Fast Quantification of Uncertainty and Robustness with Variational Bayes

Tamara Broderick
ITT Career Development
Assistant Professor,
MIT

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

Uncertainty & robustness quantification

- Bayesian inference

Uncertainty & robustness quantification

- Bayesian inference
  - Complex, modular models

Uncertainty & robustness quantification

- Bayesian inference
  - Complex, modular models; posterior distribution

Uncertainty & robustness quantification

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

Uncertainty & robustness quantification

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

Uncertainty & robustness quantification

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

Uncertainty & robustness quantification

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Challenge: Express prior beliefs in a distribution

Uncertainty & robustness quantification

- Bayesian inference $p(\theta|x) \propto_{\theta} p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Challenge: Express prior beliefs in a distribution
  - Time-consuming

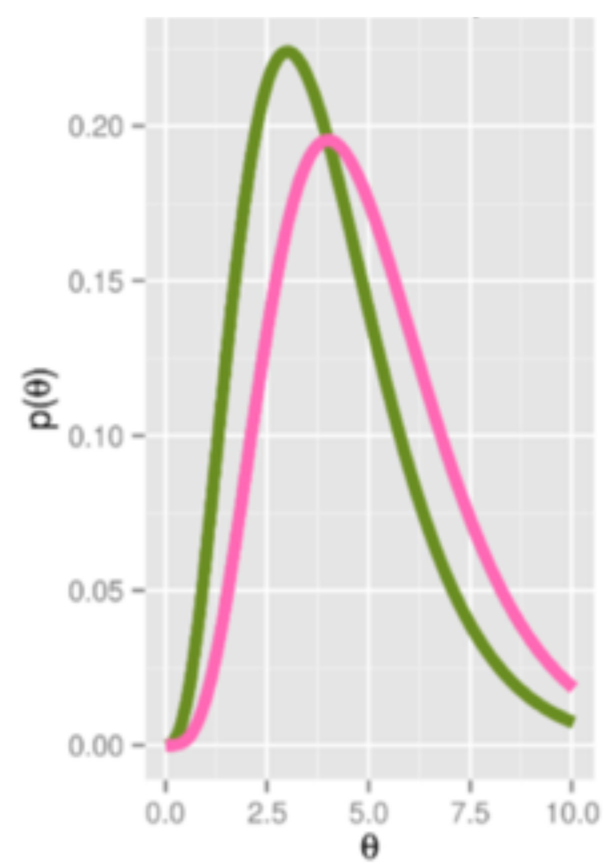
Uncertainty & robustness quantification

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

Uncertainty & robustness quantification

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

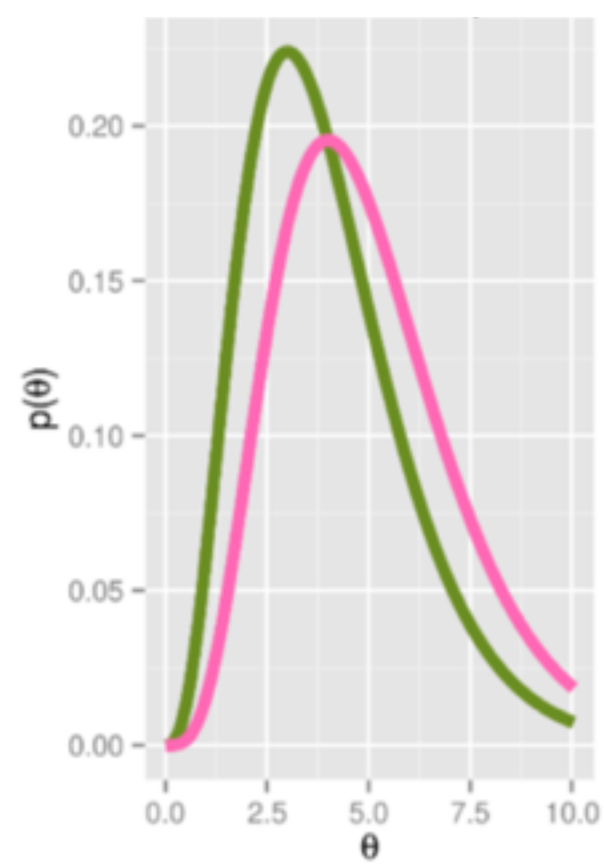
Some reasonable priors

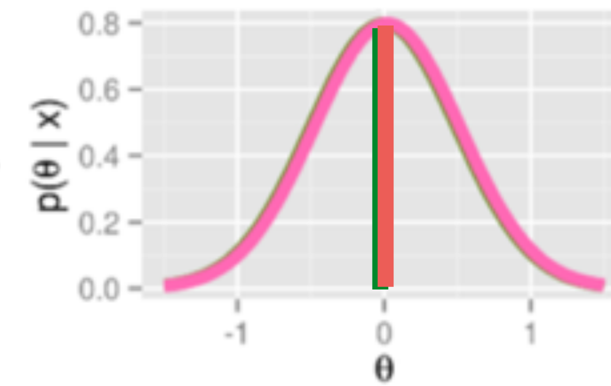


Uncertainty & robustness quantification

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

Some reasonable priors



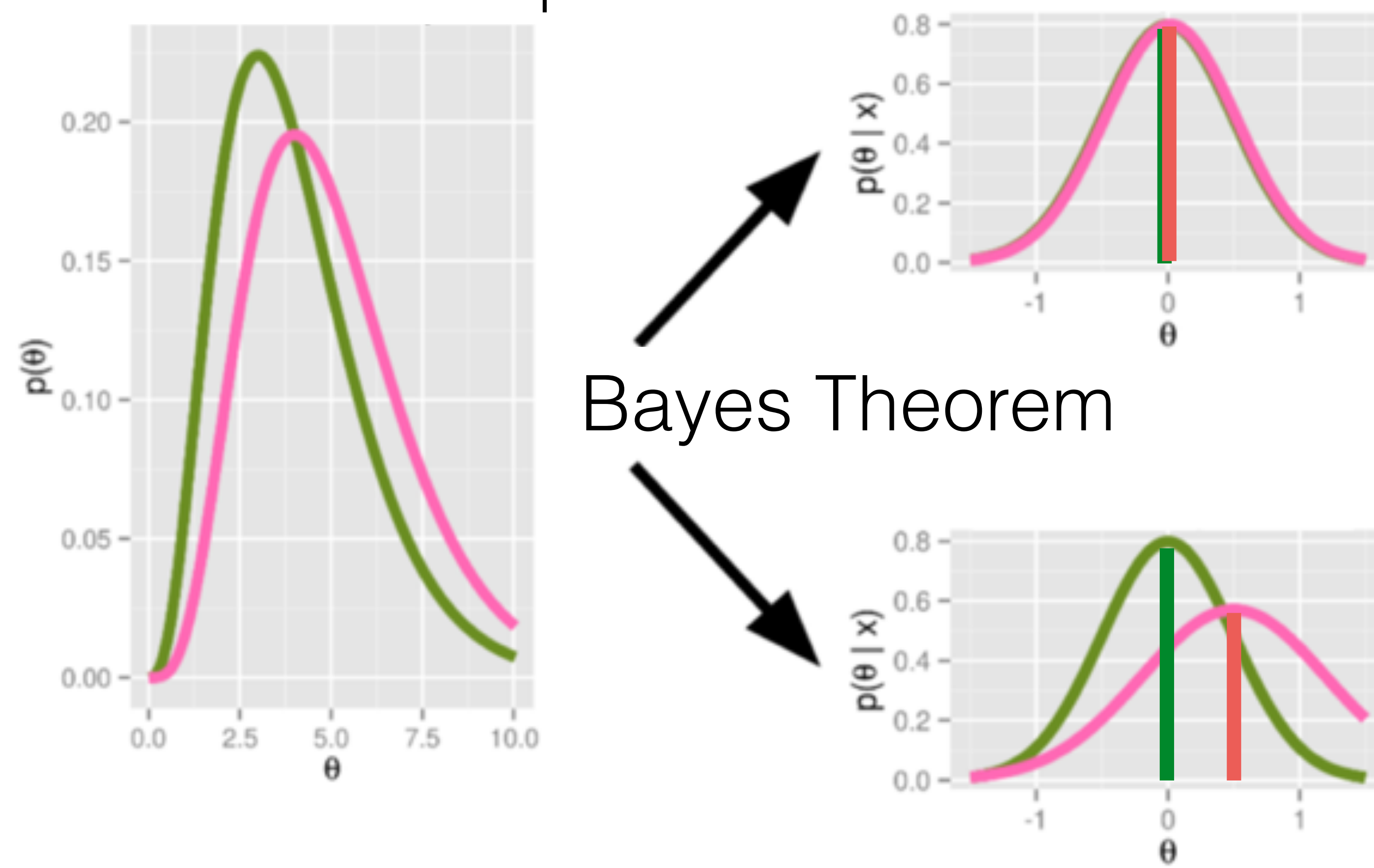


Bayes Theorem

Uncertainty & robustness quantification

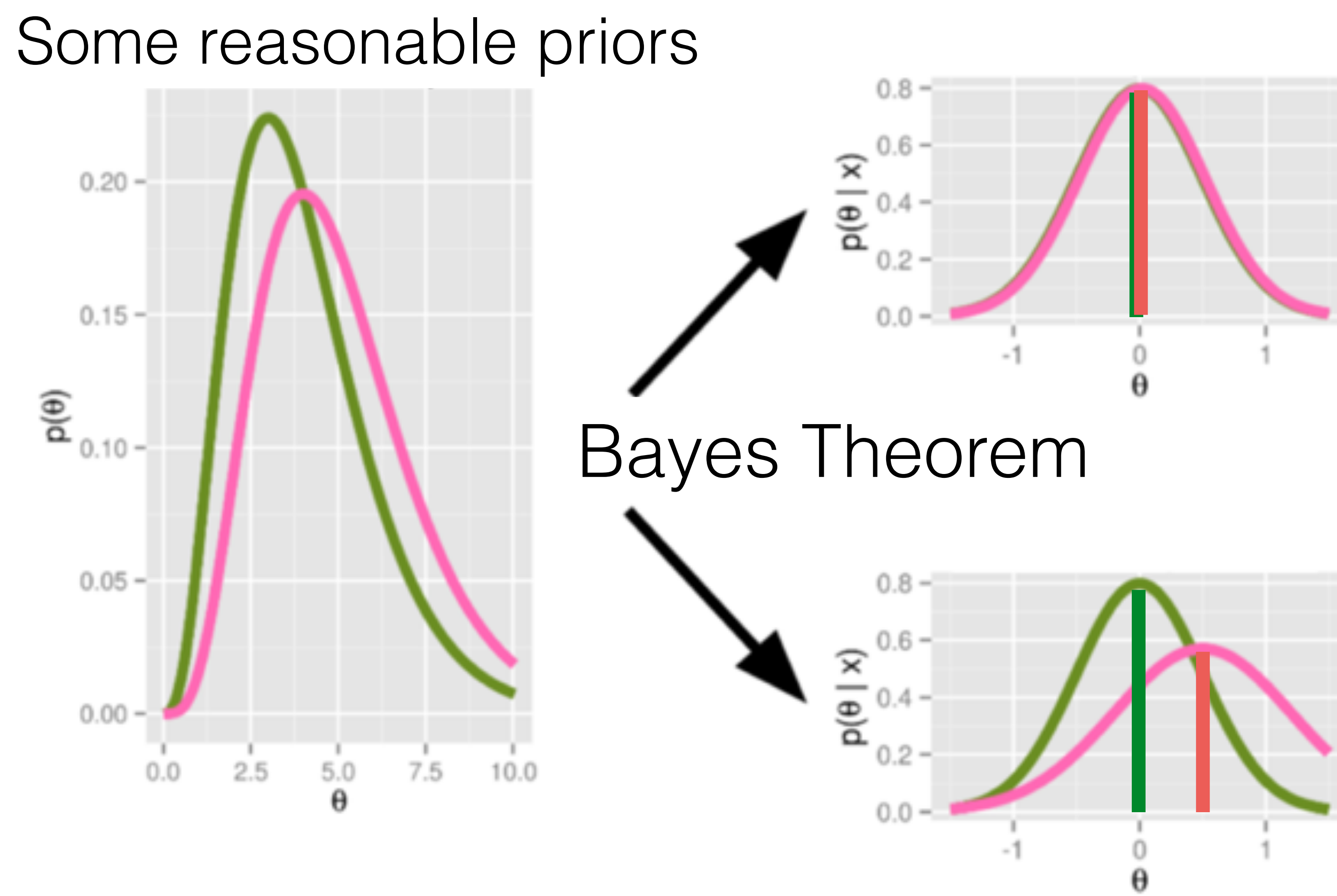
- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Challenge: Express prior beliefs in a distribution
  - Time-consuming; subjective

Some reasonable priors



Uncertainty & robustness quantification

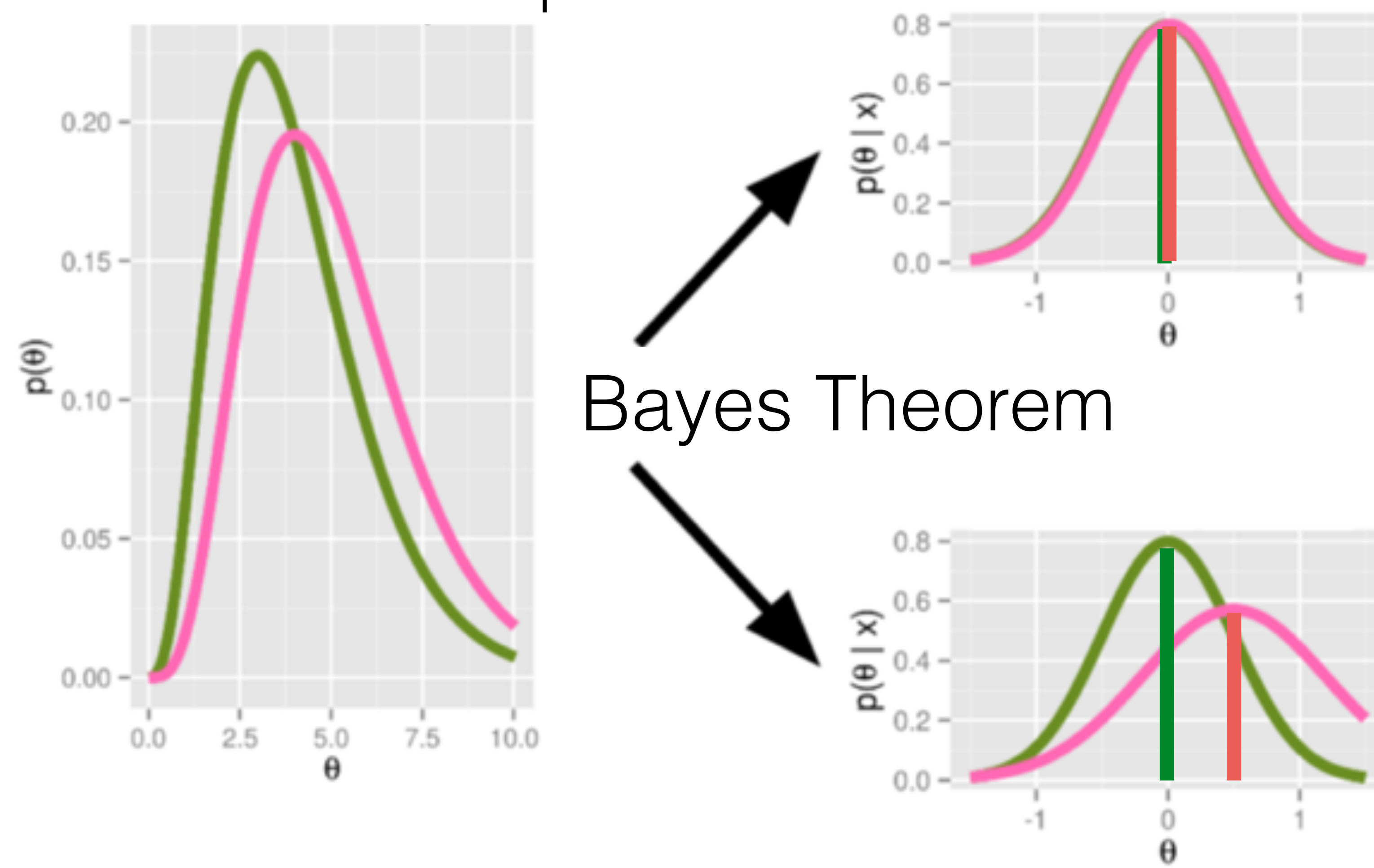
- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Challenge: Express prior beliefs in a distribution
  - Time-consuming; subjective; complex models



Uncertainty & robustness quantification

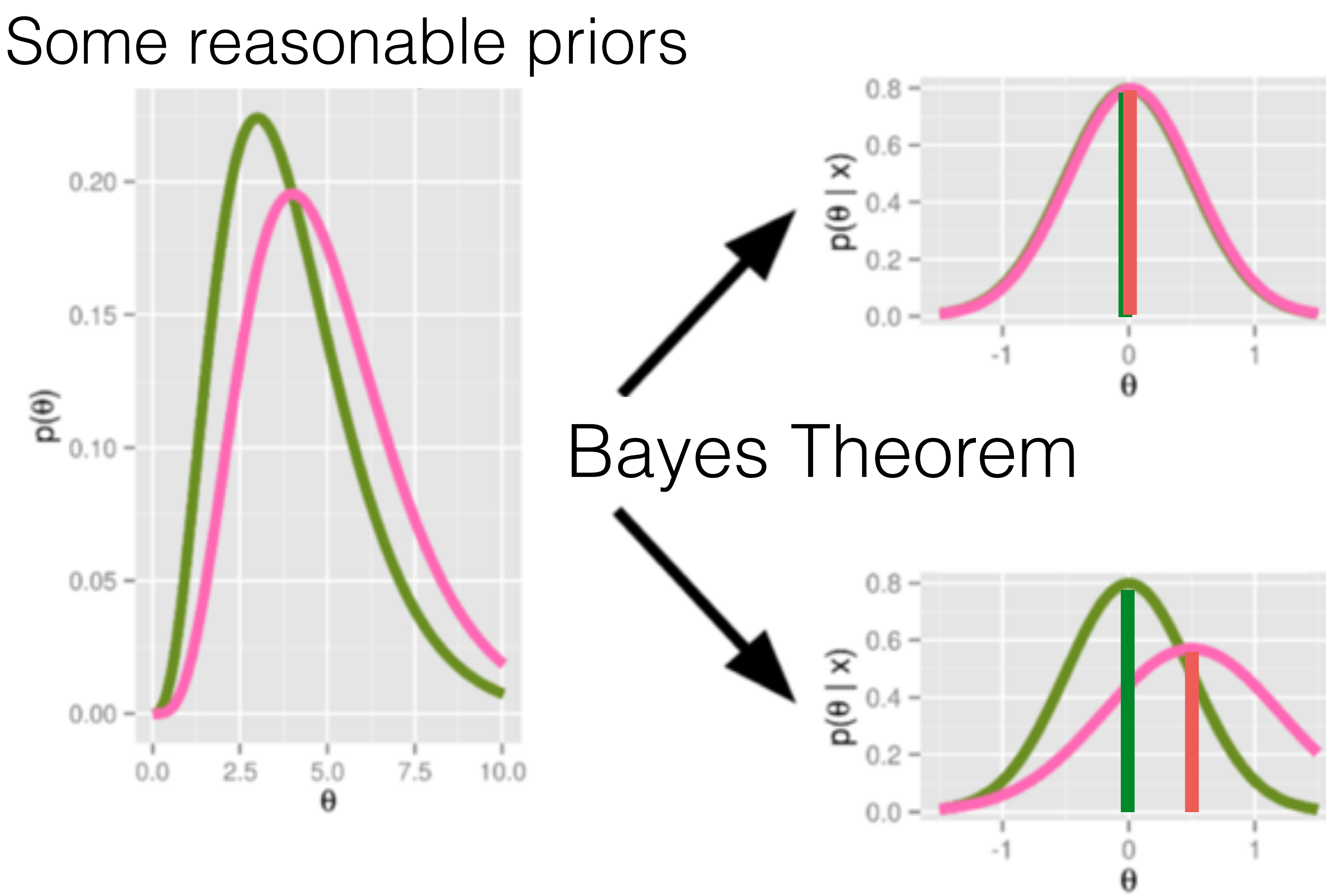
- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Challenge: Express prior beliefs in a distribution
  - Time-consuming; subjective; complex models
- Challenge: Approximating the posterior can be computationally expensive

Some reasonable priors



Uncertainty & robustness quantification

- Bayesian inference $p(\theta|x) \propto_\theta p(x|\theta)p(\theta)$
  - Complex, modular models; posterior distribution
- Challenge: Express prior beliefs in a distribution
  - Time-consuming; subjective; complex models
- Challenge: Approximating the posterior can be computationally expensive



- We propose: *linear response variational Bayes*

Roadmap

Roadmap

- Variational Bayes as an alternative to MCMC

Roadmap

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

Roadmap

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

Roadmap

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

Roadmap

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

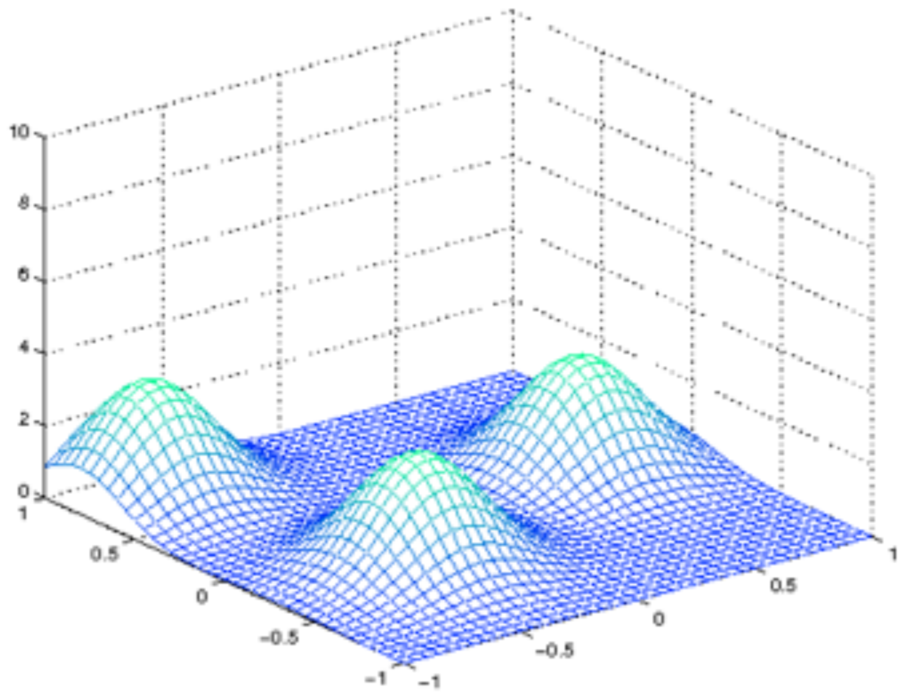
- Big idea: derivatives/perturbations are relatively easy in VB

Variational Bayes

Variational Bayes

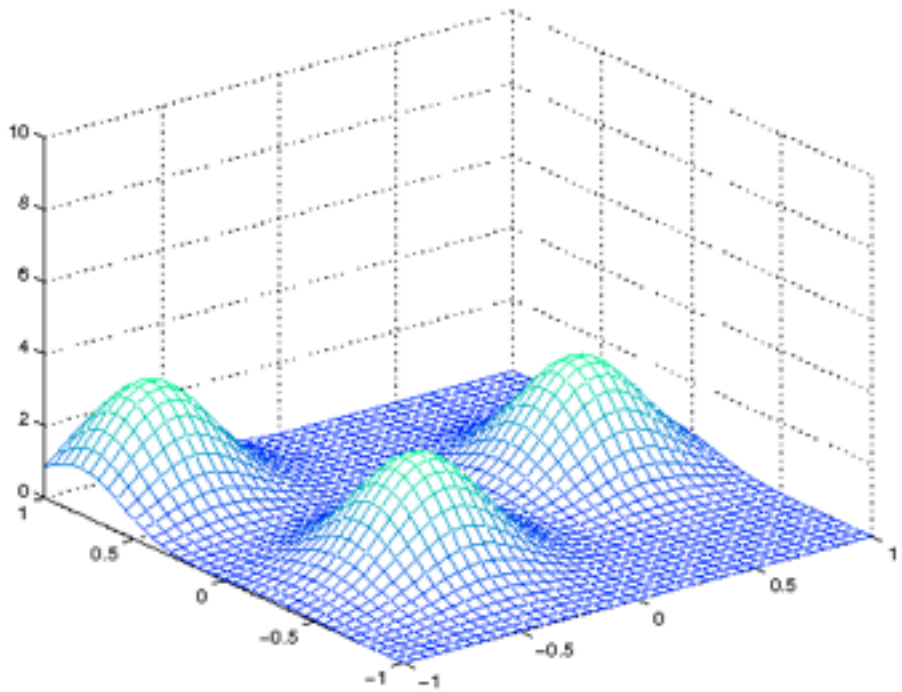
- Variational Bayes (VB)

Variational Bayes



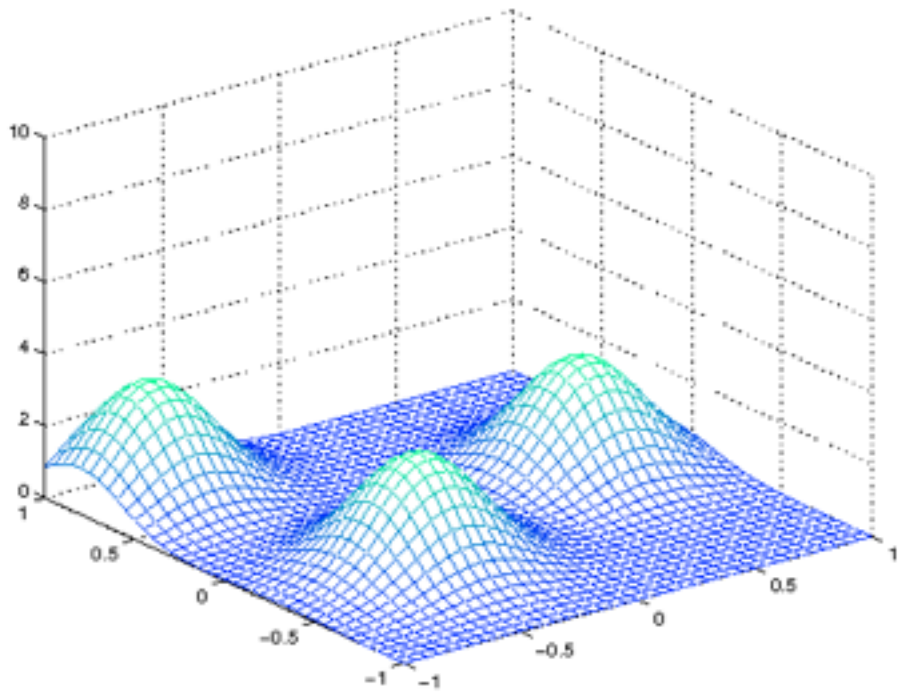
- Variational Bayes (VB)

Variational Bayes

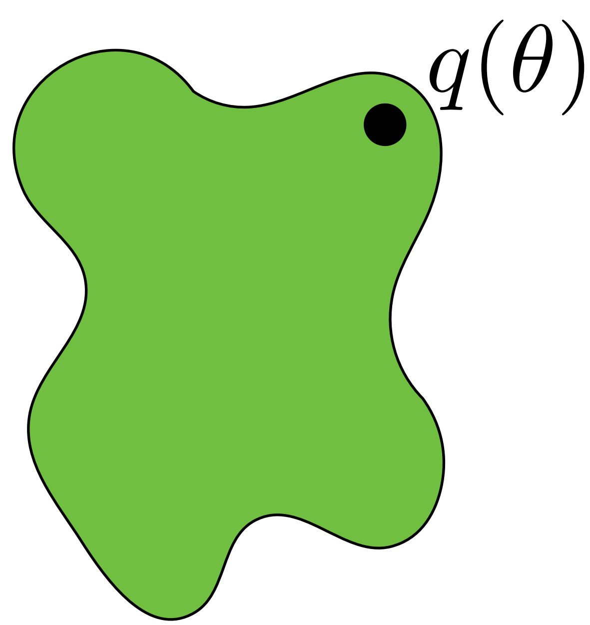


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$

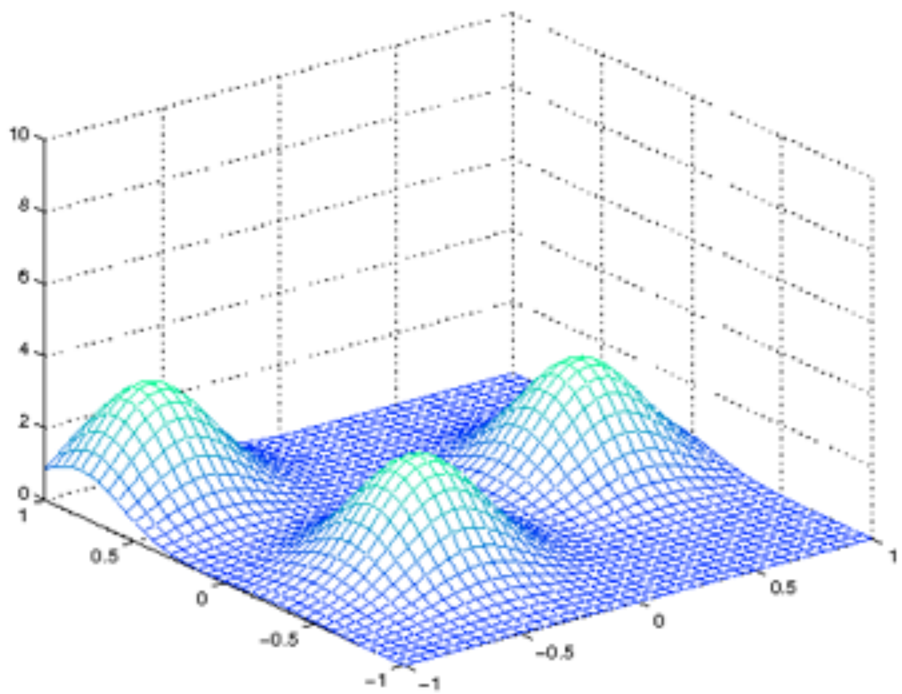
Variational Bayes



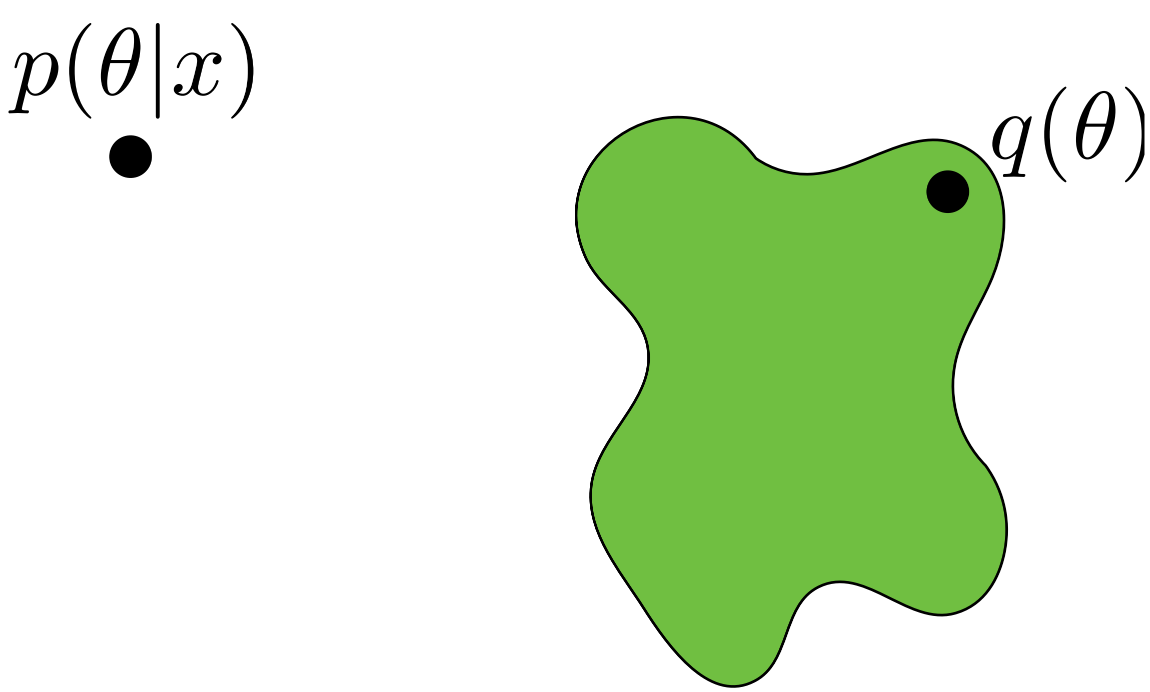
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$



