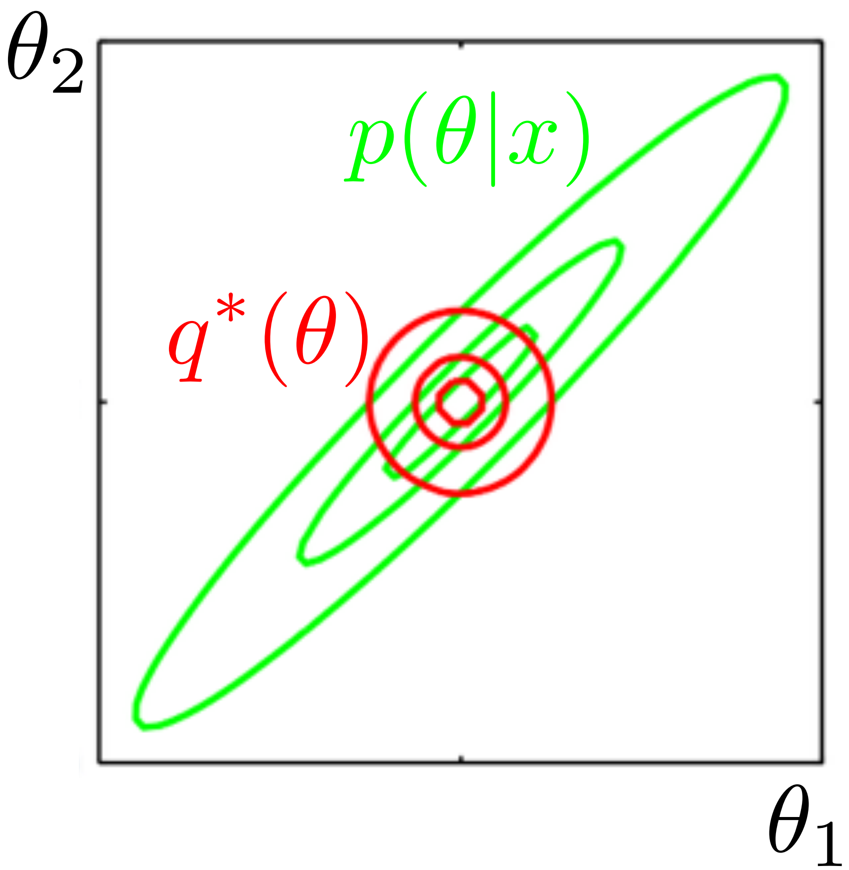
$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)

- No covariance estimates



[Bishop 2006]

What about uncertainty?

- Variational Bayes

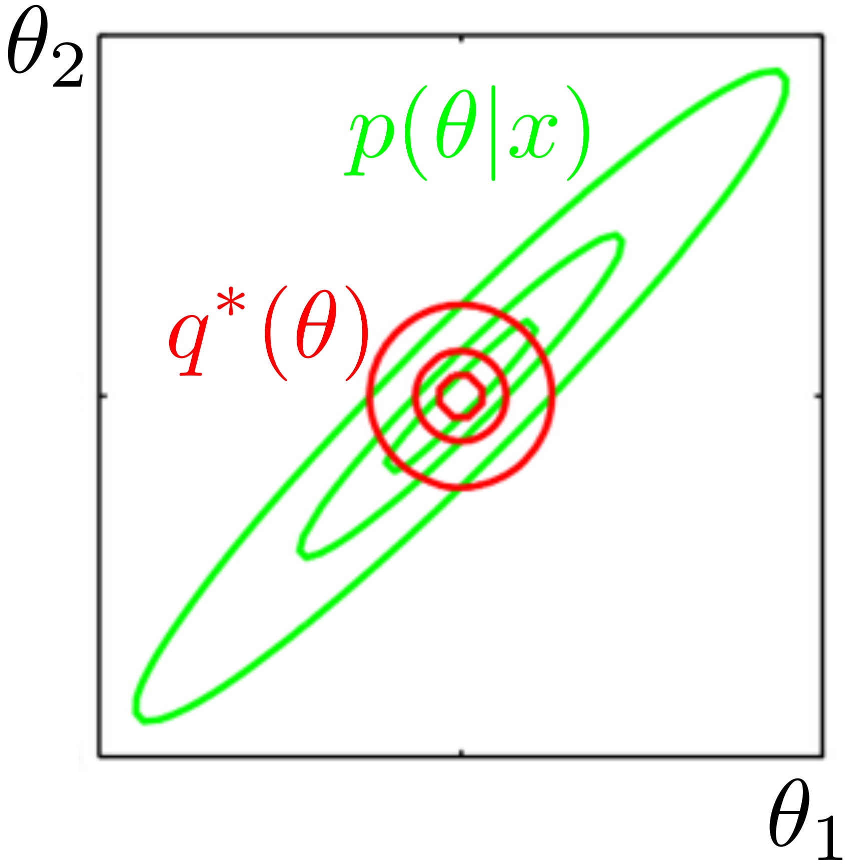
$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)

- No covariance estimates



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

What about uncertainty?

- Variational Bayes

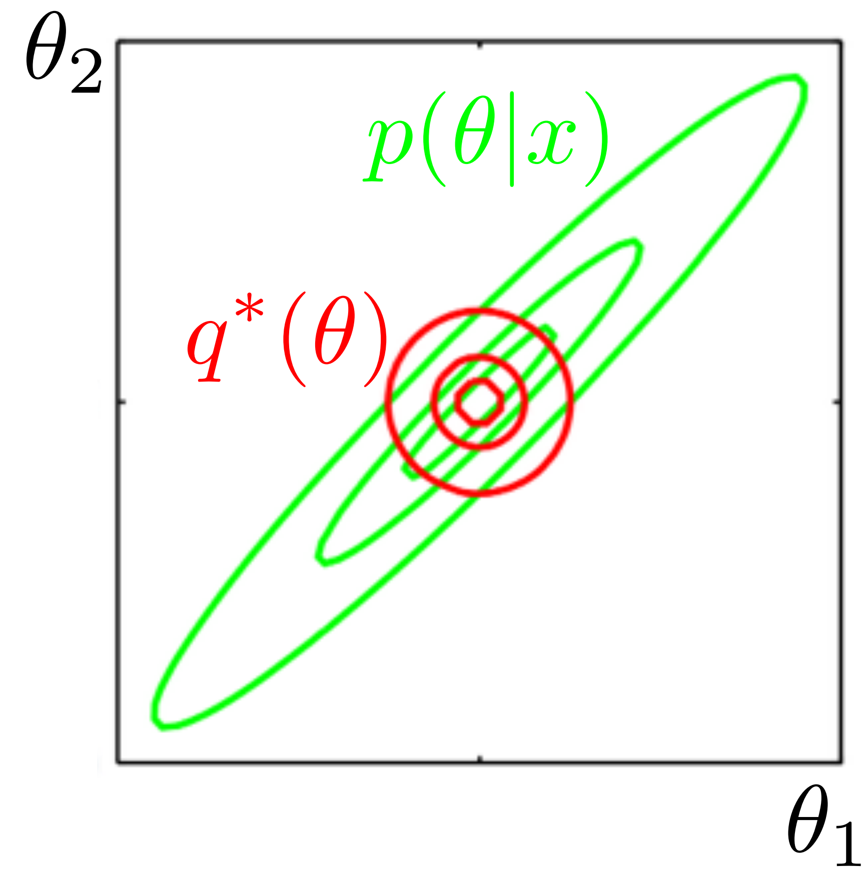
$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)

- No covariance estimates



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

Linear response

Linear response

- Cumulant-generating function

Linear response

- Cumulant-generating function

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

Linear response

- Cumulant-generating function

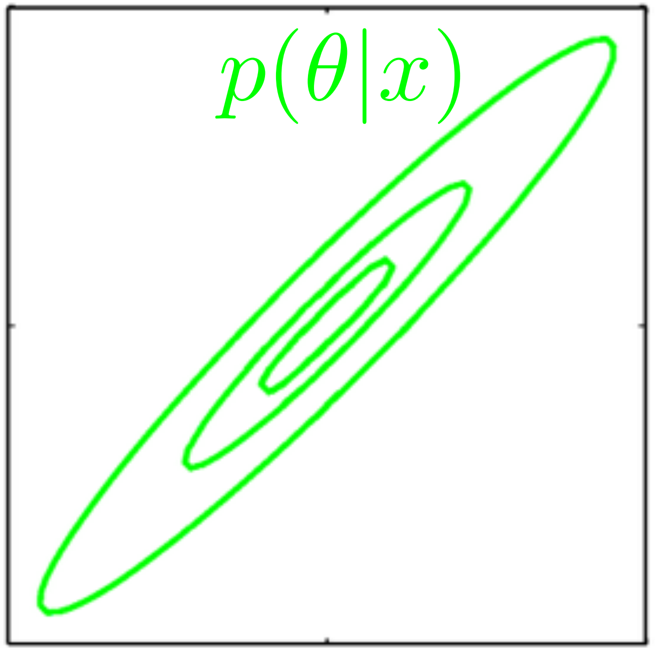
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

Linear response

- Cumulant-generating function

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

- True posterior covariance



[Bishop 2006]

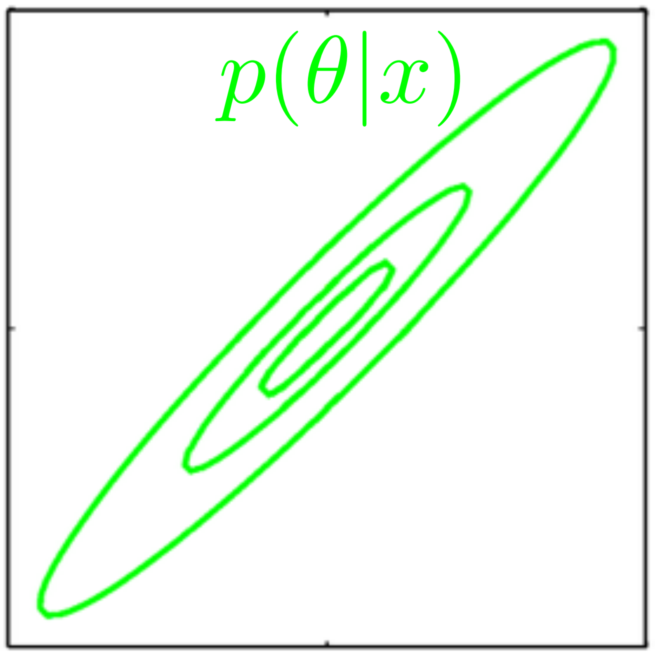
Linear response

- Cumulant-generating function

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

- True posterior covariance

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



[Bishop 2006]

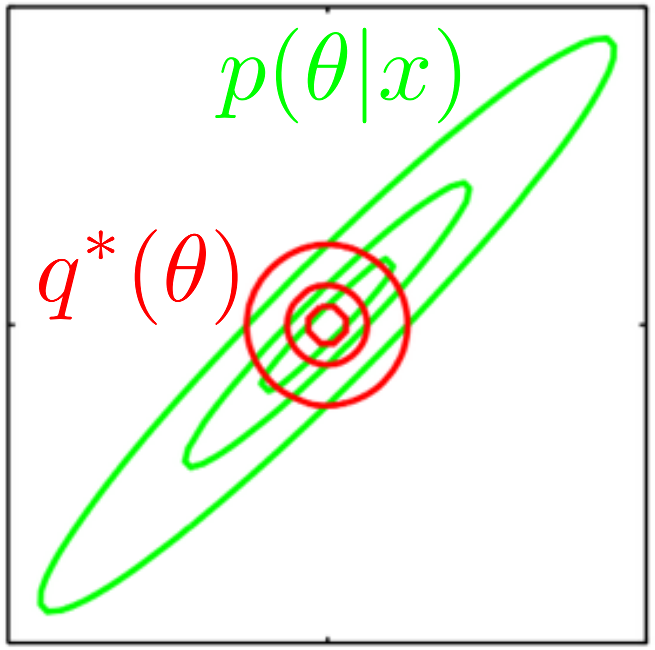
Linear response

- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

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



[Bishop 2006]

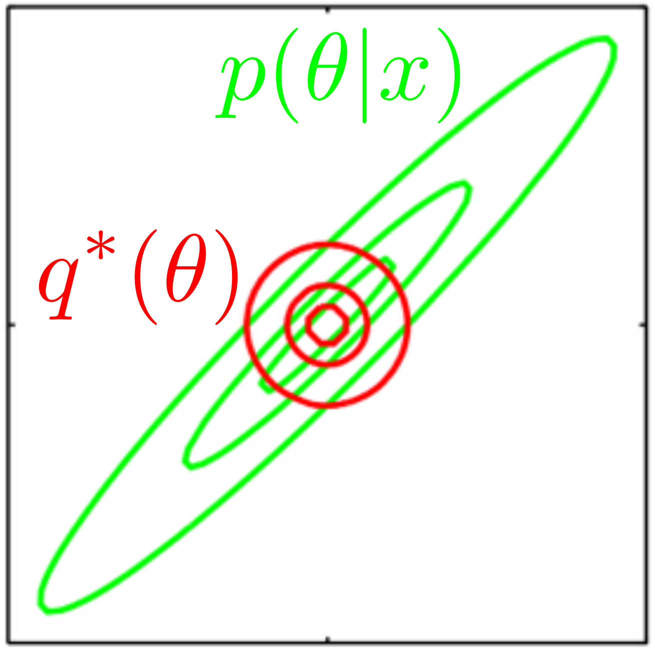
Linear response

- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

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



[Bishop 2006]

Linear response

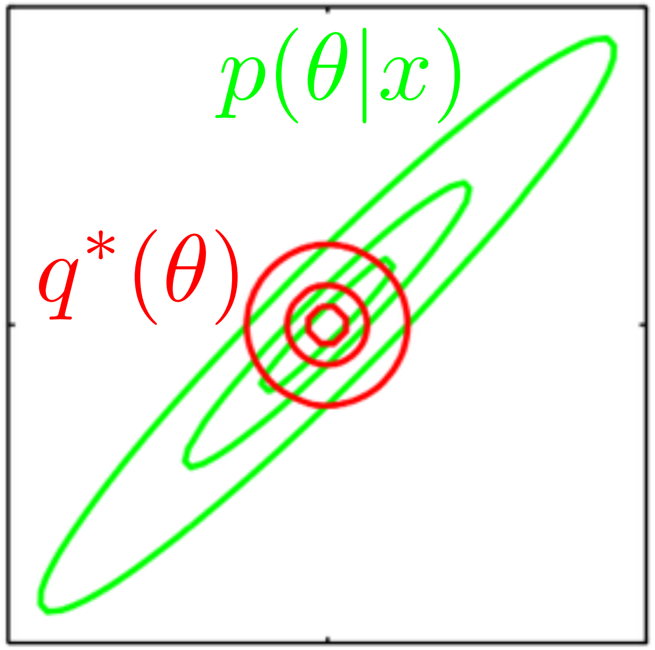
- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

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

- "Linear response"



[Bishop 2006]

Linear response

- Cumulant-generating function

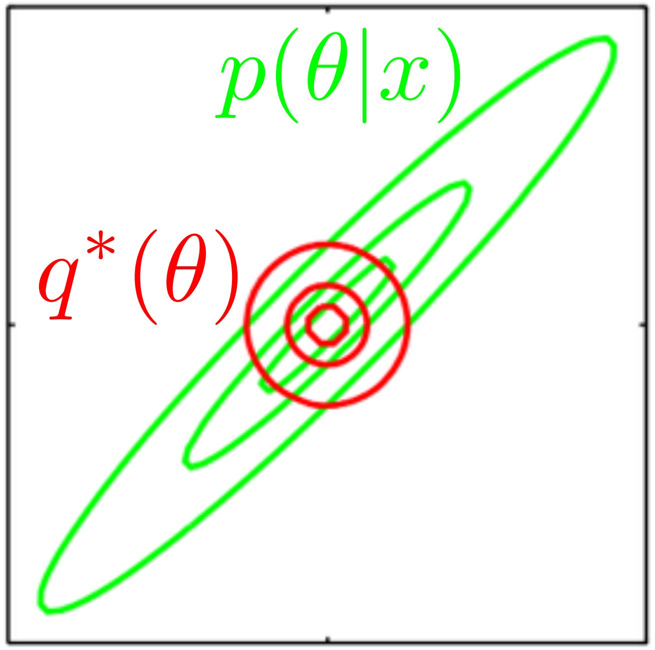
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

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

- "Linear response"

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



[Bishop 2006]

Linear response

- Cumulant-generating function

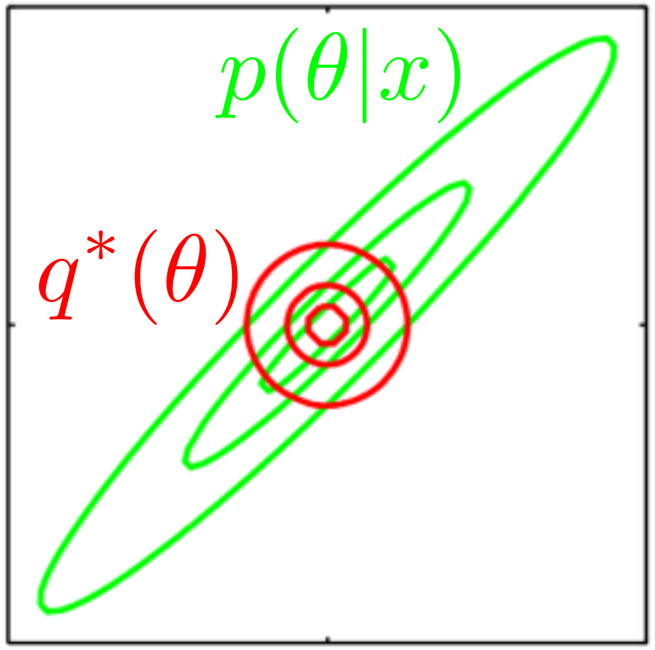
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

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

- "Linear response"

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



[Bishop 2006]

Linear response

- Cumulant-generating function

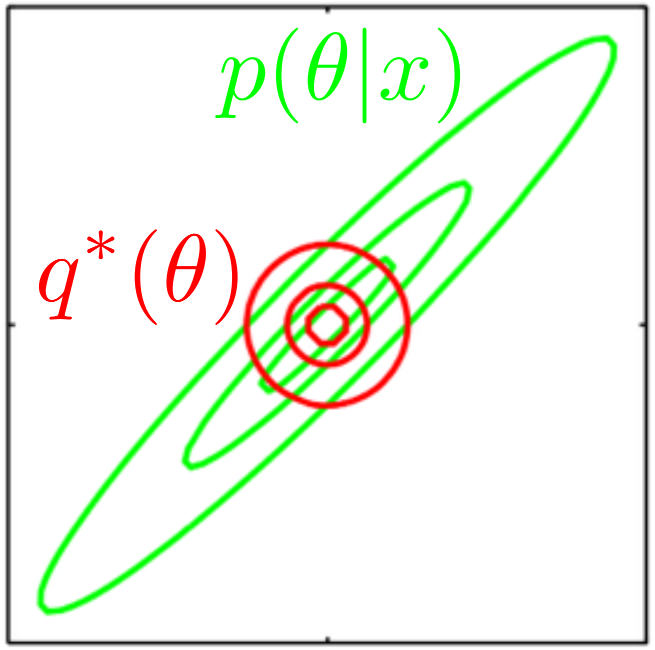
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

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

- "Linear response"

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



[Bishop 2006]

Linear response

- Cumulant-generating function

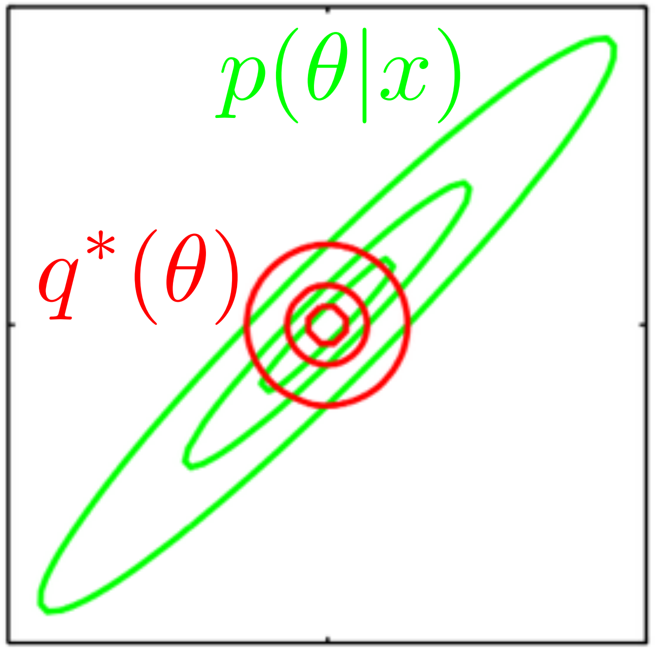
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

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

- "Linear response"

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



[Bishop 2006]