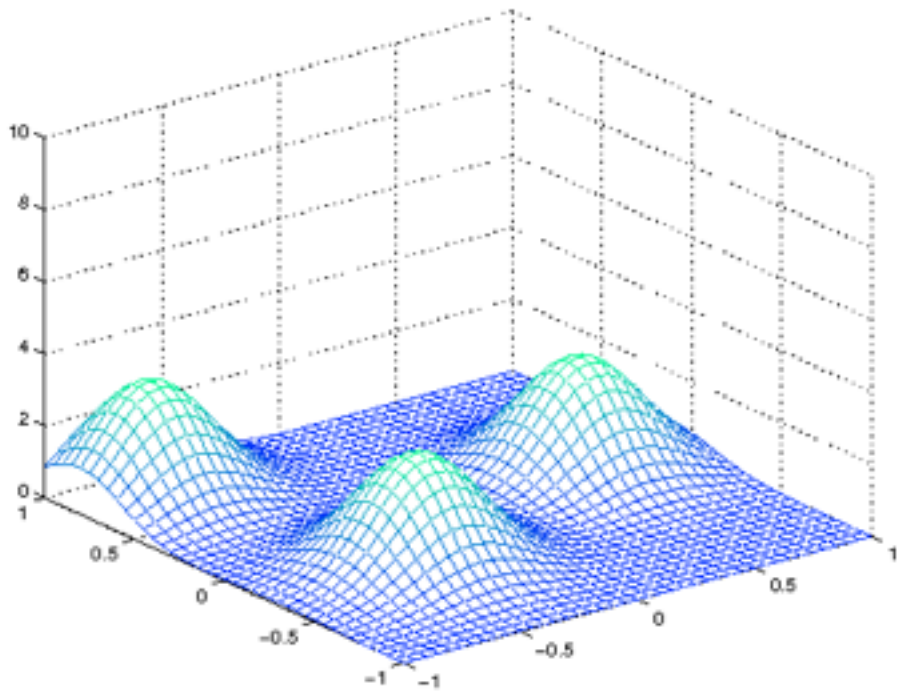
Variational Bayes

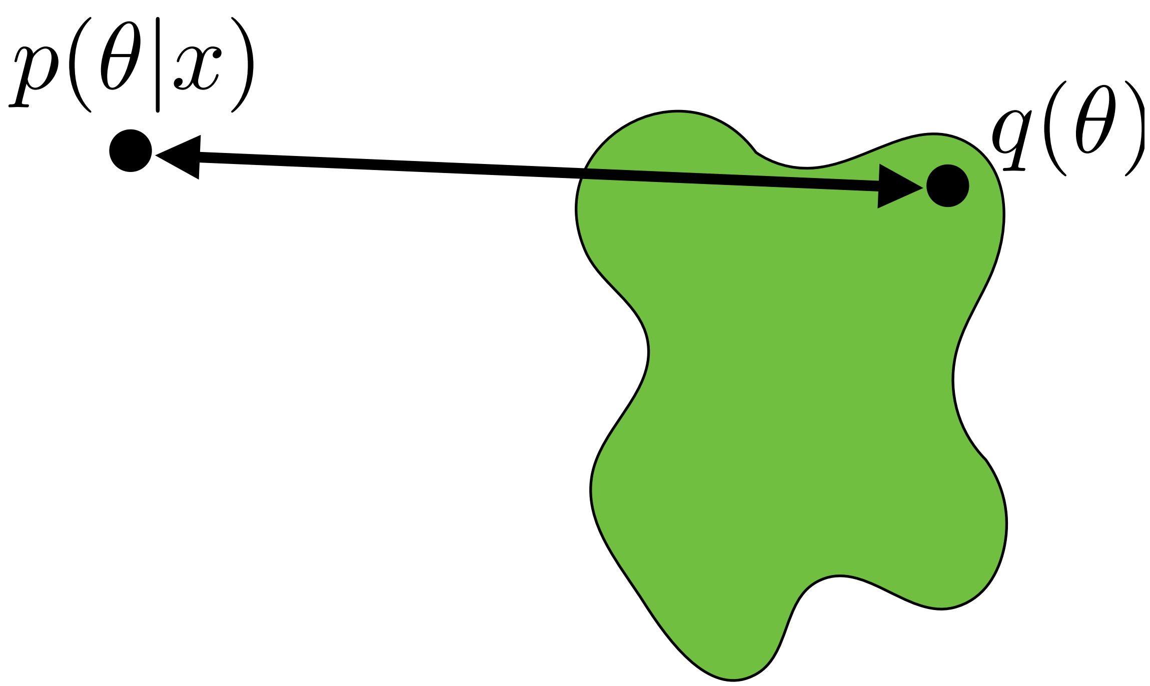


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$



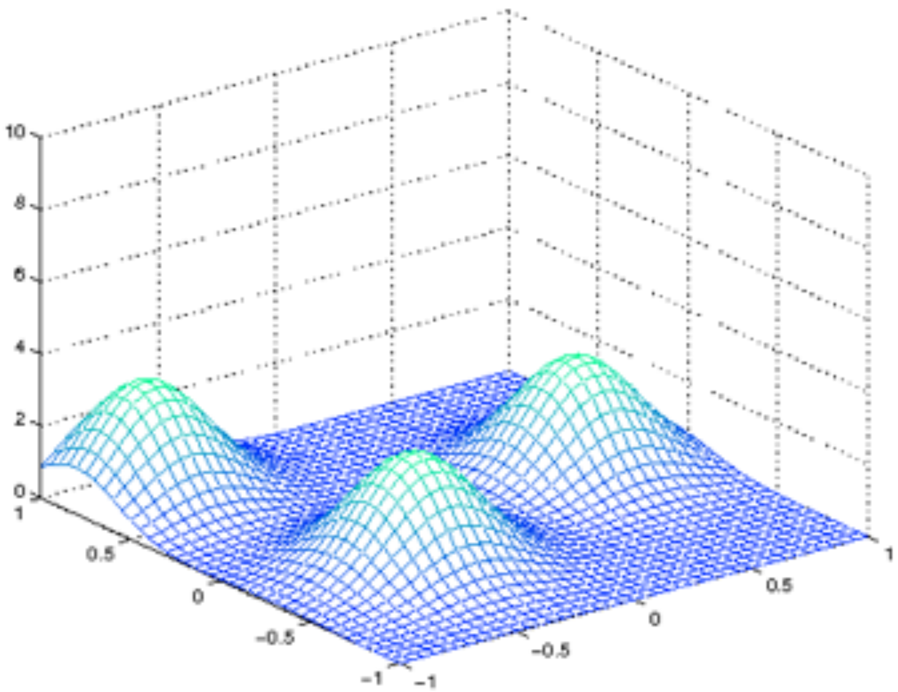
Variational Bayes



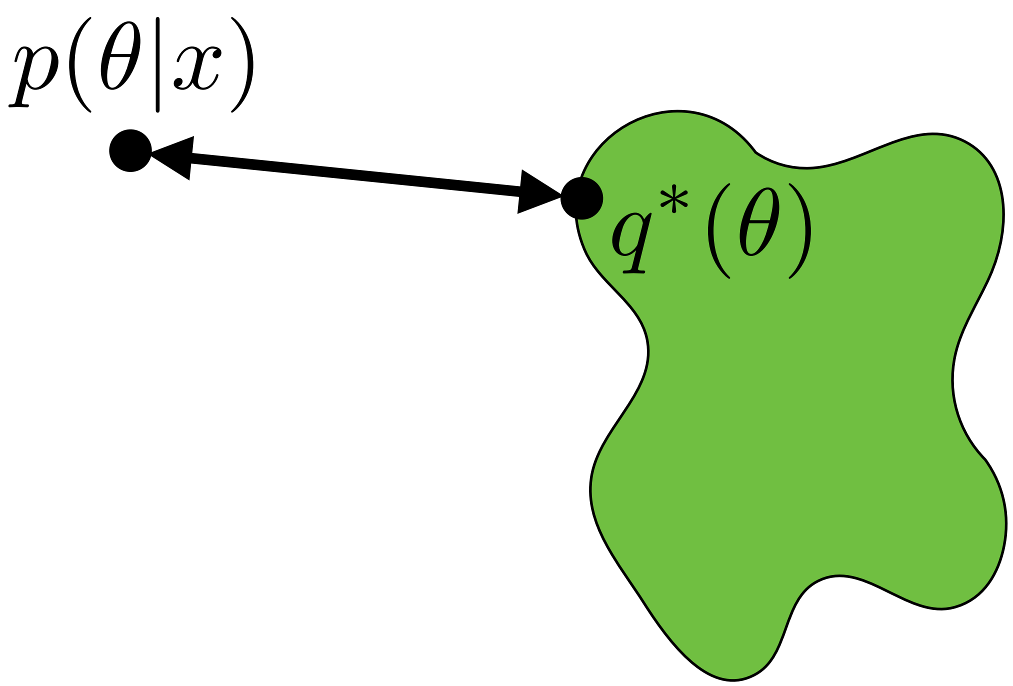


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$

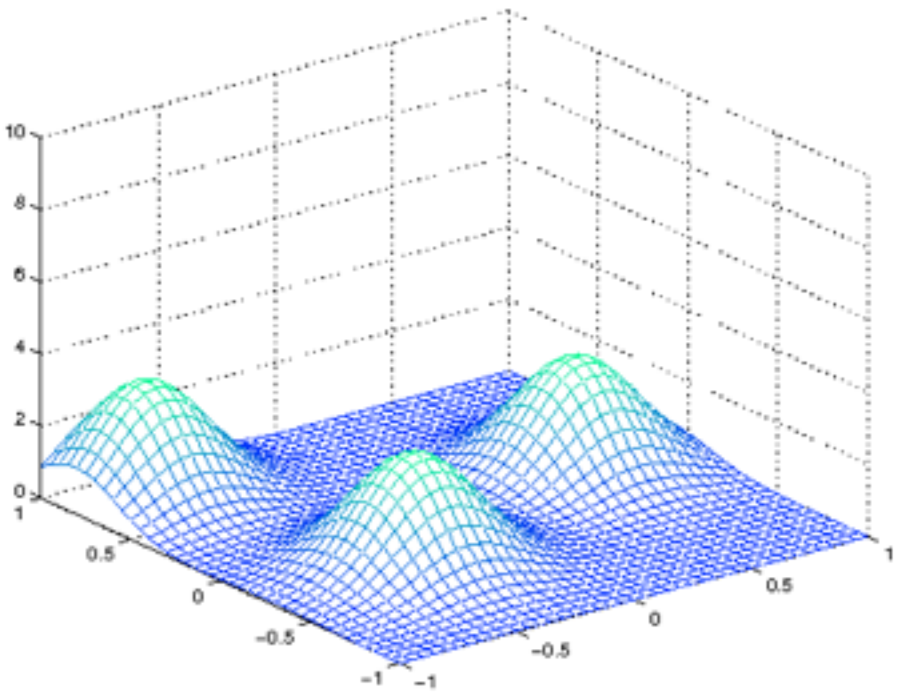
Variational Bayes

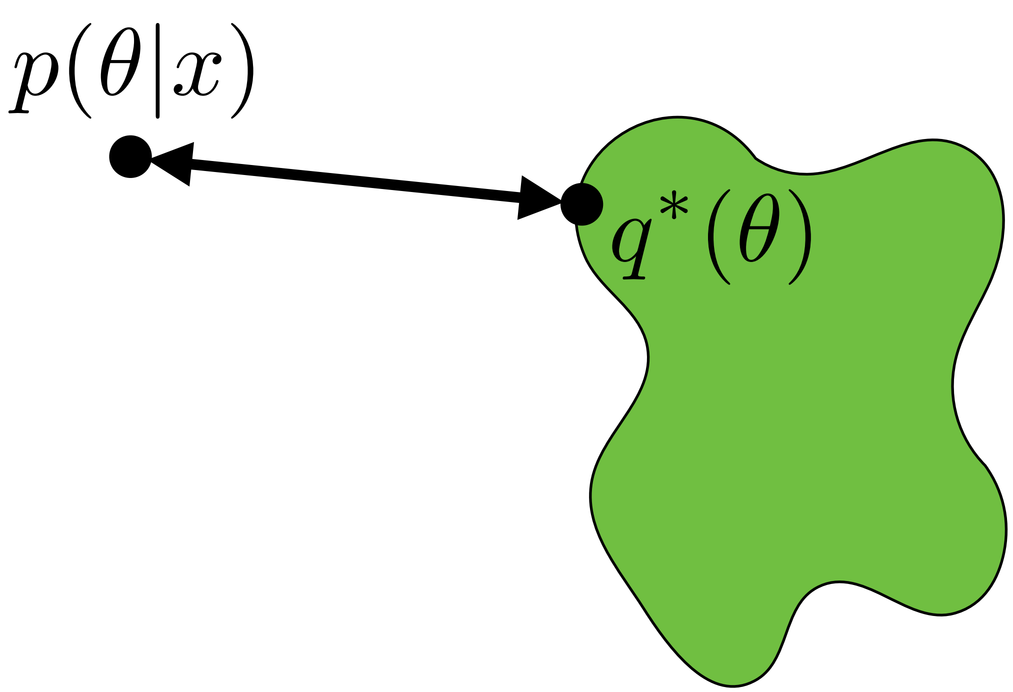


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$



Variational Bayes

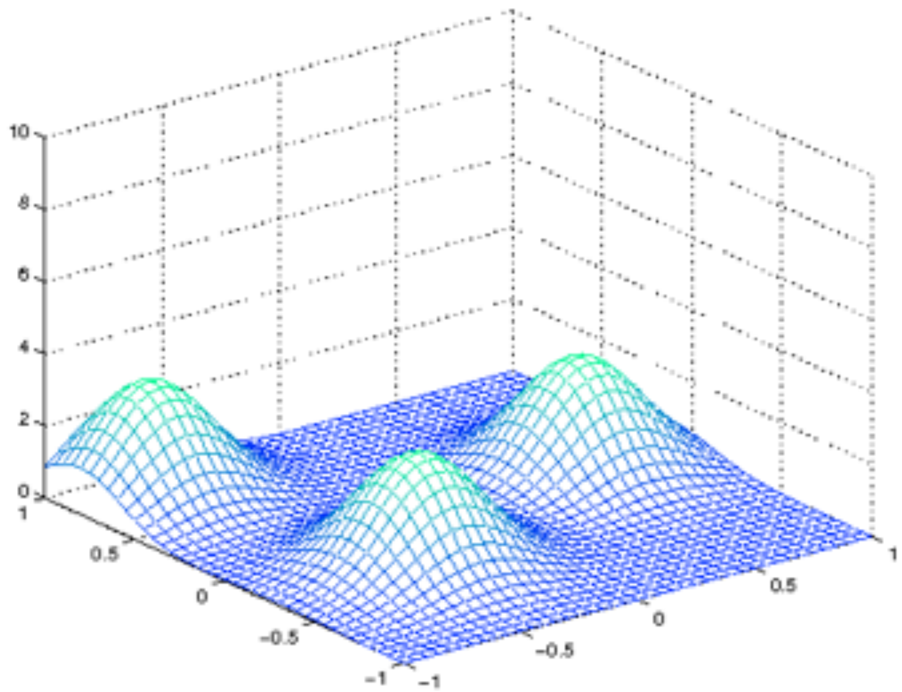




- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

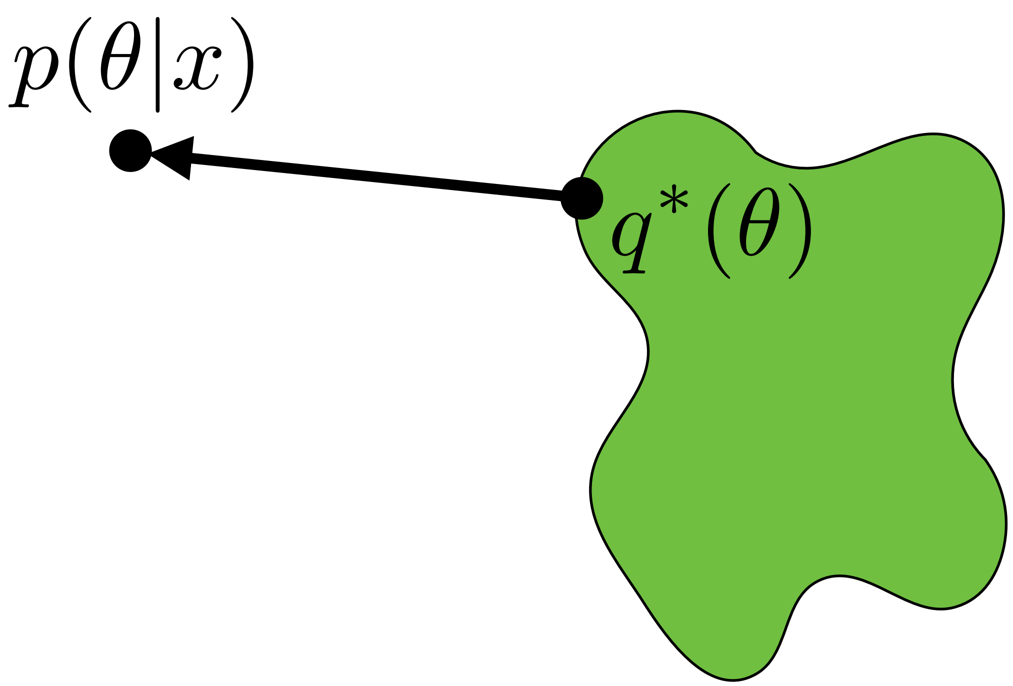
$$KL(q||p(\cdot|x))$$

Variational Bayes

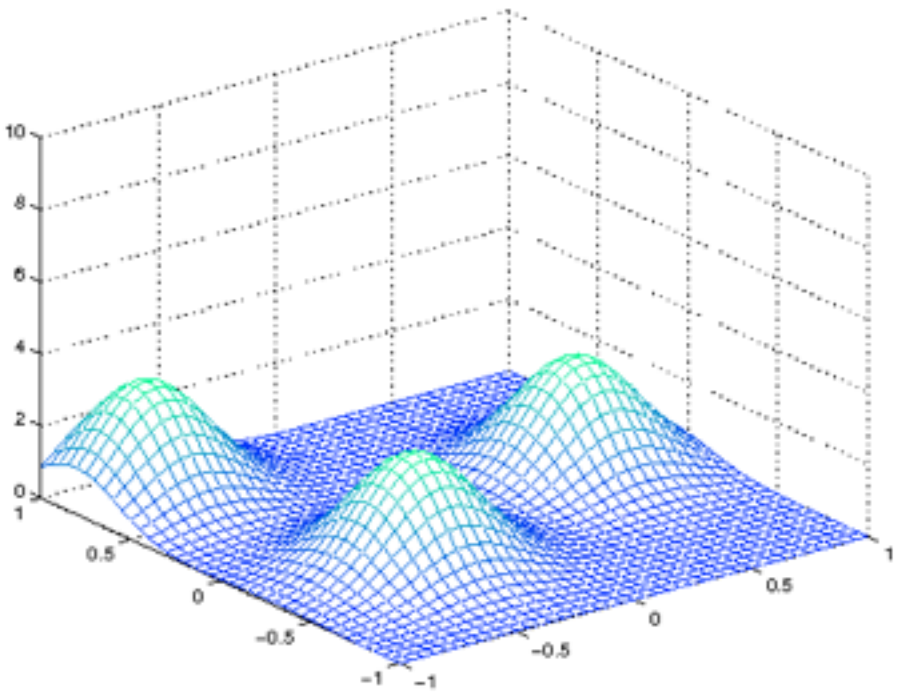


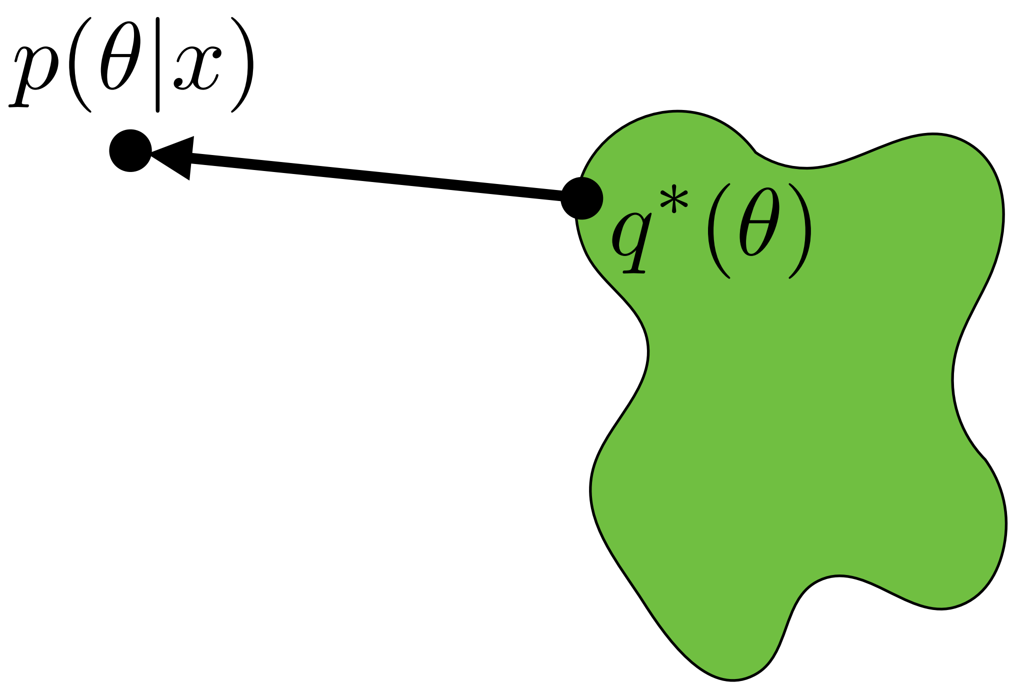
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

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



Variational Bayes



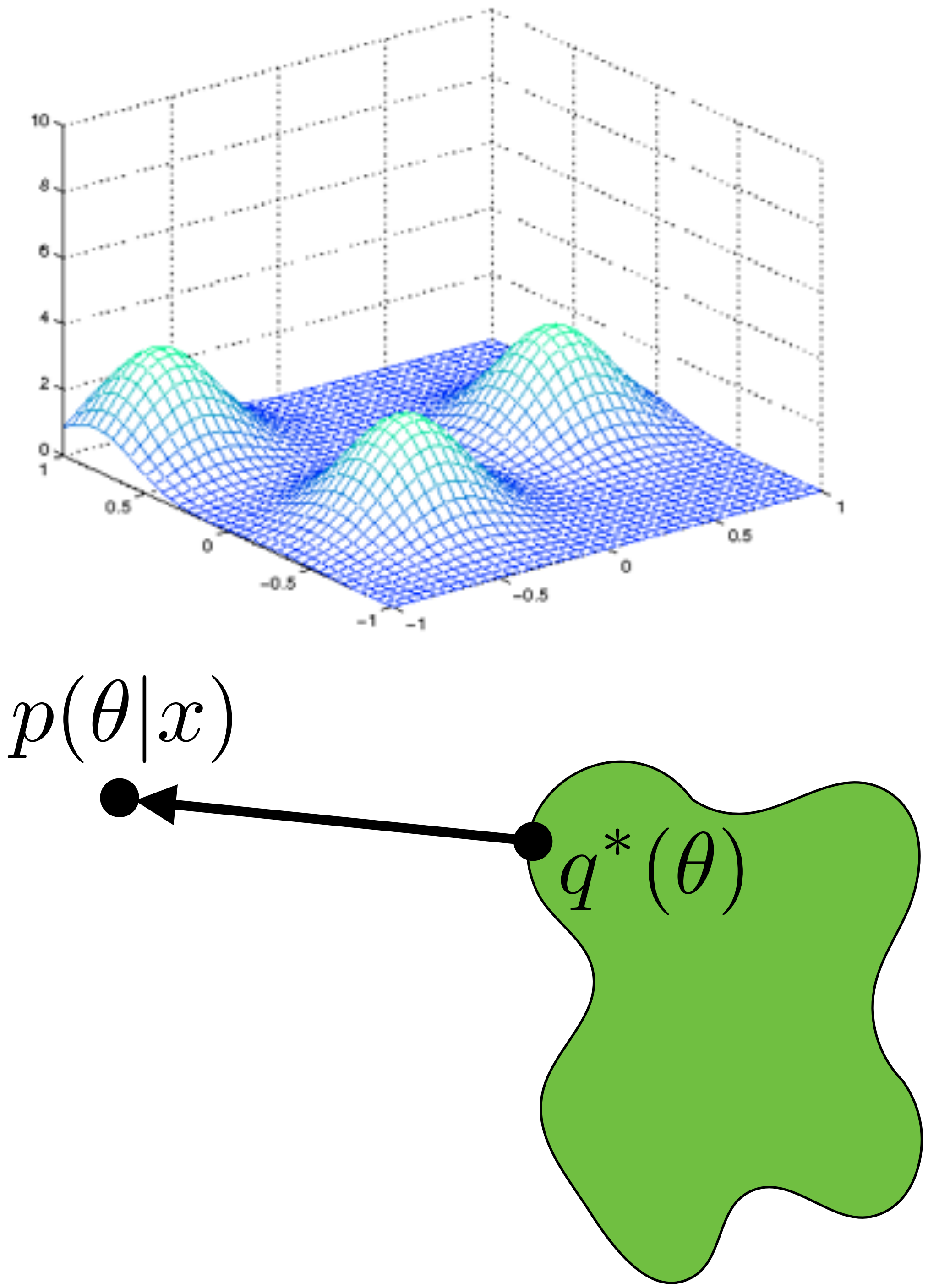


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

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

- VB practical success

Variational Bayes

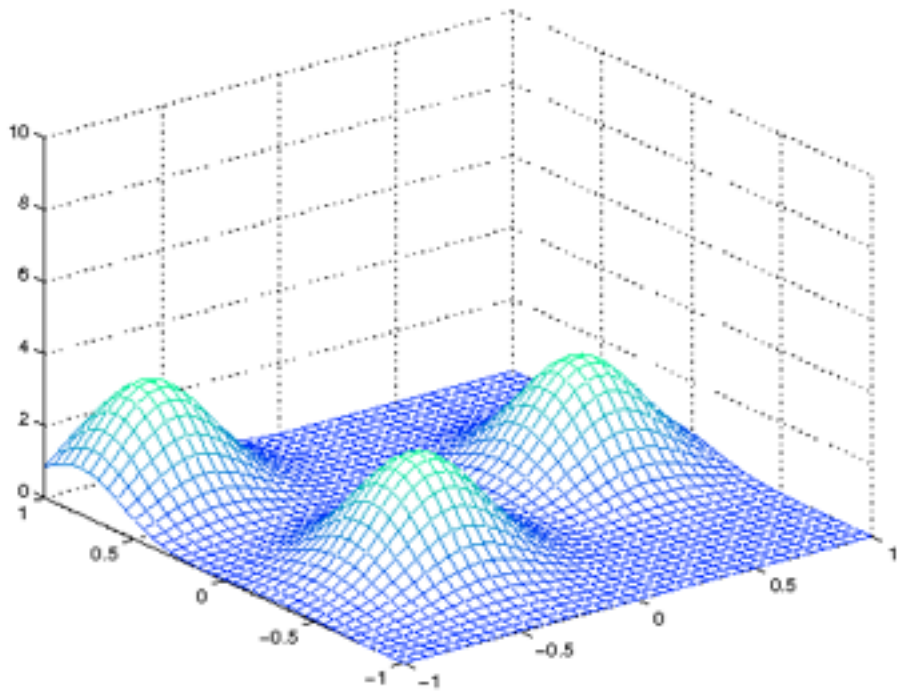


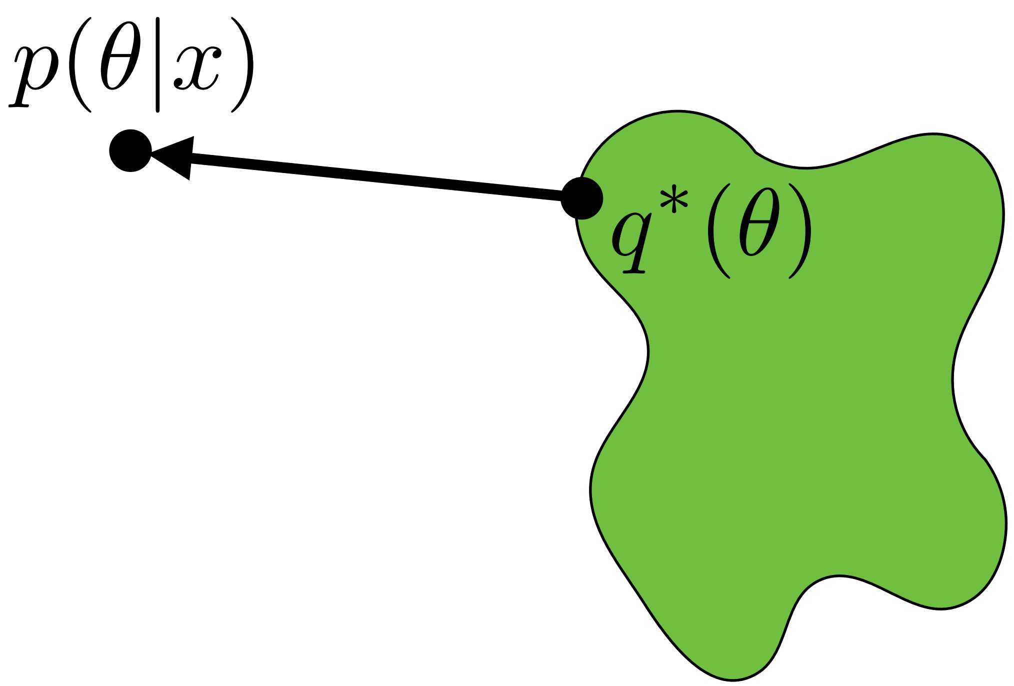
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

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

- VB practical success
  - point estimates and prediction

Variational Bayes





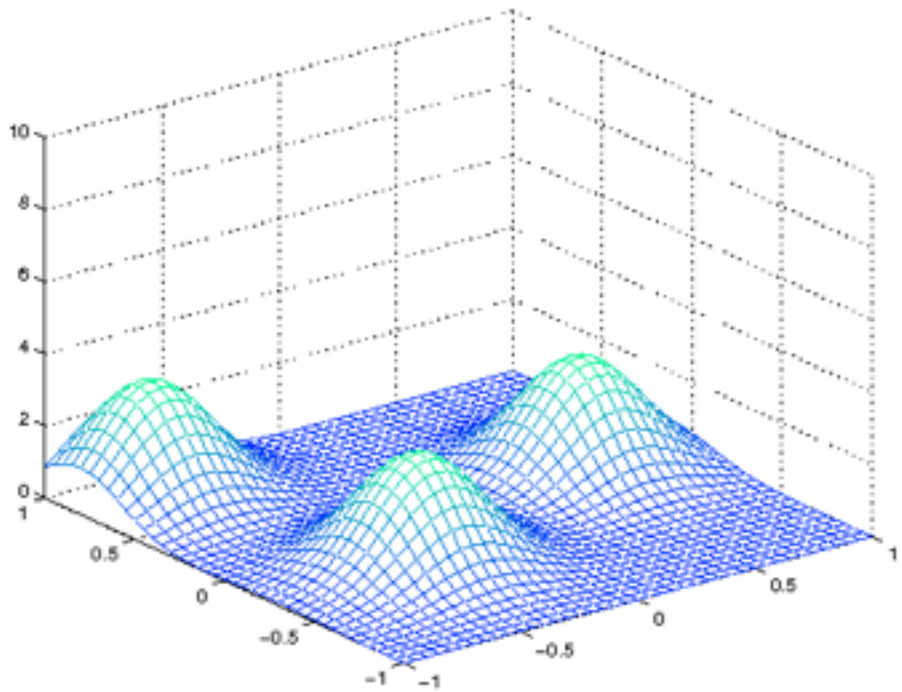
- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Liebler (KL) divergence:

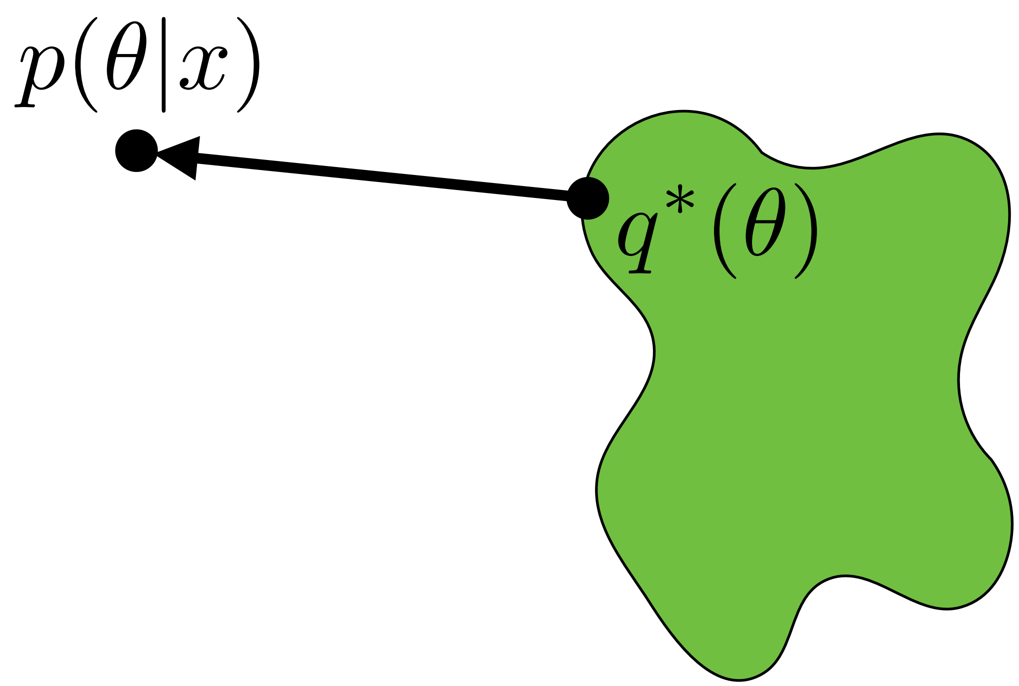
$$KL(q||p(\cdot|x))$$

- 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]

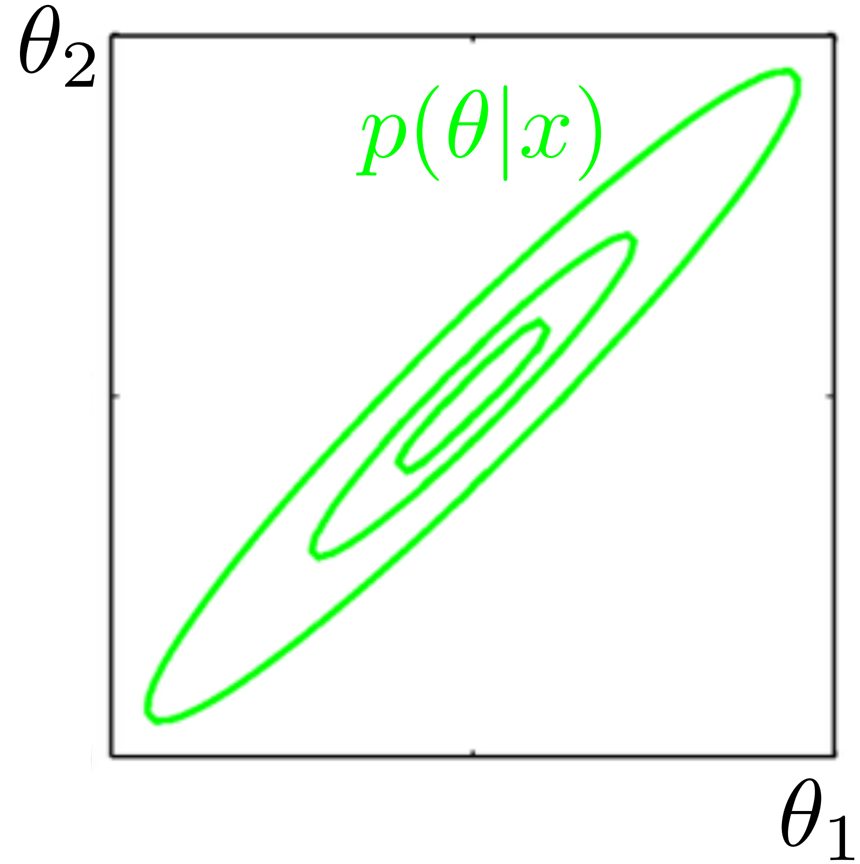
What about uncertainty?

What about uncertainty?

- Variational Bayes

What about uncertainty?

- Variational Bayes

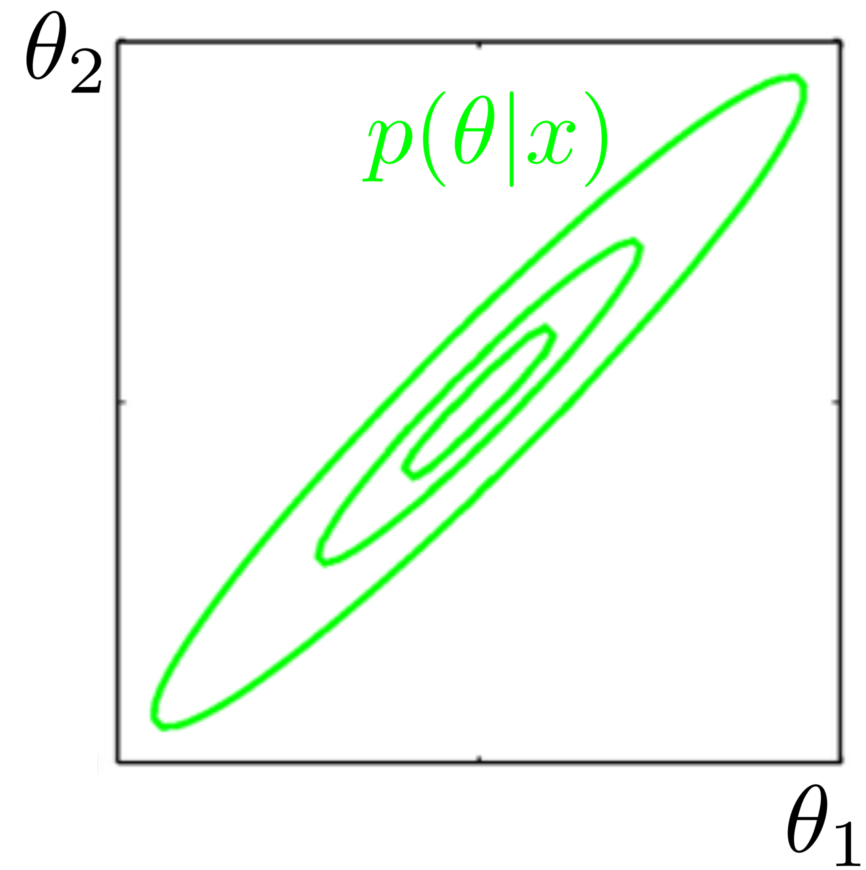


[Bishop 2006]

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]

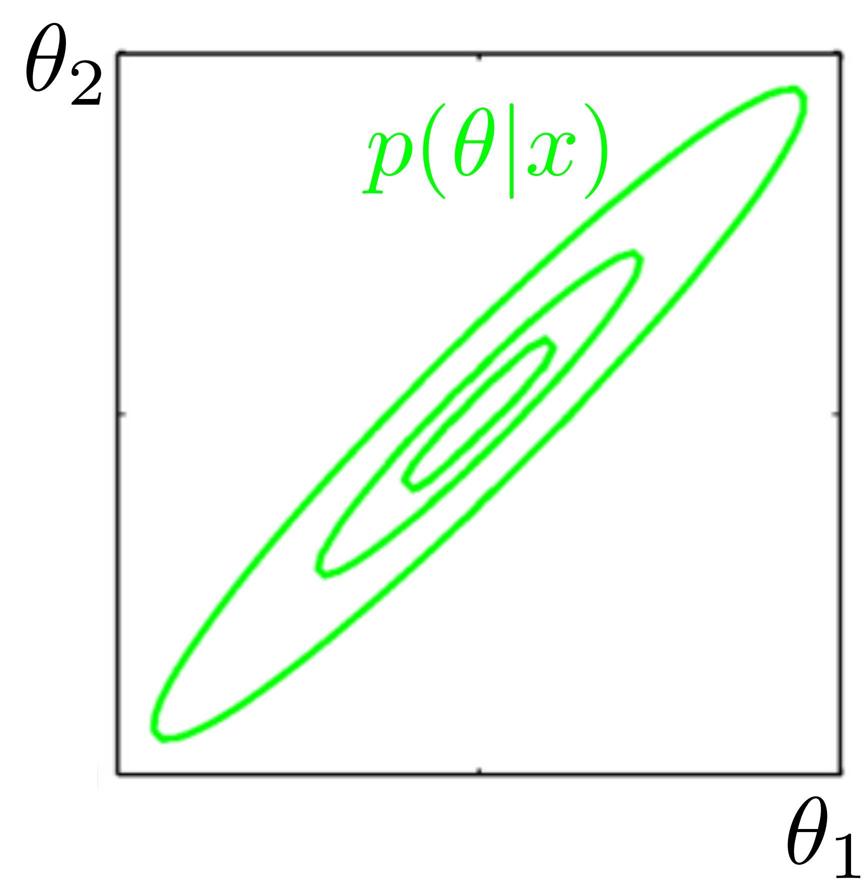
What about uncertainty?

- 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]

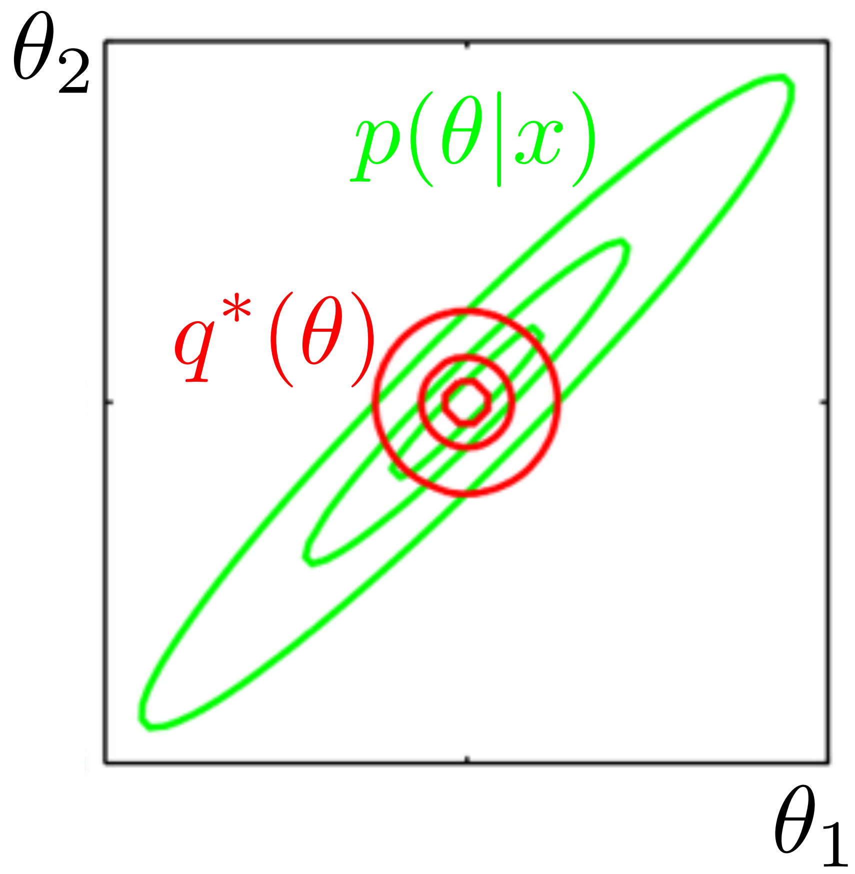
What about uncertainty?

- 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?

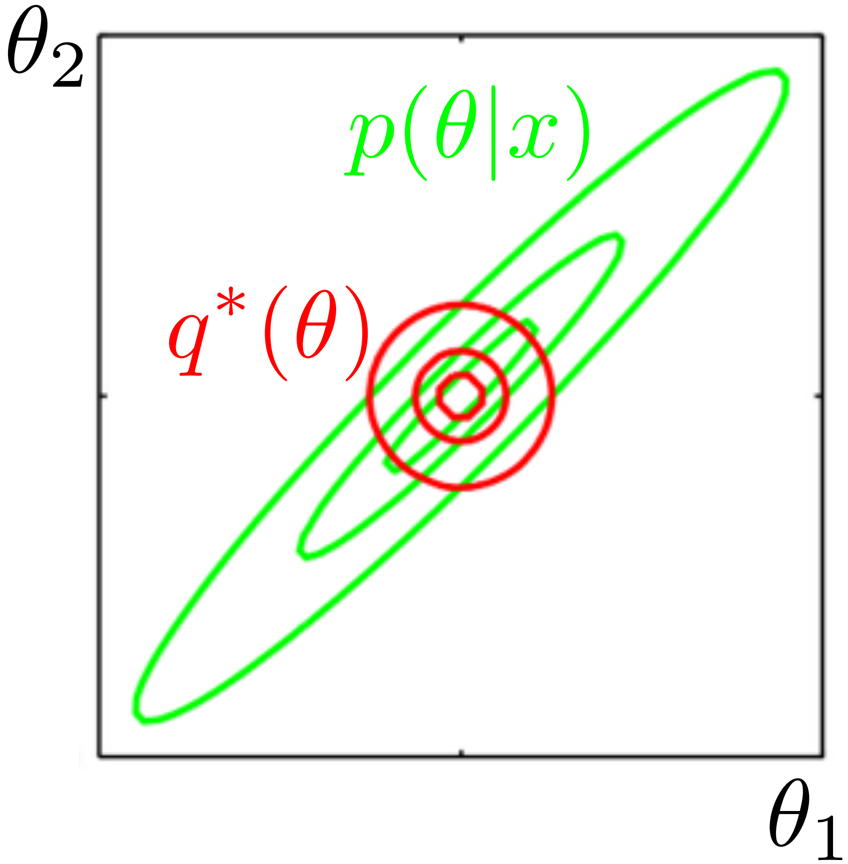
- Variational Bayes

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

- 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

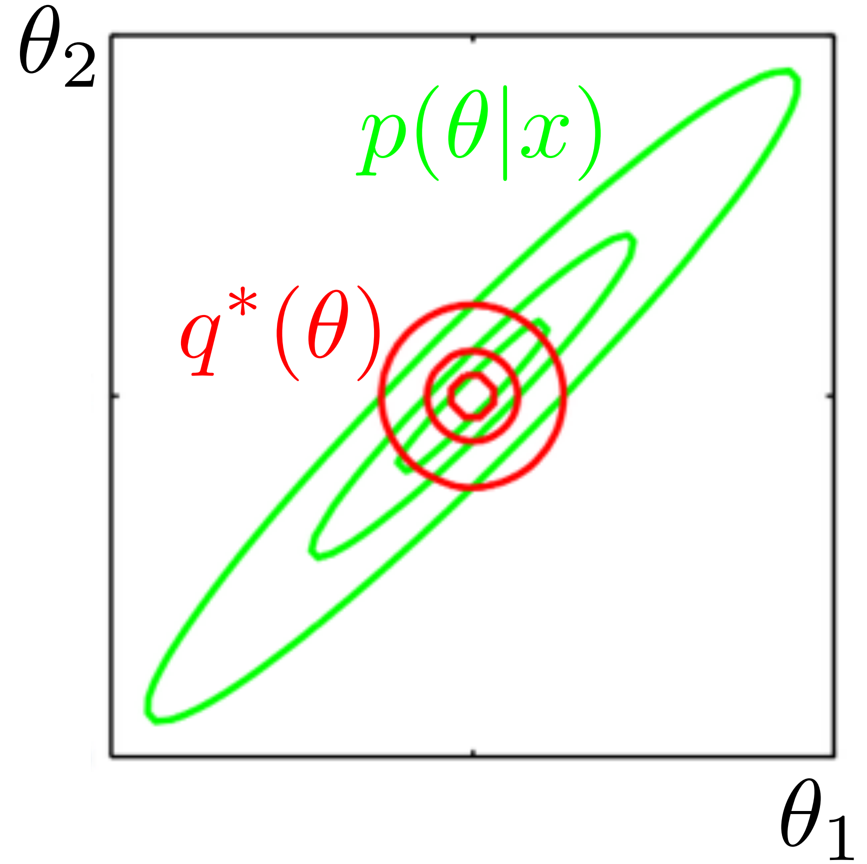
$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)

- No covariance estimates



[Bishop 2006]

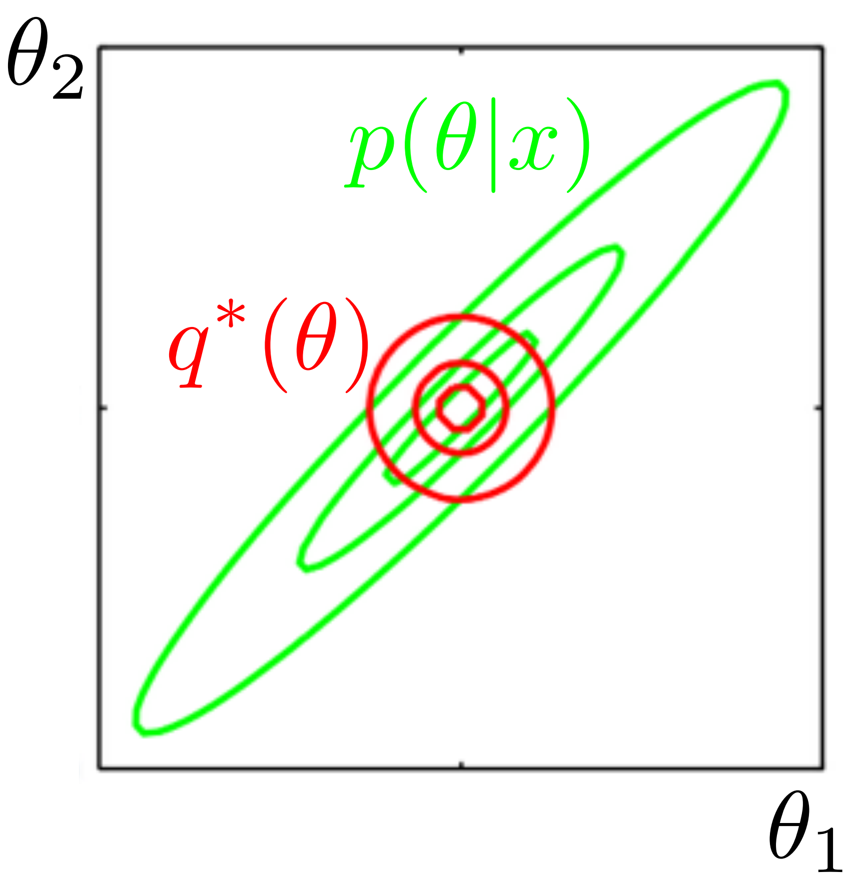
What about uncertainty?

- 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?

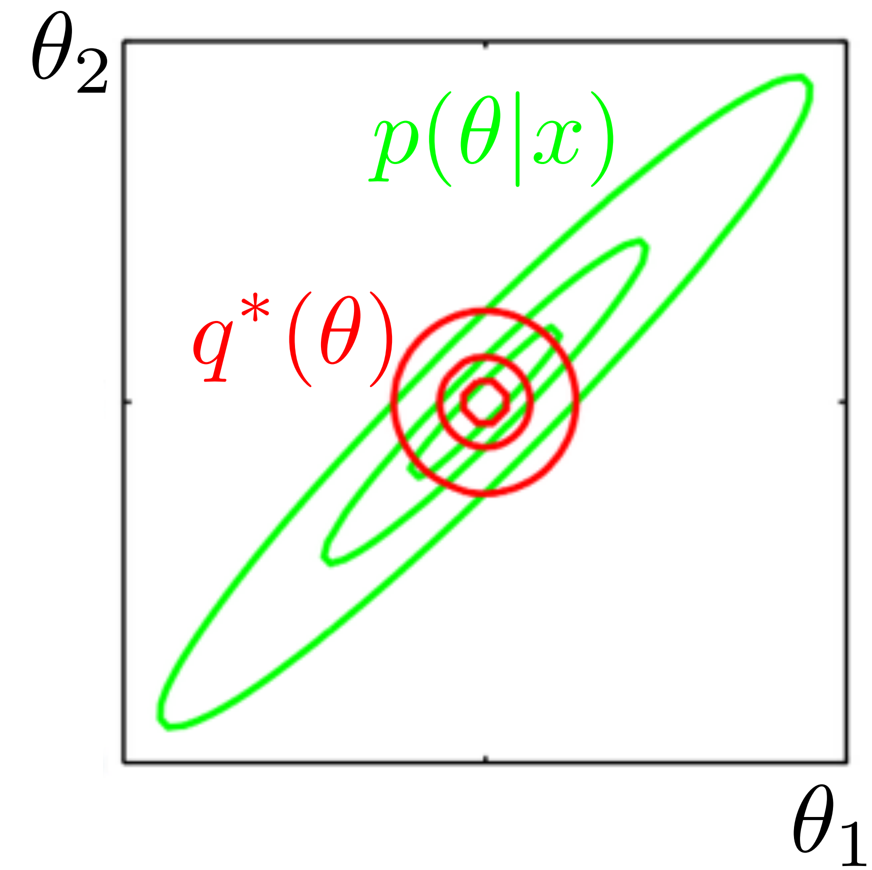
- Variational Bayes

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

- 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]
[Fosdick 2013; 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

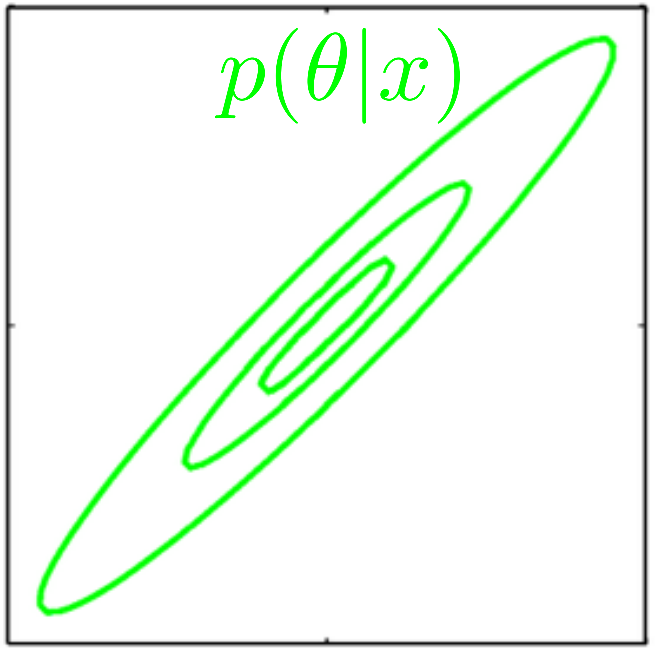
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

Linear response

- Cumulant-generating function

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

- True posterior covariance



[Bishop 2006]

Linear response

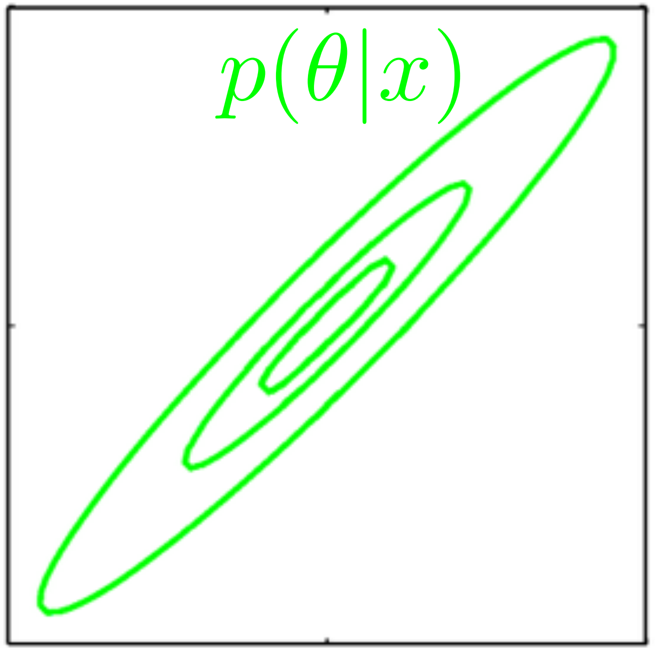
- Cumulant-generating function

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

$$\text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- True posterior covariance

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



[Bishop 2006]

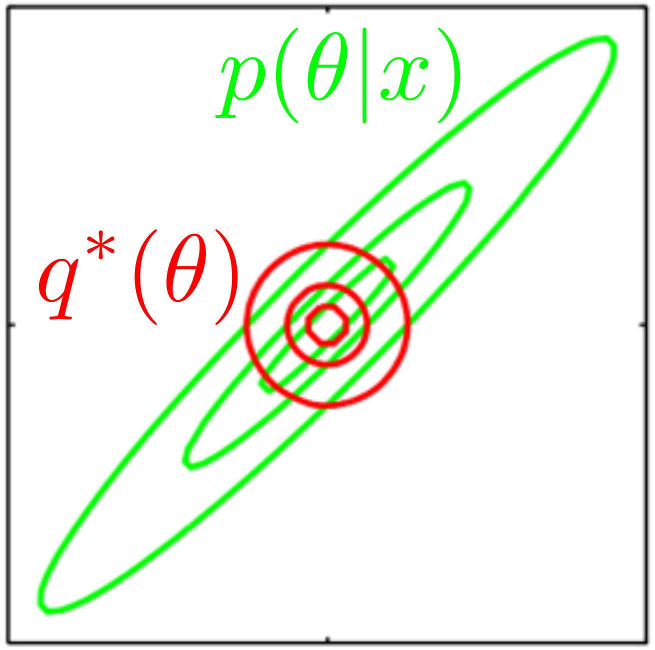
Linear response

- 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]

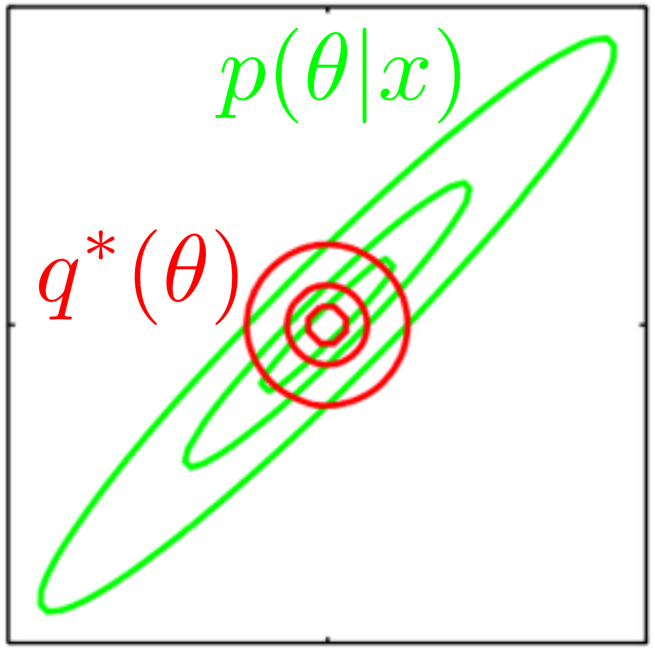
Linear response

- 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

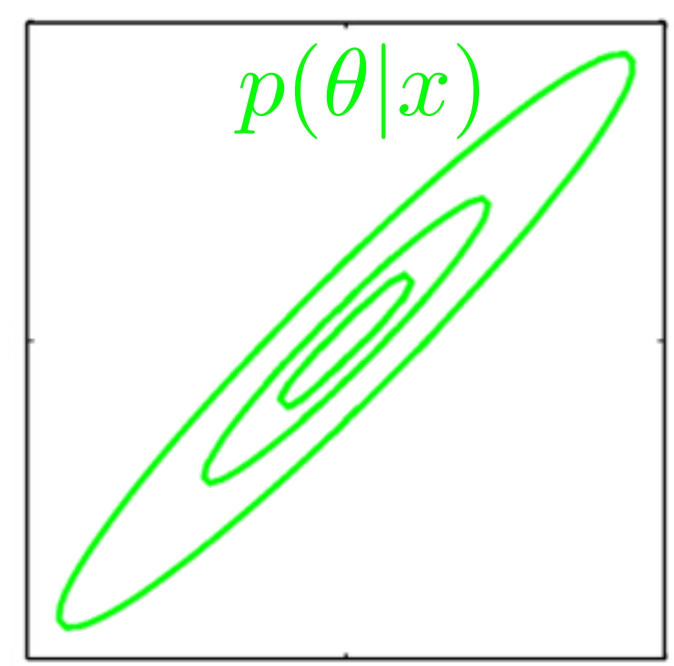
- Cumulant-generating function

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

- 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

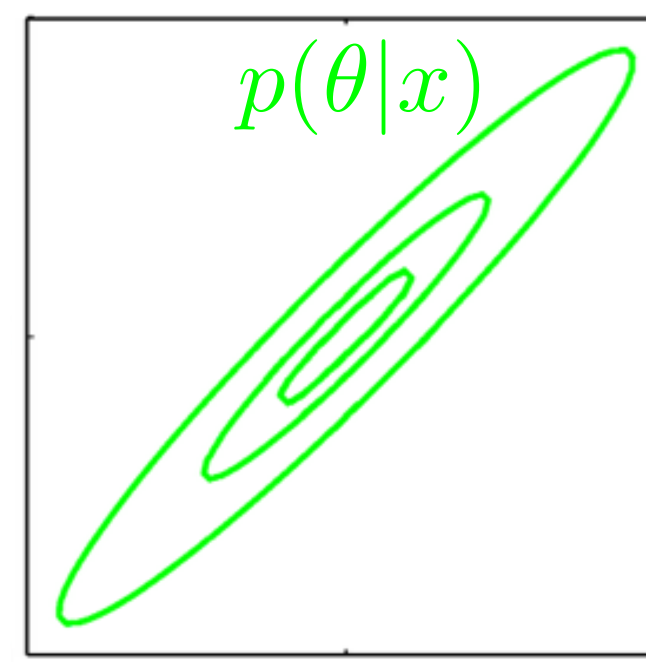
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- 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