Linear response

- Cumulant-generating function

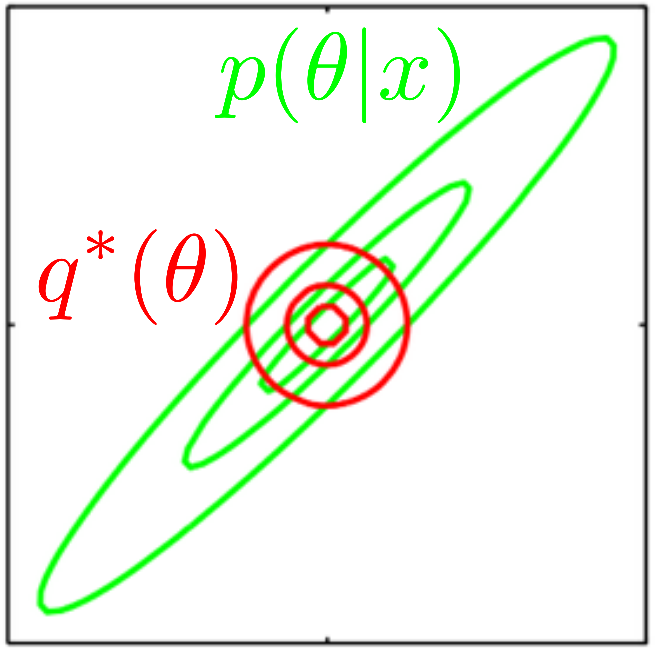
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

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

- "Linear response"

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



[Bishop 2006]

Linear response

- Cumulant-generating function

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

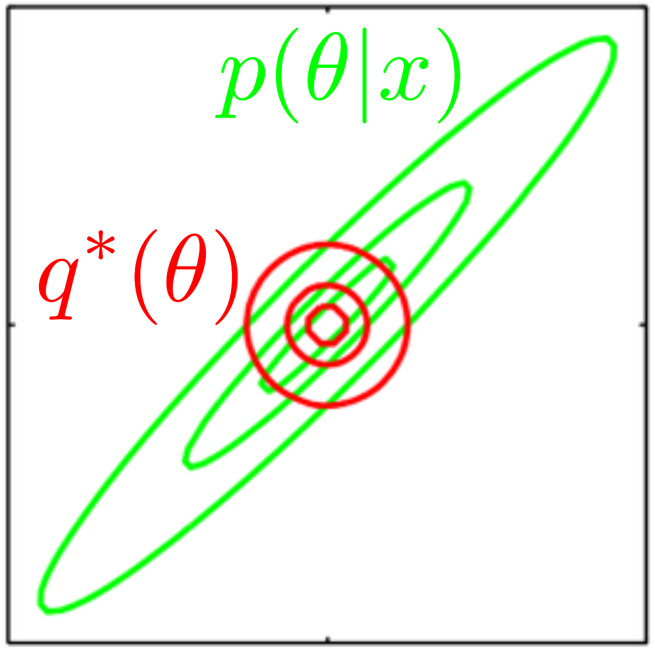
- True posterior covariance vs MFVB covariance

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

- "Linear response"

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

- The LRVB approximation



[Bishop 2006]

Linear response

- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

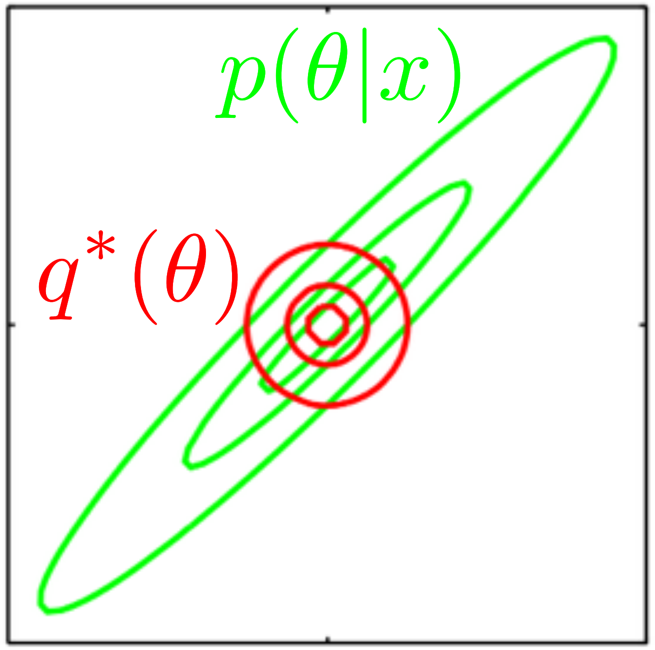
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

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

- The LRVB approximation

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



[Bishop 2006]

Linear response

- Cumulant-generating function

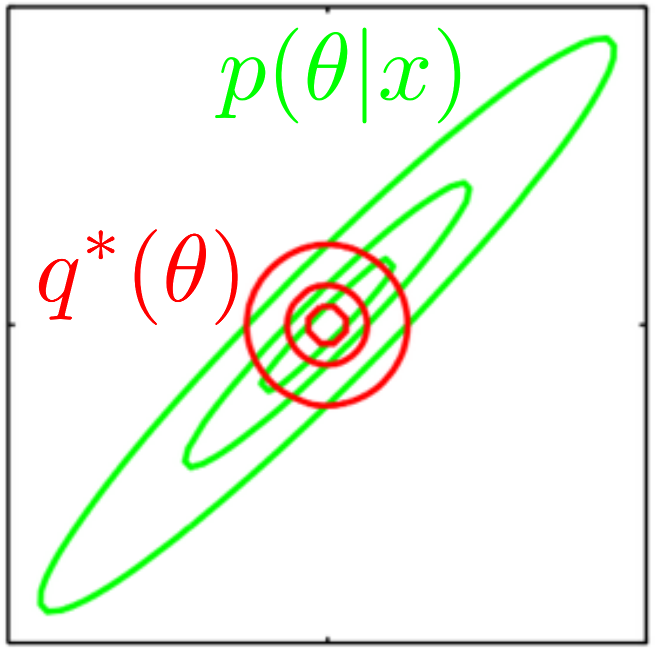
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance vs MFVB covariance

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

- "Linear response"

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



- The LRVB approximation

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

[Bishop 2006]

Linear response

- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

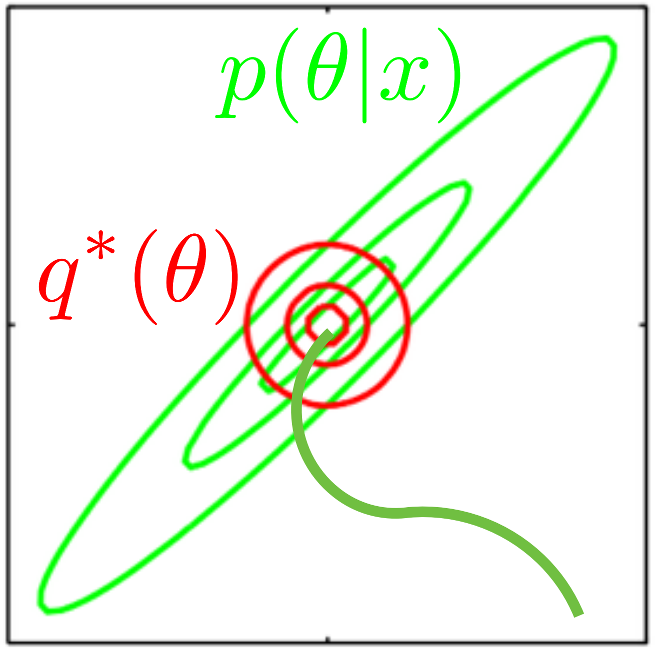
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

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

- The LRVB approximation

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



[Bishop 2006]

Linear response

- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

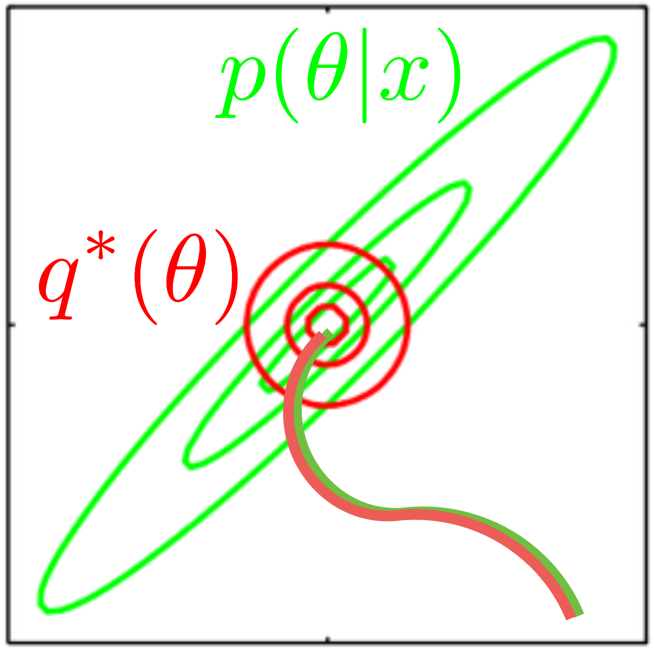
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

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

- The LRVB approximation

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



[Bishop 2006]

Linear response

- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

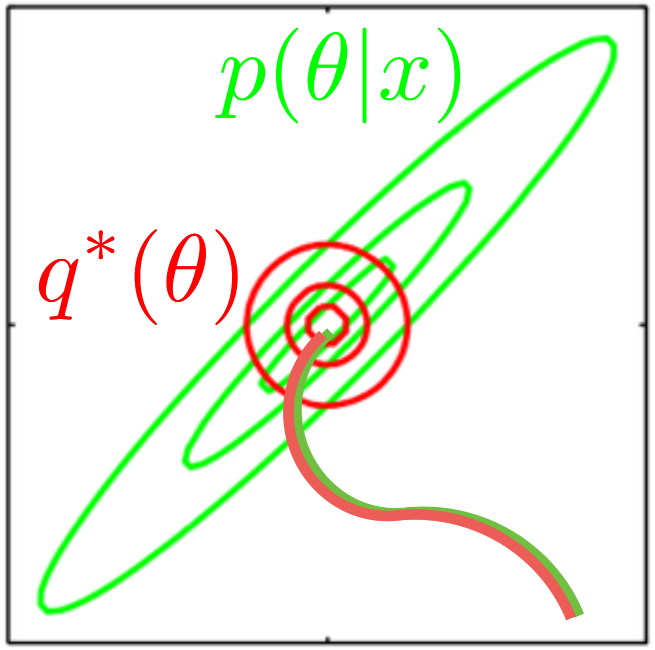
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

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

- The LRVB approximation

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



[Bishop 2006]

Linear response

- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

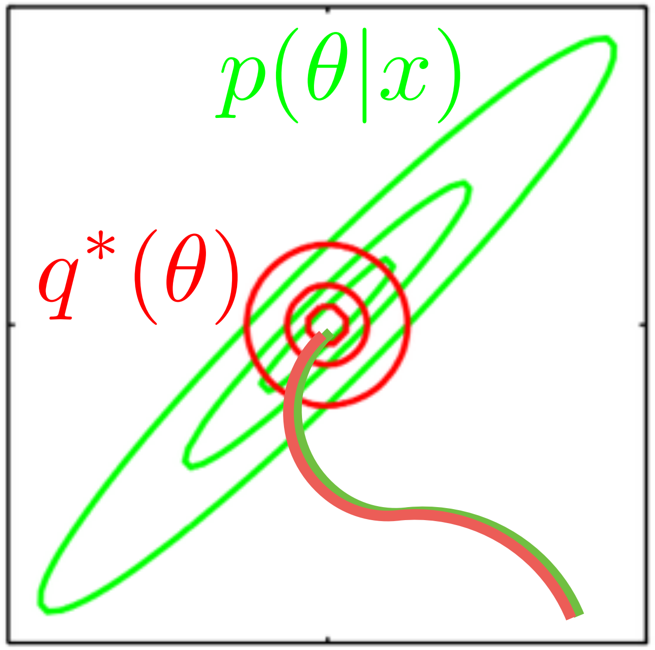
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

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

- The LRVB approximation

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



[Bishop 2006]

LRVB estimator

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

LRVB estimator

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

LRVB estimator

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

LRVB estimator

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

$$\hat{\Sigma} =$$

LRVB estimator

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

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

LRVB estimator

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

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

LRVB estimator

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

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

- Symmetric and positive definite at local min of KL

LRVB estimator

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

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

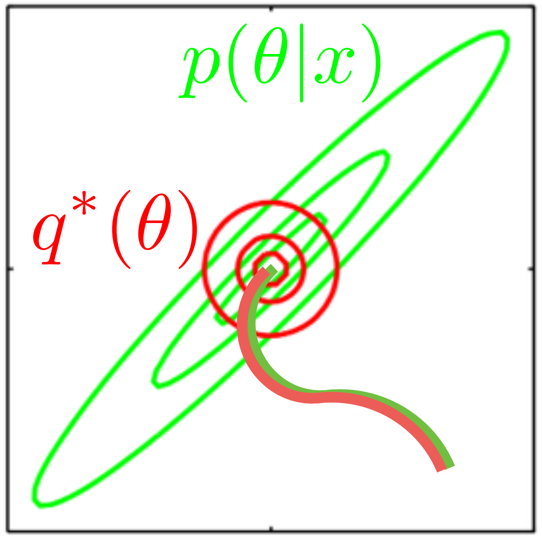
- Symmetric and positive definite at local min of KL
- The LRVB assumption: $\mathbb{E}_{p_t} \theta \approx \mathbb{E}_{q_t^*} \theta$

LRVB estimator

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

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

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



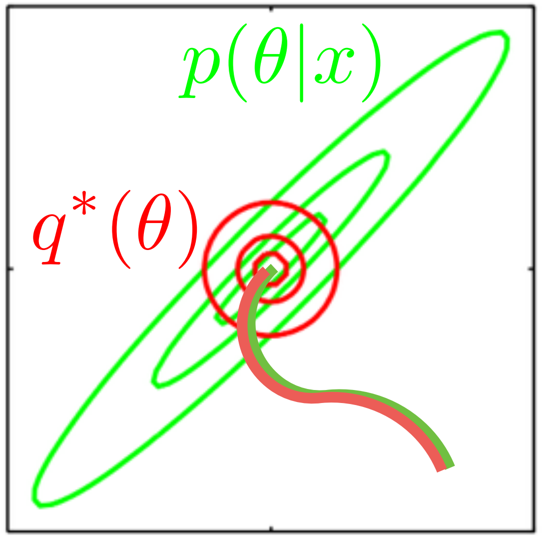
[Bishop 2006]

LRVB estimator

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

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

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



[Bishop 2006]

Microcredit Experiment

Microcredit Experiment

- Simplified from Meager (2015)

Microcredit Experiment

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

Microcredit Experiment

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

Microcredit Experiment

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

Microcredit Experiment

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

profit → y_{kn}

Microcredit Experiment

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

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

Microcredit Experiment

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

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

Microcredit Experiment

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

profit

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

Microcredit Experiment

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

1 if microcredit

profit

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

Microcredit Experiment

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

1 if microcredit

profit

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

Microcredit Experiment

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

1 if microcredit

profit

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

Microcredit Experiment

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

1 if microcredit

profit

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

- Priors and hyperpriors:

Microcredit Experiment

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

profit → y_{kn}; 1 if microcredit → T_{kn}

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

- Priors and hyperpriors:

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

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

Microcredit Experiment

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

profit

1 if microcredit

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

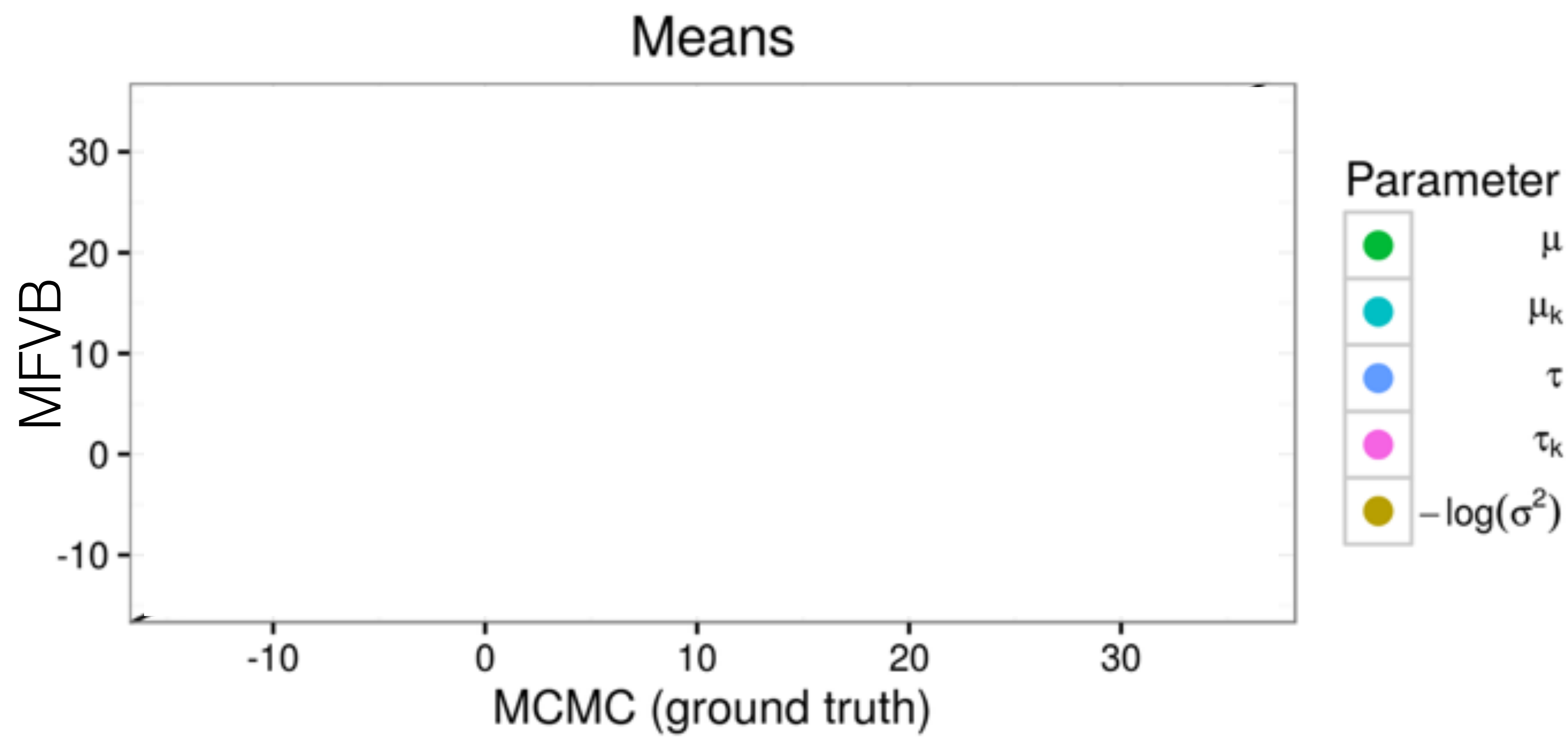
- Priors and hyperpriors:

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

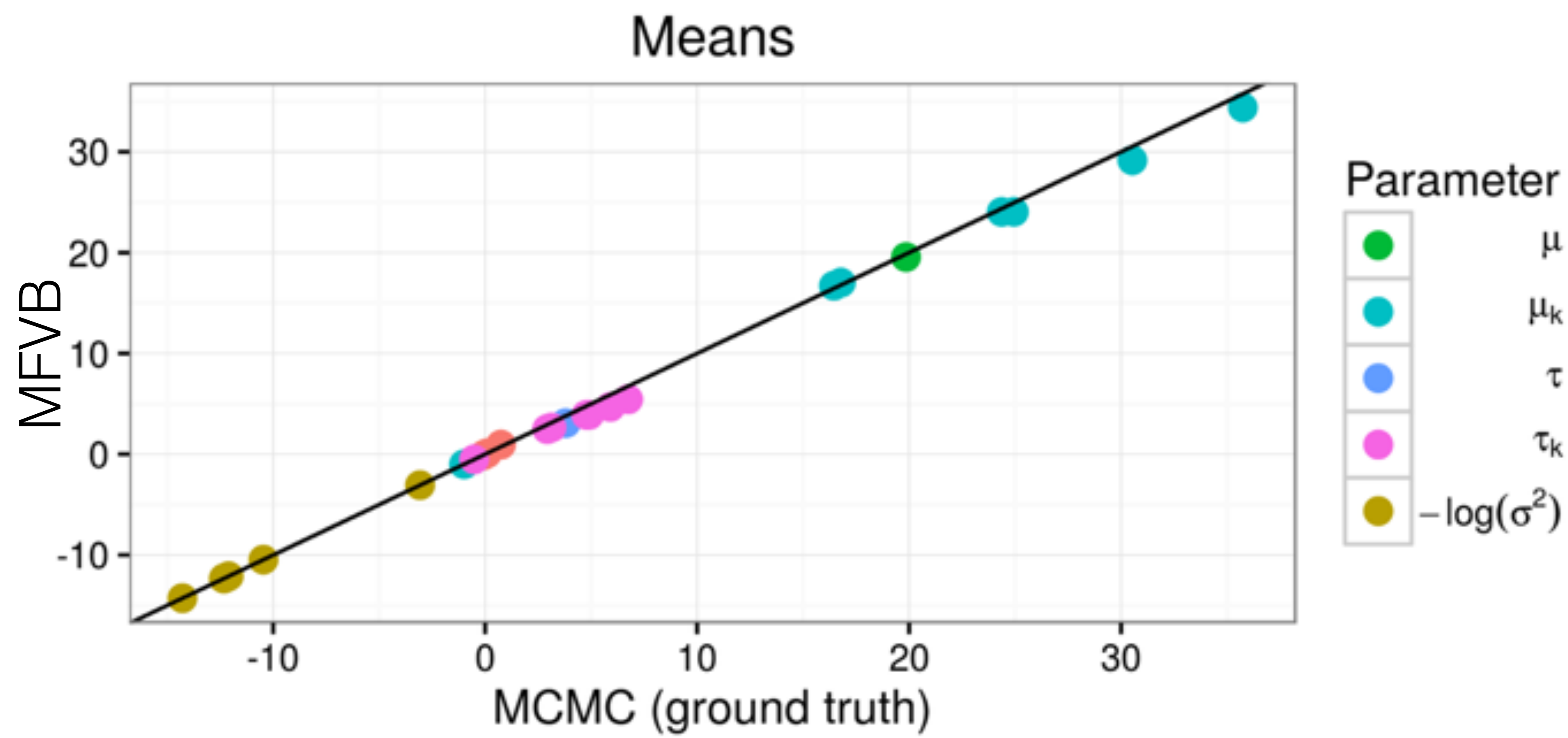
$$\sigma_k^{-2} \stackrel{iid}{\sim} \Gamma(a, b) \qquad C \sim \mathrm{Sep\&LKJ}(\eta, c, d)$$

Microcredit Experiment

Microcredit Experiment

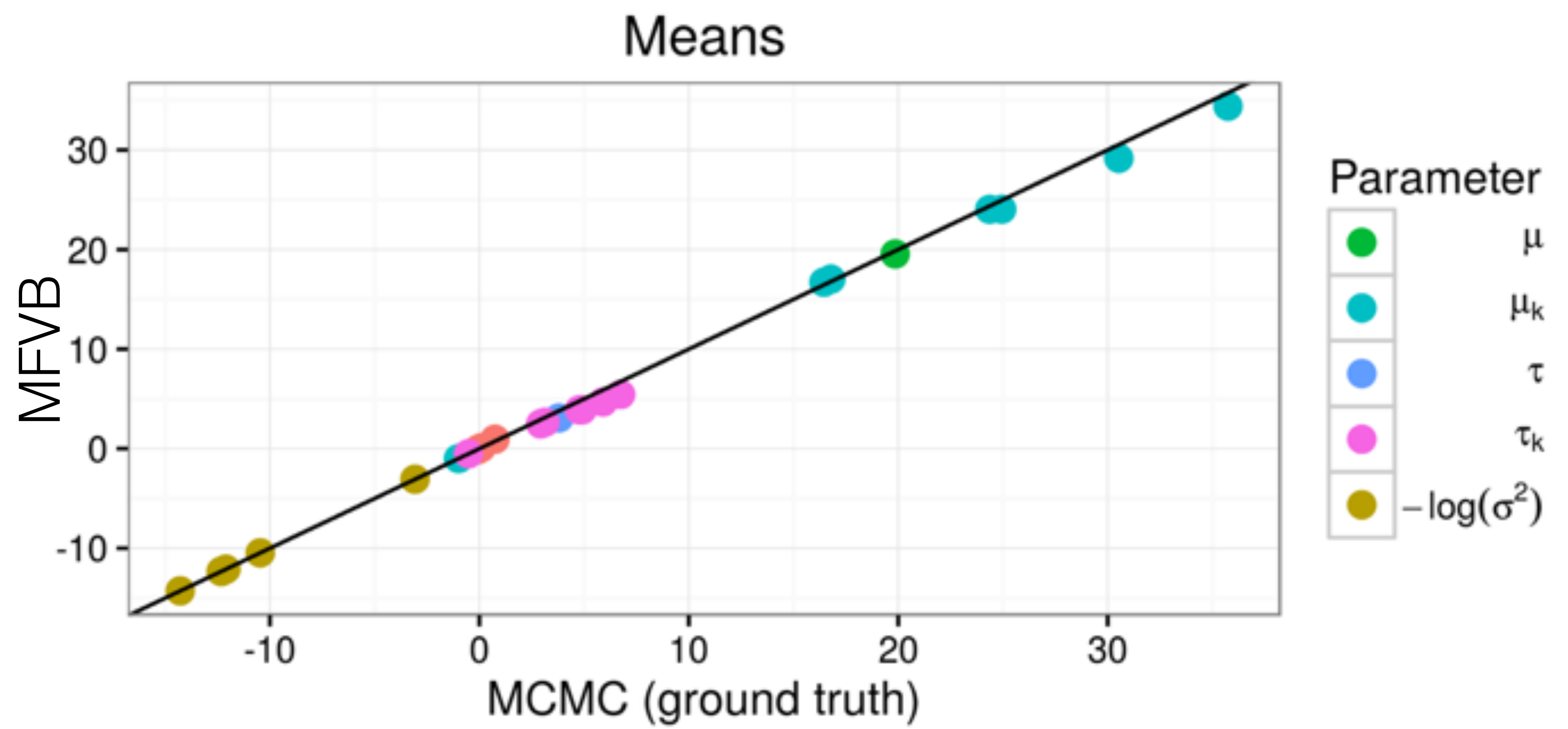


Microcredit Experiment



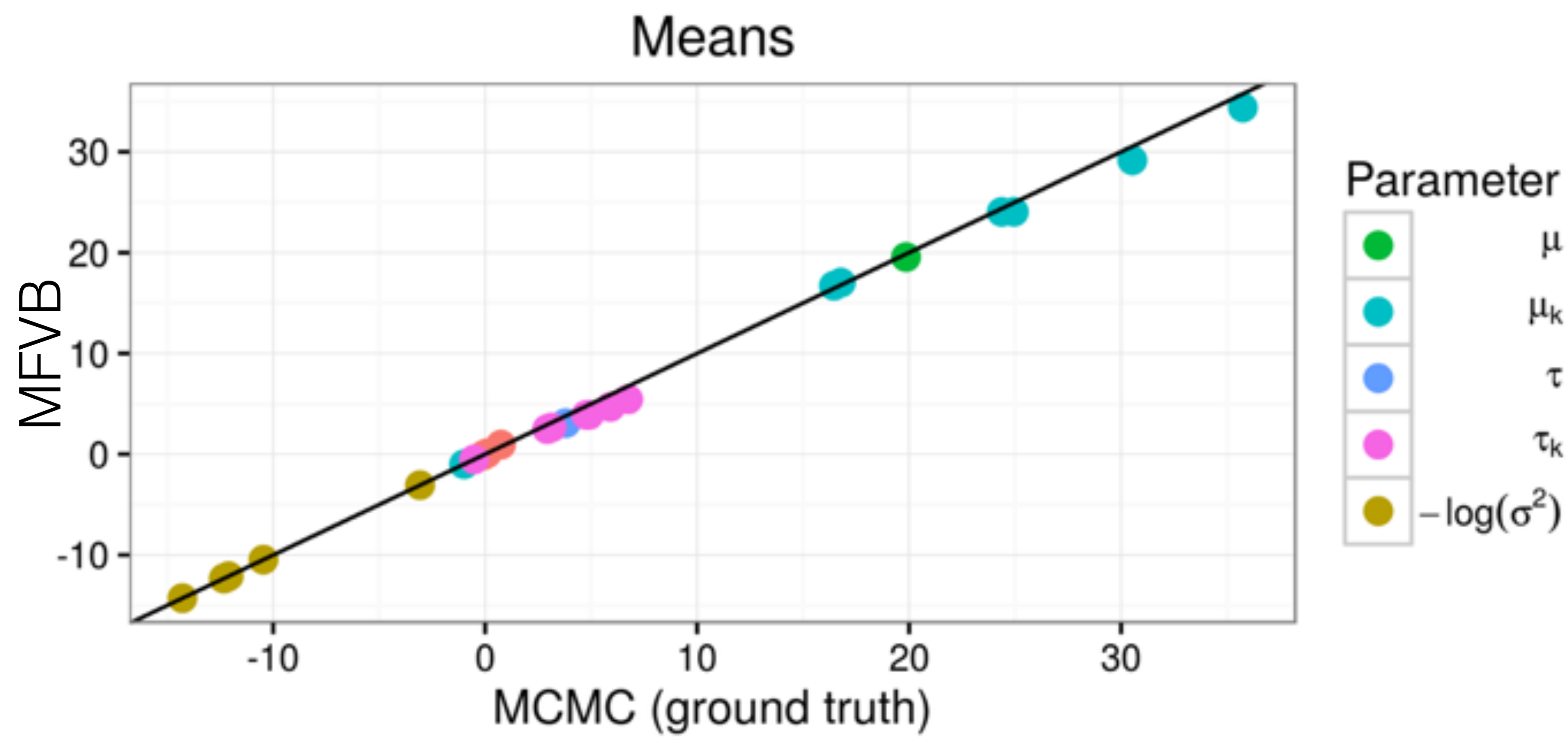
Microcredit Experiment

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



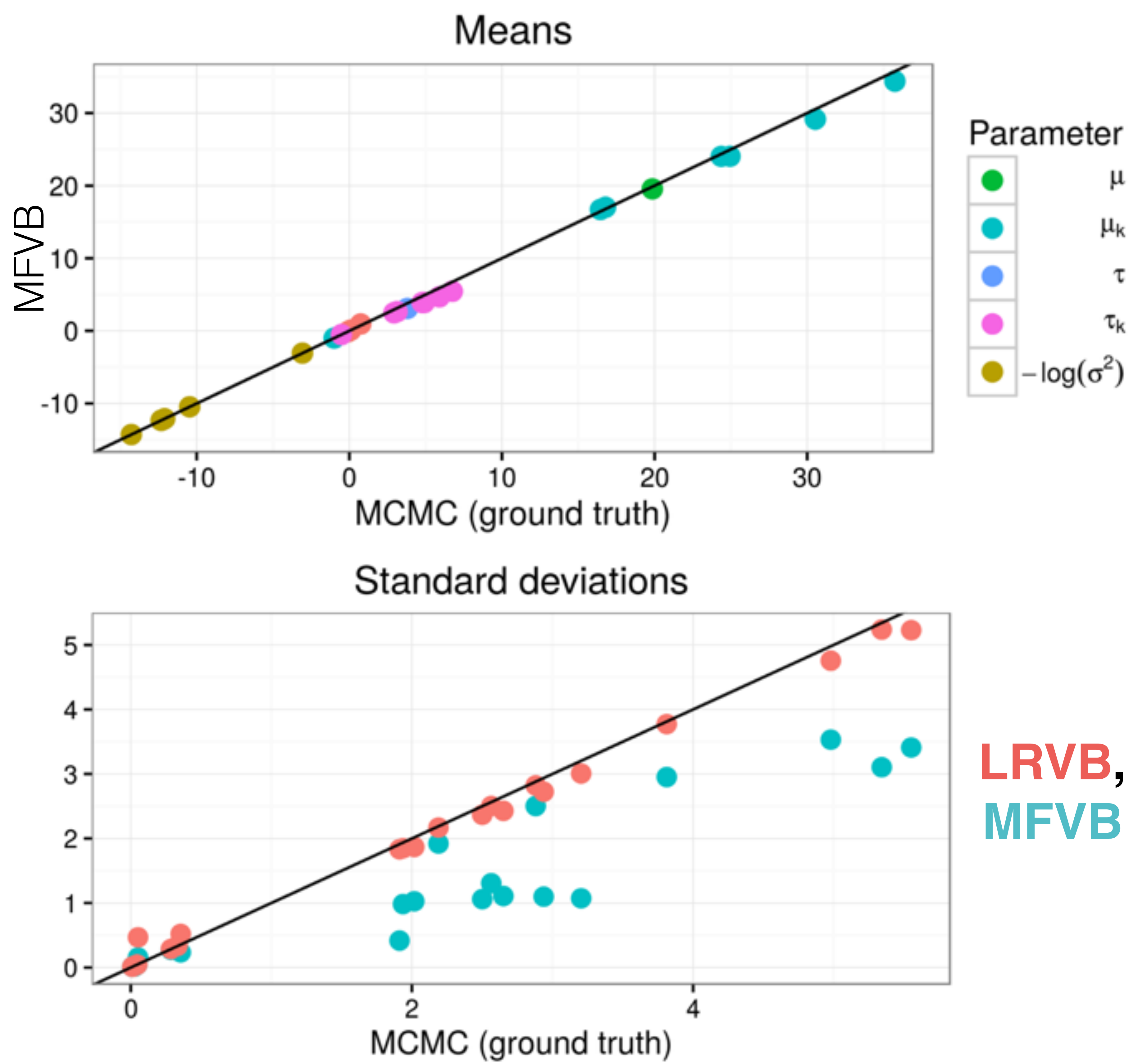
Microcredit Experiment

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



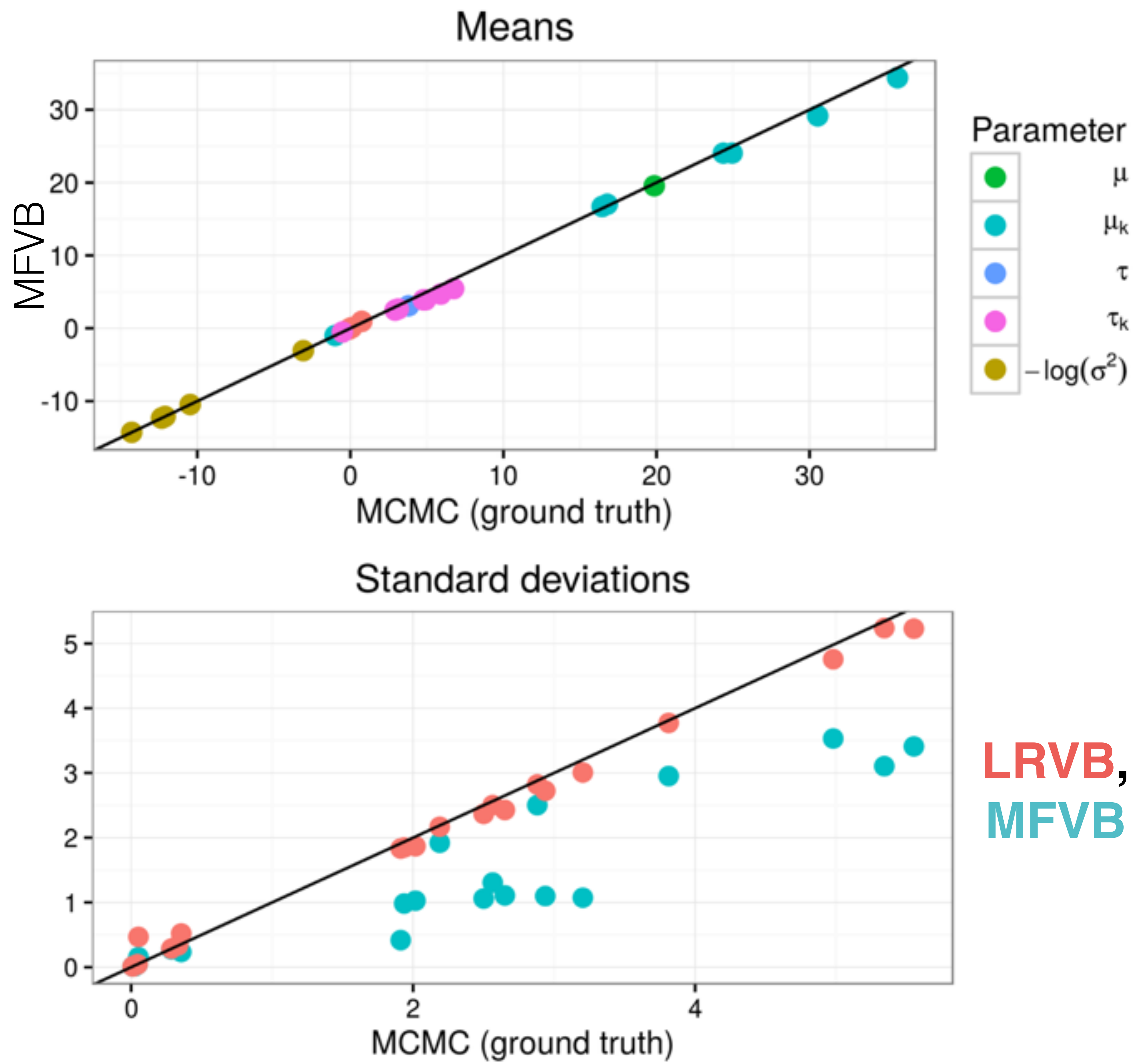
Microcredit Experiment

- *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**
- Many other models and data sets: Mixture models, generalized linear mixed models, etc



Robustness quantification

- Variational Bayes as an alternative to MCMC
- Challenges of VB
- Accurate uncertainties from VB
- Accurate robustness quantification from VB
- Big idea: derivatives/perturbations are easy in VB

Robustness quantification

- Variational Bayes as an alternative to MCMC
- Challenges of VB
- Accurate uncertainties from VB
- Accurate robustness quantification from VB
- Big idea: derivatives/perturbations are easy in VB