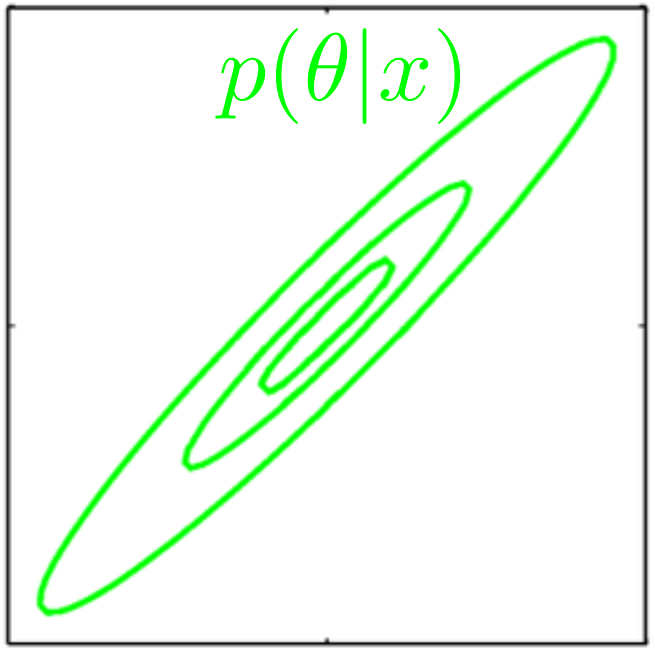
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- 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

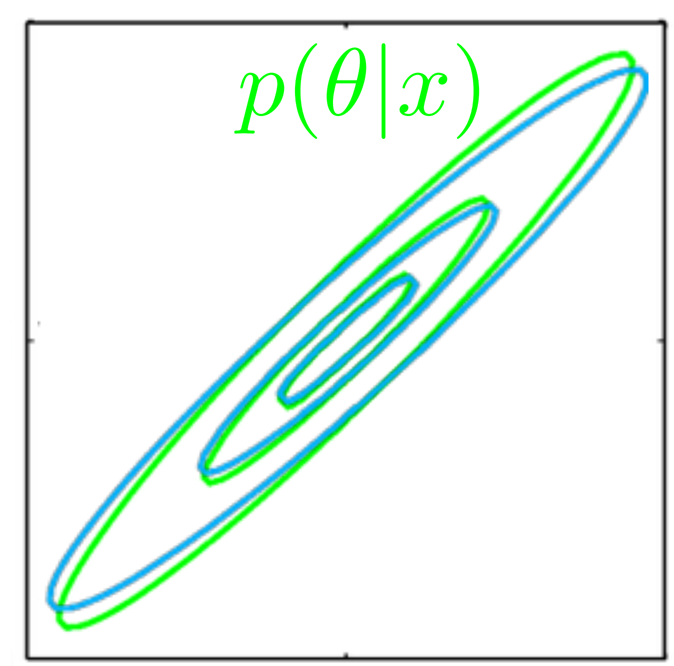
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- 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"

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



[Bishop 2006]

Linear response

- Cumulant-generating function

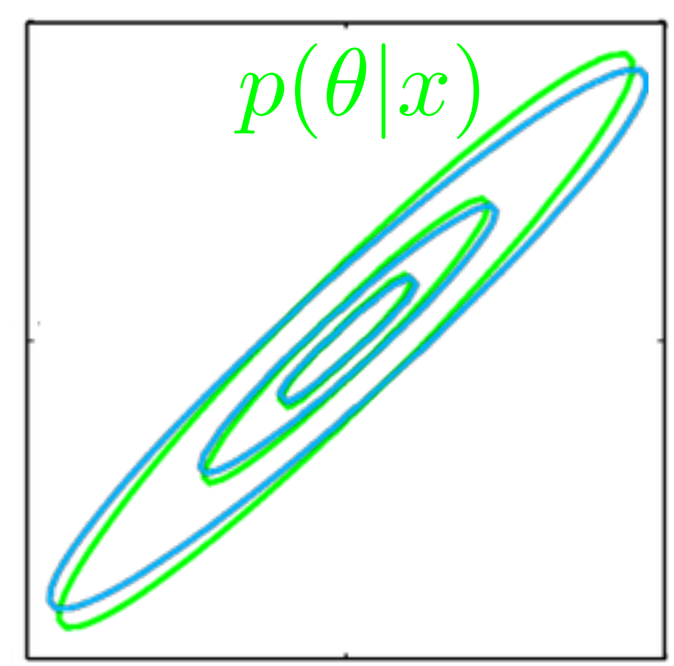
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- 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

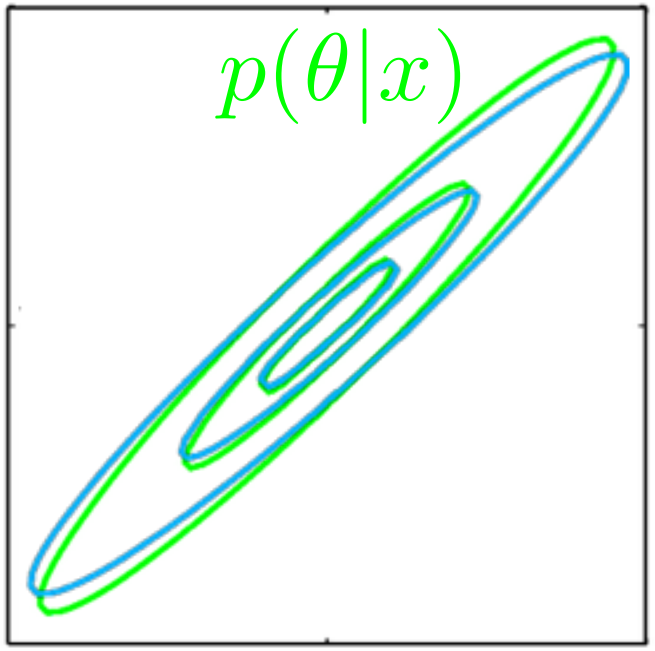
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- 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

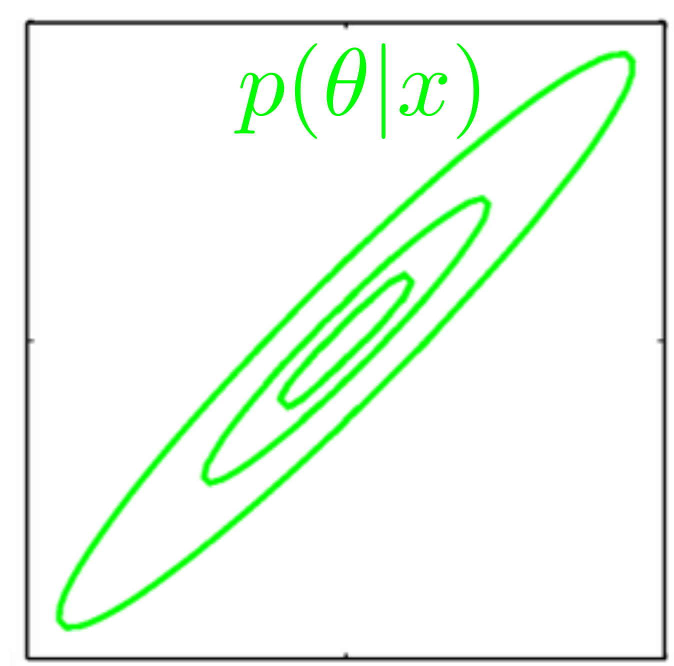
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- 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"

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



[Bishop 2006]

Linear response

- Cumulant-generating function

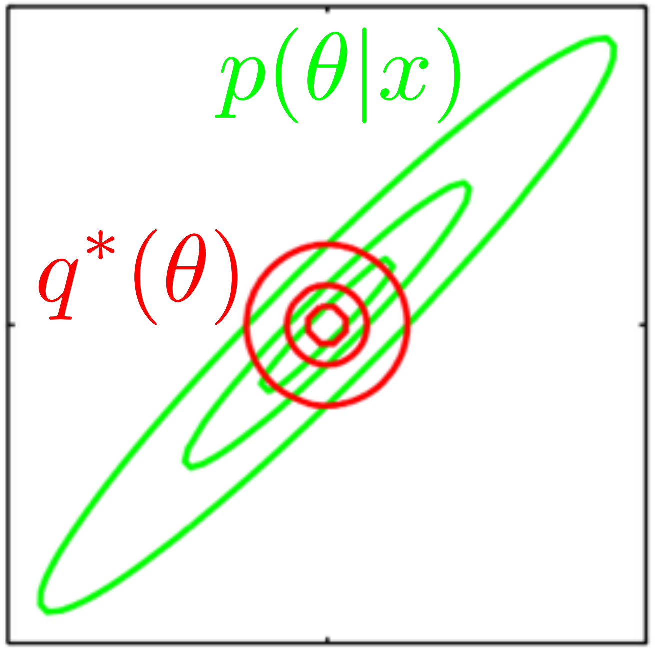
$$C(t) := \log \mathbb{E} e^{t^T \theta} \qquad \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0}$$

- 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

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

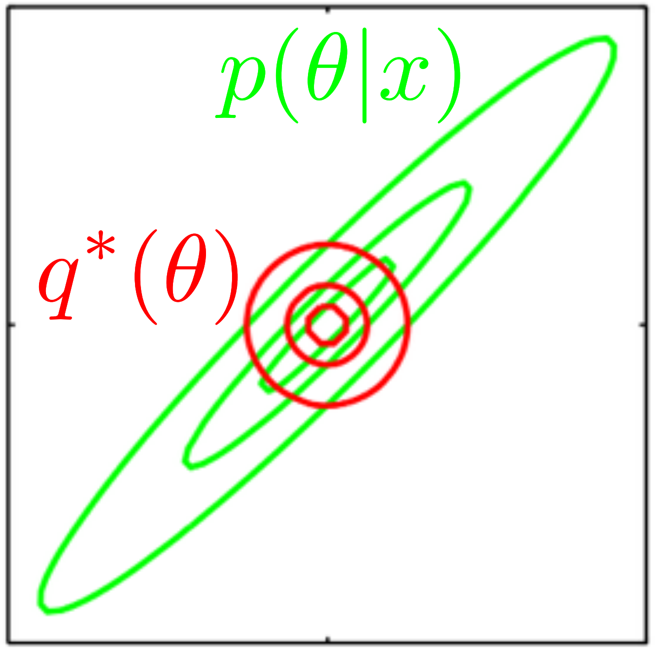
- True posterior covariance vs MFVB covariance

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

- "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

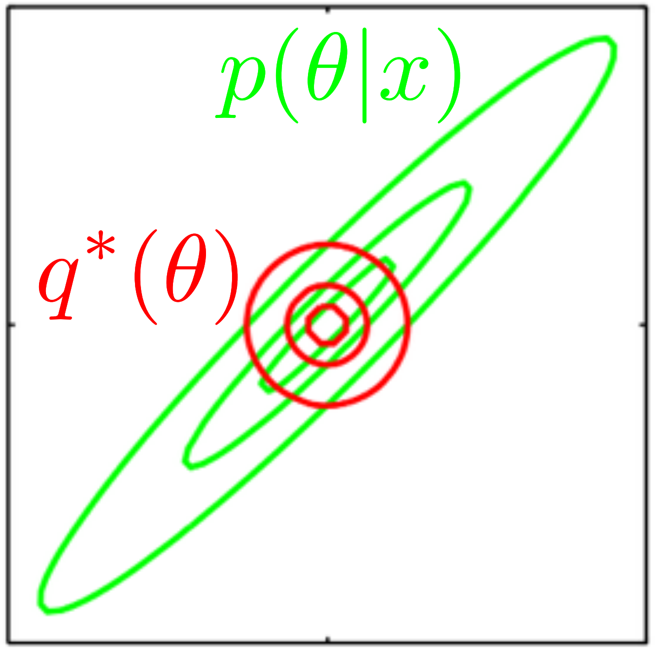
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "Linear response"

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

- The LRVB approximation

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



[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

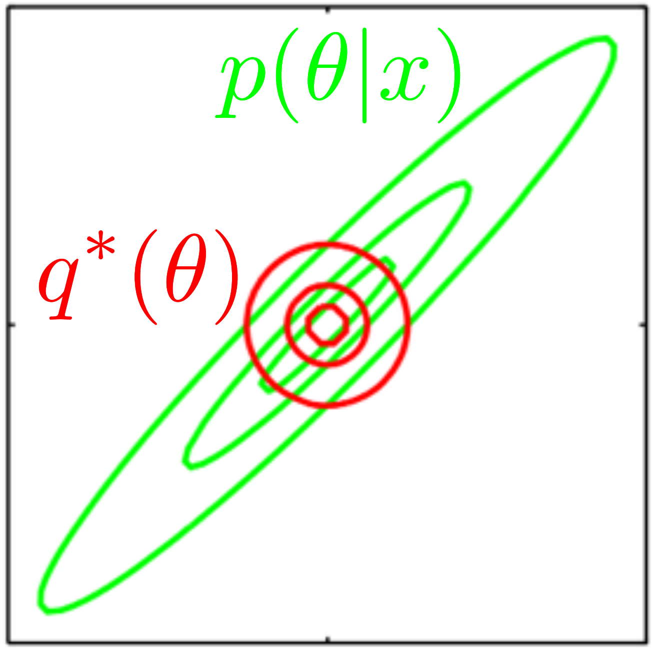
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "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

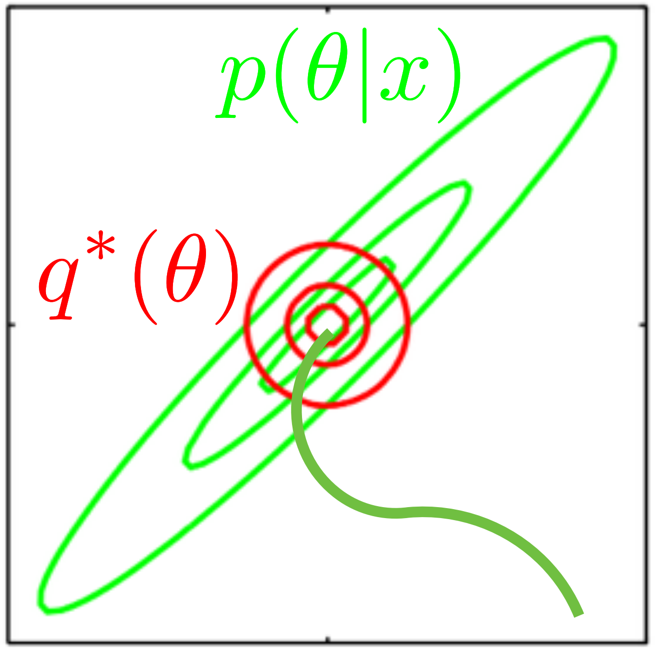
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "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

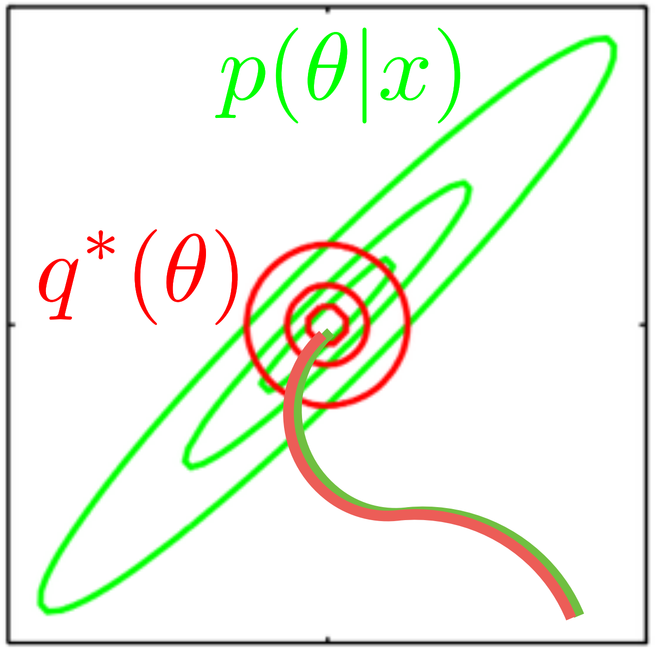
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "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

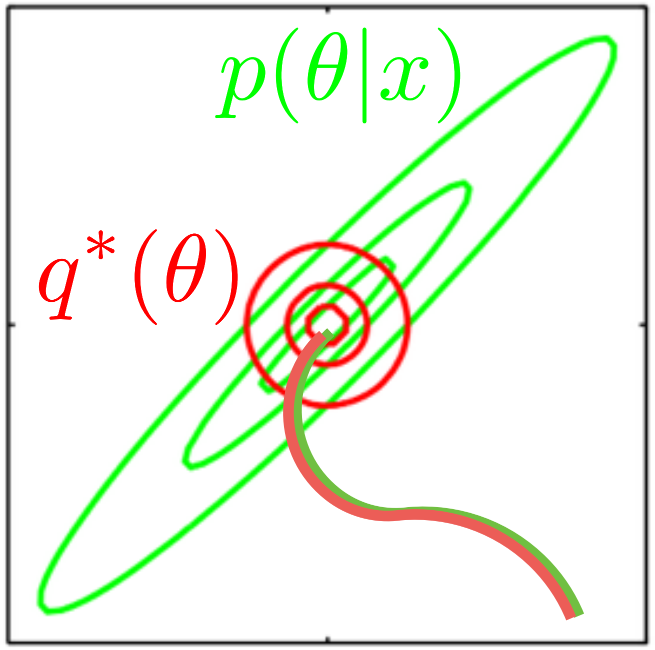
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "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

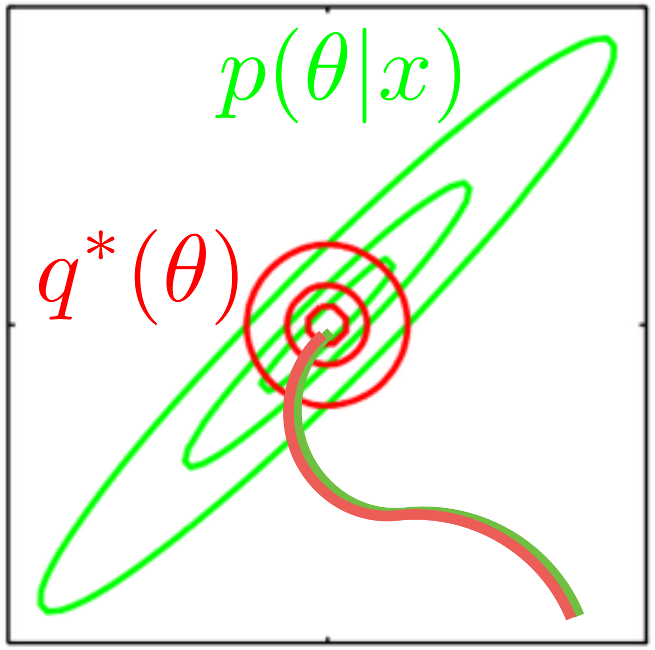
$$\Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \qquad V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0}$$

- "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

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

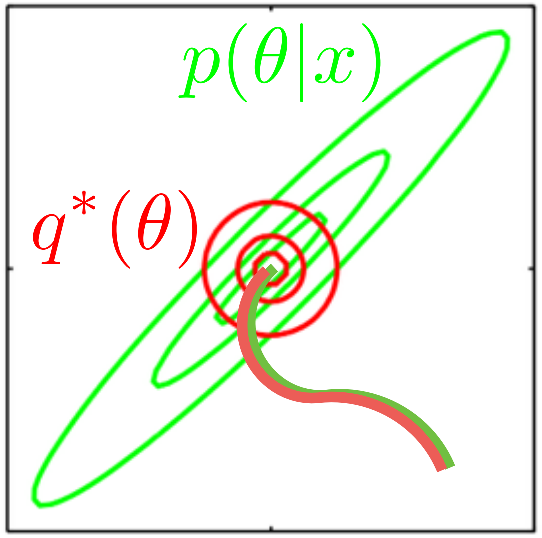
- Symmetric and positive definite at local min of KL
- The LRVB assumption: $\mathbb{E}_{p_t} \theta \approx \mathbb{E}_{q_t^*} \theta$

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: $\mathbb{E}_{p_t} \theta \approx \mathbb{E}_{q_t^*} \theta$



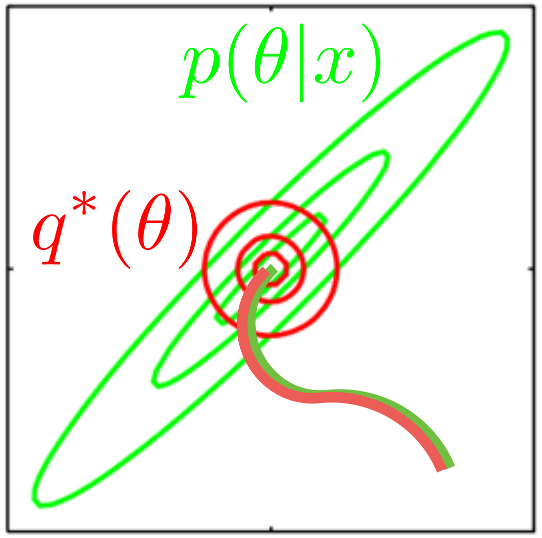
[Bishop 2006]

LRVB estimator

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

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

- 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

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

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

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

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 →

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

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

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

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

1 if microcredit

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

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

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

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

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

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

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

- Priors and hyperpriors:

Microcredit Experiment

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

1 if microcredit

profit

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

- Priors and hyperpriors:

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

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

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

- Priors and hyperpriors:

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

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

Microcredit Experiment

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

1 if microcredit

profit

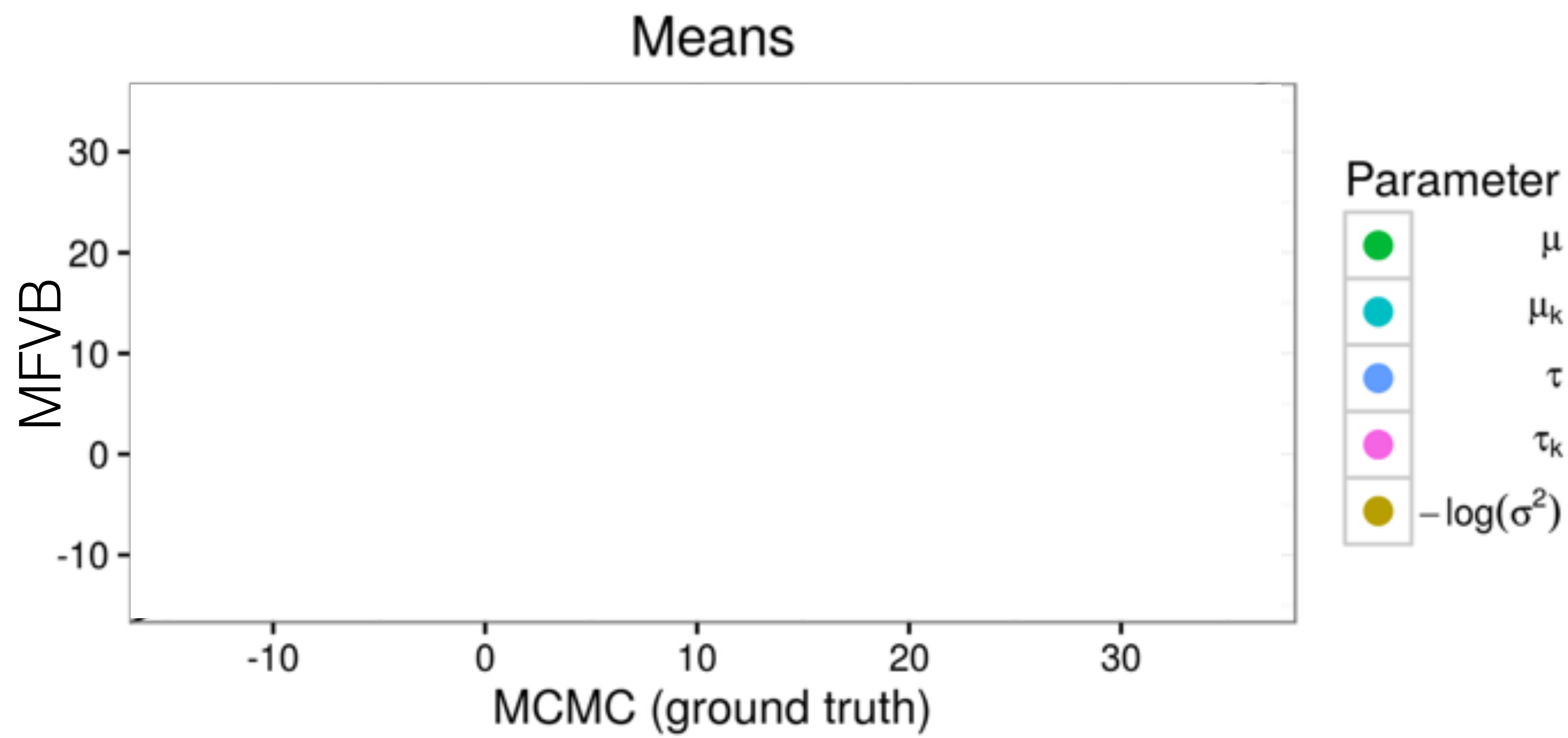
$$y_{kn} \stackrel{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

- Priors and hyperpriors:

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

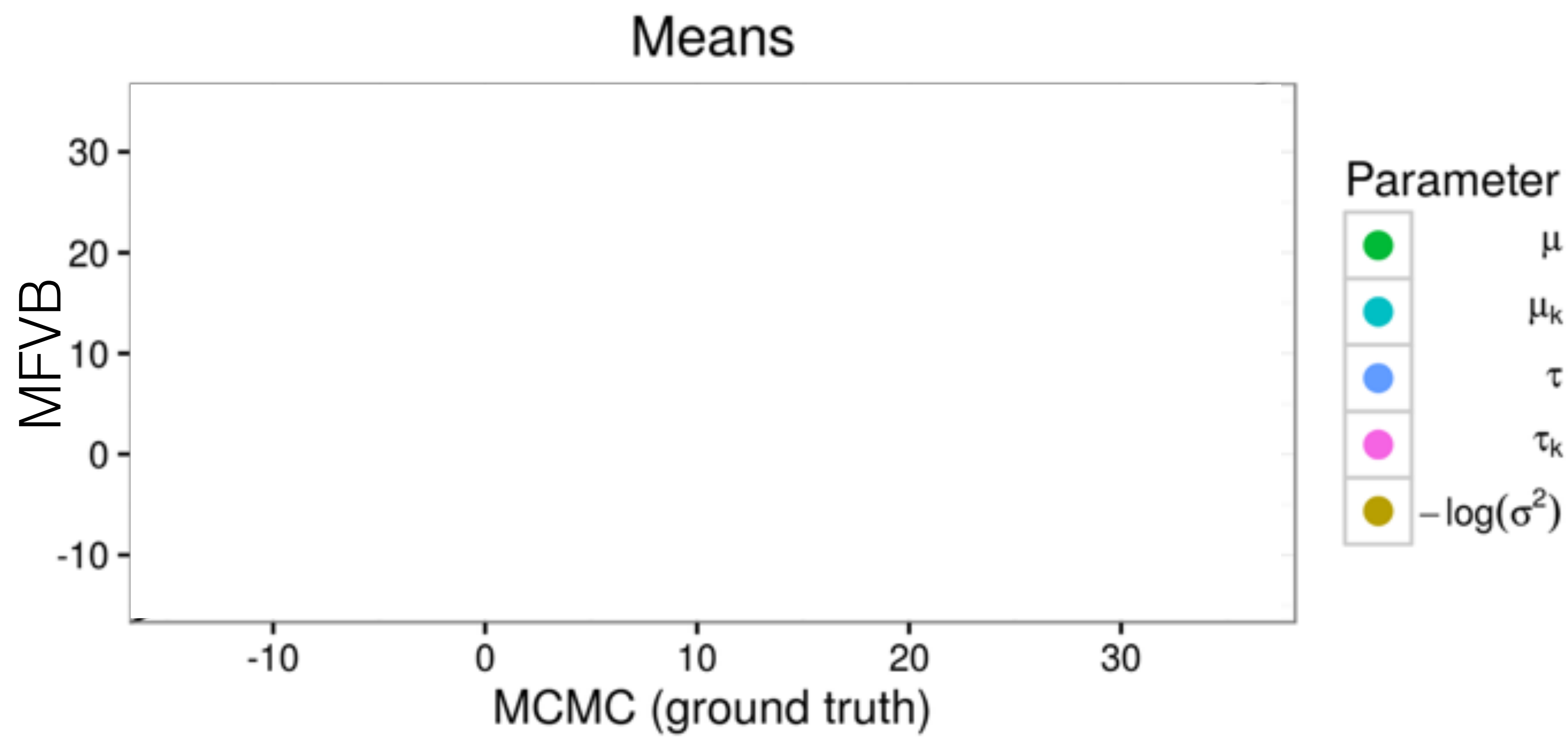
$$\sigma_k^{-2} \stackrel{iid}{\sim} \Gamma(a, b) \qquad C \sim \text{Sep\&LKJ}(\eta, c, d)$$

Microcredit Experiment



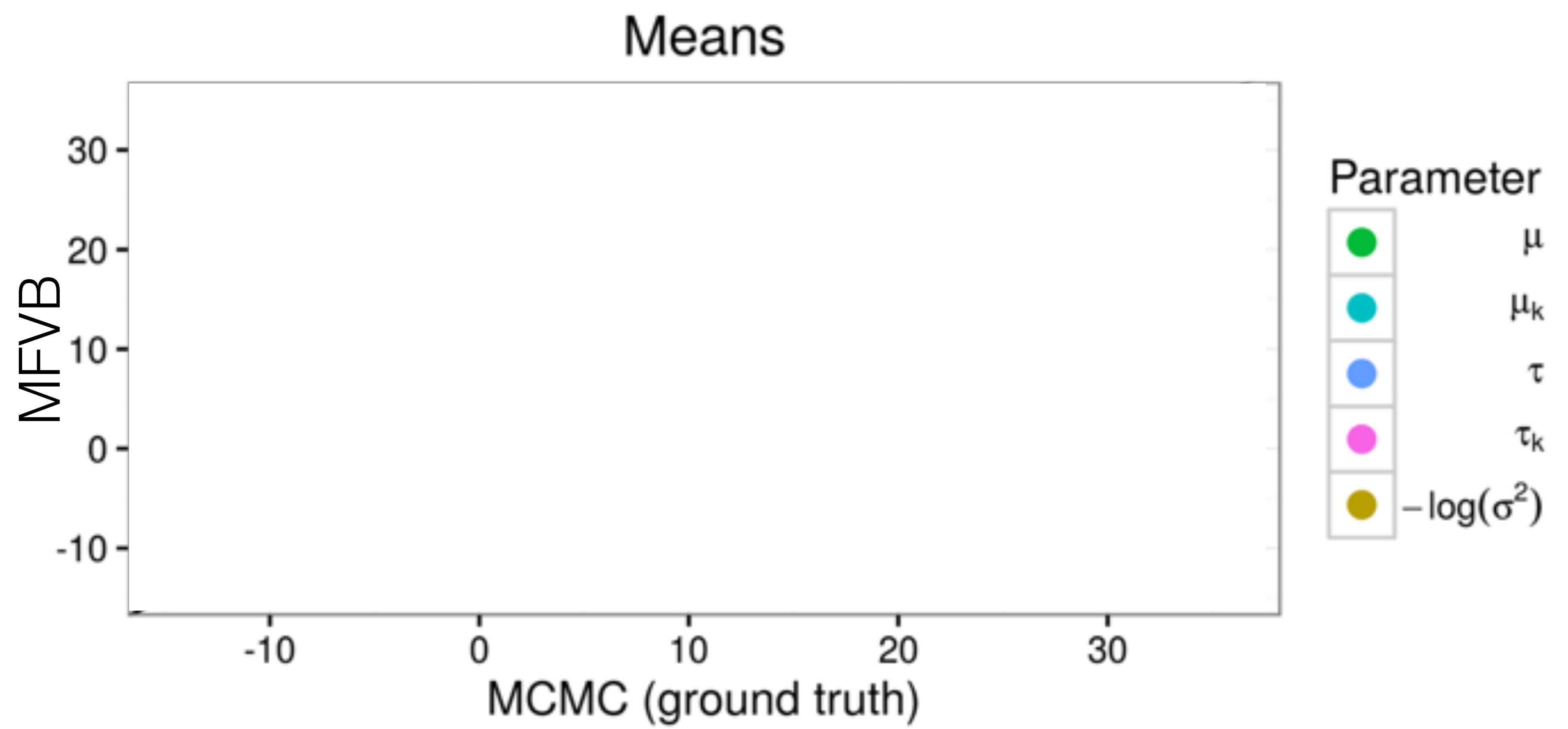
Microcredit Experiment

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



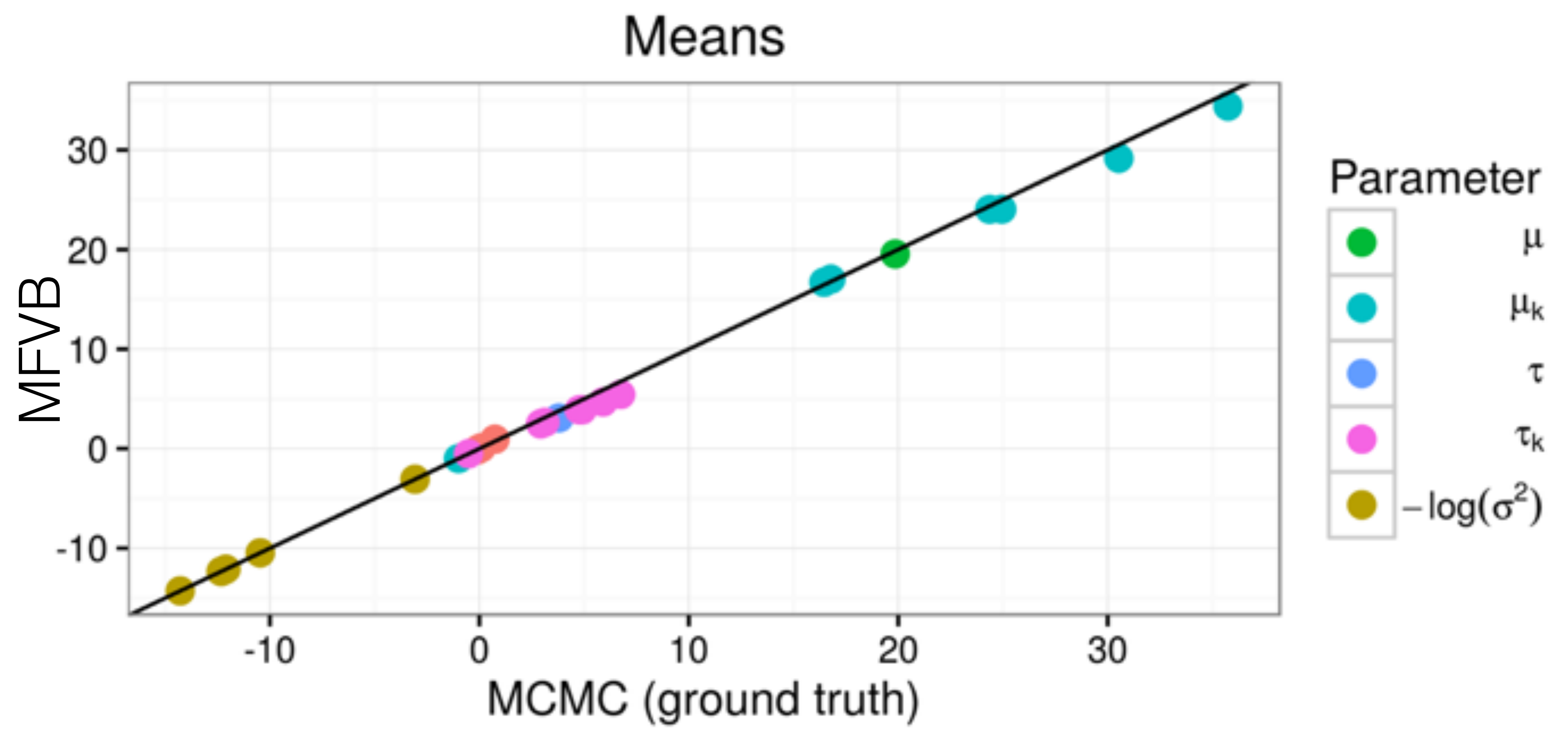
Microcredit Experiment

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



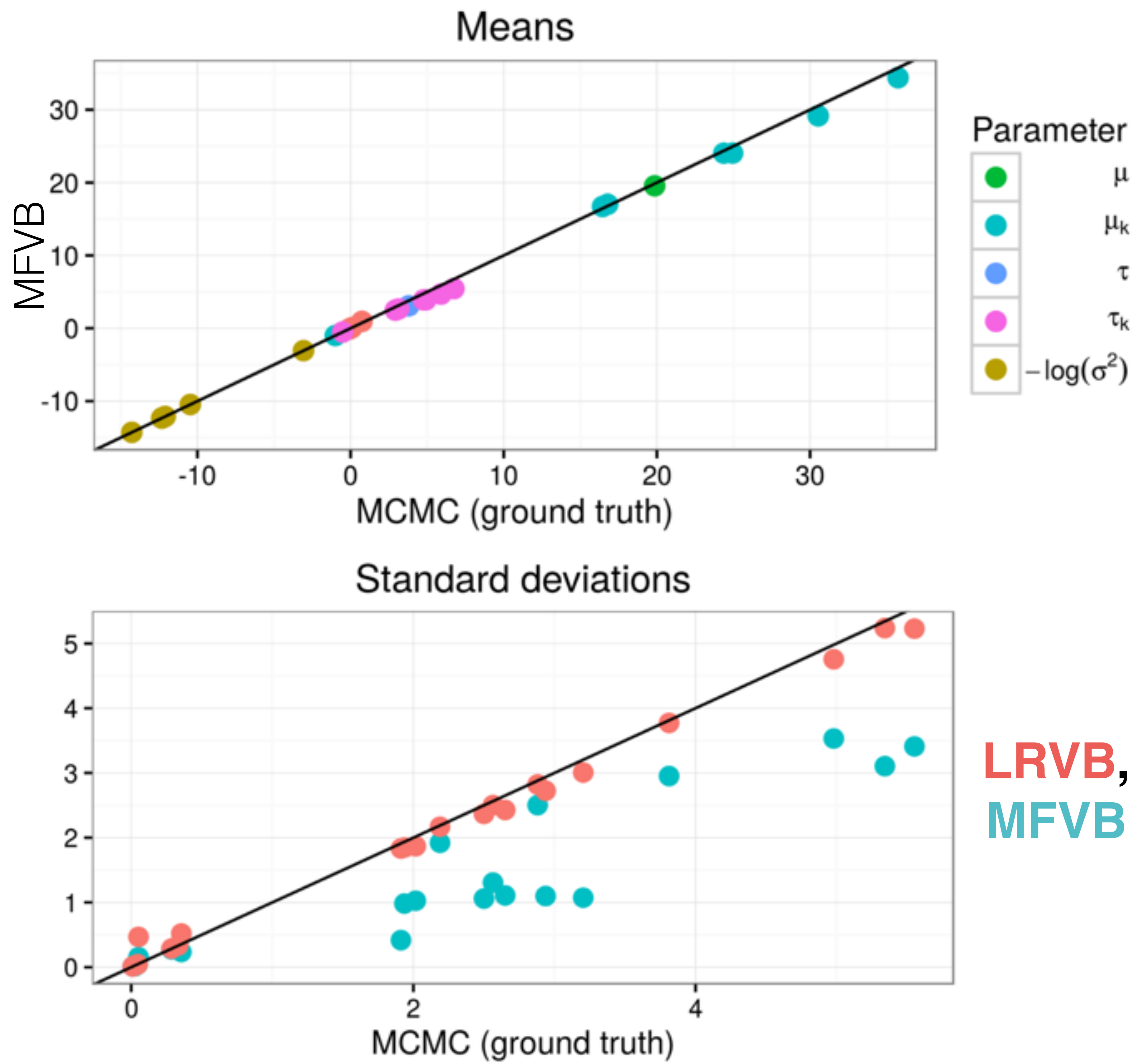
Microcredit Experiment

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



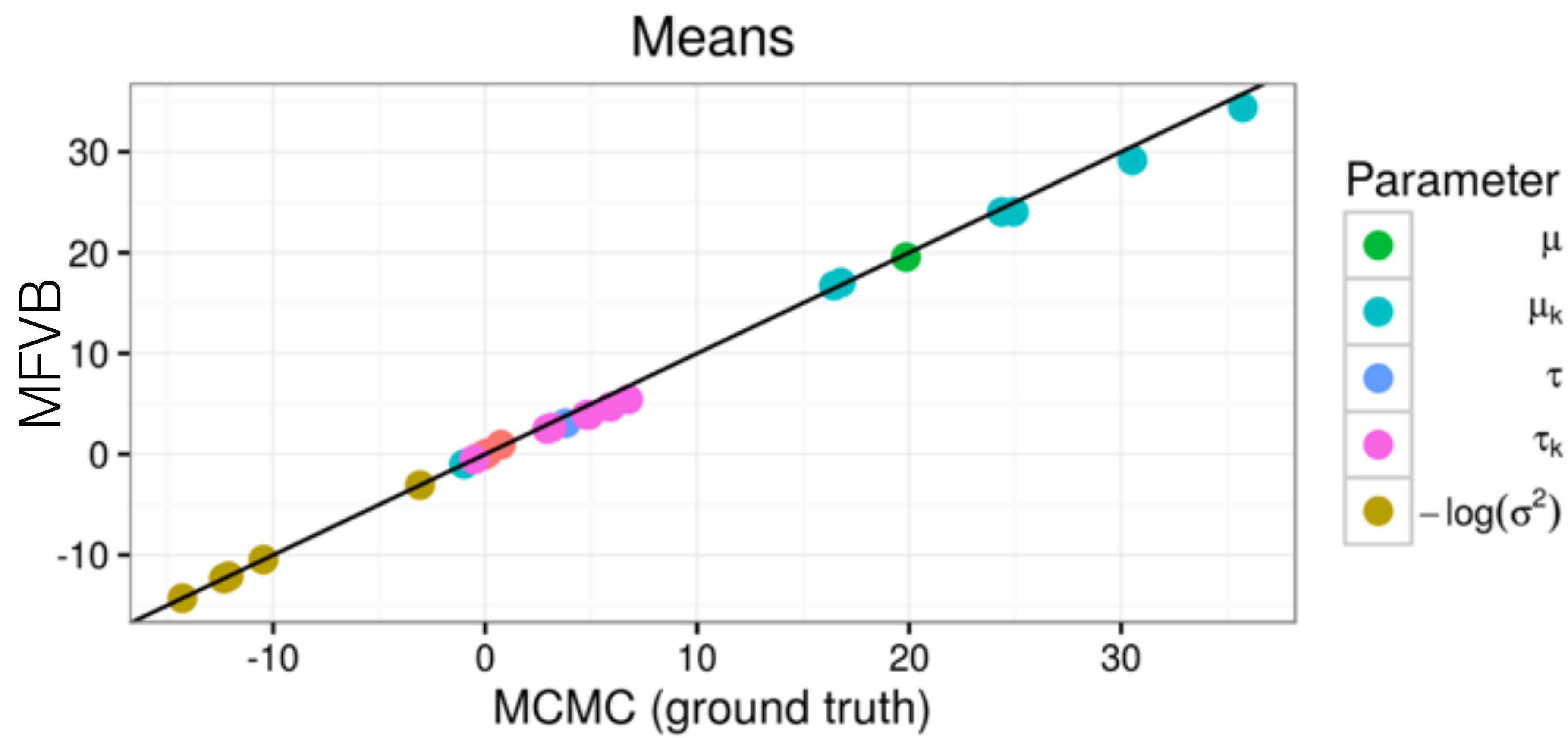
Microcredit Experiment

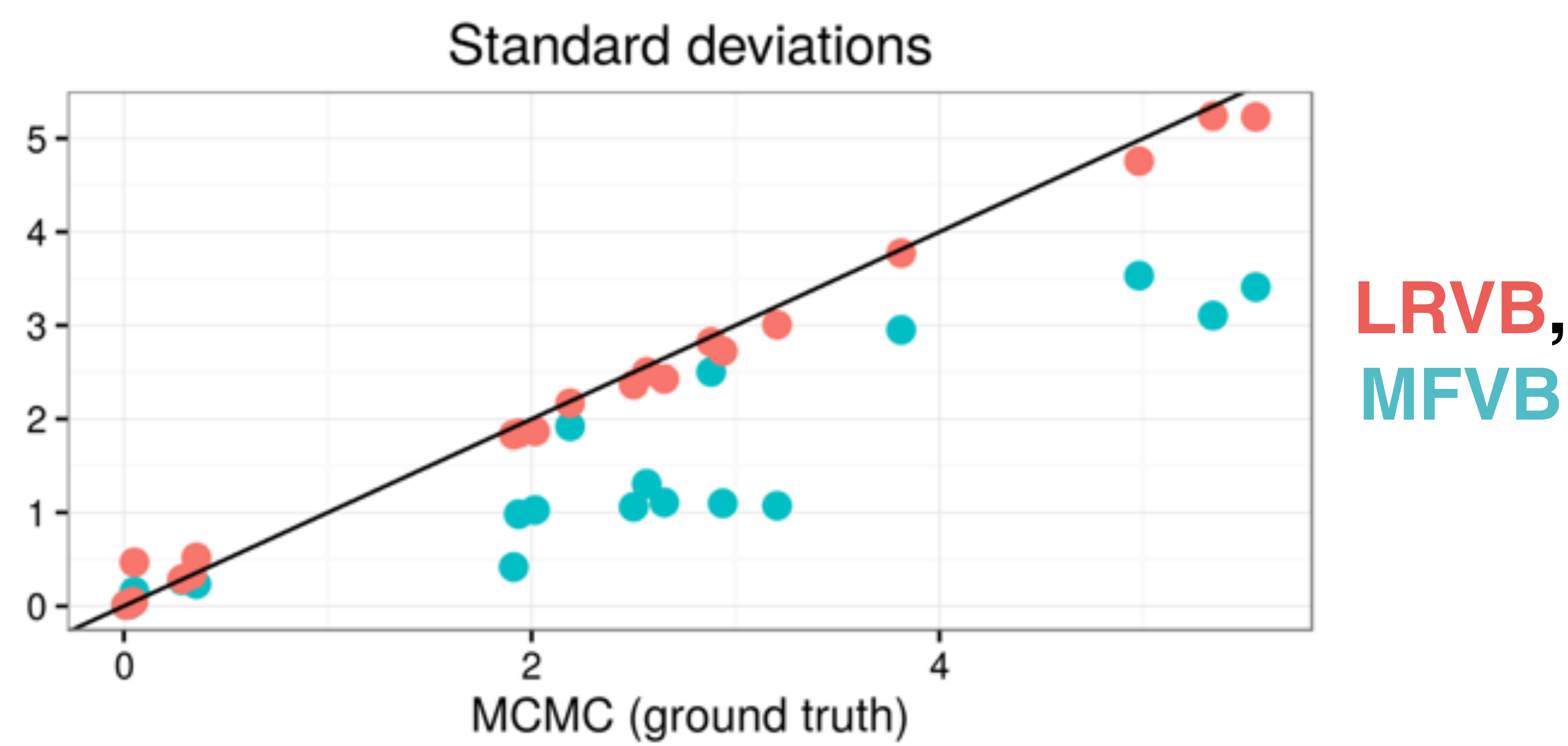
- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**



Microcredit Experiment

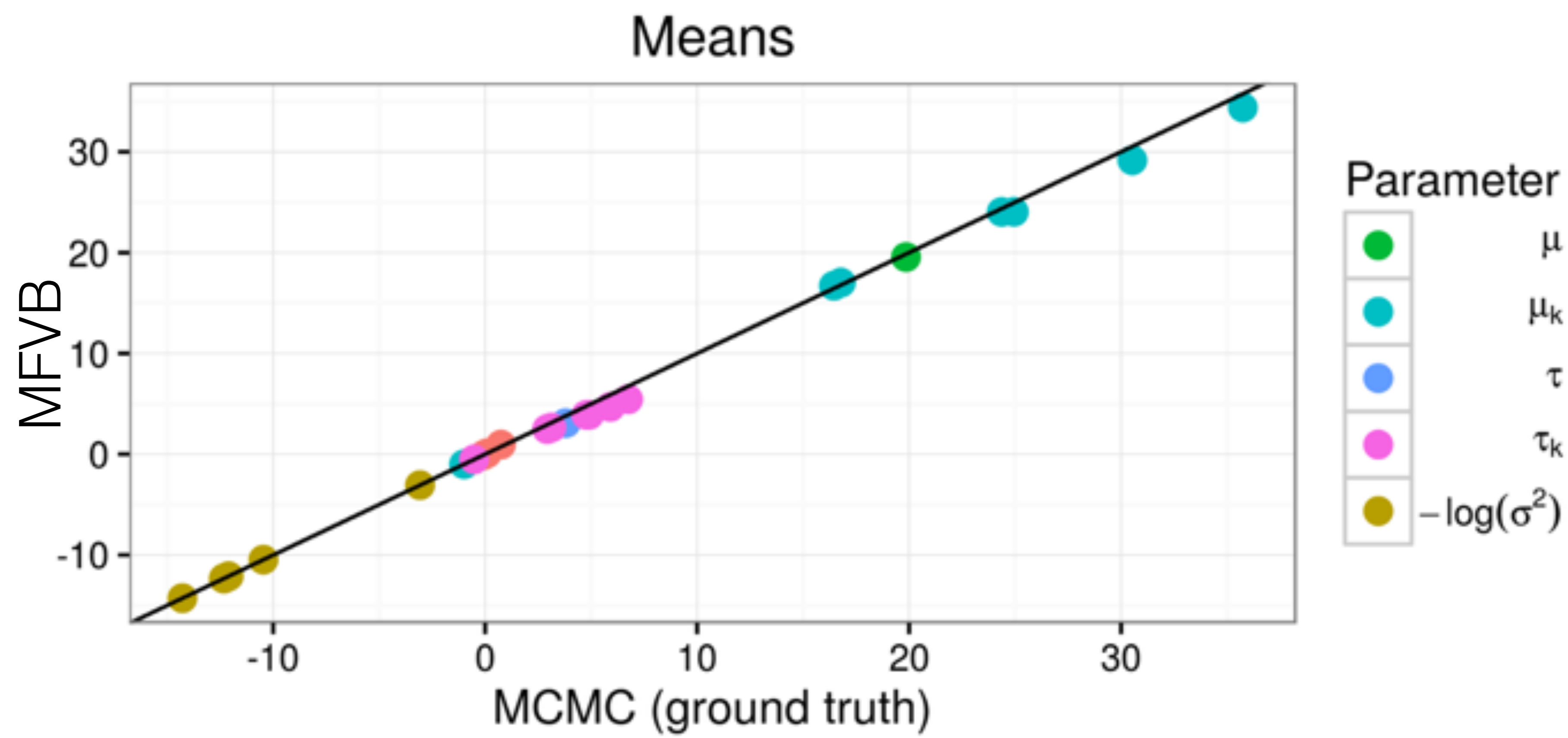
- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**
- τ mean (MFVB): 3.08 USD PPP