Robustness quantification

- Bayes Theorem

$$p(\theta|x)$$

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

Robustness quantification

- Bayes Theorem

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

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

Robustness quantification

- Bayes Theorem

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

Robustness quantification

- Bayes Theorem

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

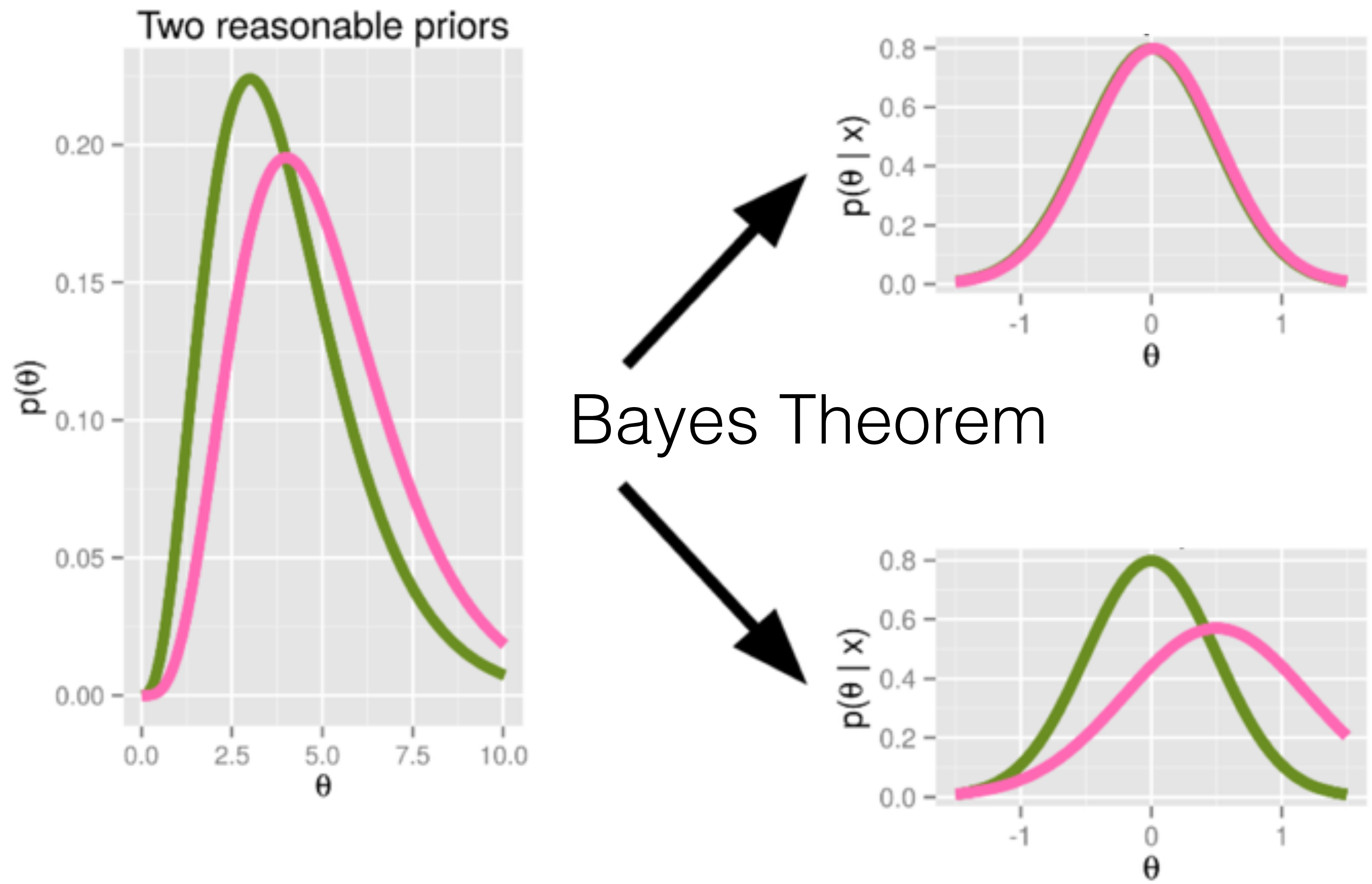
- Sensitivity

Robustness quantification

- Bayes Theorem

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

- Sensitivity



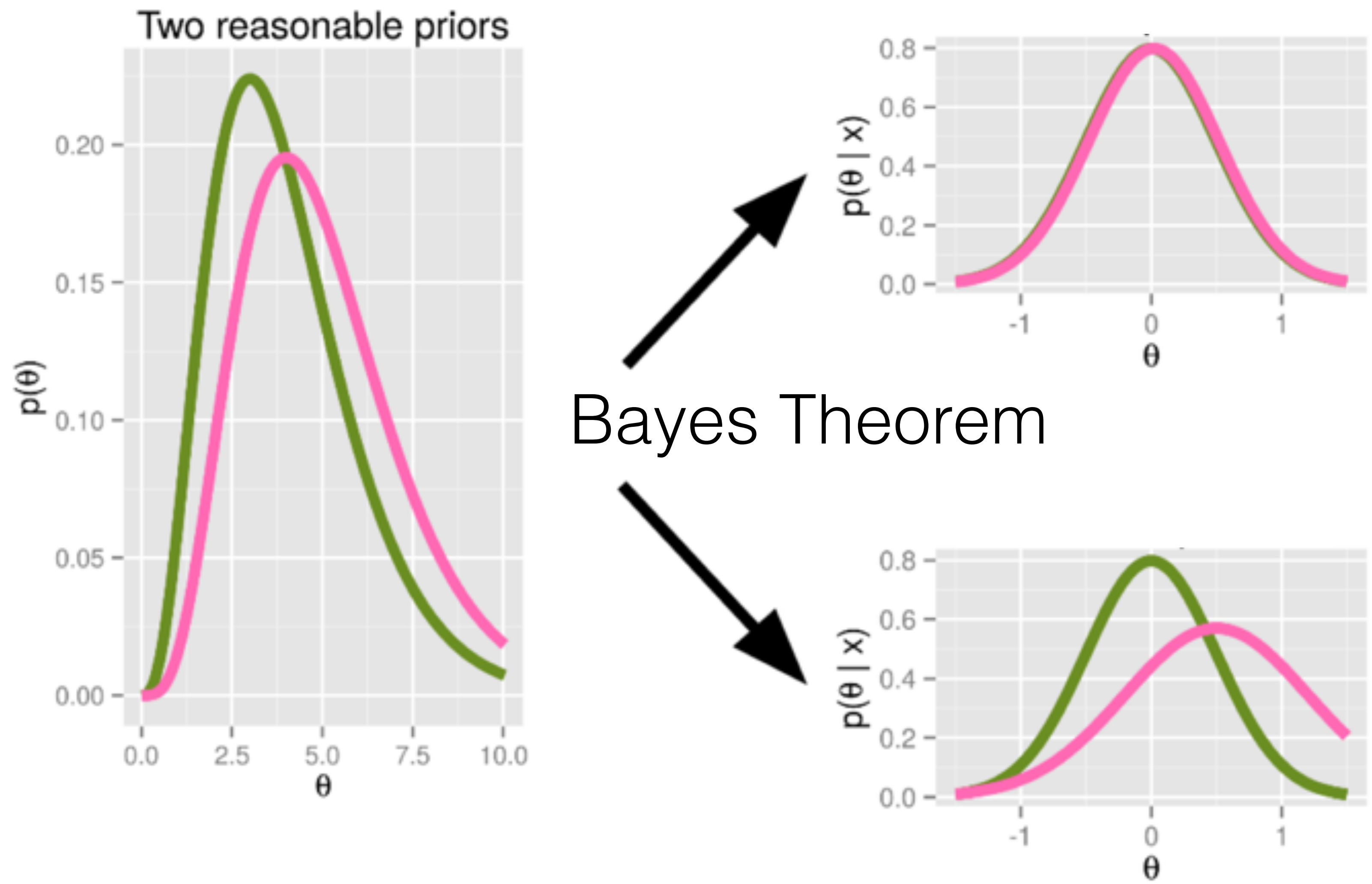
Robustness quantification

- Bayes Theorem

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

- Sensitivity

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



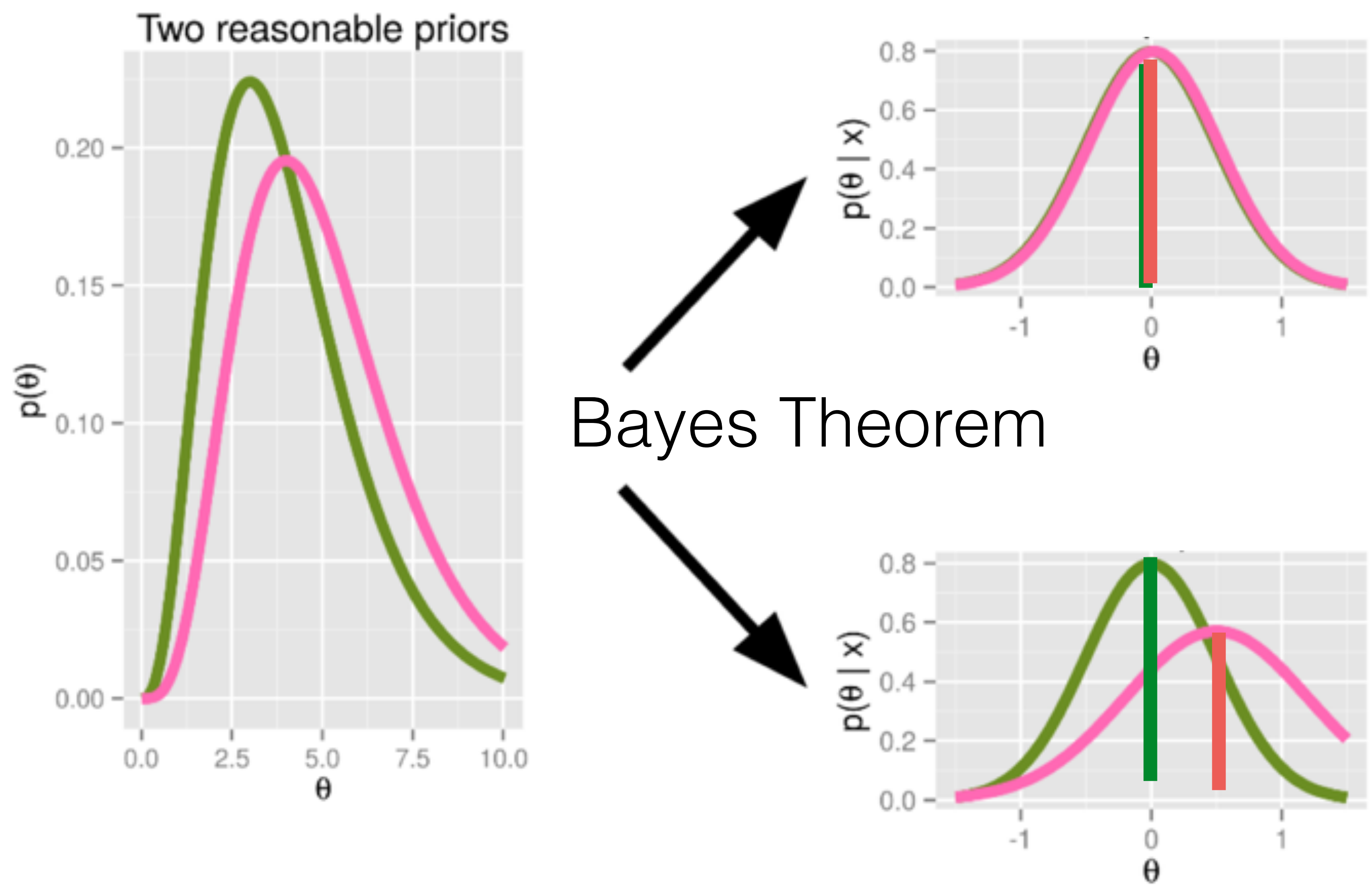
Robustness quantification

- Bayes Theorem

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

- Sensitivity

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



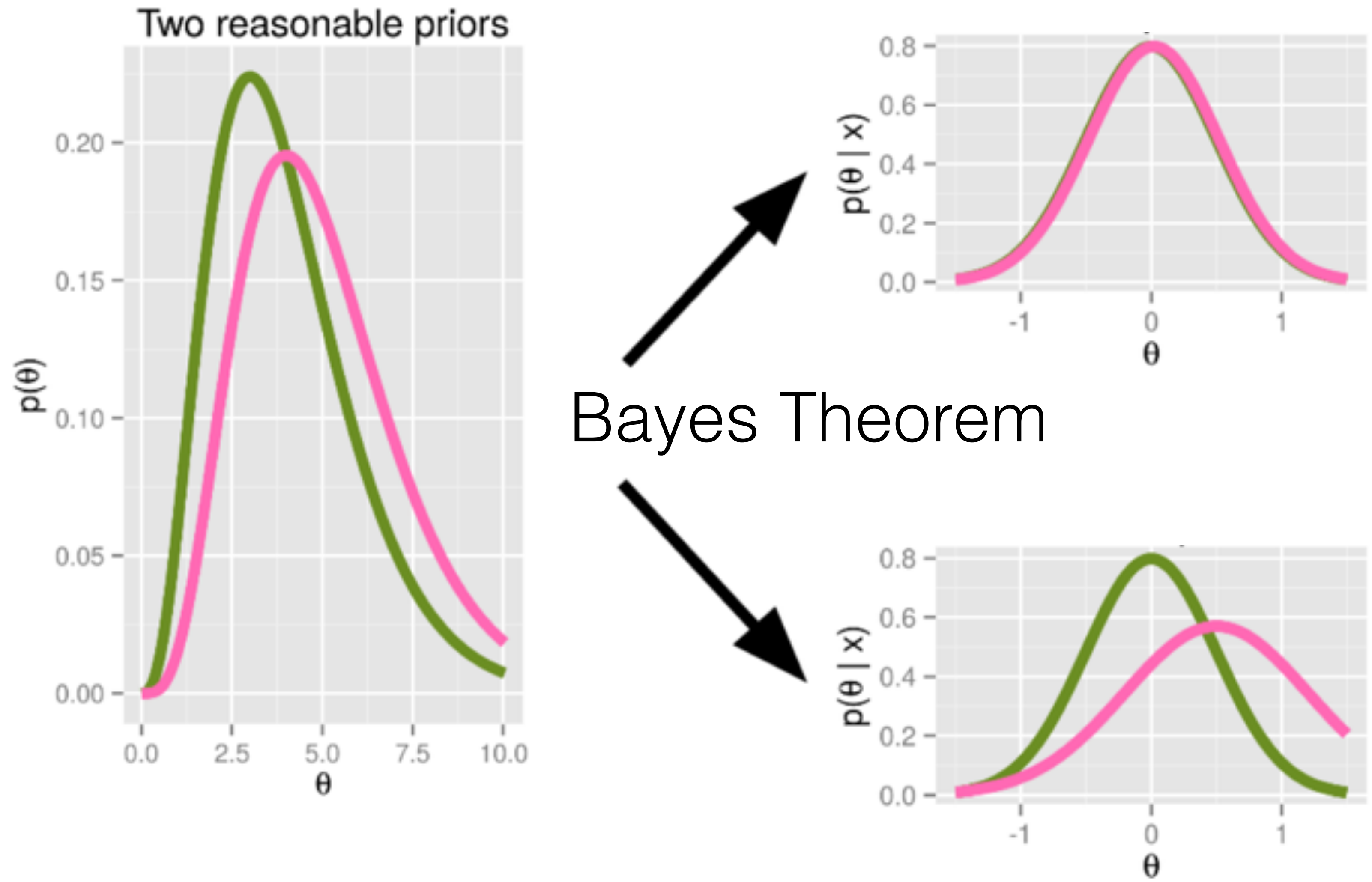
Robustness quantification

- Bayes Theorem

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

- Sensitivity

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



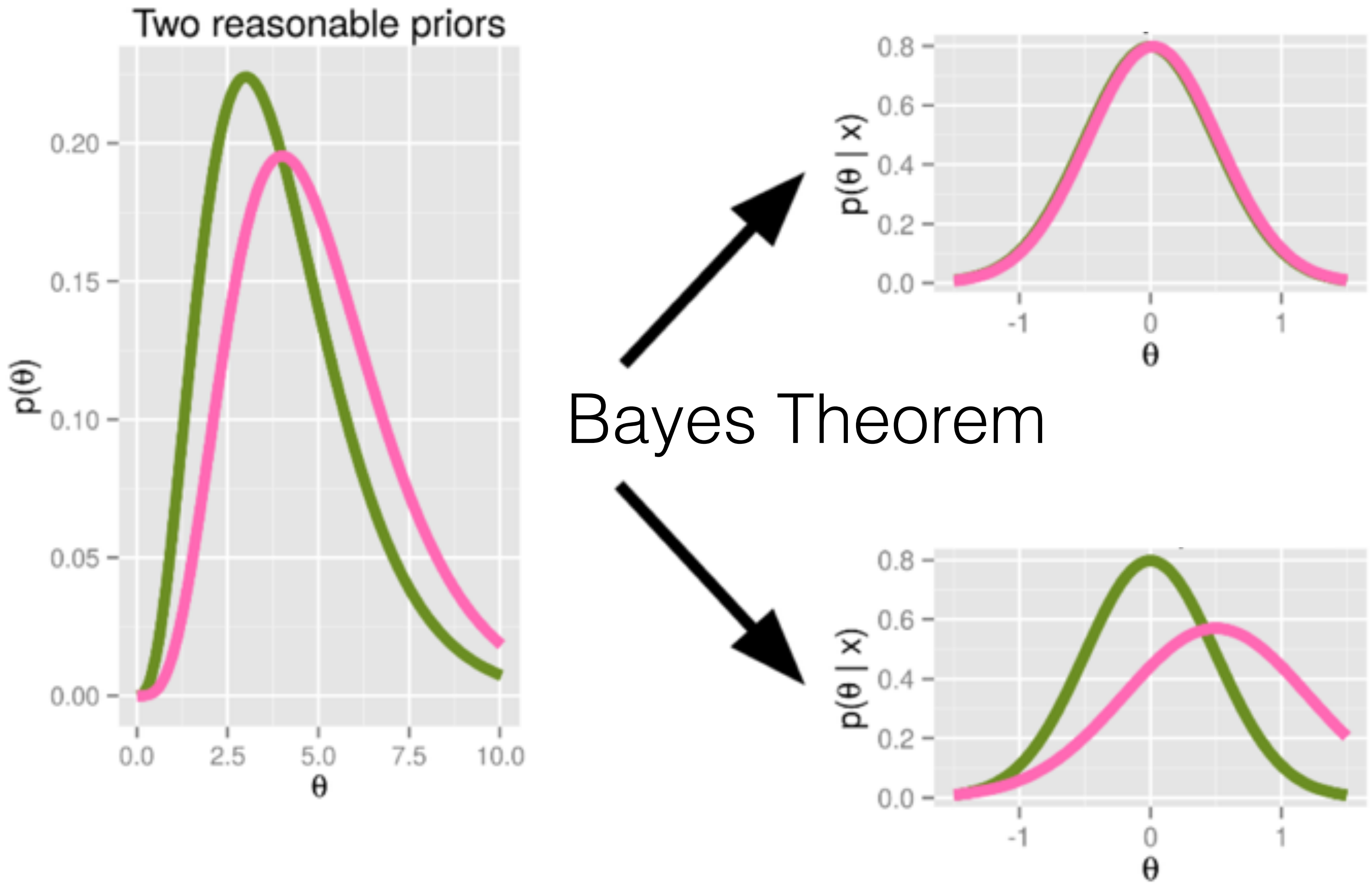
Robustness quantification

- Bayes Theorem

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

- Sensitivity

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