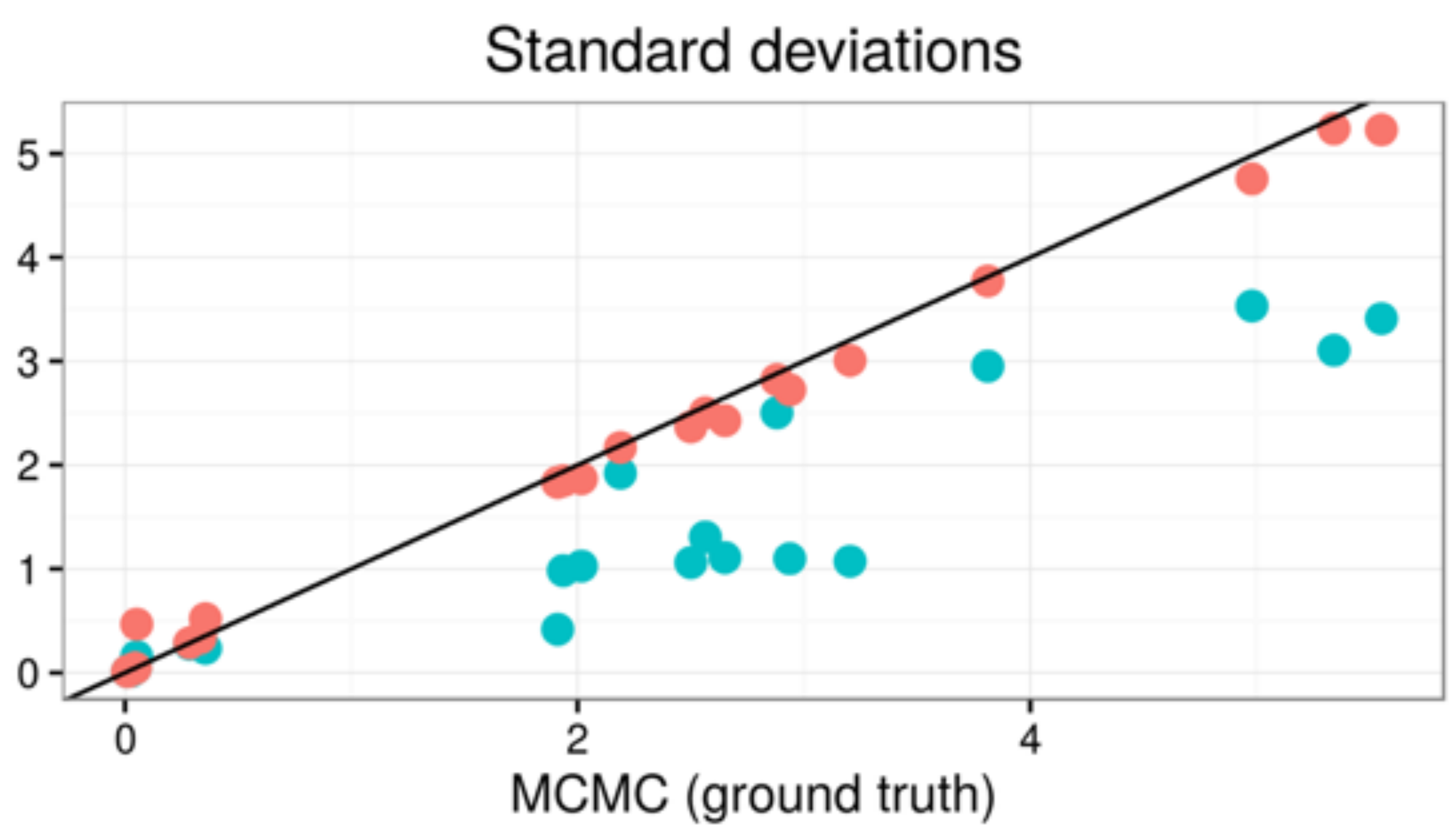




Microcredit Experiment

- *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

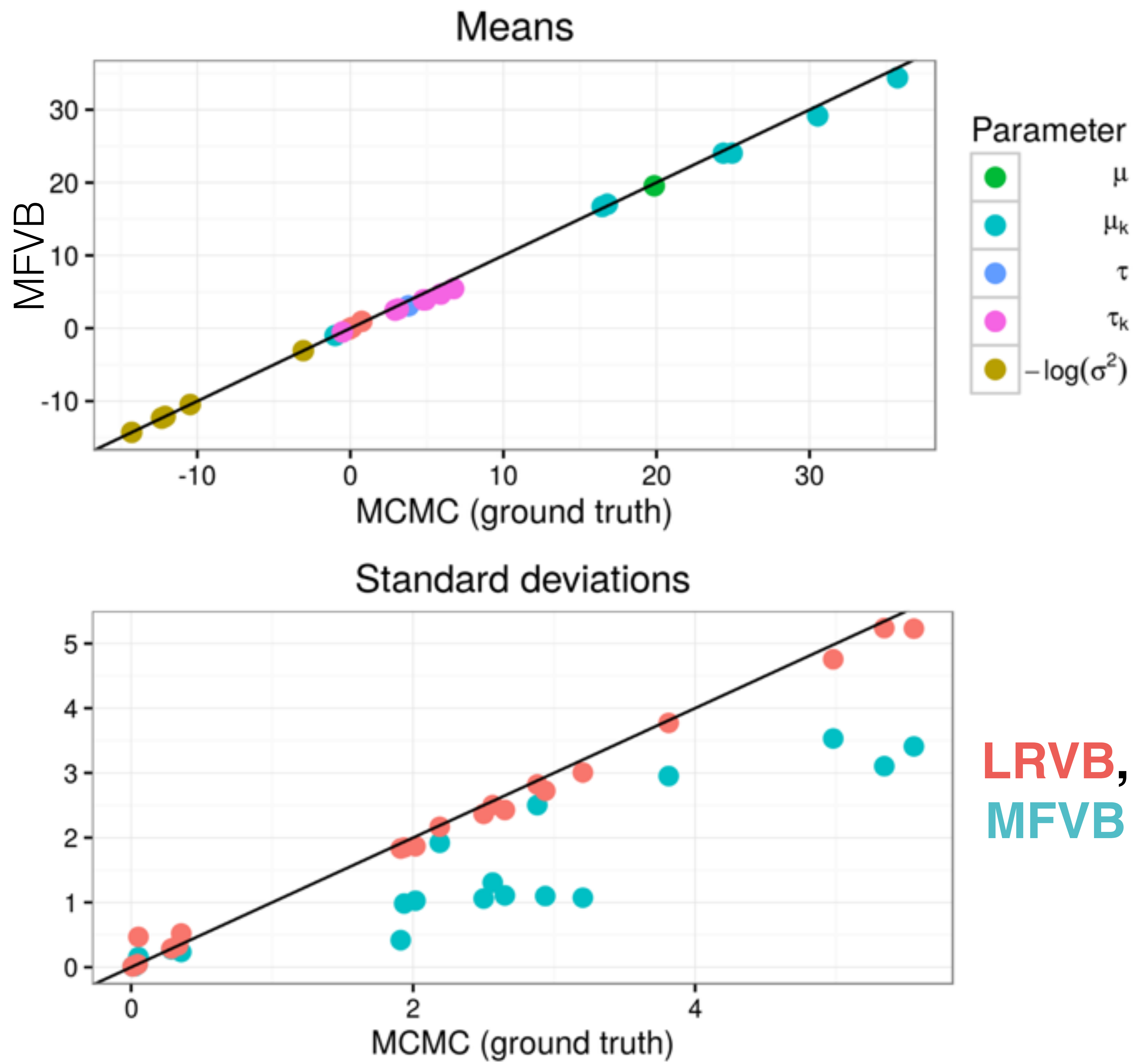




LRVB, **MFVB**

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



Experiments

Experiments

- Gaussian mixture model

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

with conjugate priors on π, μ, Λ

Experiments

- Gaussian mixture model

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

with conjugate priors on π, μ, Λ

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R `bayesm` package

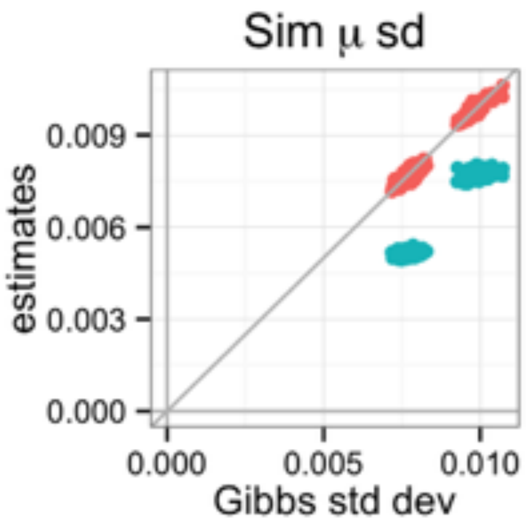
Experiments

- Gaussian mixture model

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

with conjugate priors on π, μ, Λ

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R `bayesm` package



LRVB, **MFVB**

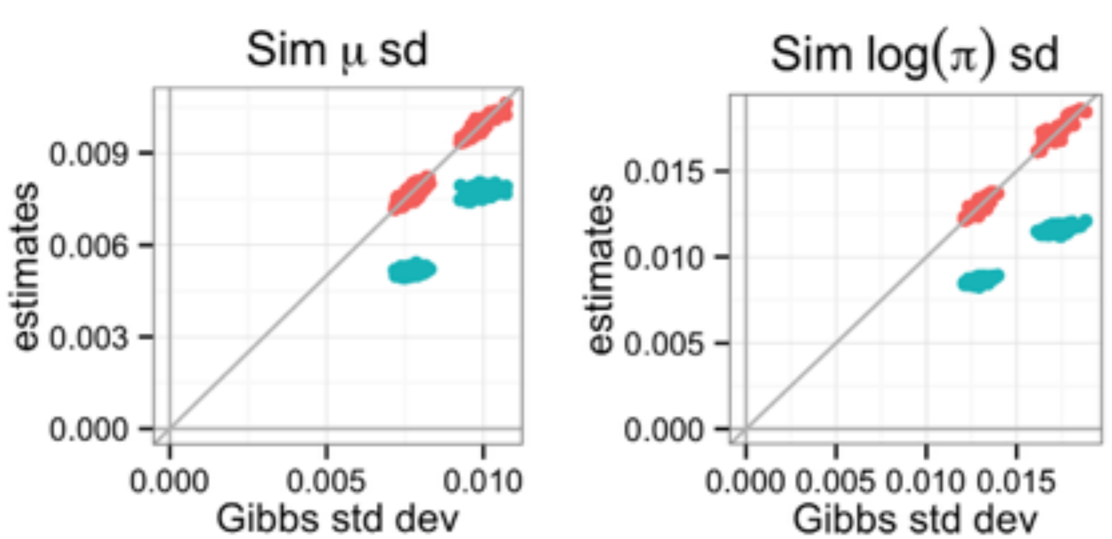
Experiments

- Gaussian mixture model

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

with conjugate priors on π, μ, Λ

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R `bayesm` package



LRVB, **MFVB**

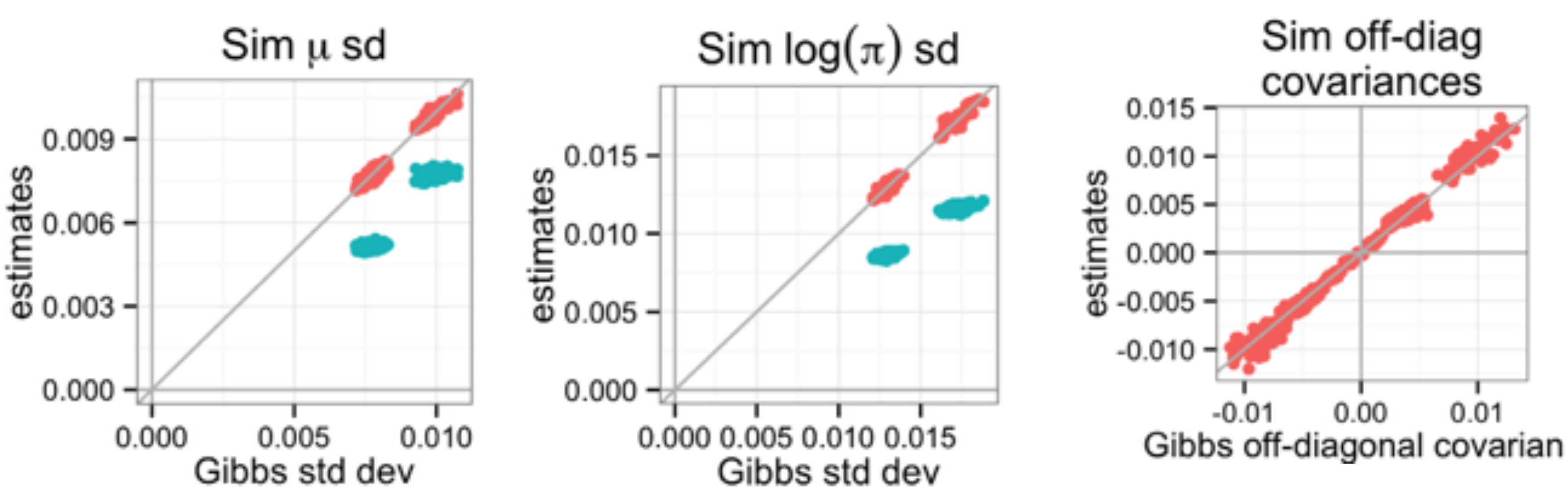
Experiments

- Gaussian mixture model

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

with conjugate priors on π, μ, Λ

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R `bayesm` package



LRVB, **MFVB**

Experiments

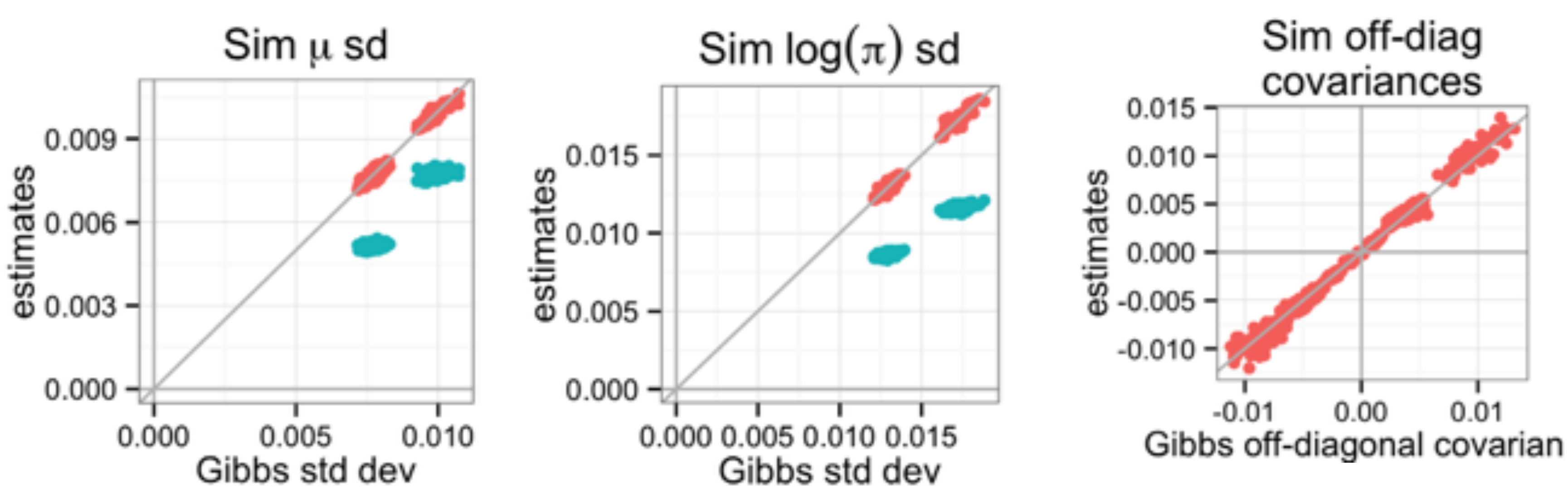
- Gaussian mixture model

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

with conjugate priors on π, μ, Λ

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R `bayesm` package

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



LRVB, **MFVB**

Experiments

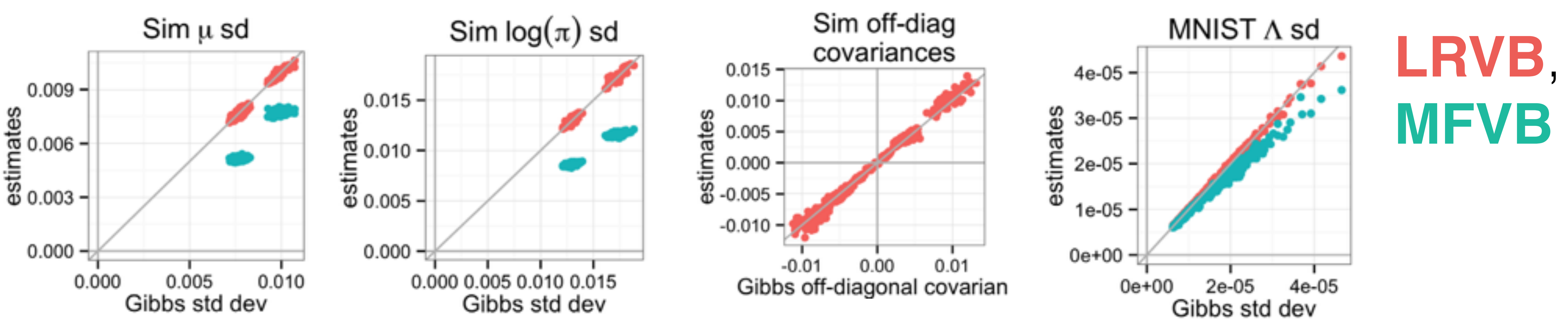
- Gaussian mixture model

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

  with conjugate priors on π, μ, Λ

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R `bayesm` package

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



Experiments

Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

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

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

Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

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

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

- 100 simulated data sets, 500 data points each, R `MCMCglmm` package

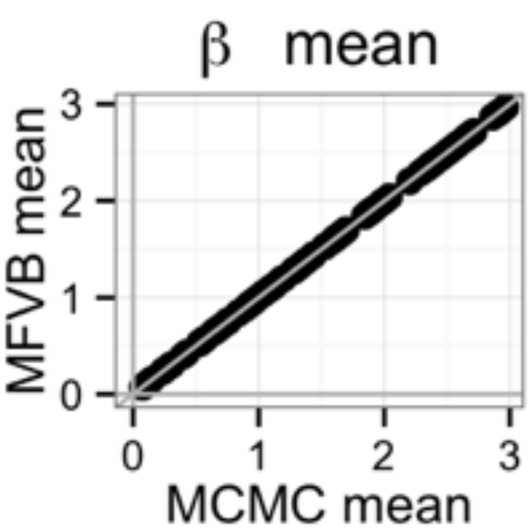
Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

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

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

- 100 simulated data sets, 500 data points each, R `MCMCglmm` package



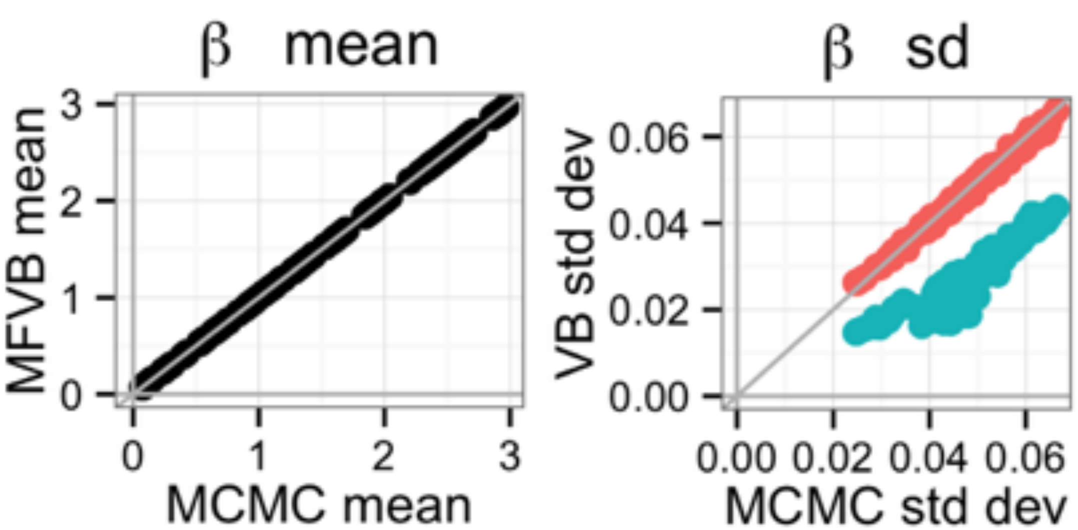
Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

$$z_n|\beta,\tau \overset{indep}{\sim} \mathcal{N}\left(z_n|\beta x_n,\tau^{-1}\right), \quad y_n|z_n \overset{indep}{\sim} \text{Poisson}\left(y_n|\exp(z_n)\right),$$
$$\beta \sim \mathcal{N}(\beta|0,\sigma_\beta^2), \quad \tau \sim \text{Gamma}(\tau|\alpha_\tau,\beta_\tau)$$

- 100 simulated data sets, 500 data points each, R `MCMCglmm` package

LRVB, **MFVB**



Experiments

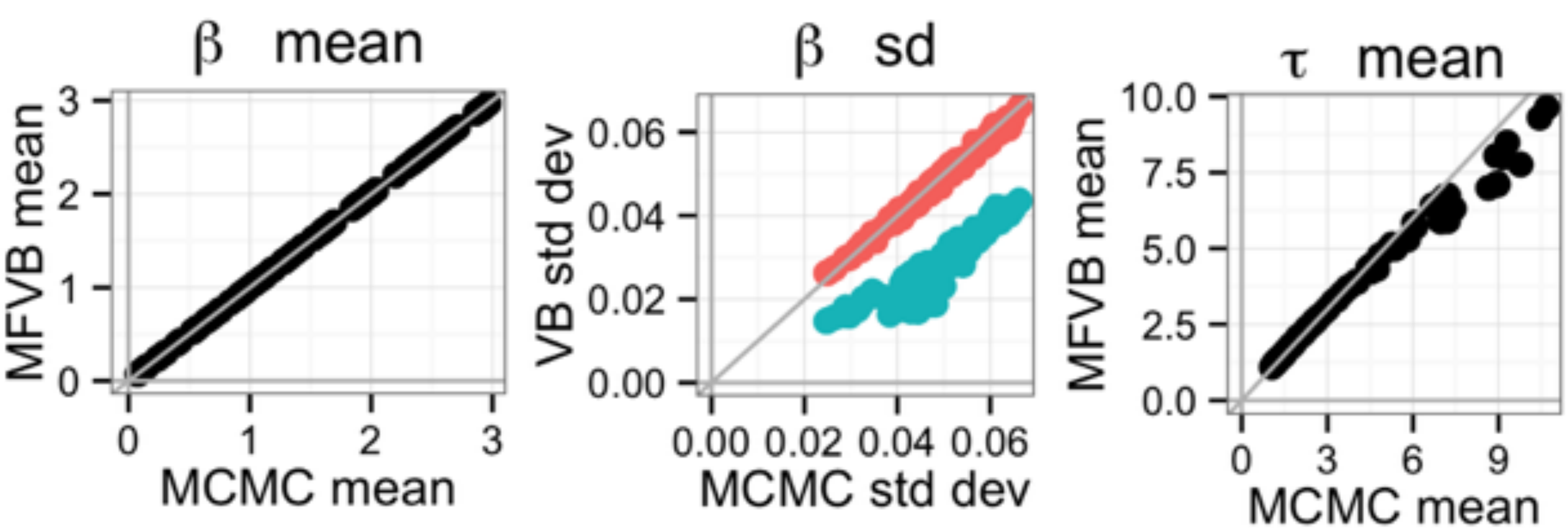
- Non-conjugate normal-Poisson generalized linear mixed model

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

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

- 100 simulated data sets, 500 data points each, R `MCMCglmm` package

LRVB, **MFVB**



Experiments

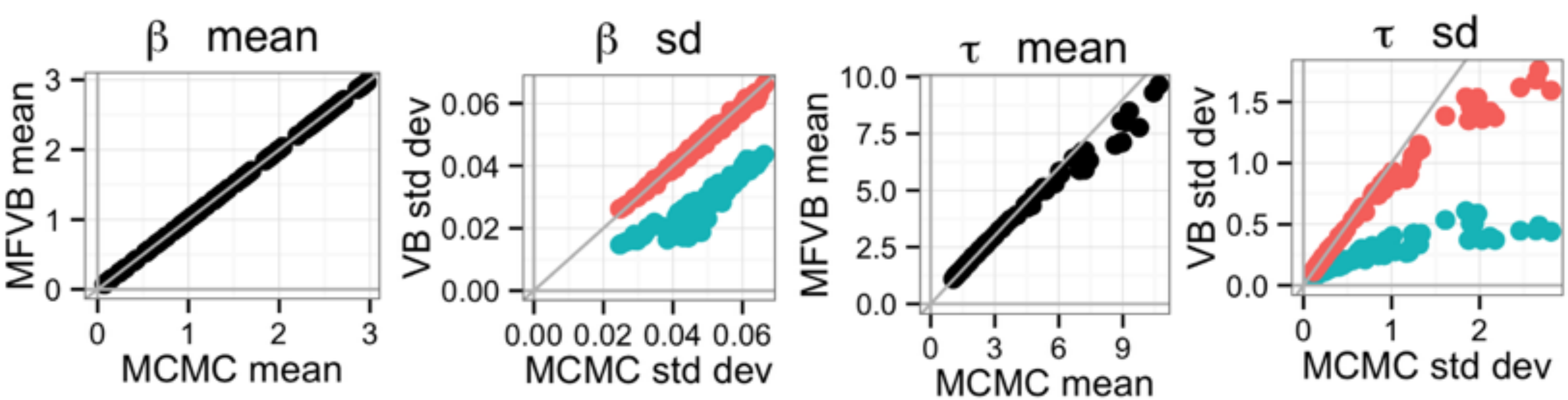
- Non-conjugate normal-Poisson generalized linear mixed model

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

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

- 100 simulated data sets, 500 data points each, R `MCMCglmm` package

LRVB, **MFVB**

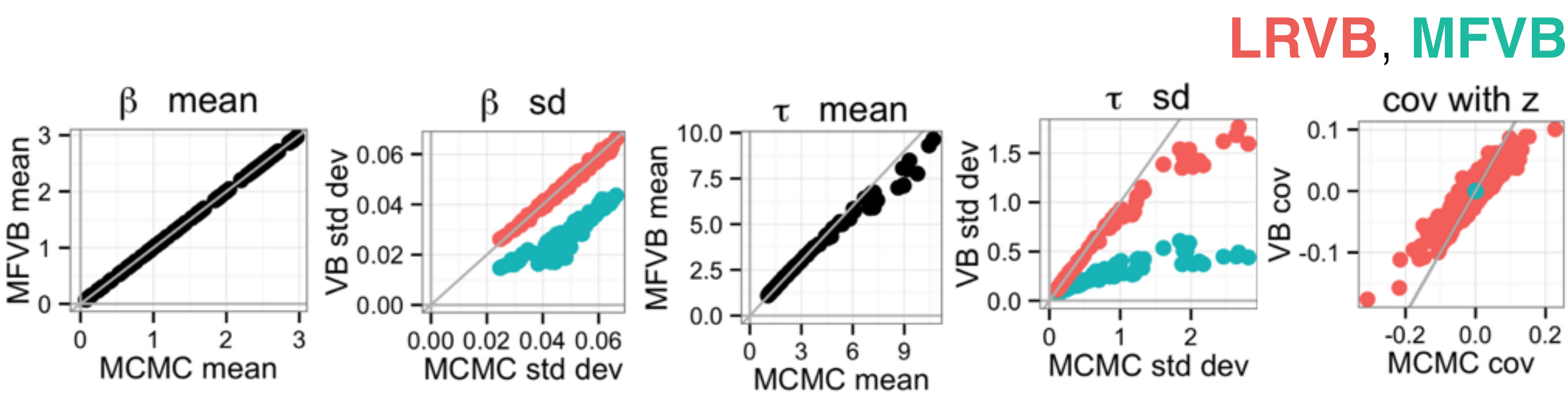


Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

$$z_n|\beta,\tau \overset{indep}{\sim} \mathcal{N}\left(z_n|\beta x_n, \tau^{-1}\right), \quad y_n|z_n \overset{indep}{\sim} \text{Poisson}\left(y_n|\exp(z_n)\right),$$
$$\beta \sim \mathcal{N}(\beta|0,\sigma_\beta^2), \quad \tau \sim \text{Gamma}(\tau|\alpha_\tau,\beta_\tau)$$

- 100 simulated data sets, 500 data points each, R `MCMCglmm` package



Scaling the matrix inverse

Scaling the matrix inverse

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

Scaling the matrix inverse

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

Scaling the matrix inverse

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

- Decomposition of parameter vector

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

Scaling the matrix inverse

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

- Decomposition of parameter vector

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

Scaling the matrix inverse

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

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

- Schur complement

Scaling the matrix inverse

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

- Decomposition of parameter vector

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

- Schur complement

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

Scaling the matrix inverse

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

- Decomposition of parameter vector

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

- Schur complement

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

Scaling the matrix inverse

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

- Decomposition of parameter vector

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

- Schur complement

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

Scaling the matrix inverse

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

- Decomposition of parameter vector

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

- Schur complement

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

- Sparsity patterns

Scaling the matrix inverse

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

- Decomposition of parameter vector

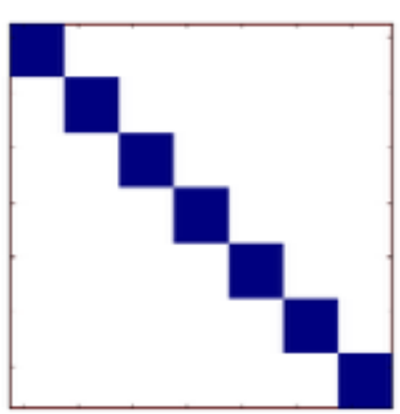
$$\theta = (\alpha^T, z^T)^T \qquad H = \begin{pmatrix} H_\alpha & H_{\alpha z} \\ H_{z\alpha} & H_z \end{pmatrix}$$

- Schur complement

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

- Sparsity patterns

V



Scaling the matrix inverse

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

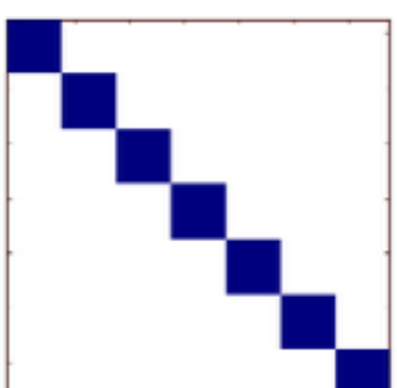
- Decomposition of parameter vector

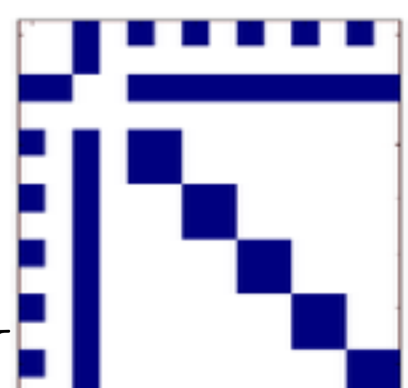
$$\theta = (\alpha^T, z^T)^T \qquad H = \begin{pmatrix} H_\alpha & H_{\alpha z} \\ H_{z\alpha} & H_z \end{pmatrix}$$

- Schur complement

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

- Sparsity patterns

V 

H 

Scaling the matrix inverse

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

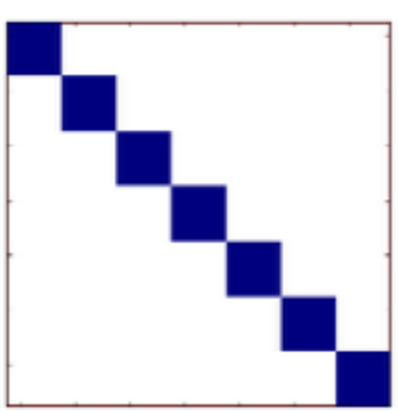
- Decomposition of parameter vector

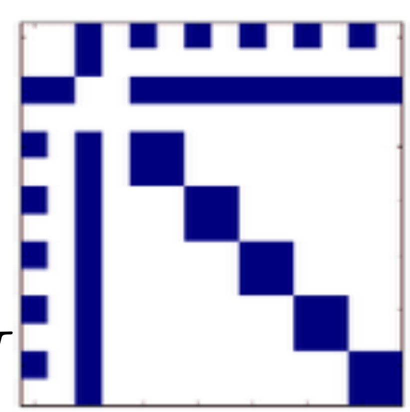
$$\theta = (\alpha^T, z^T)^T \qquad H = \begin{pmatrix} H_\alpha & H_{\alpha z} \\ H_{z\alpha} & H_z \end{pmatrix}$$

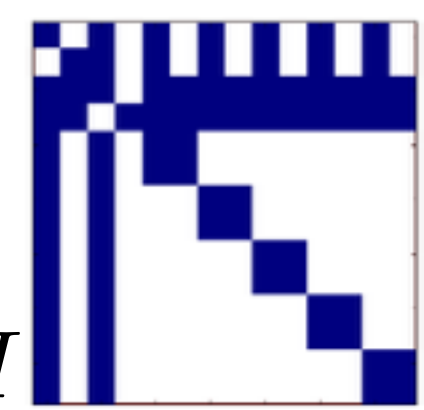
- Schur complement

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

- Sparsity patterns

V 

H 

$I - VH$ 

Experiments

Experiments

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

Experiments

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

Experiments

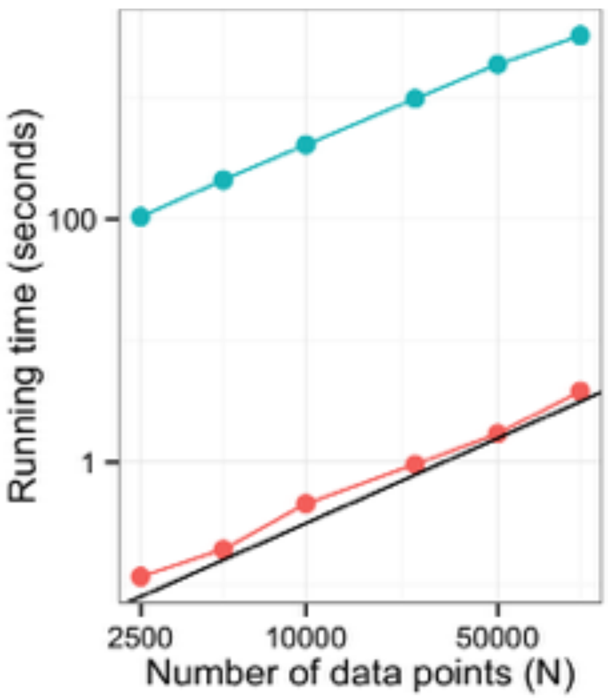
- Scaling: Gaussian mixture model (*K* components, *P* dimensions, *N* data points)
- The number of parameters in μ, π, Λ grows as $O(KP^2)$
- The number of parameters in z grows as $O(KN)$

Experiments

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

Experiments

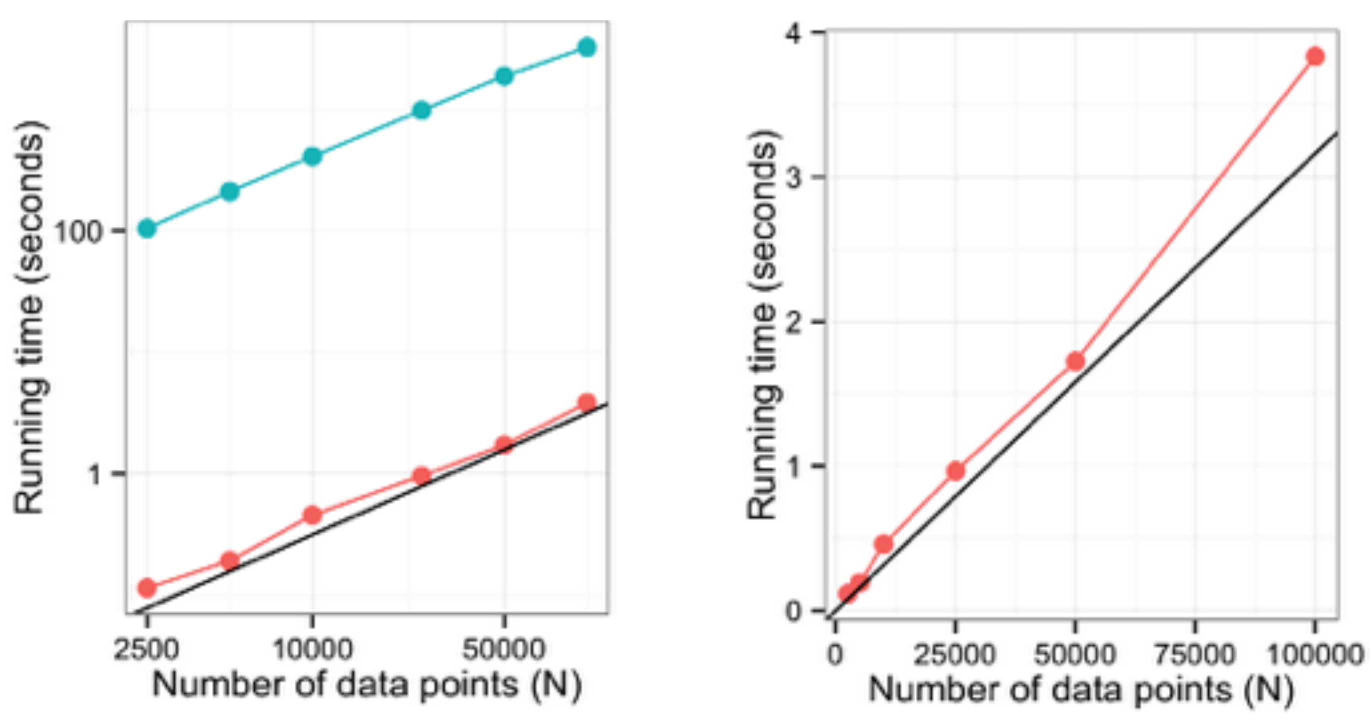
- Scaling: Gaussian mixture model (*K* components, *P* dimensions, *N* data points)
- The number of parameters in μ, π, Λ grows as $O(KP^2)$
- The number of parameters in z grows as $O(KN)$
- Worst case scaling: $O(K^3), O(P^6), O(N)$



LRVB, **Gibbs**

Experiments

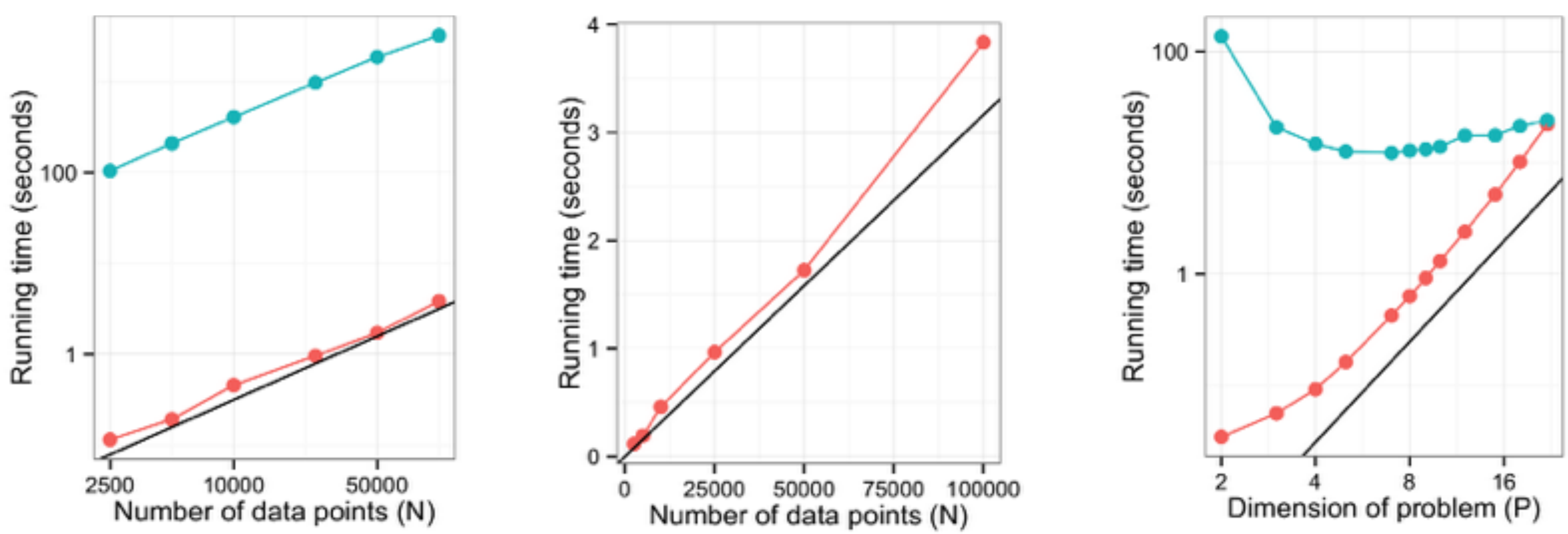
- Scaling: Gaussian mixture model (*K* components, *P* dimensions, *N* data points)
- The number of parameters in μ, π, Λ grows as $O(KP^2)$
- The number of parameters in z grows as $O(KN)$
- Worst case scaling: $O(K^3), O(P^6), O(N)$



LRVB,
Gibbs

Experiments

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



LRVB, **Gibbs**

Roadmap

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

Roadmap

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

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

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

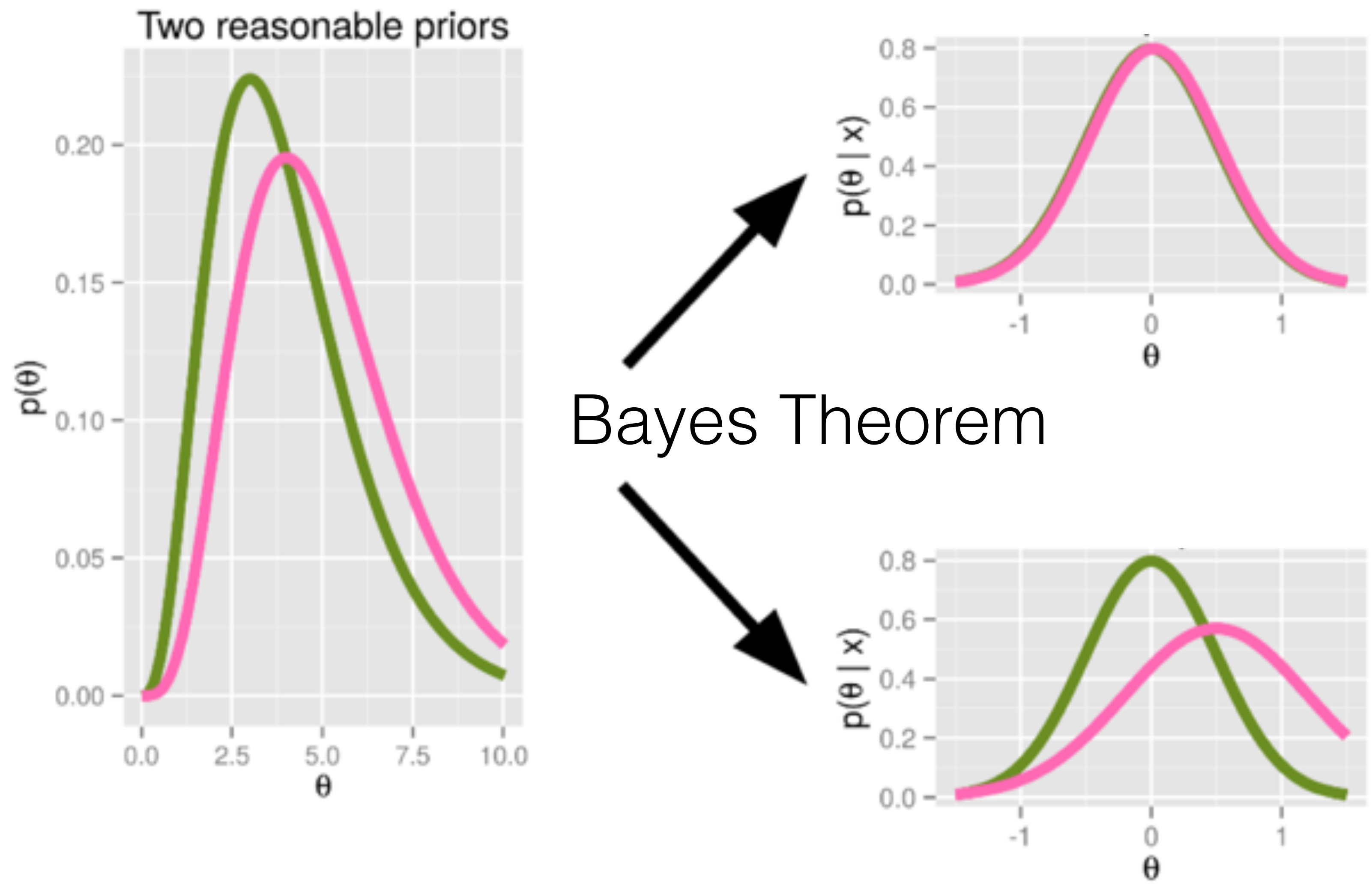
- Sensitivity

Robustness quantification

- Bayes Theorem

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

- Sensitivity



Robustness quantification

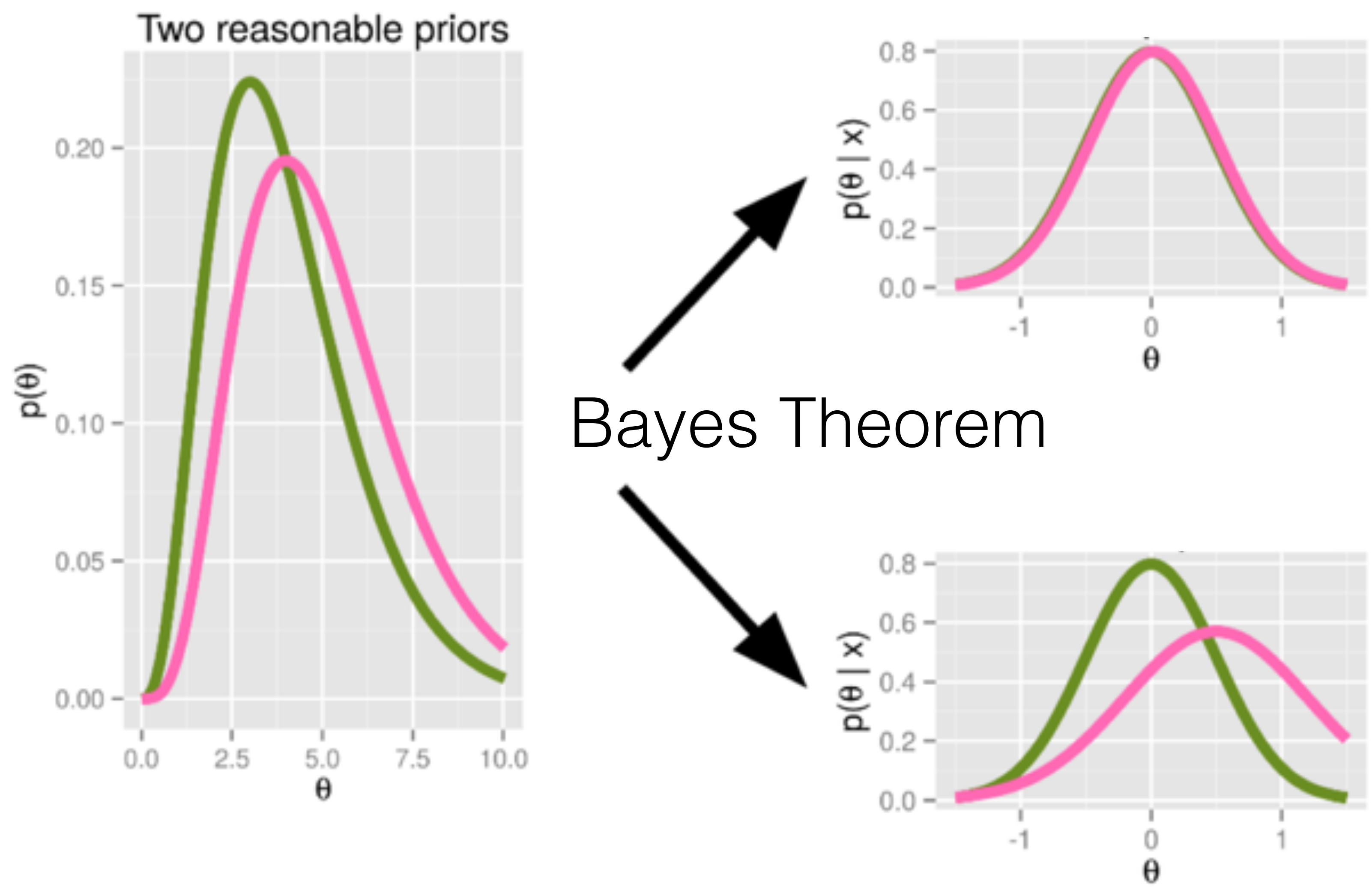
- Bayes Theorem

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

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

- Sensitivity

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



Robustness quantification

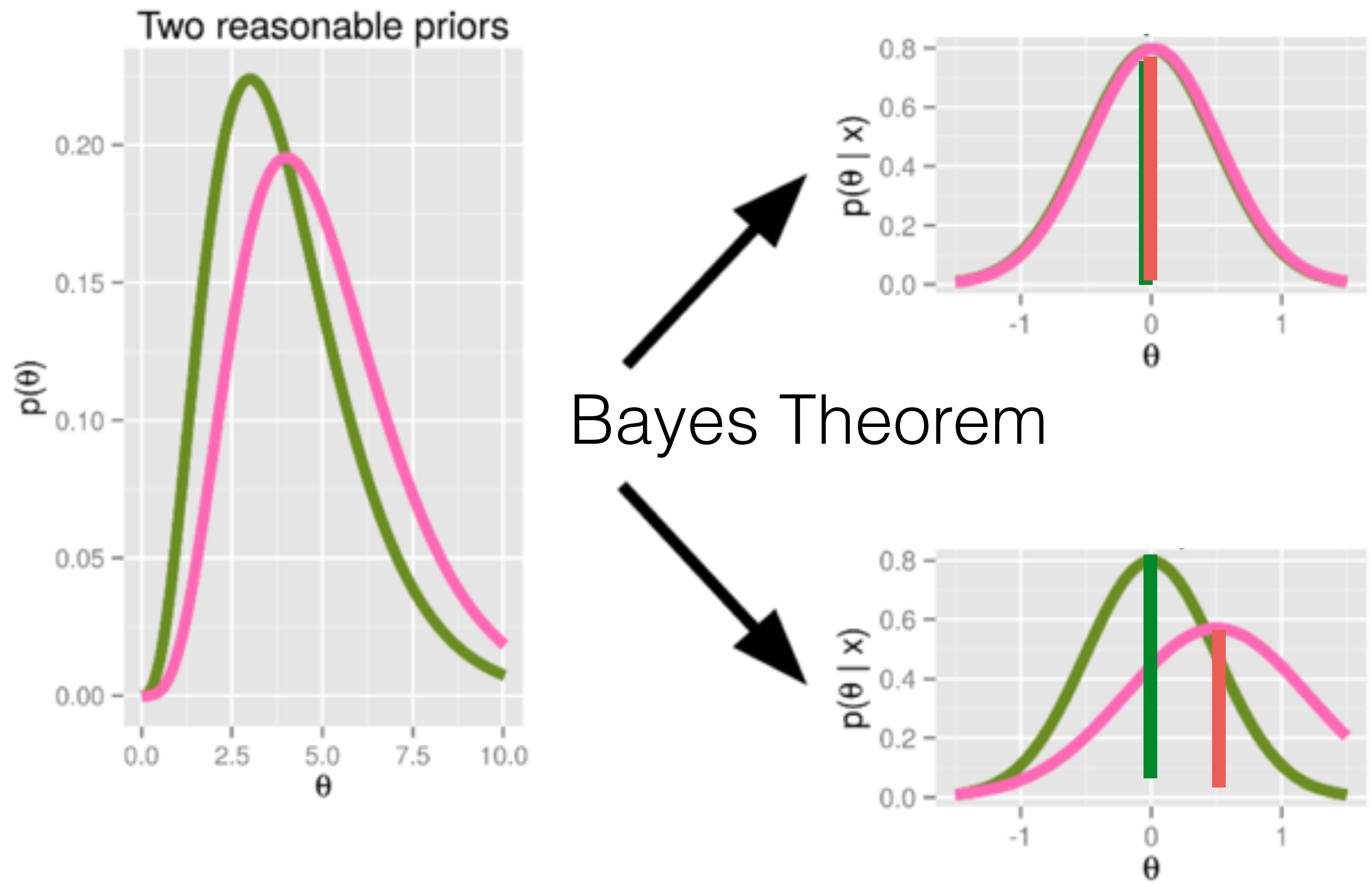
- Bayes Theorem

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

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

- Sensitivity

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



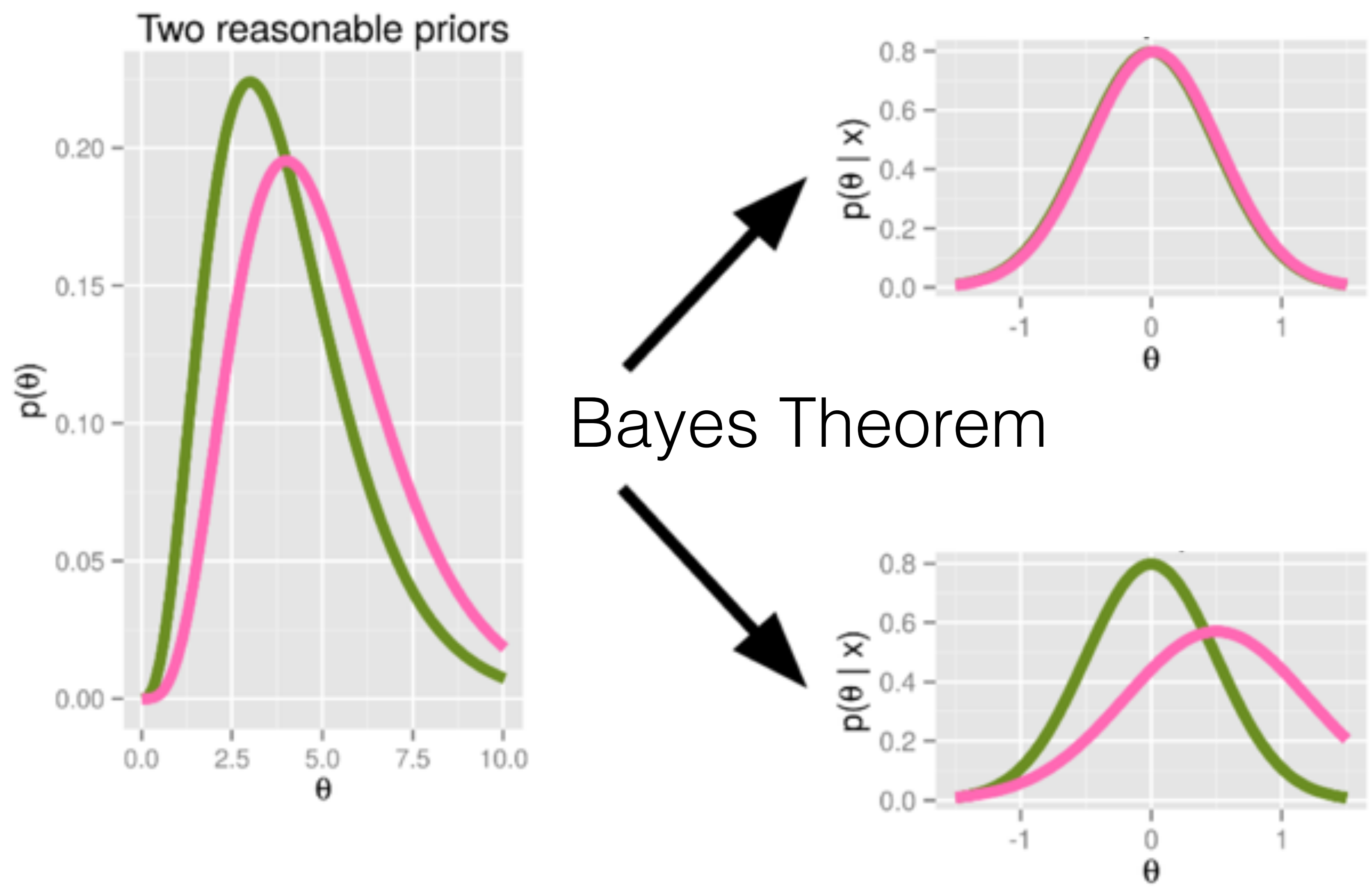
Robustness quantification

- Bayes Theorem

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

- Sensitivity

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



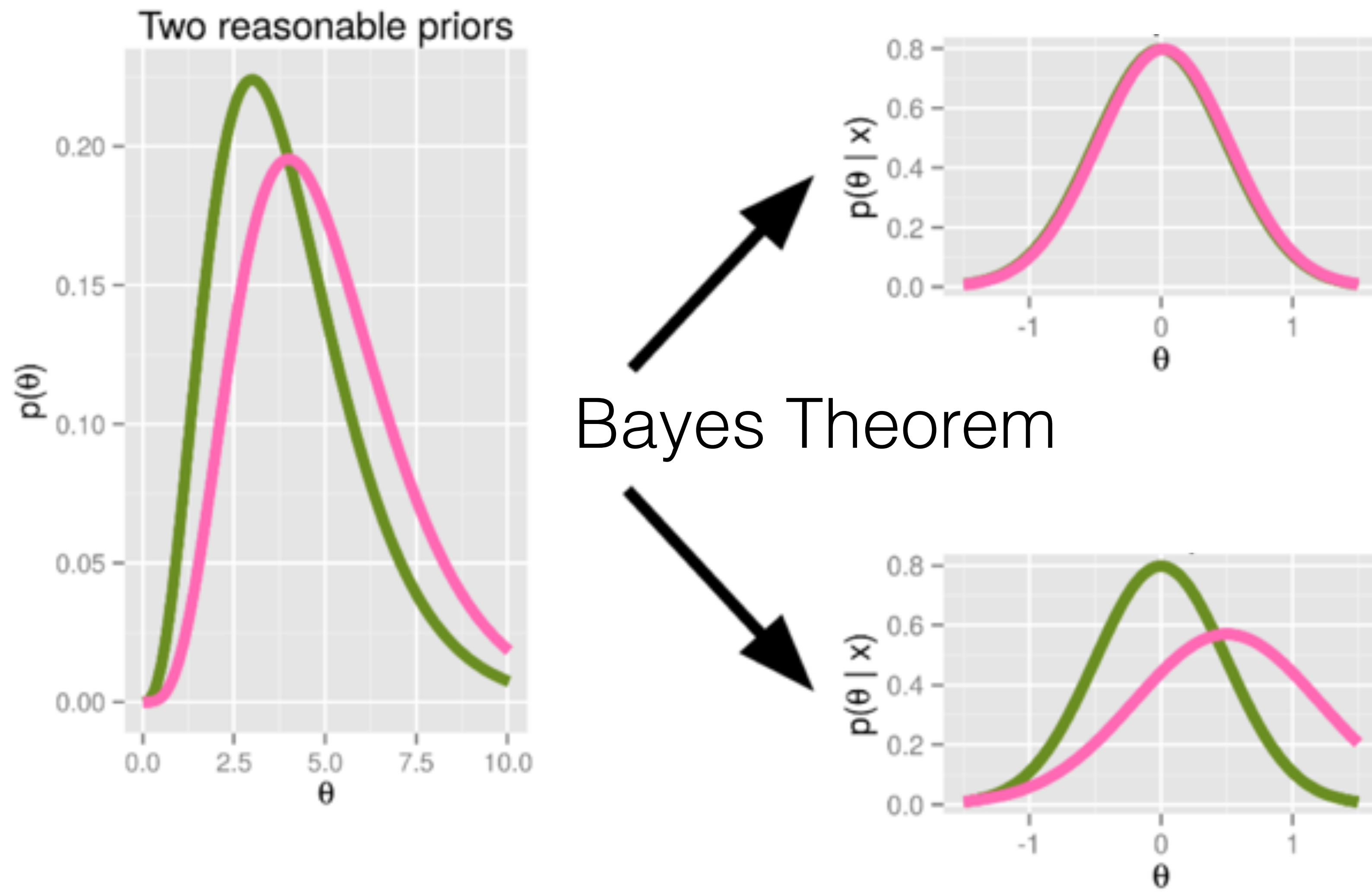
Robustness quantification

- Bayes Theorem

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

- Sensitivity

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



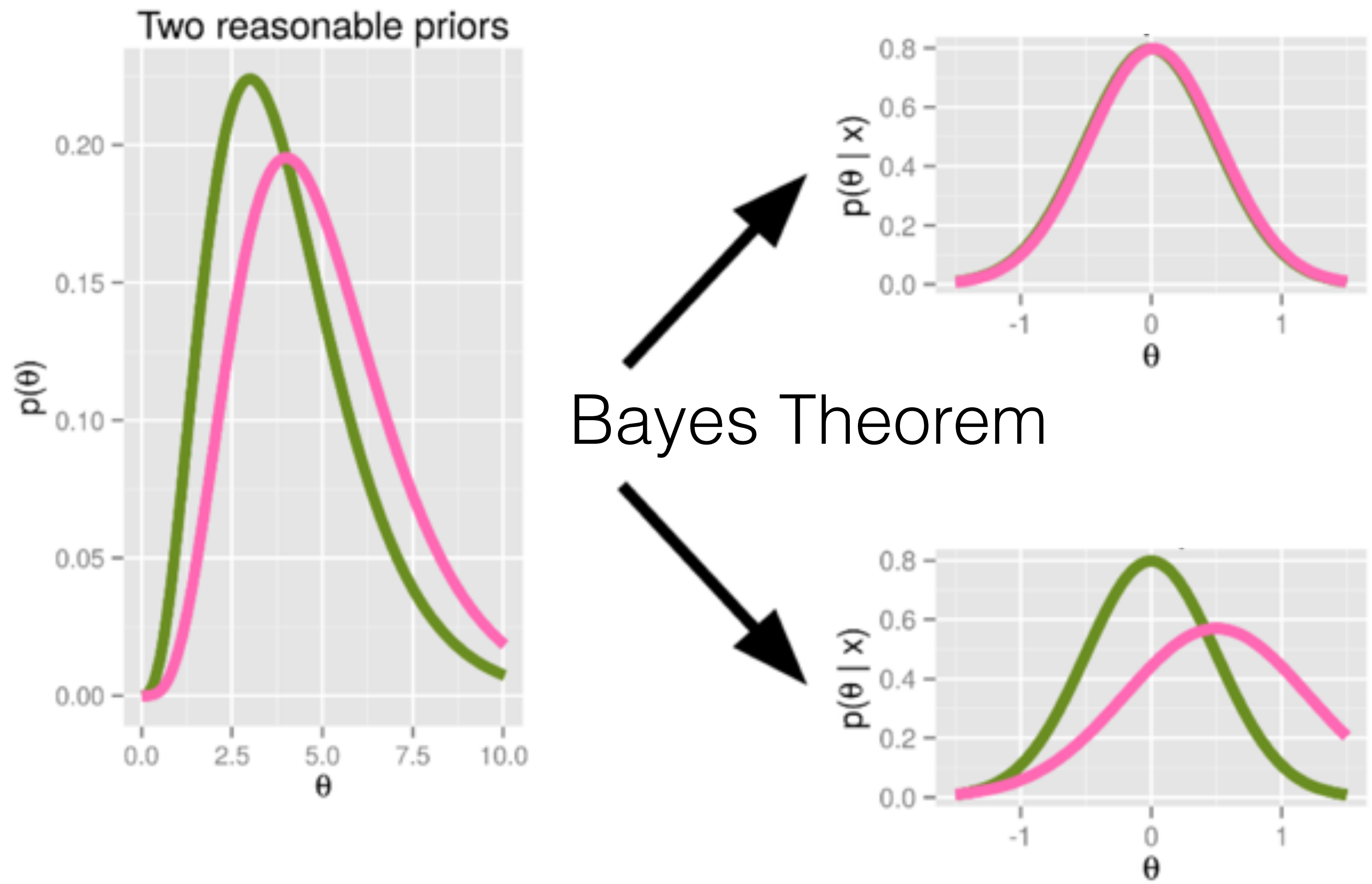
Robustness quantification

- Bayes Theorem