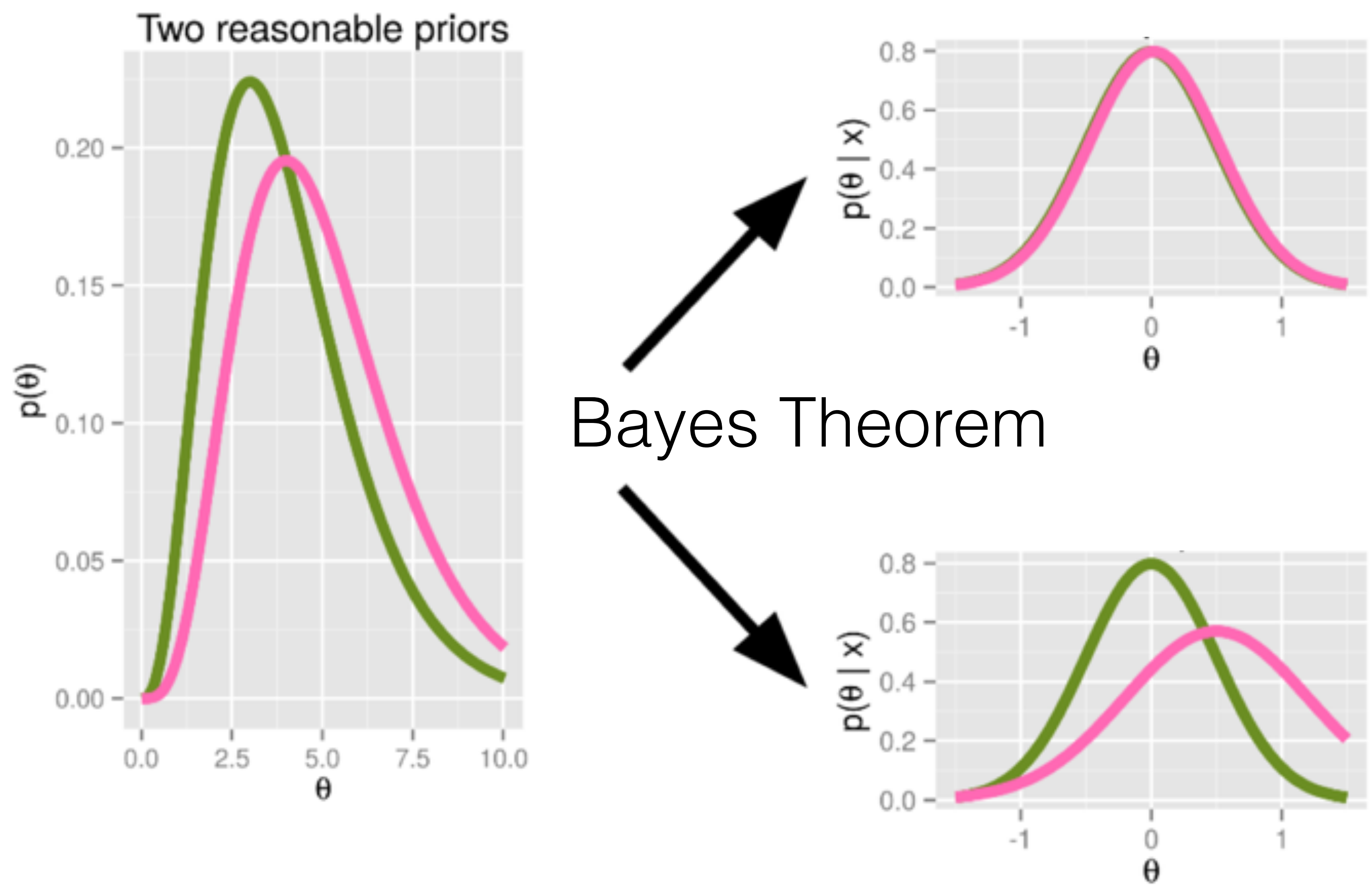
Robustness quantification

- Bayes Theorem

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

- Sensitivity

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



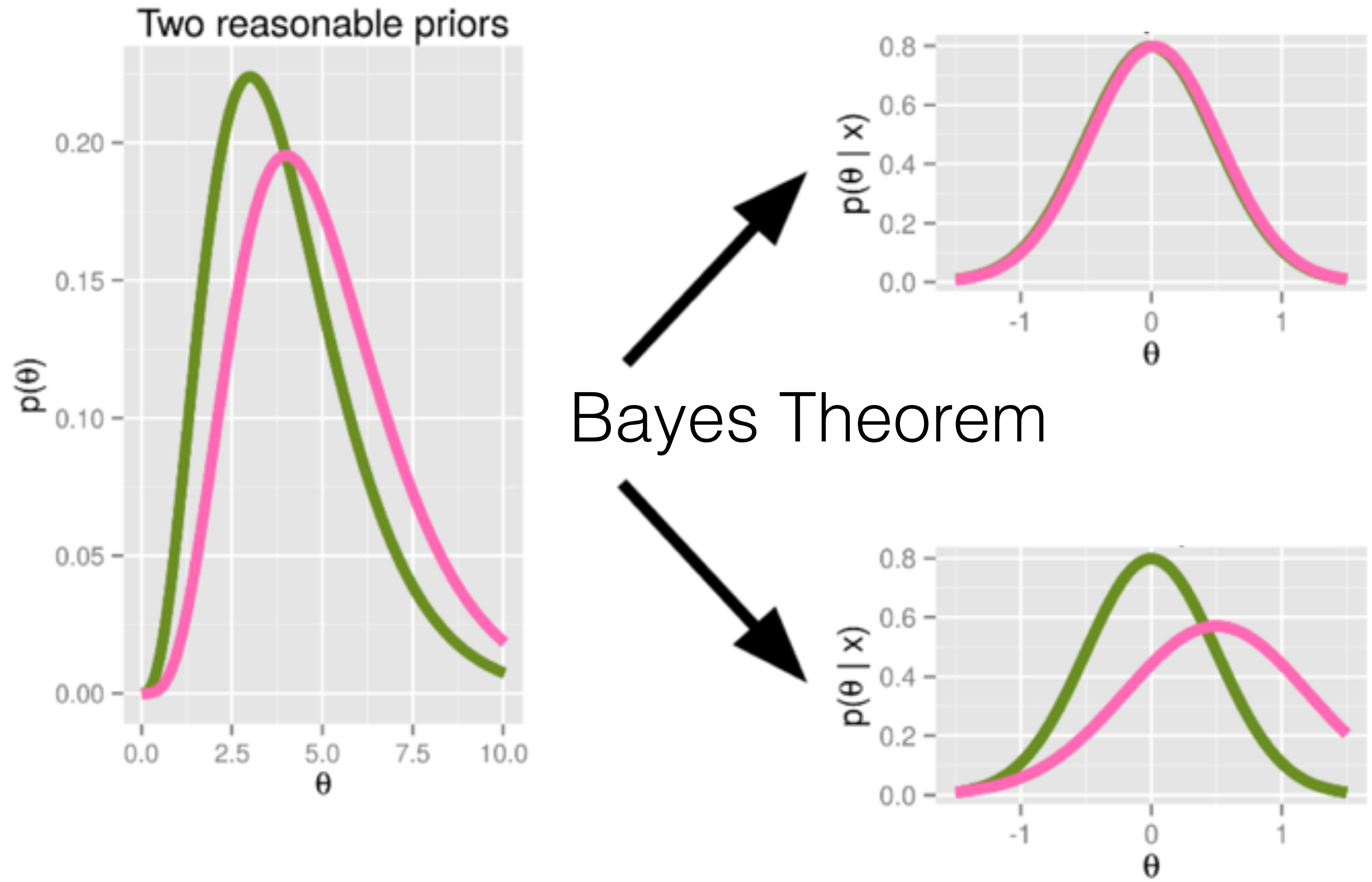
Robustness quantification

- Bayes Theorem

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

- Sensitivity

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

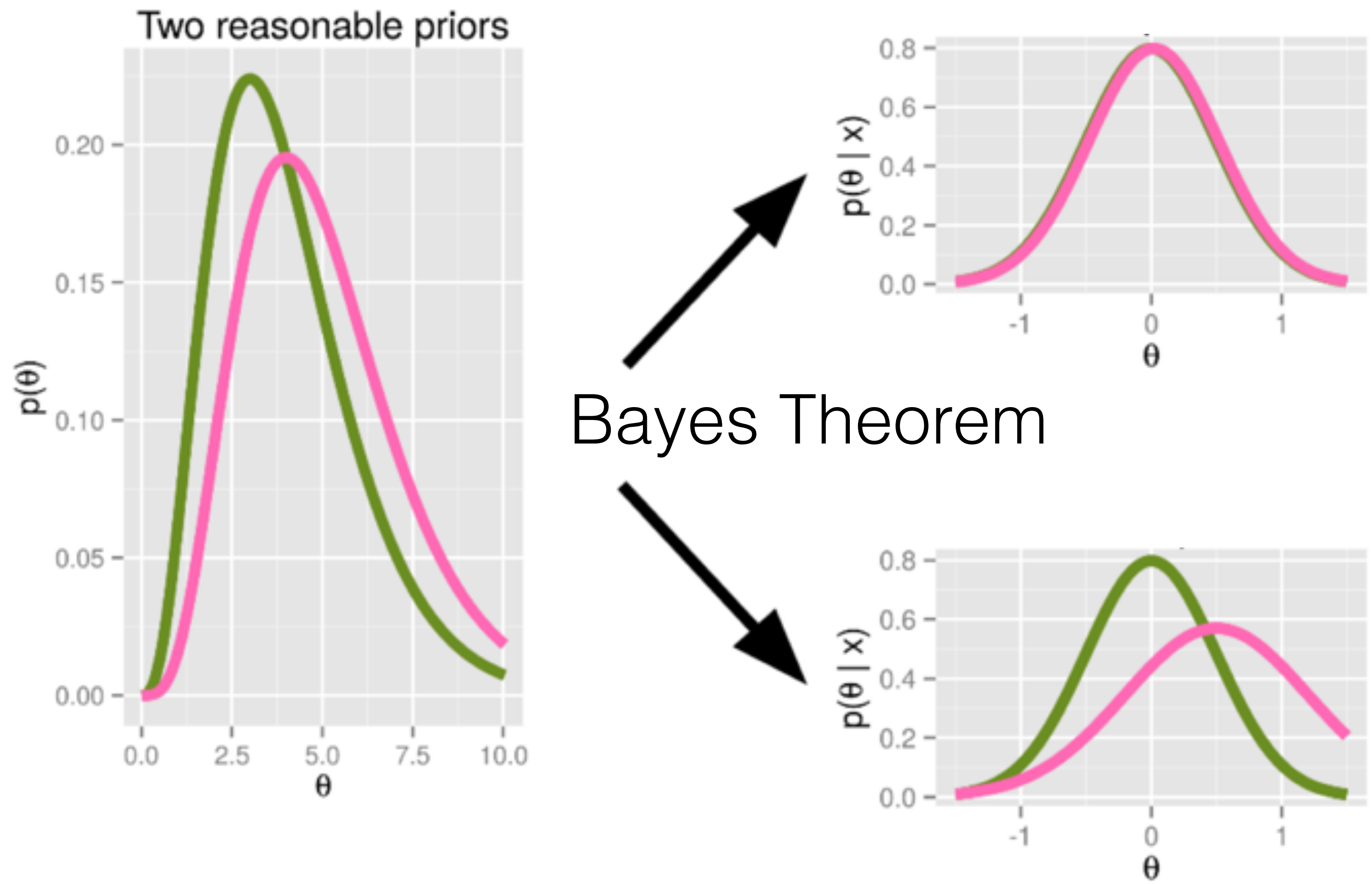


← LRVB estimator

Robustness quantification

- Bayes Theorem

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



- Sensitivity

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

← LRVB estimator

- When q^*_α in exponential family

Robustness quantification

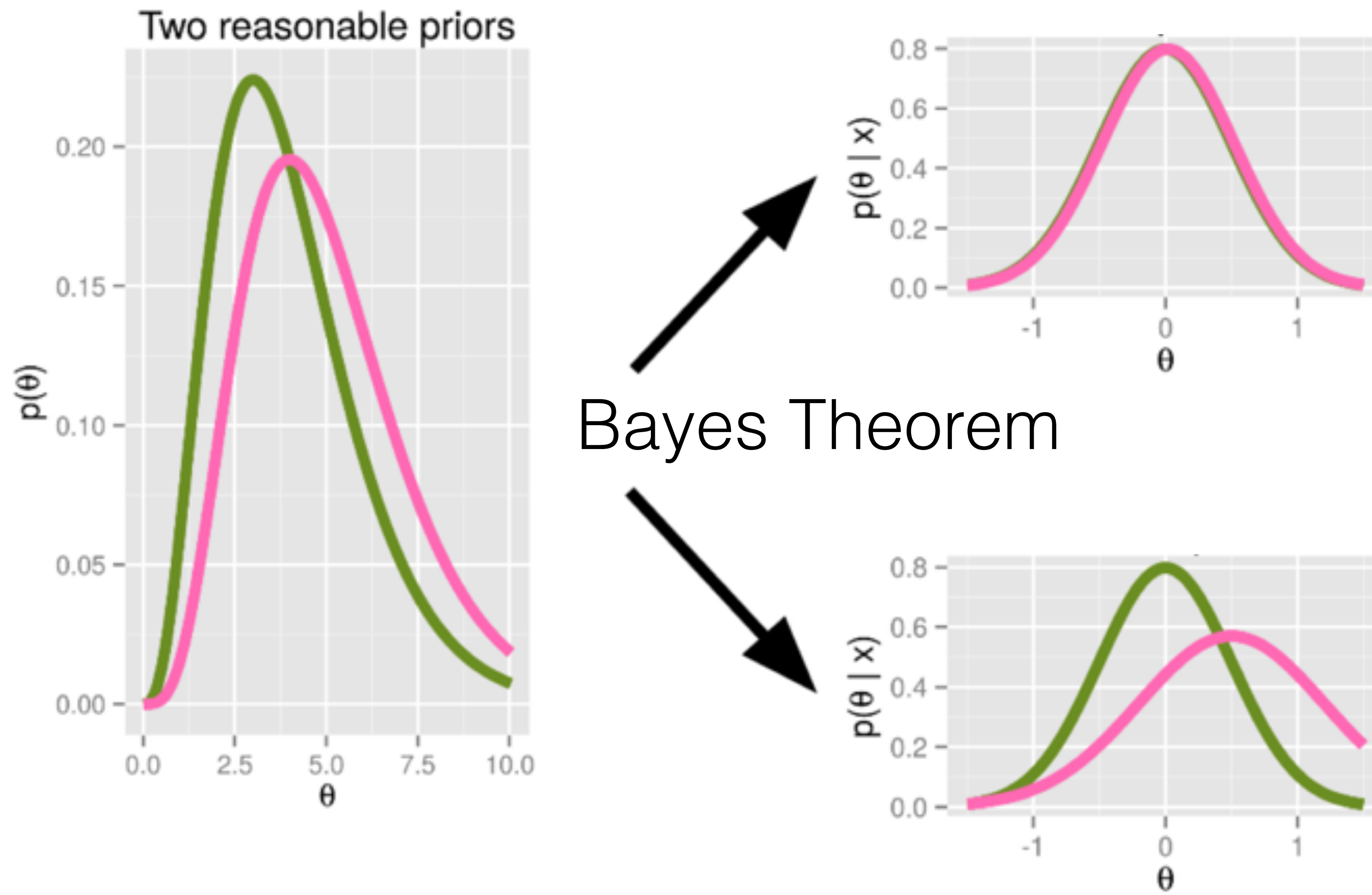
- Bayes Theorem

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

- Sensitivity

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

← LRVB estimator



- When q^*_α in exponential family

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

Microcredit Experiment

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

profit → 1 if microcredit →

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

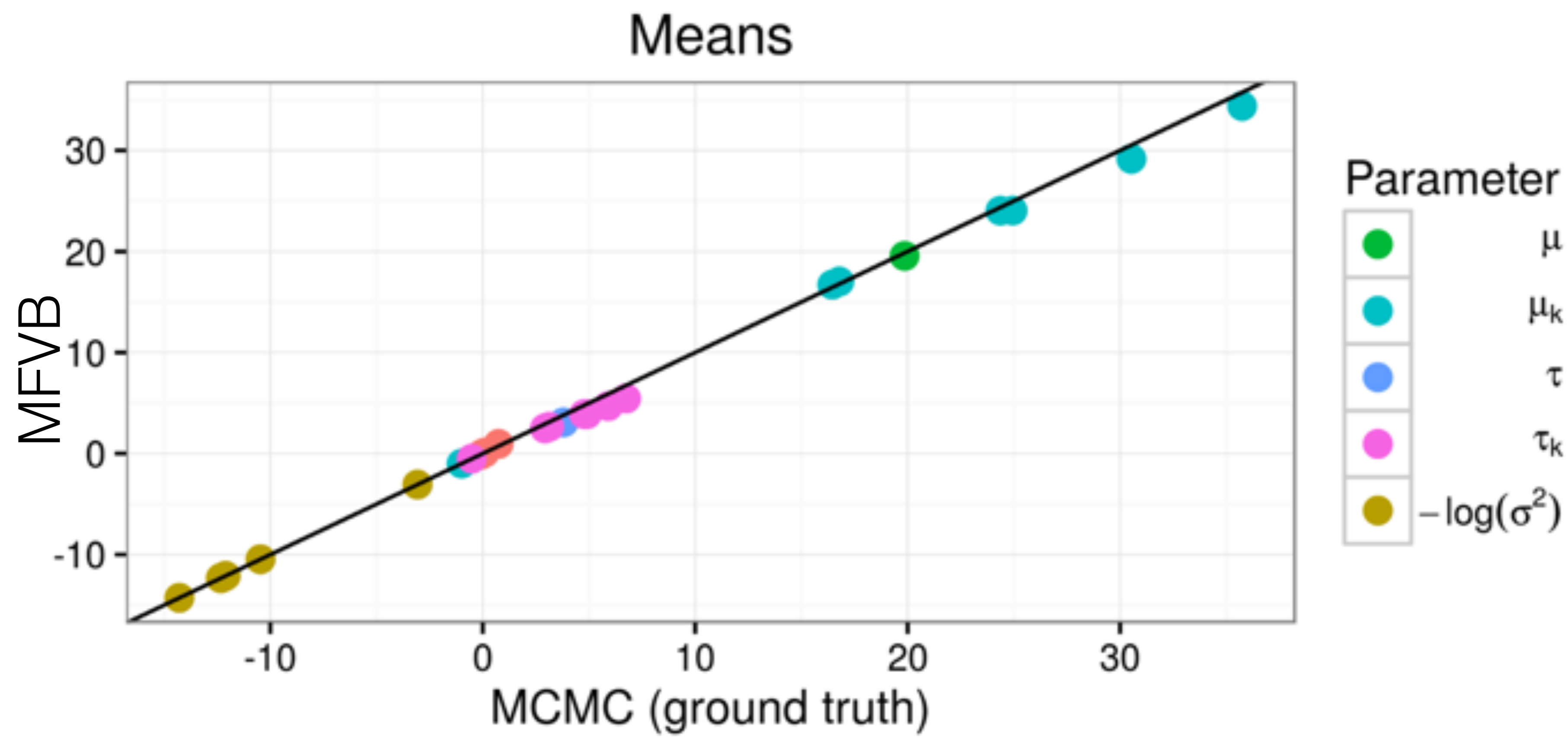
- Priors and hyperpriors:

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

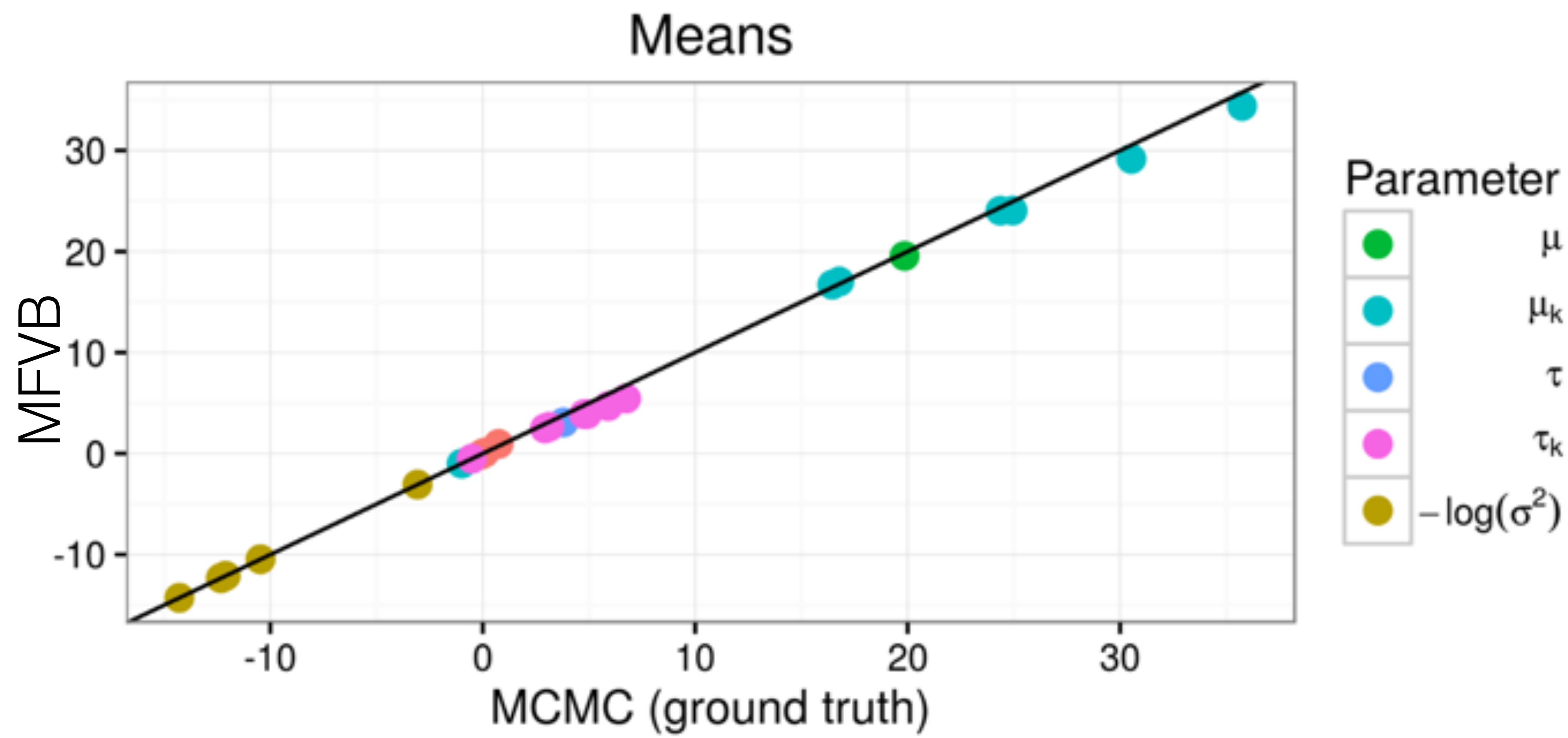
$$\sigma_k^{-2} \overset{iid}{\sim} \Gamma(a, b) \qquad C \sim \text{Sep\&LKJ}(\eta, c, d)$$

Microcredit Experiment

Microcredit Experiment



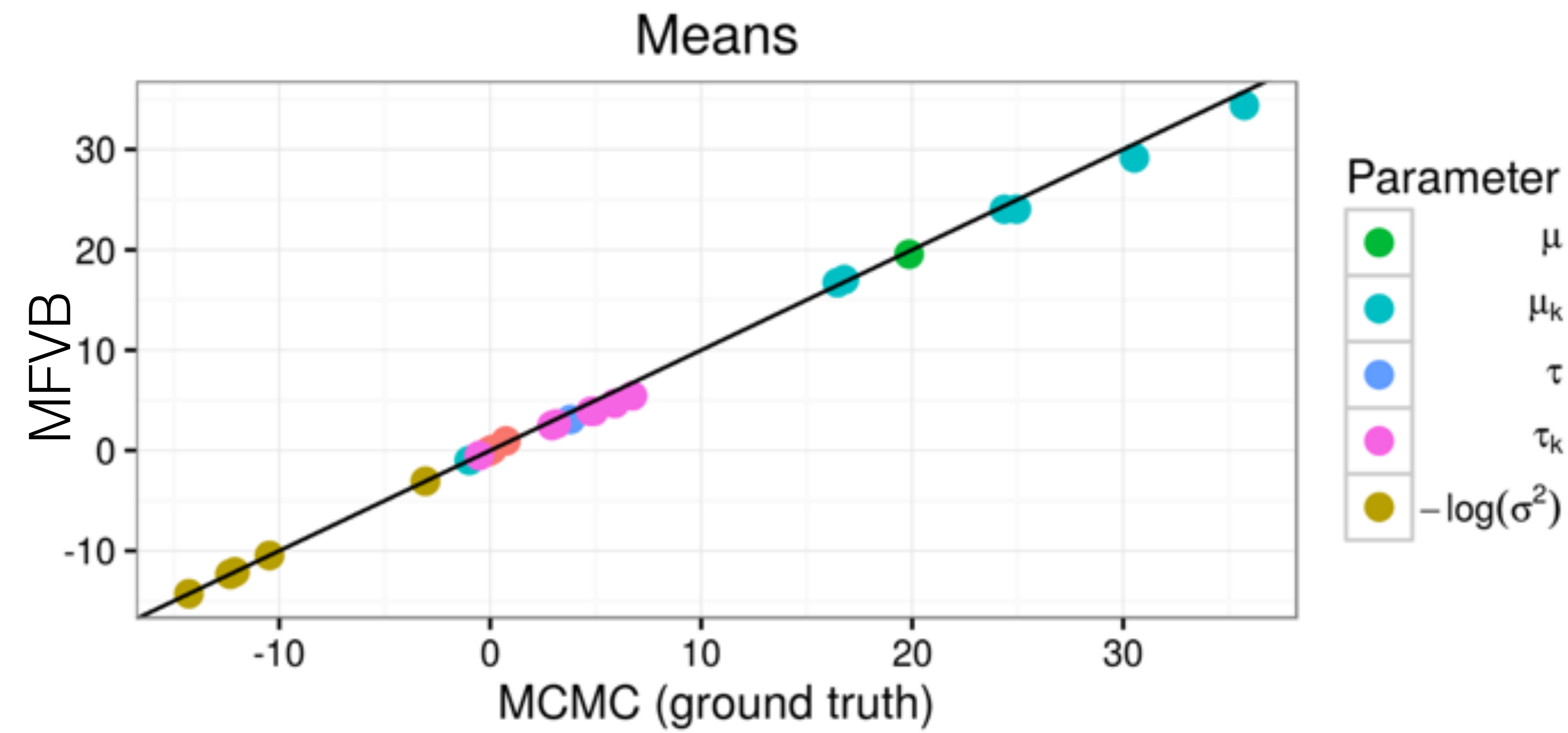
Microcredit Experiment

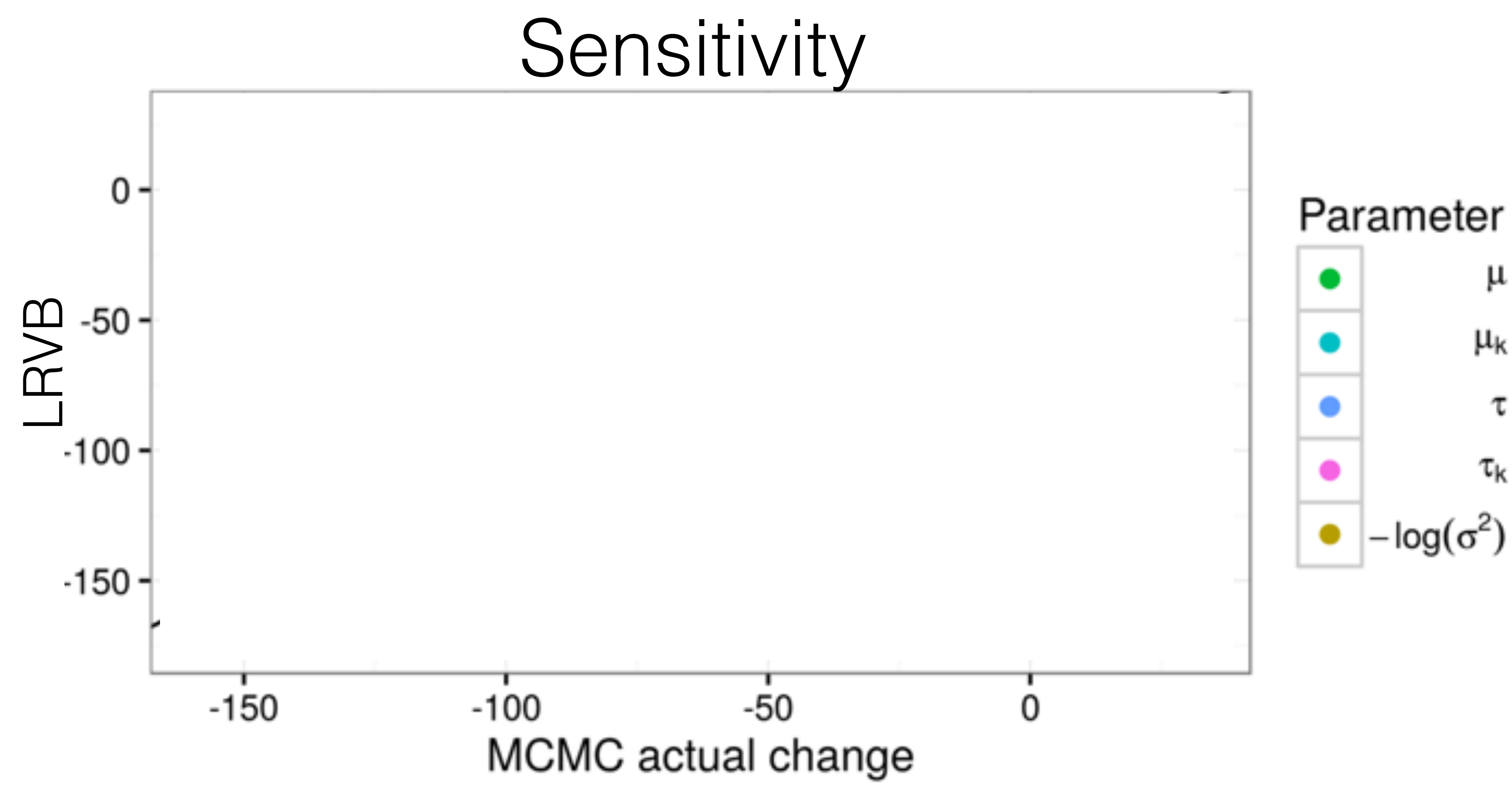


- Perturb Λ_{11}: 0.03 ➔ 0.04

Microcredit Experiment

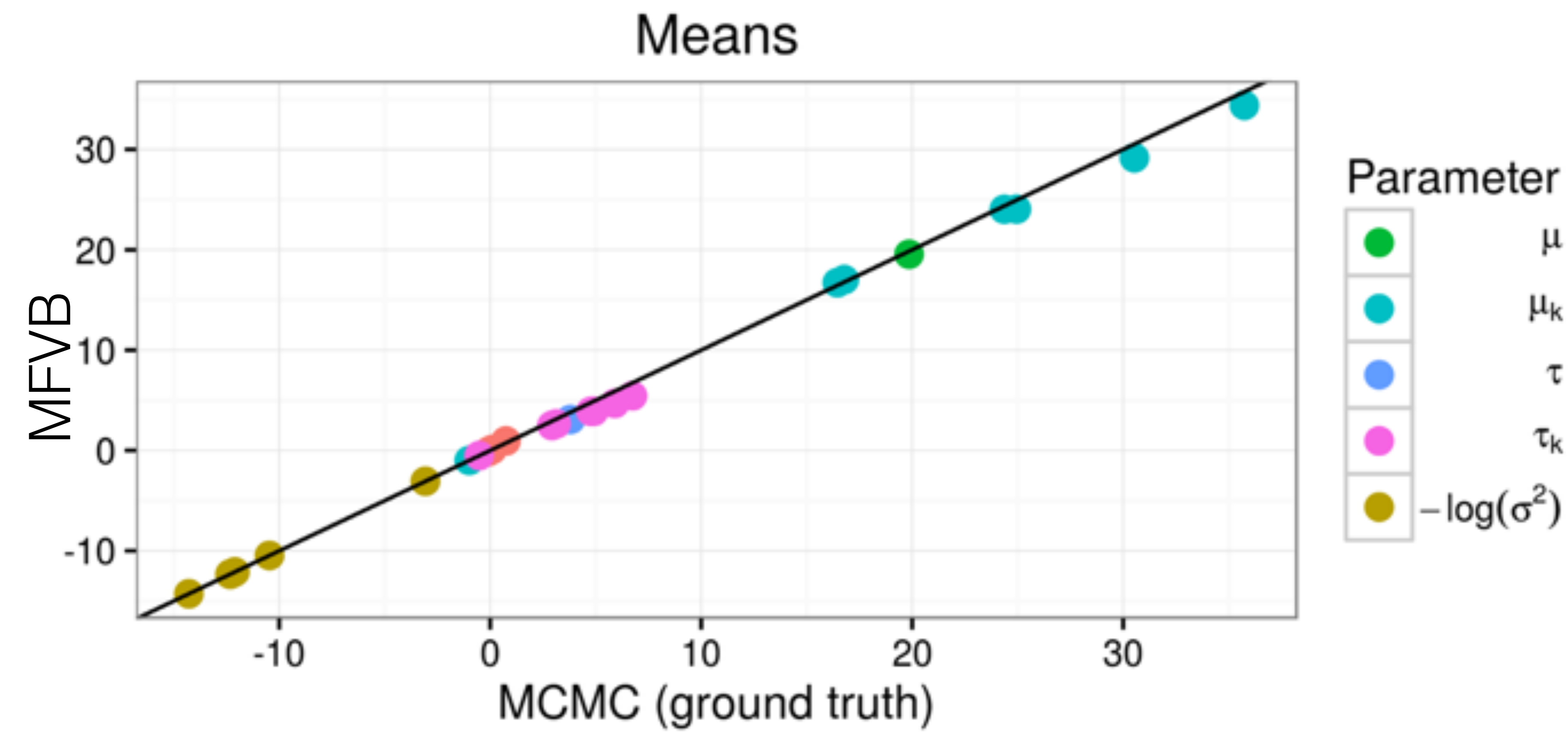
- Perturb Λ_{11}: 0.03 ➔ 0.04

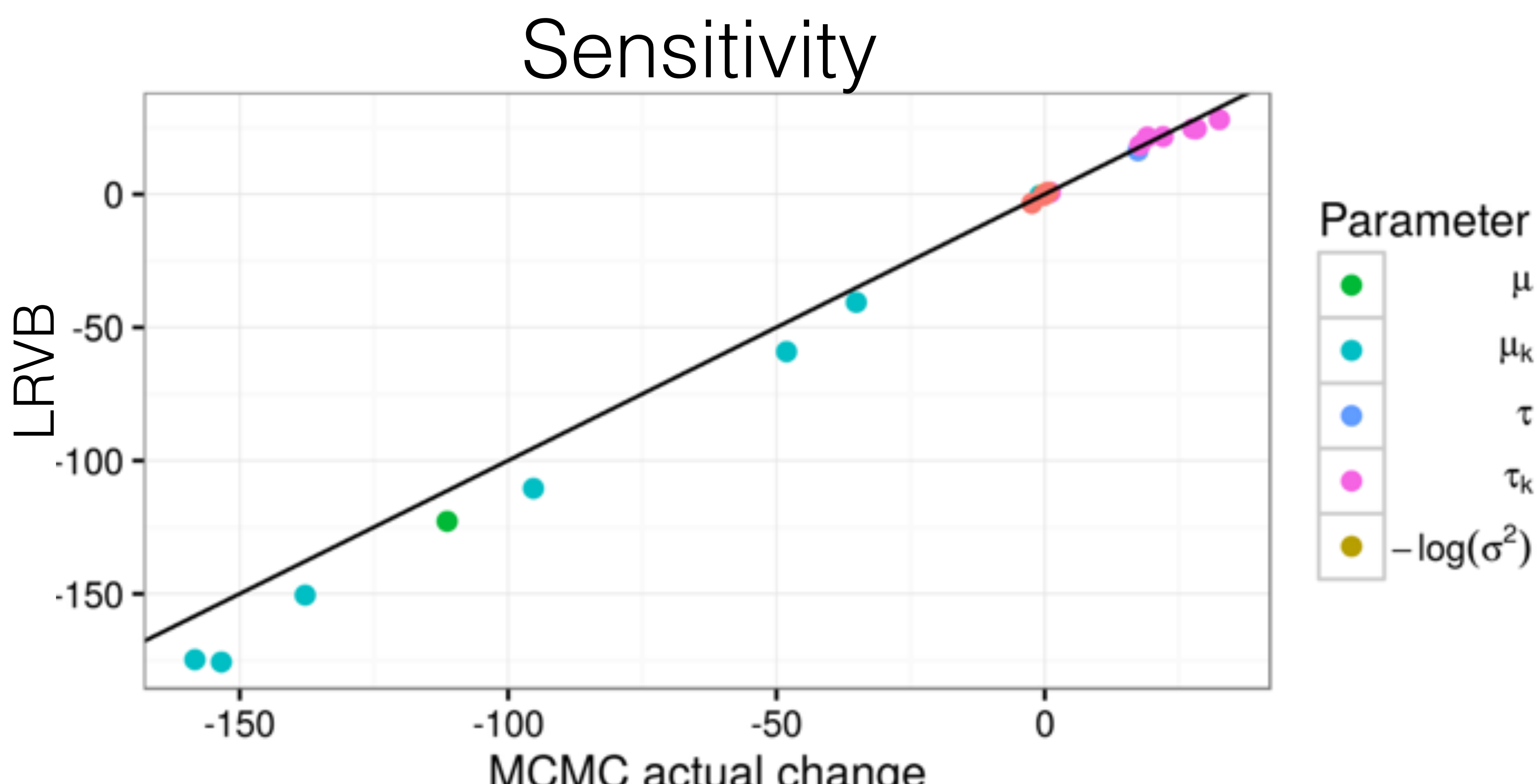




Microcredit Experiment

- Perturb Λ_{11}: 0.03 ➔ 0.04





Microcredit Experiment

Microcredit Experiment

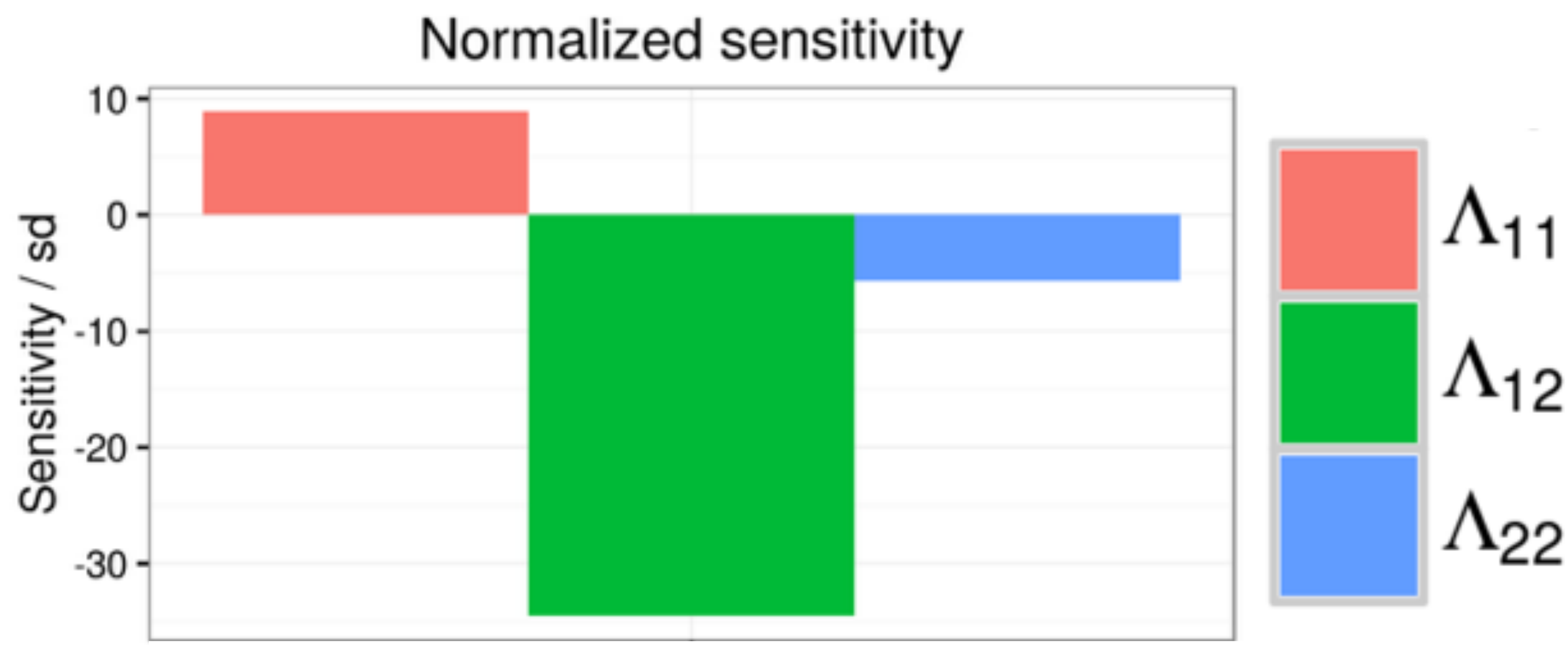
- Sensitivity of the expected microcredit effect (τ)

Microcredit Experiment

- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of standard deviations in τ

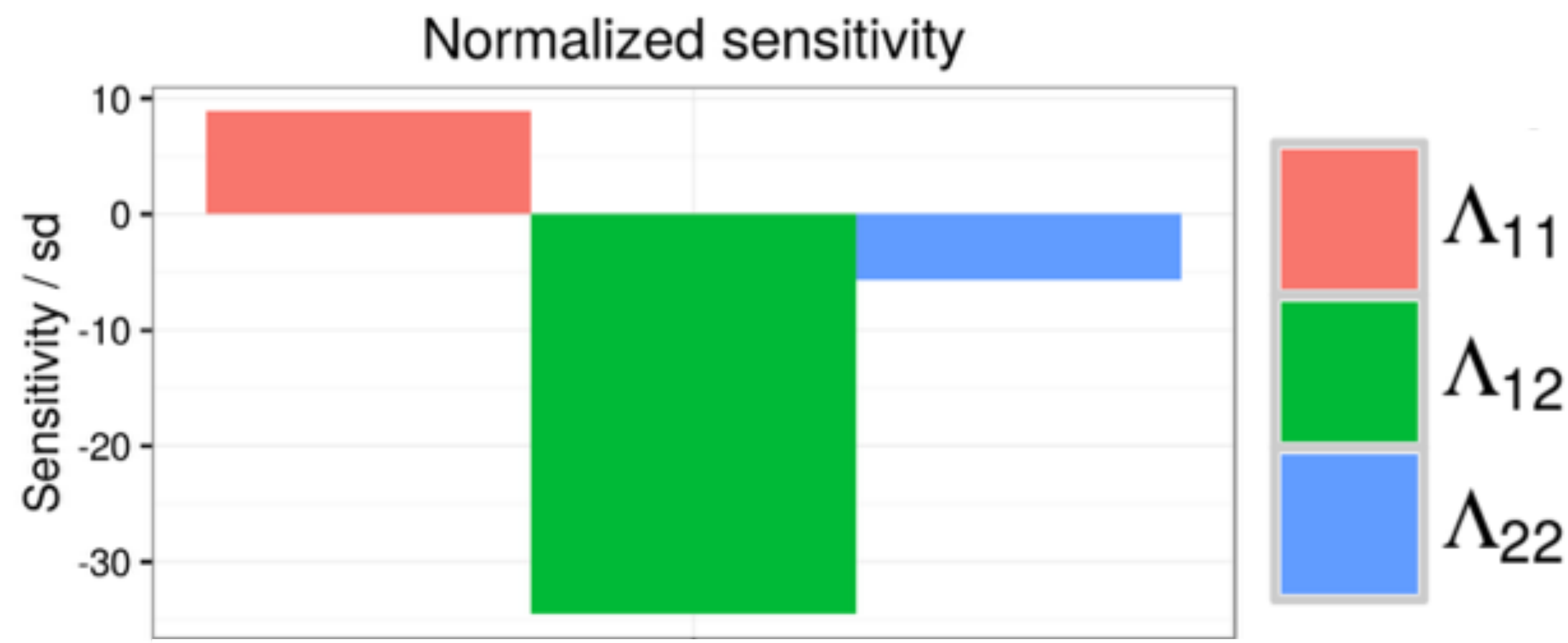
Microcredit Experiment

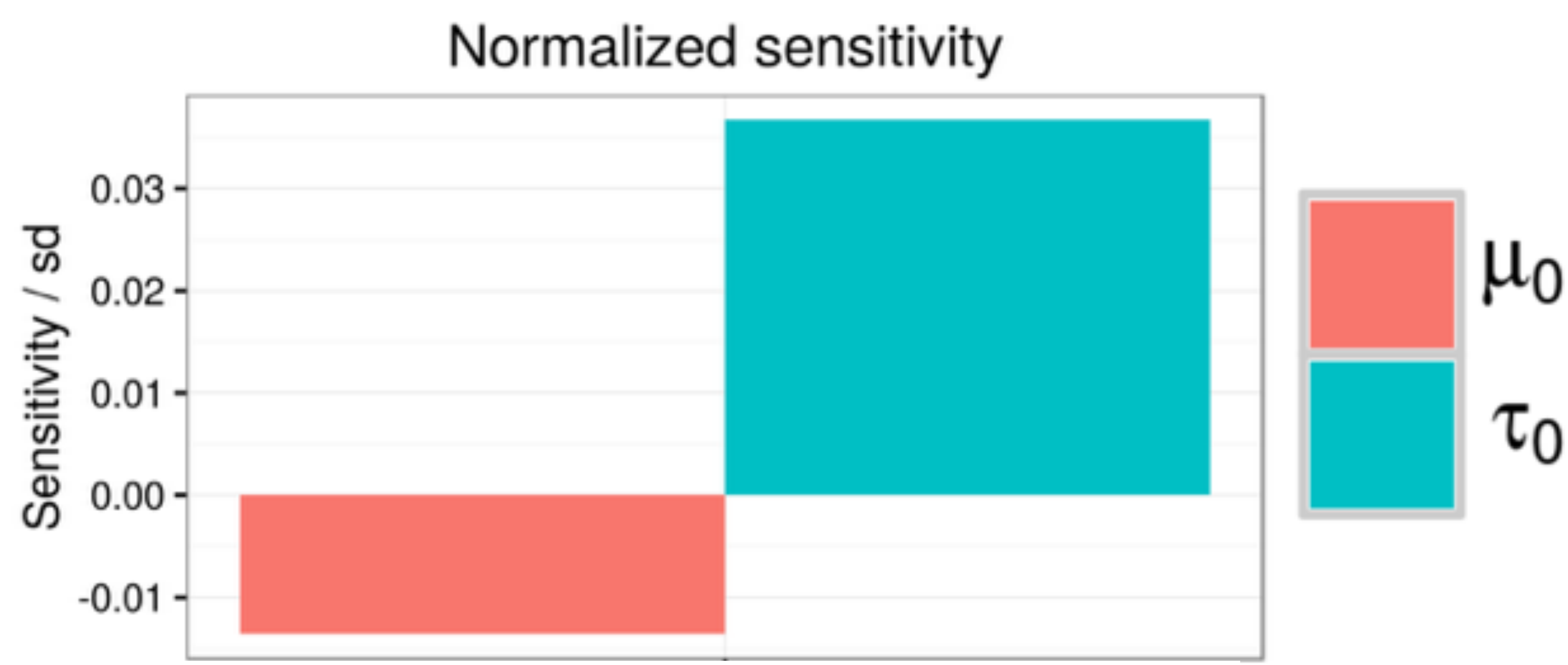
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of standard deviations in τ



Microcredit Experiment

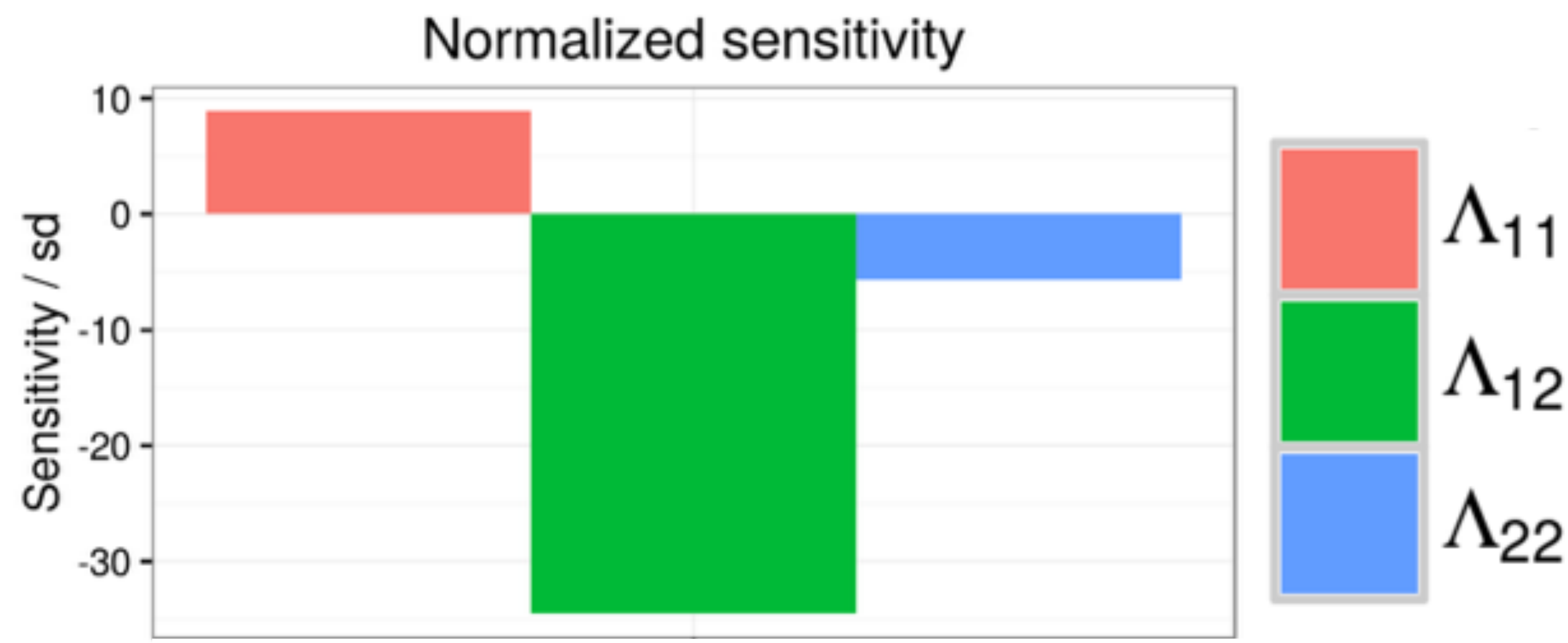
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of standard deviations in τ

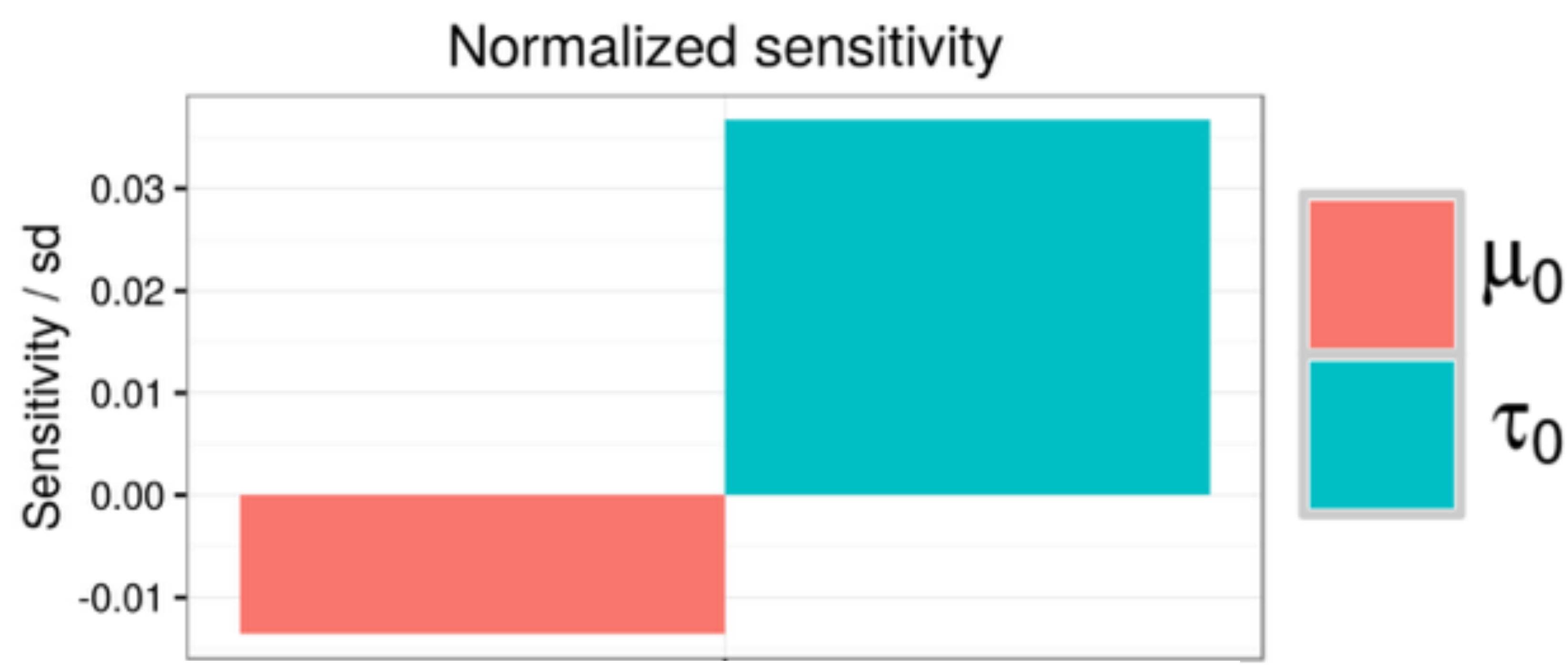




Microcredit Experiment

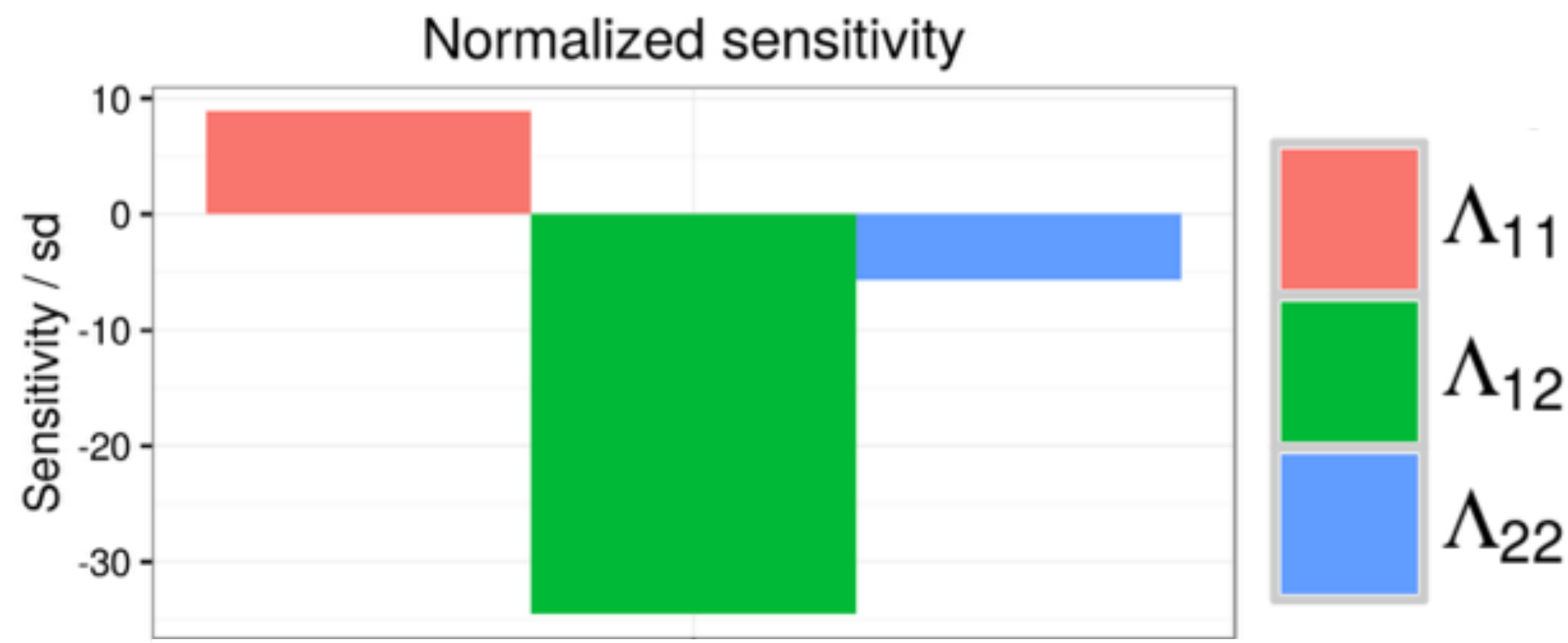
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of standard deviations in τ
- E.g.

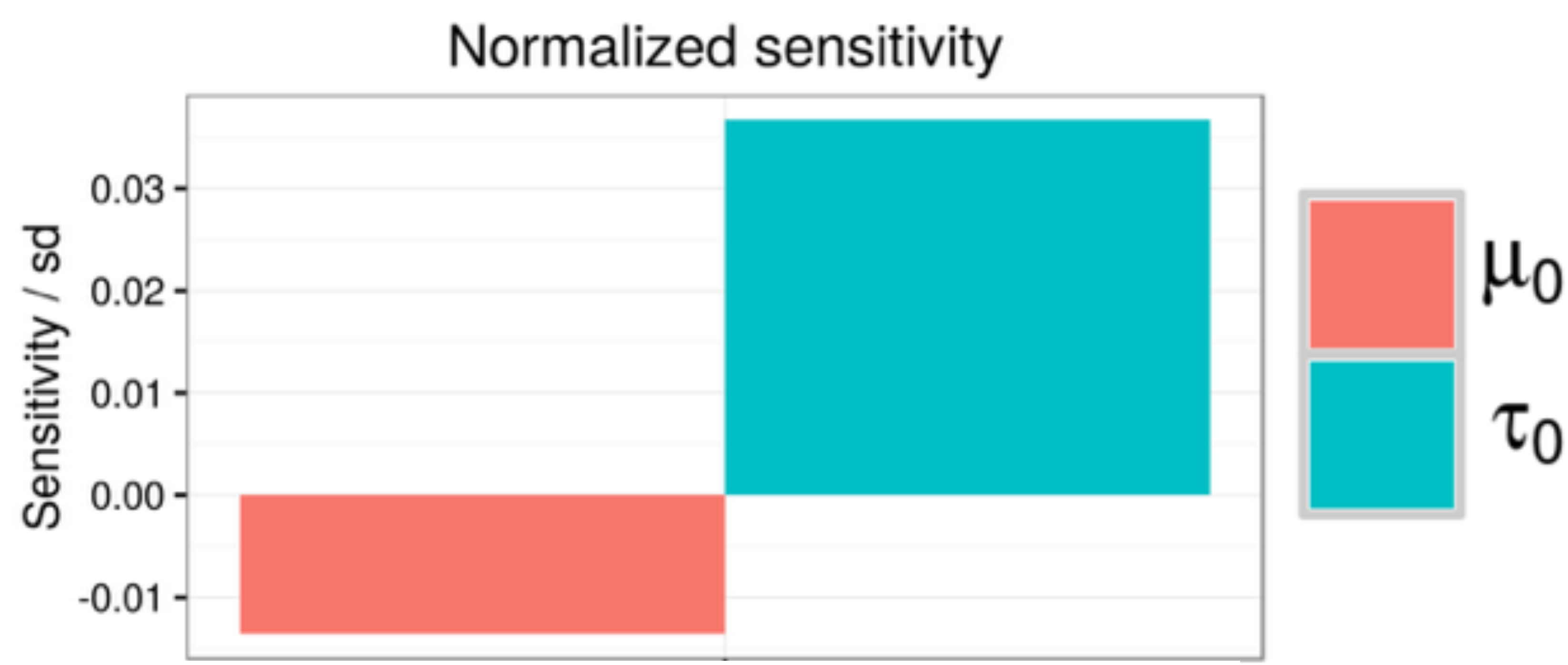




Microcredit Experiment

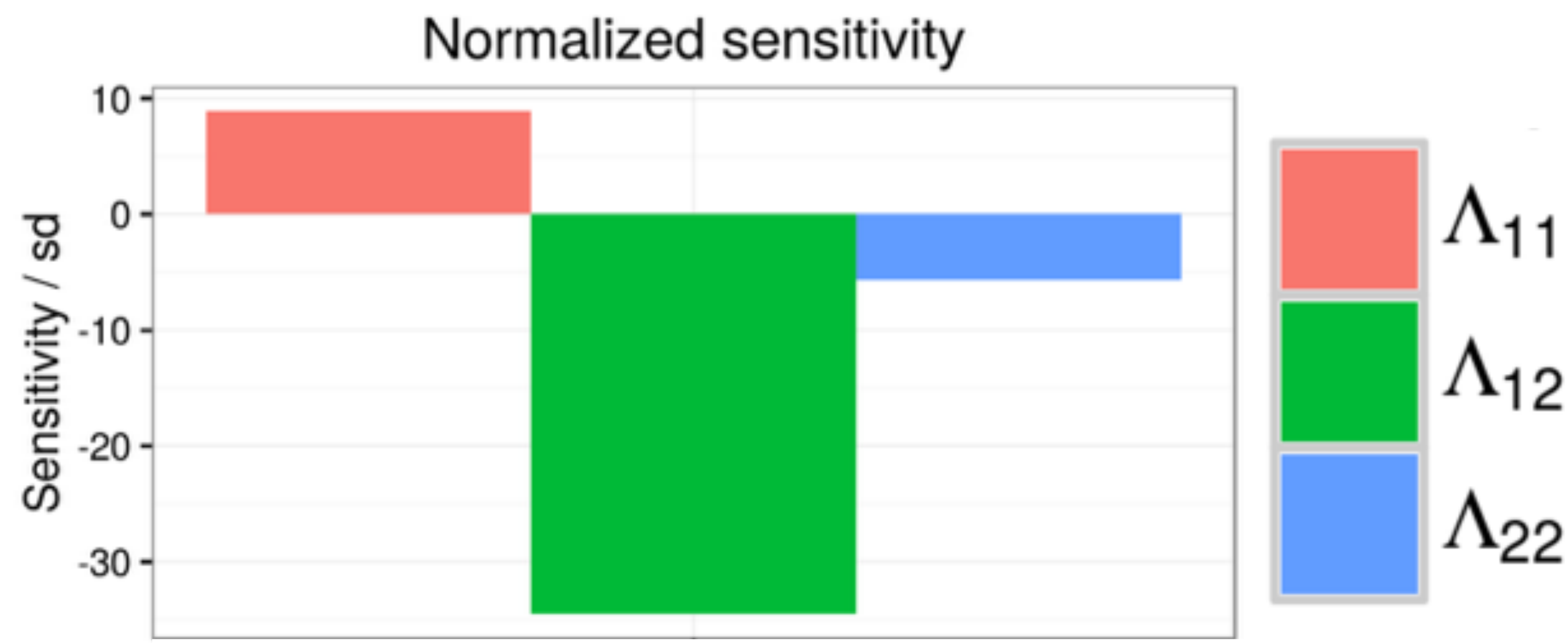
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of standard deviations in τ
- E.g. $\mathrm{StdDev}_q \tau = 1.8$

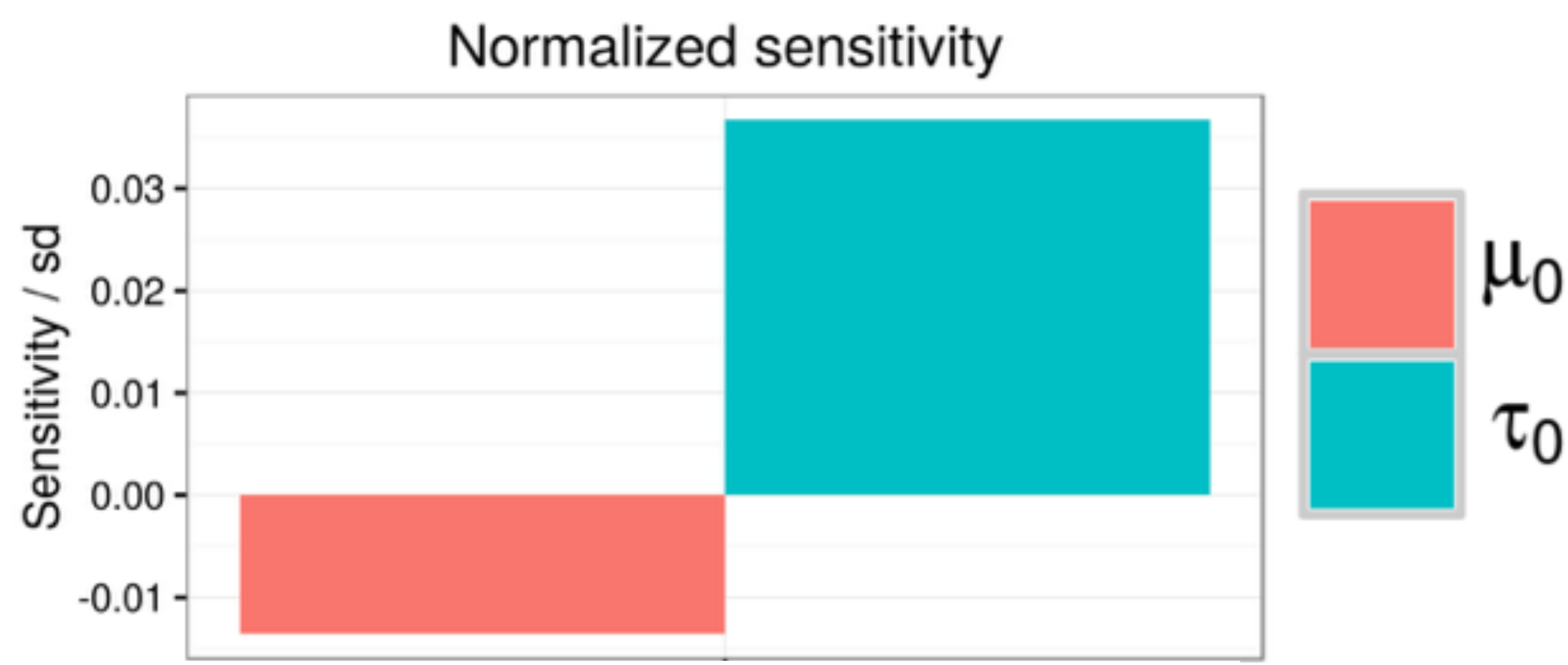




Microcredit Experiment

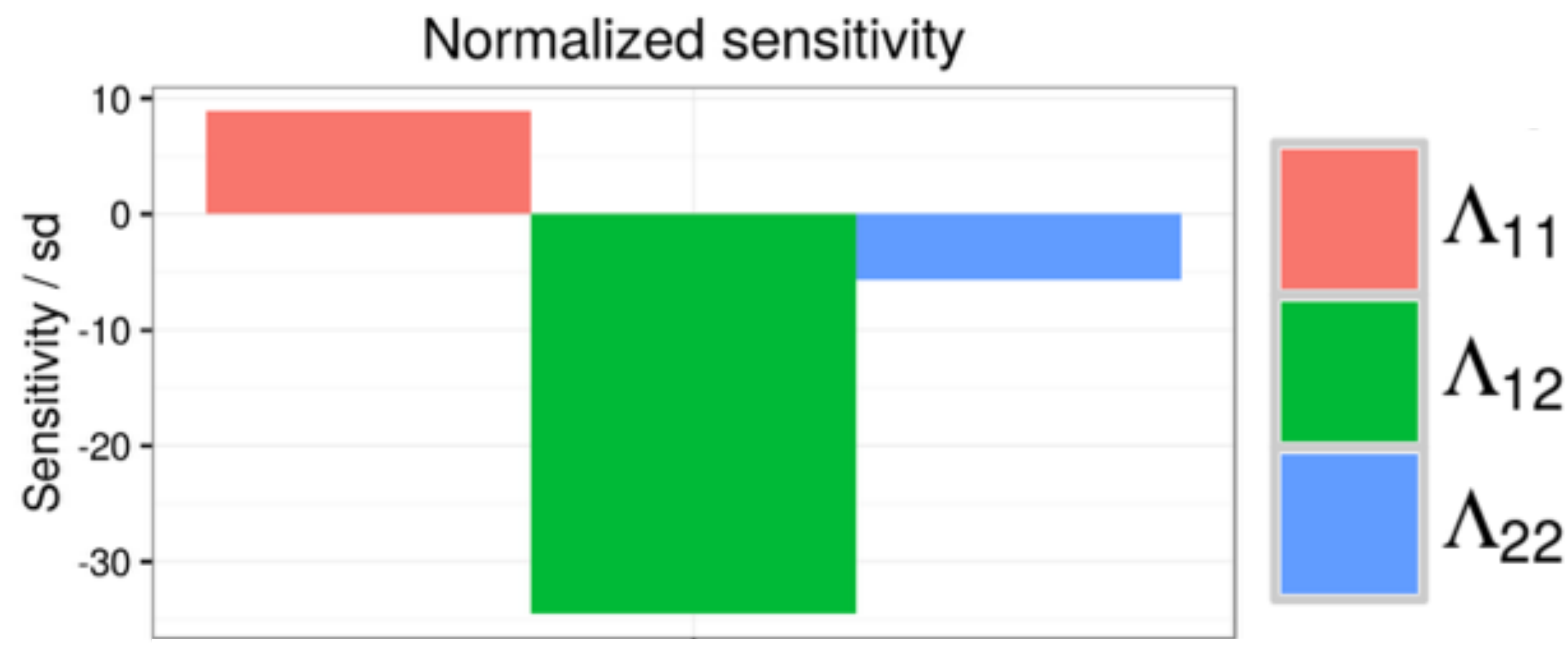
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of standard deviations in τ
- E.g. $\mathrm{StdDev}_q \tau = 1.8$ $\mathbb{E}_q \tau = 3.7$

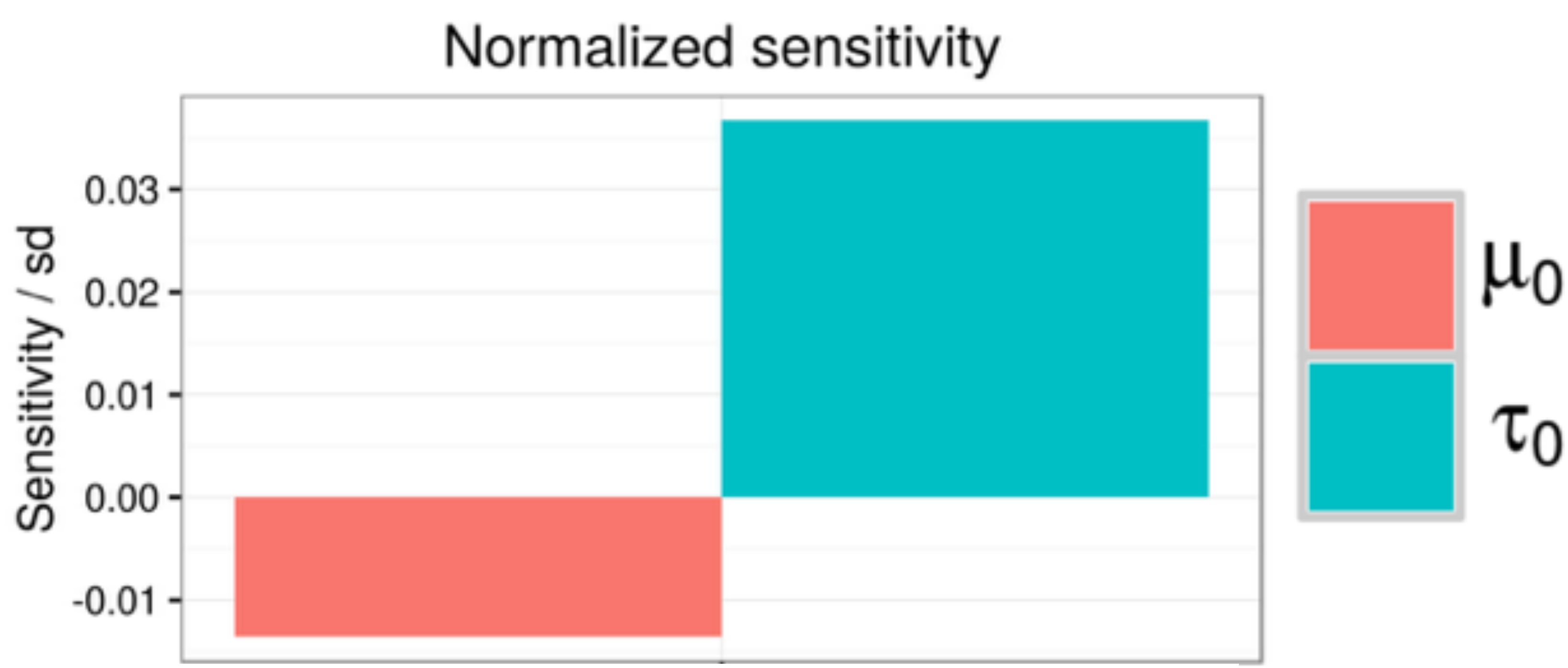




Microcredit Experiment

- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of standard deviations in τ
- E.g.
  $\mathrm{StdDev}_q \tau = 1.8$
  $\mathbb{E}_q \tau = 3.7$
  $= 2.06 * \mathrm{StdDev}_q \tau$



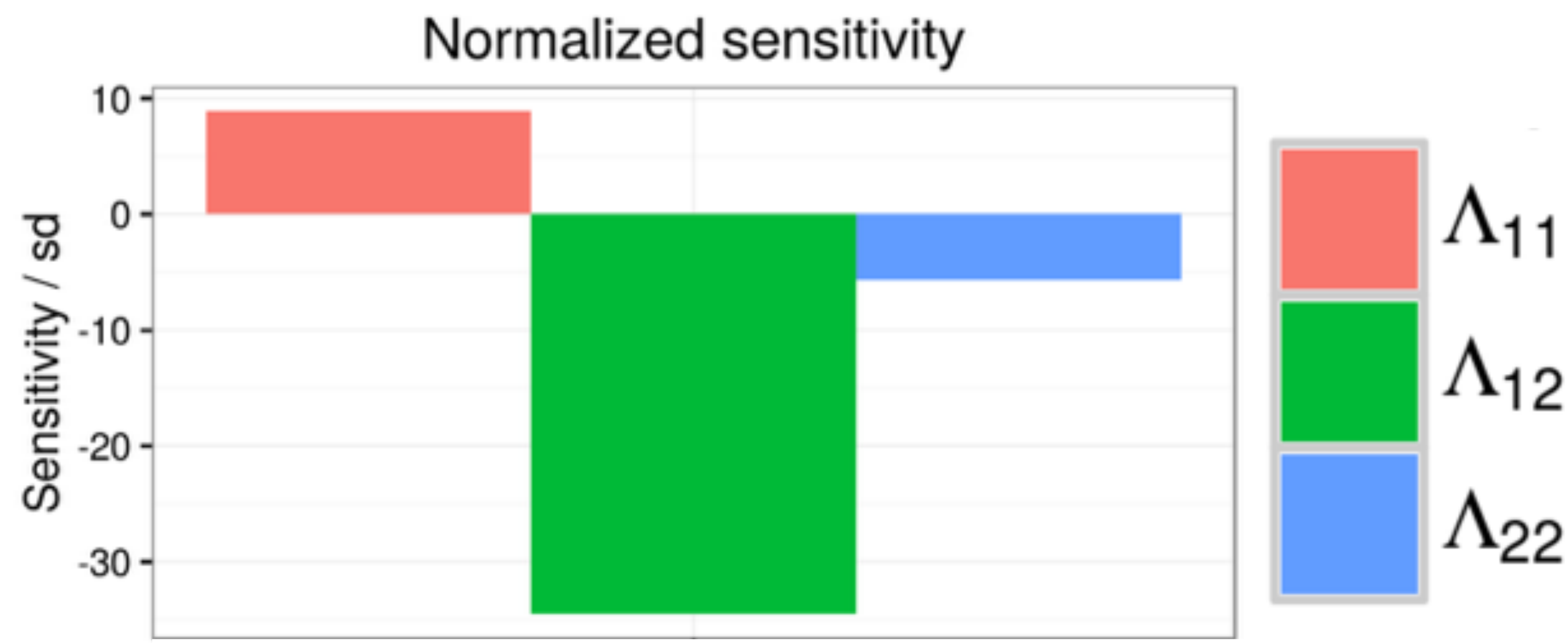


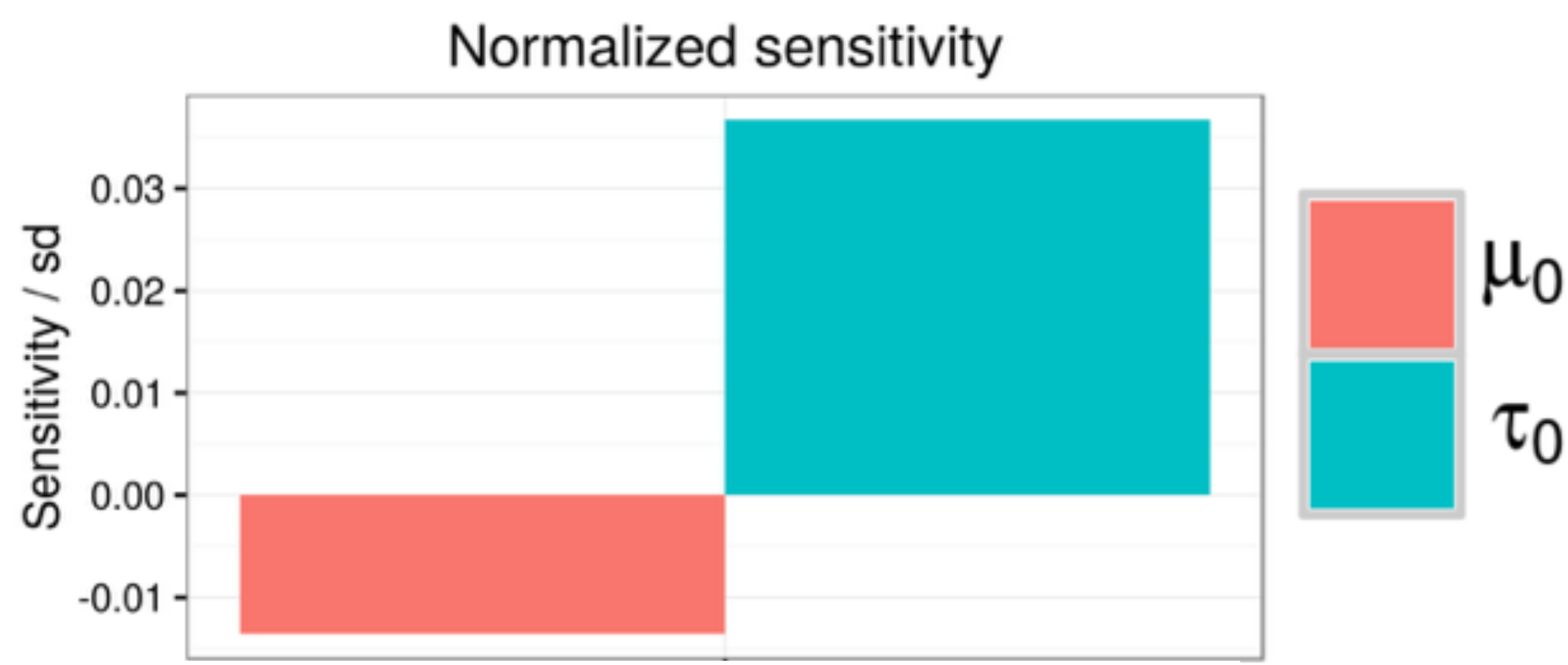
Microcredit Experiment

- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of standard deviations in τ
- E.g.

$$\mathrm{StdDev}_q \tau = 1.8$$
$$\mathbb{E}_q \tau = 3.7$$
$$= 2.06 * \mathrm{StdDev}_q \tau$$

$$\Lambda_{12} += 0.03$$

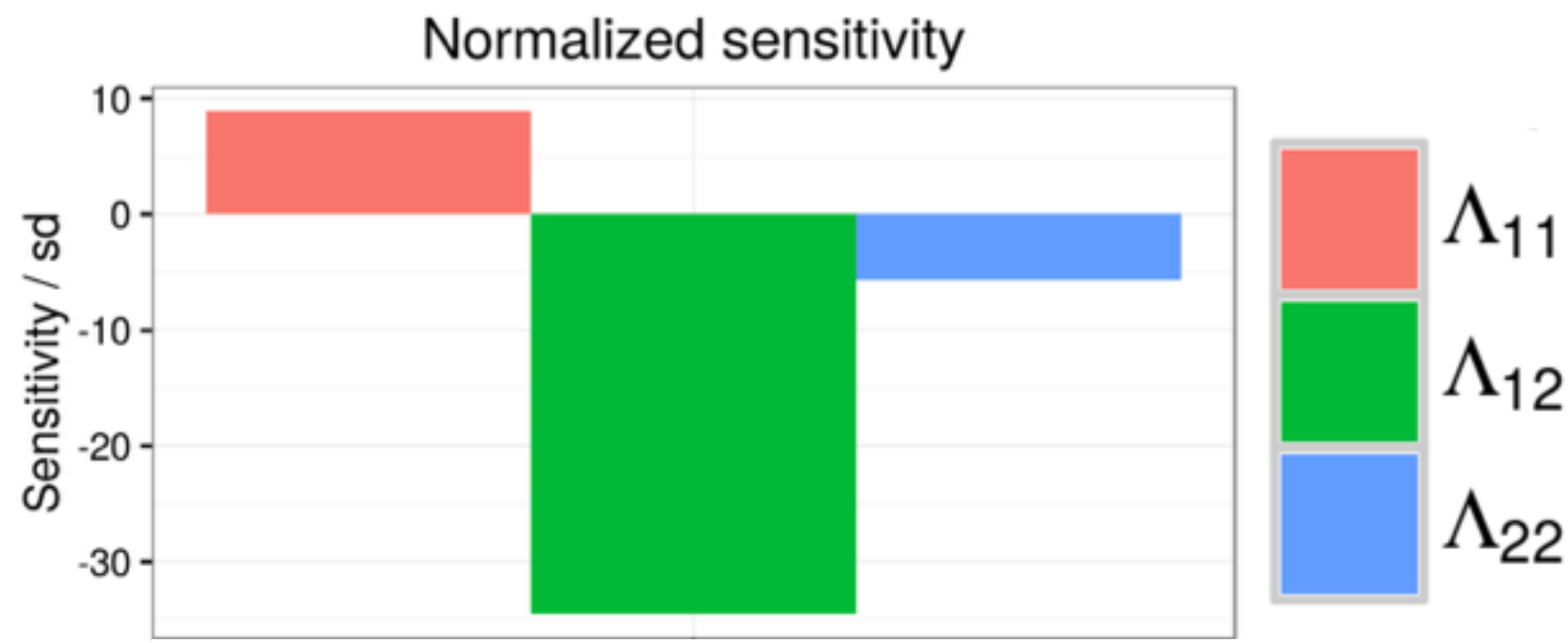


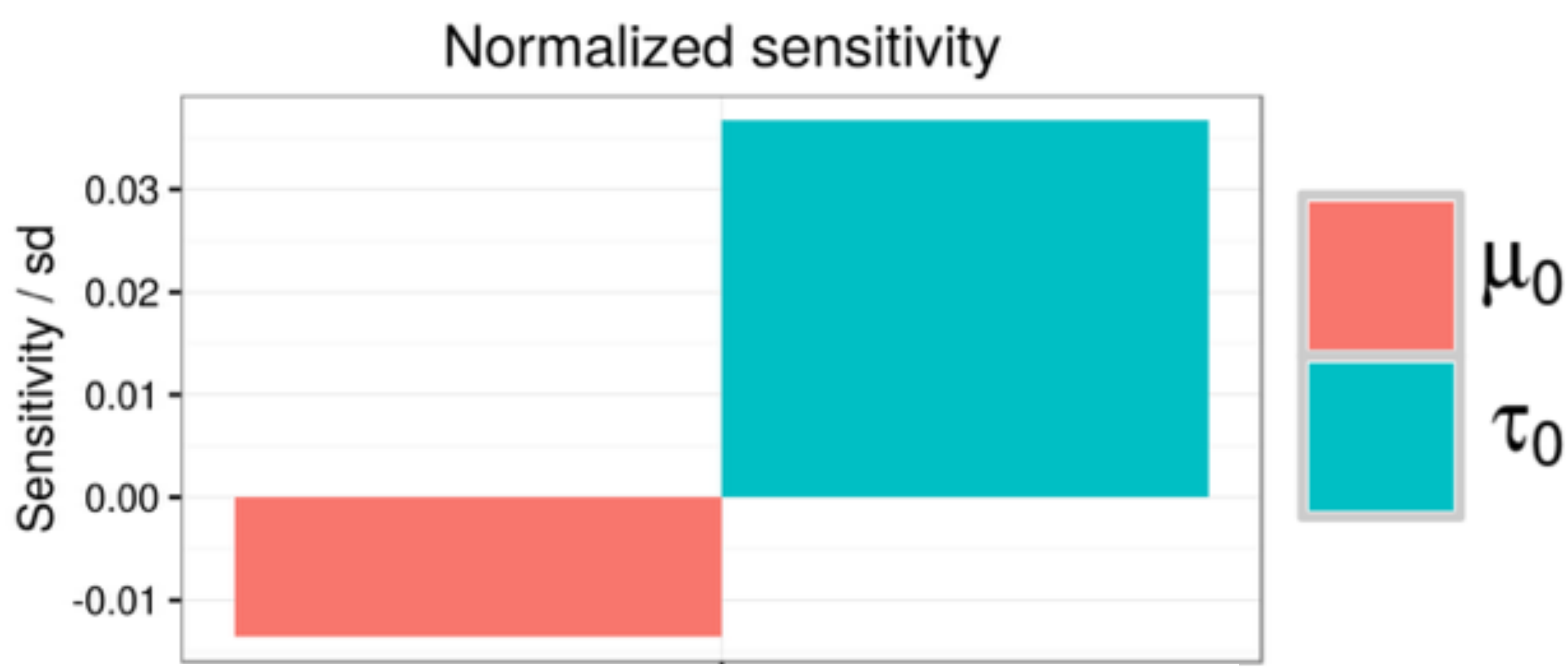


Microcredit Experiment

- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of standard deviations in τ
- E.g.
  $\mathrm{StdDev}_q \tau = 1.8$
  $\mathbb{E}_q \tau = 3.7$
  $= 2.06 * \mathrm{StdDev}_q \tau$

  $\Lambda_{12} += 0.03$
  $\mathbb{E}_q \tau < 1.0 * \mathrm{StdDev}_q \tau$





Conclusion

- We provide *linear response variational Bayes*: supplements MFVB for fast & accurate covariance estimate
- More from LRVB: fast & accurate robustness quantification
- Interested in your data and models:
  - Sensitivity to prior perturbations
  - Sensitivity to data perturbations

References

T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan. Streaming variational Bayes. *NIPS*, 2013.

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, and MI Jordan. Robust Inference with Variational Bayes. *NIPS AABI Workshop*, 2015. ArXiv: 1512.02578.
https://github.com/rgiordan/MicrocreditLRVB

J. Huggins, T. Campbell, and T. Broderick. Core sets for scalable Bayesian logistic regression. *Under review*. ArXiv:1605.06423.

References

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

CM Bishop. *Pattern Recognition and Machine Learning*.

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

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 randomised experiments. ArXiv:1506.06669, 2015.

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

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