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

- Sensitivity

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



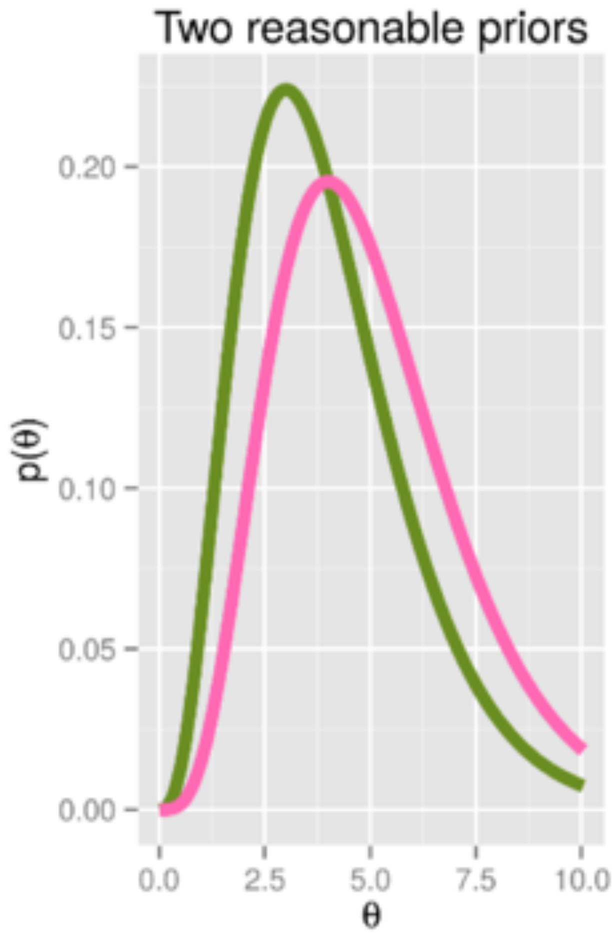
Robustness quantification

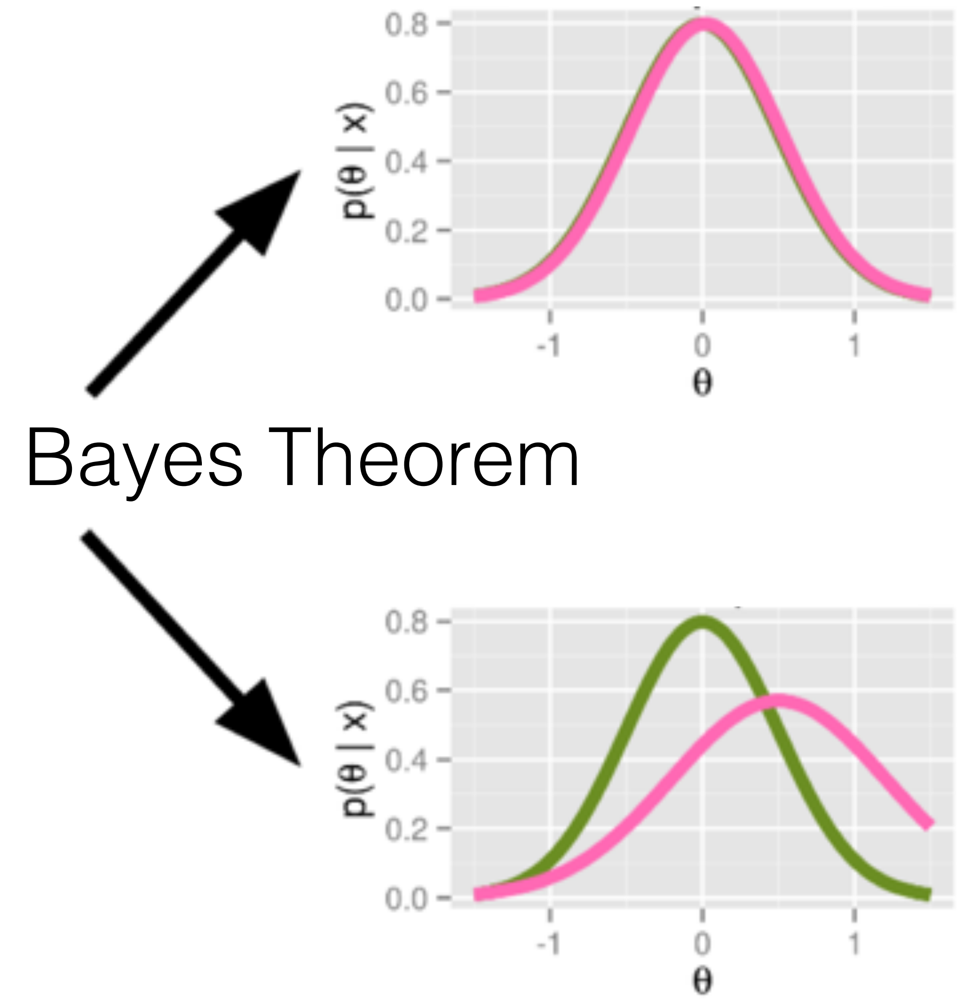
- Bayes Theorem

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

- 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

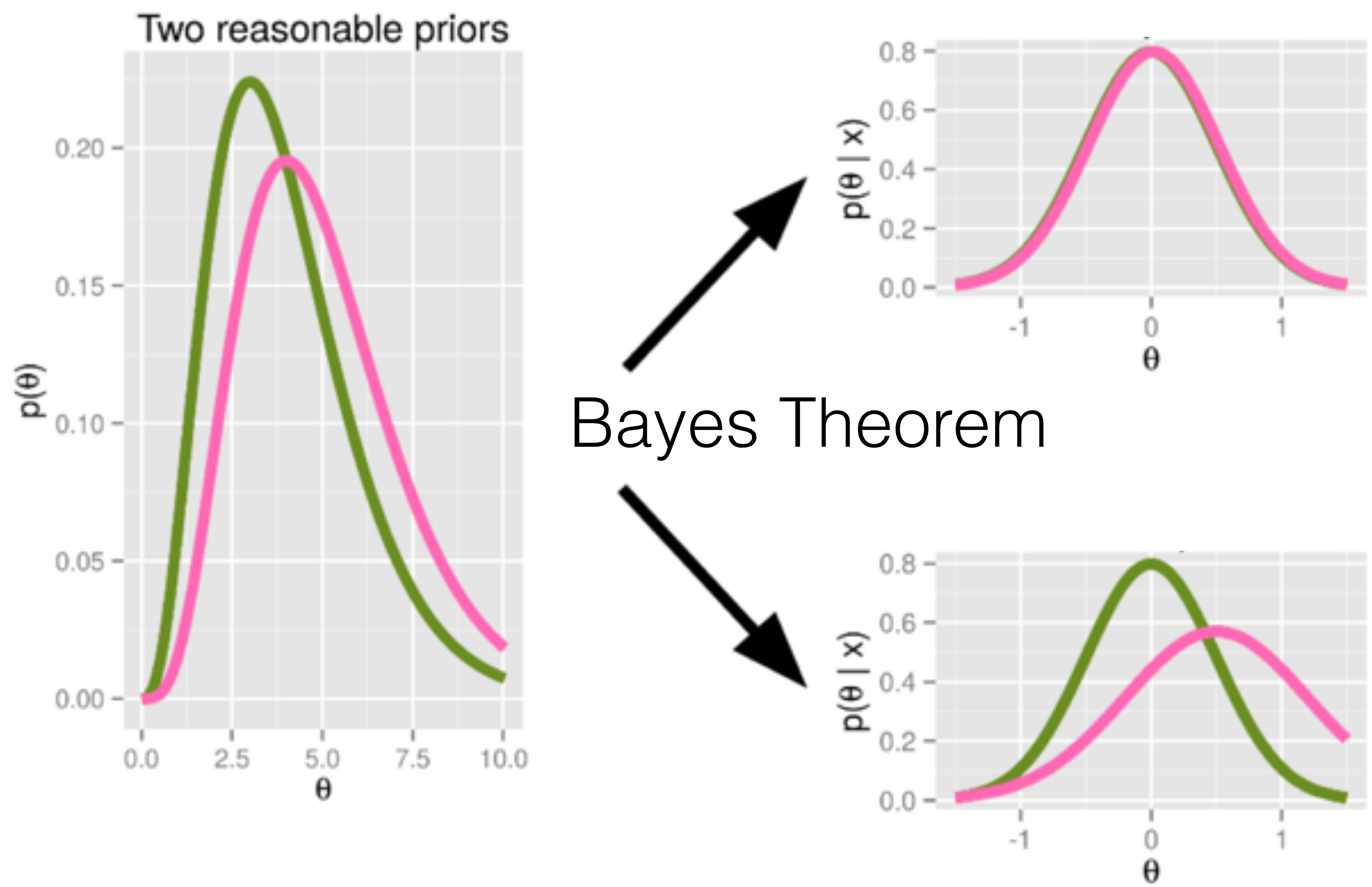
- Bayes Theorem

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

- Sensitivity

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

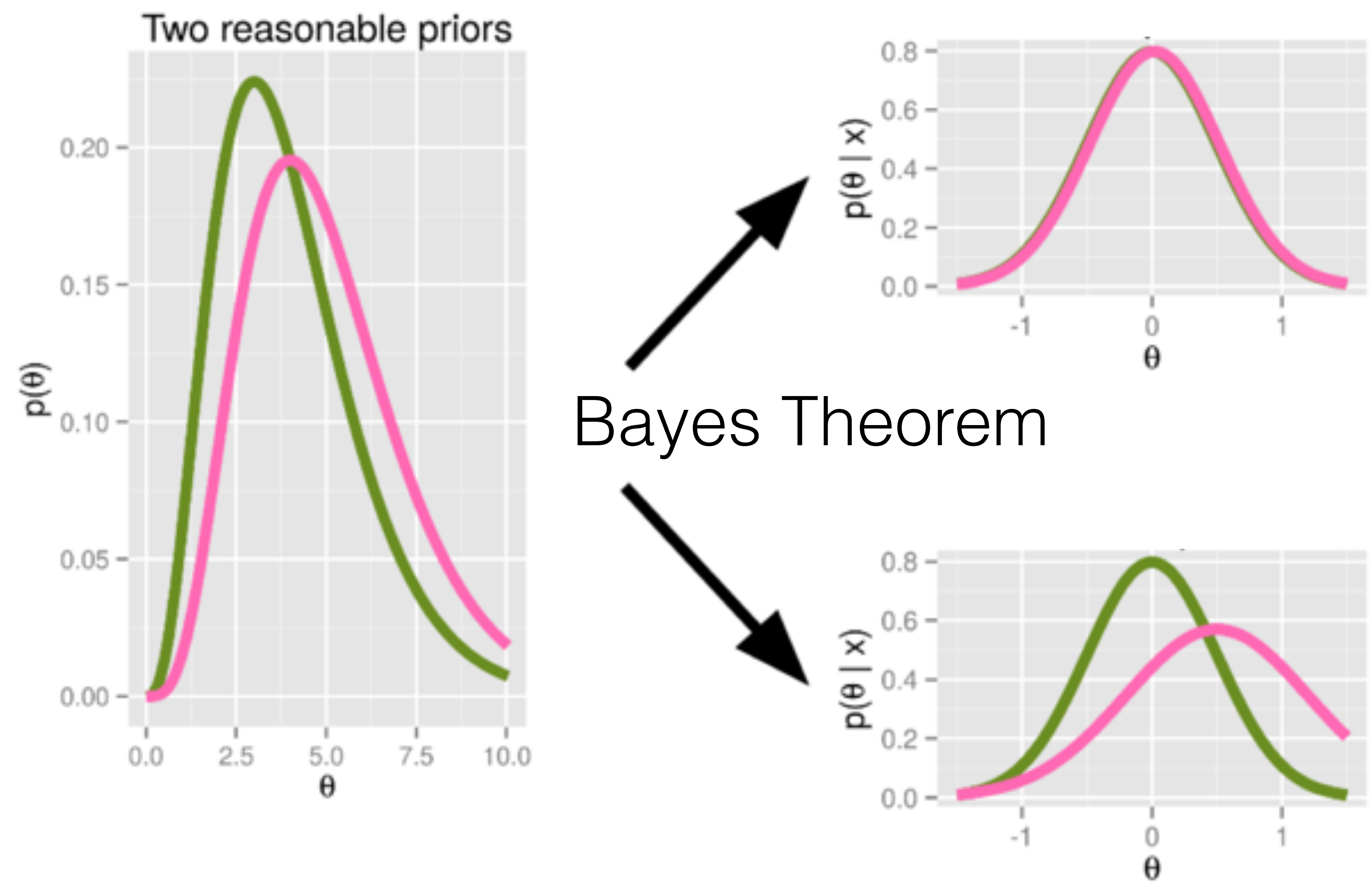
← LRVB estimator



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

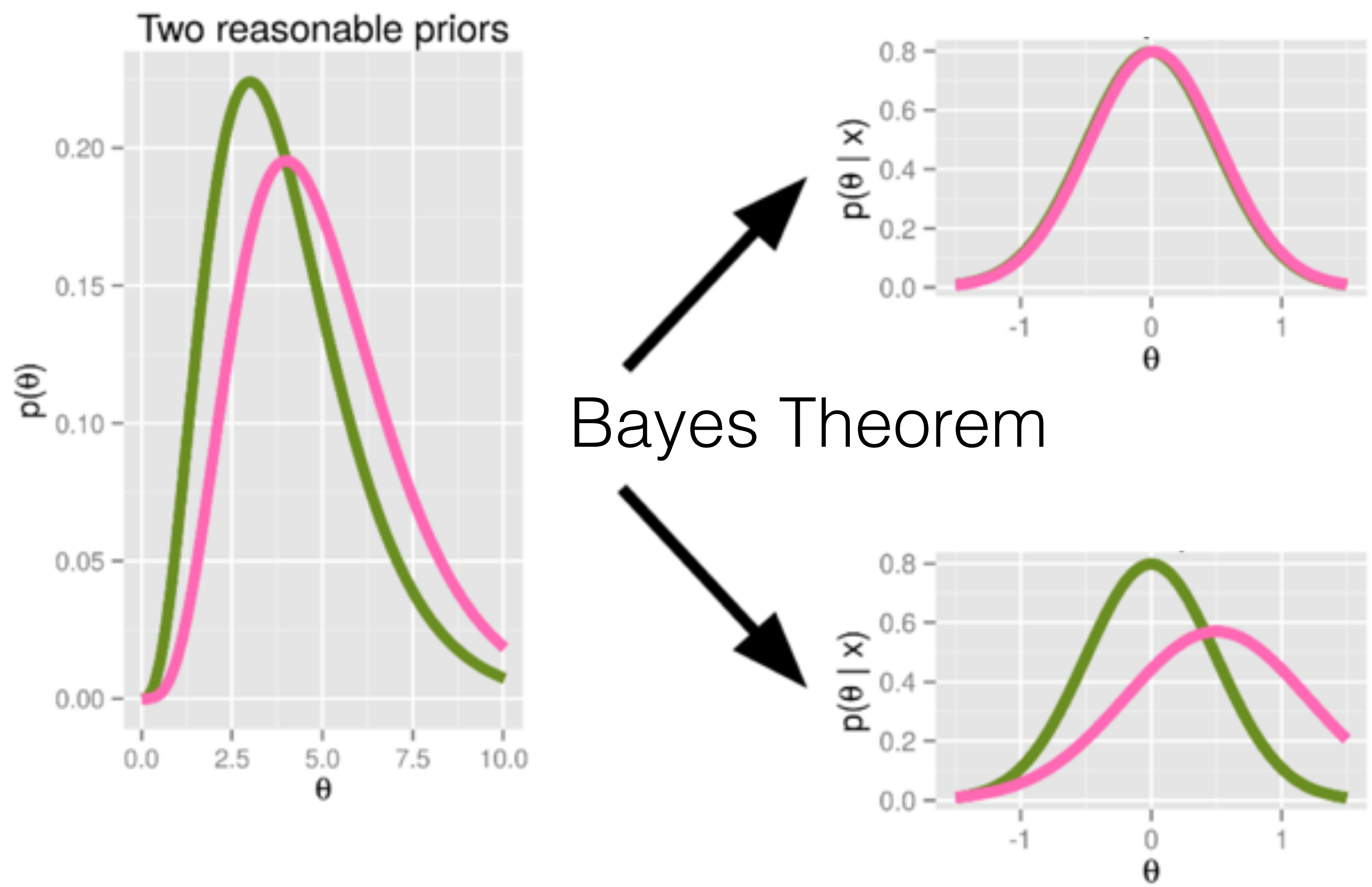
Robustness quantification

- 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 →

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

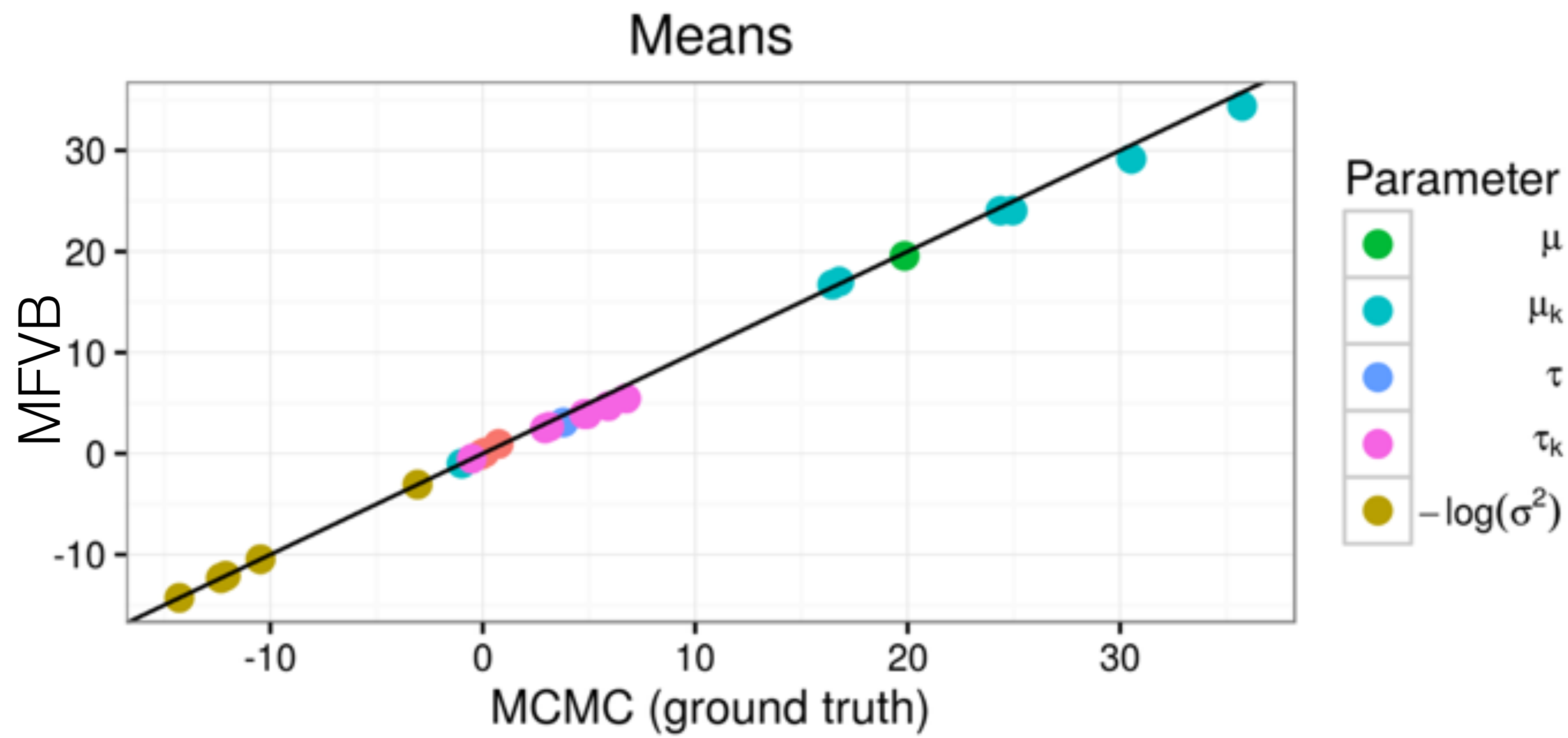
- Priors and hyperpriors:

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

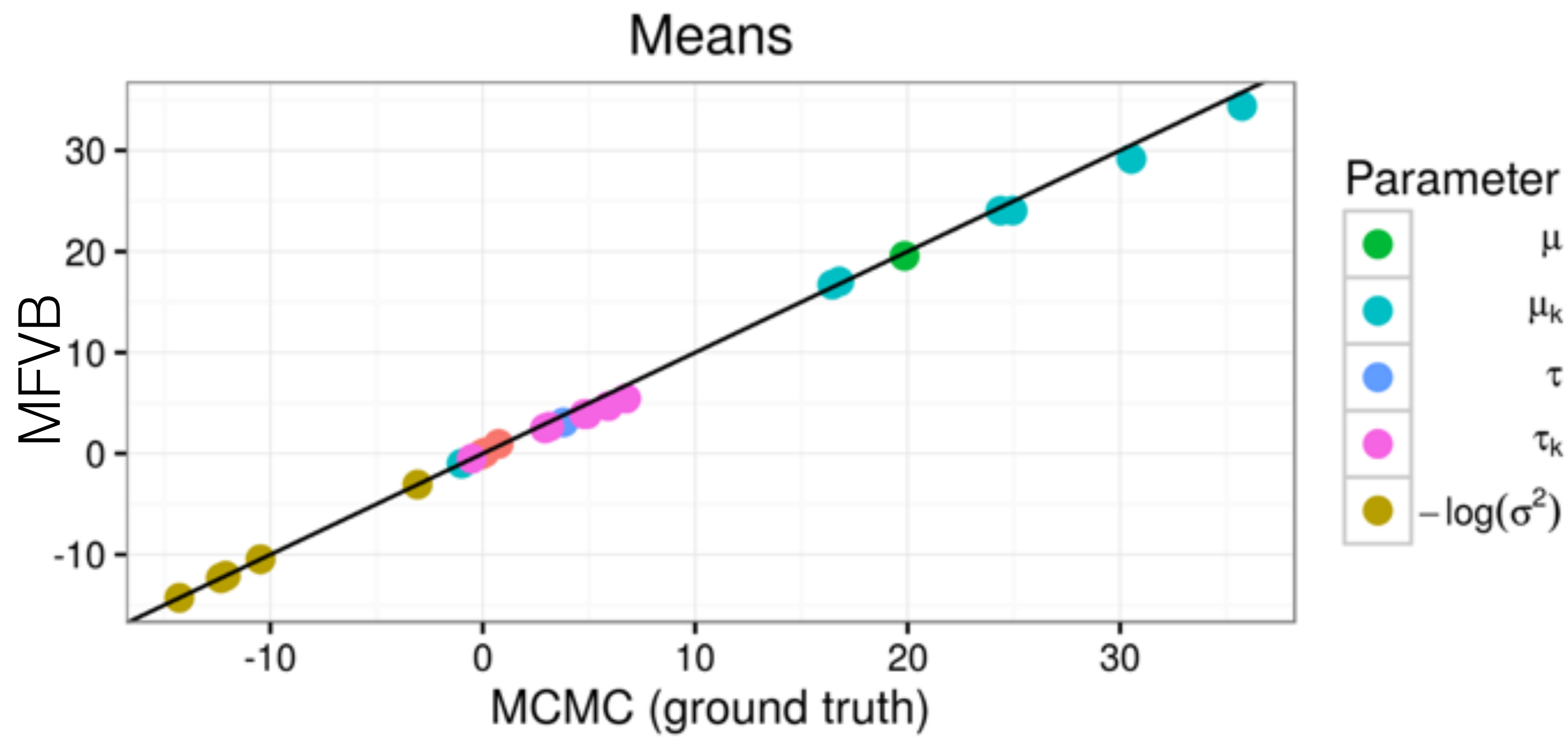
$$\sigma_k^{-2} \overset{iid}{\sim} \Gamma(a, b) \qquad C \sim \mathrm{Sep\&LKJ}(\eta, c, d)$$

Microcredit Experiment

Microcredit Experiment



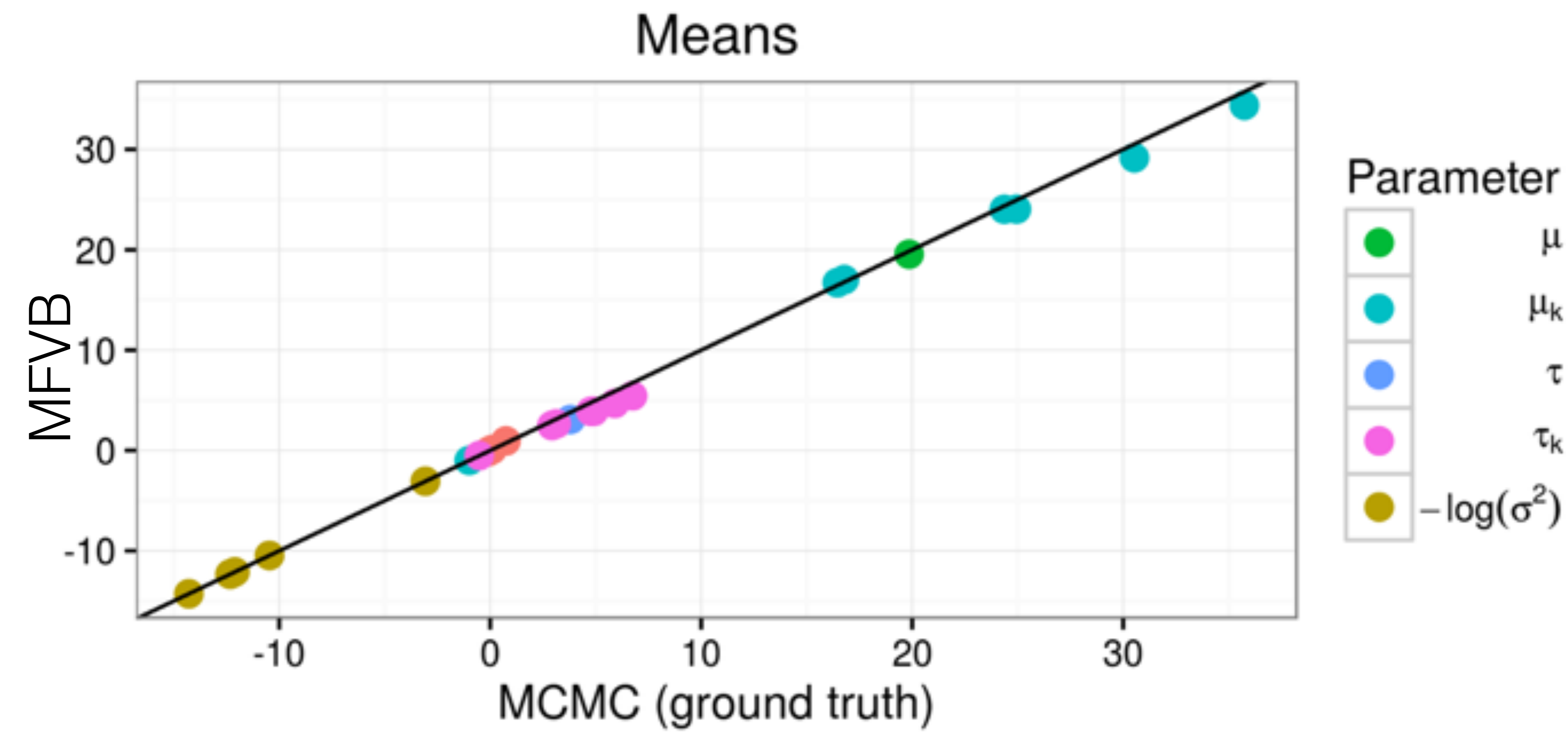
Microcredit Experiment

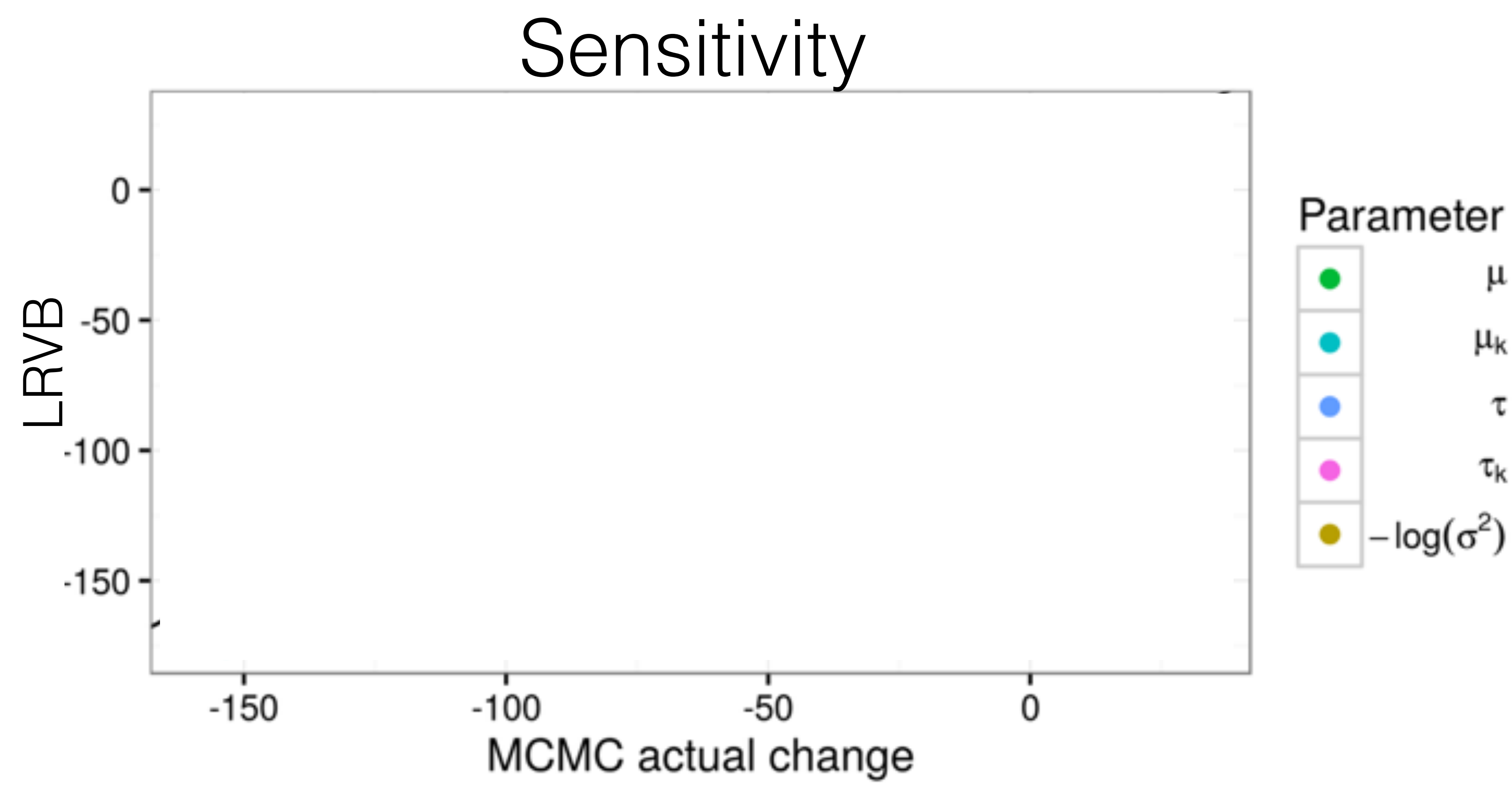


- Perturb Λ_{11}: 0.03 ➔ 0.04

Microcredit Experiment

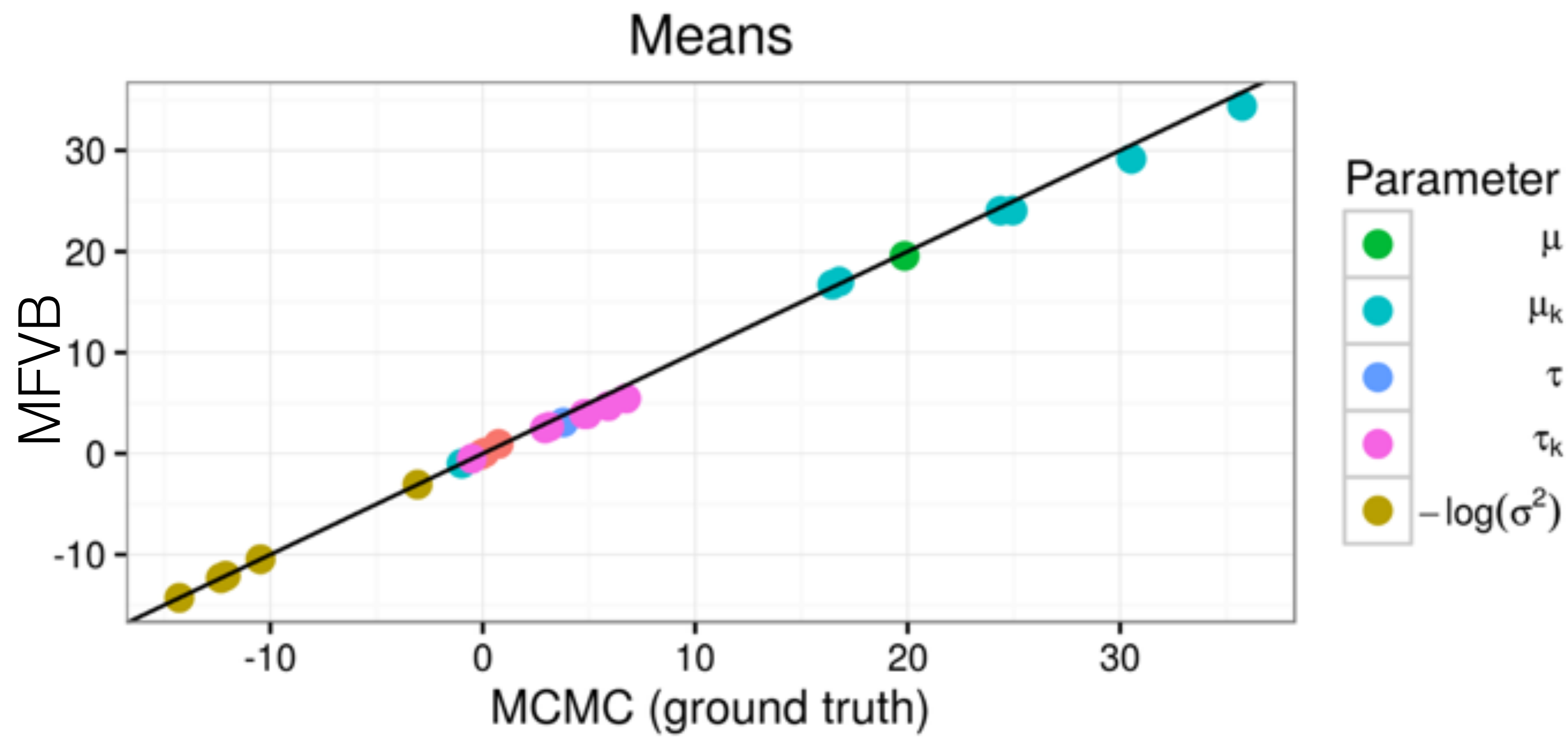
- Perturb Λ_{11}: 0.03 ➔ 0.04

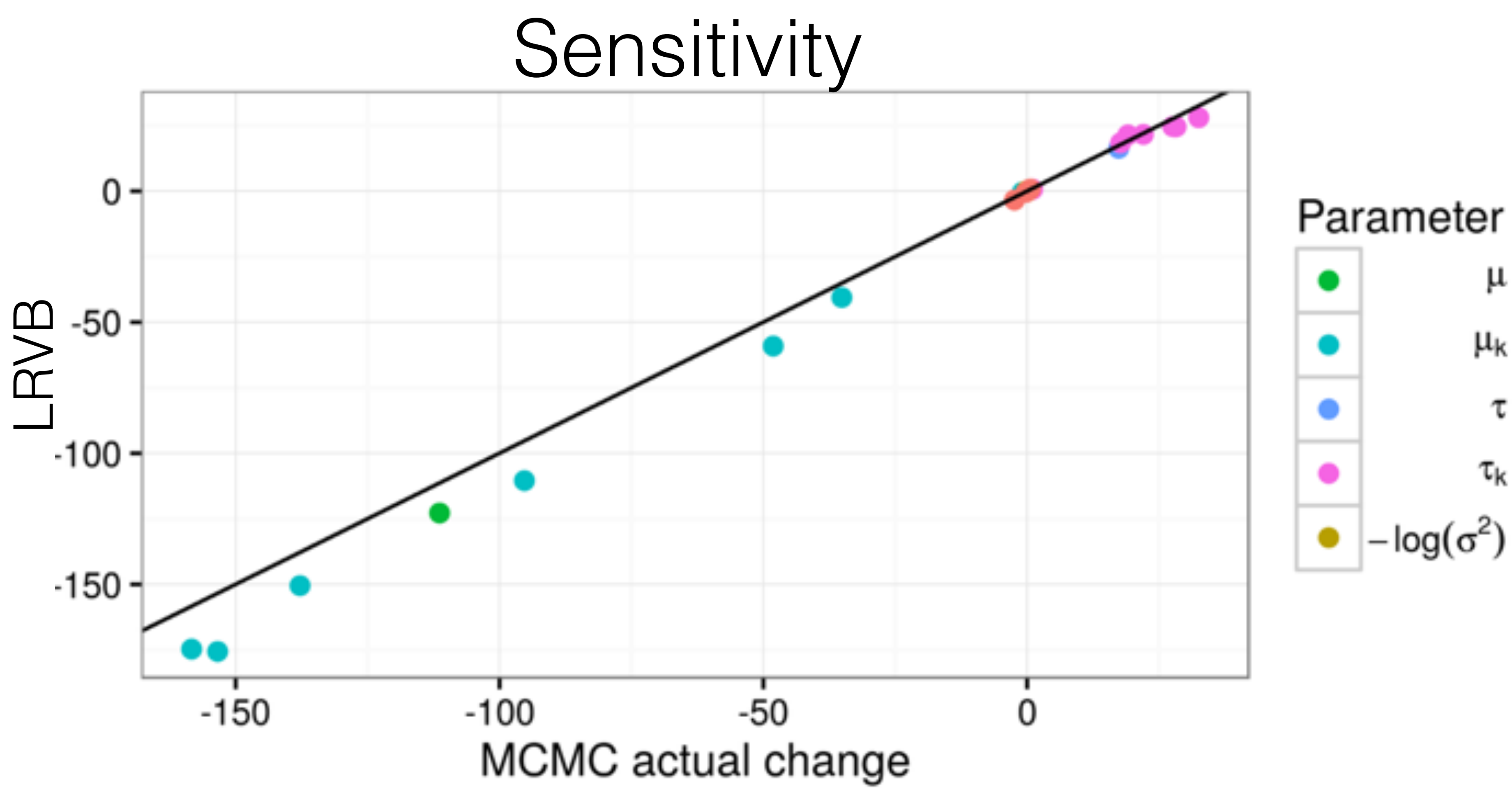




Microcredit Experiment

- Perturb Λ_{11}: 0.03 ➔ 0.04





Microcredit Experiment

Microcredit Experiment

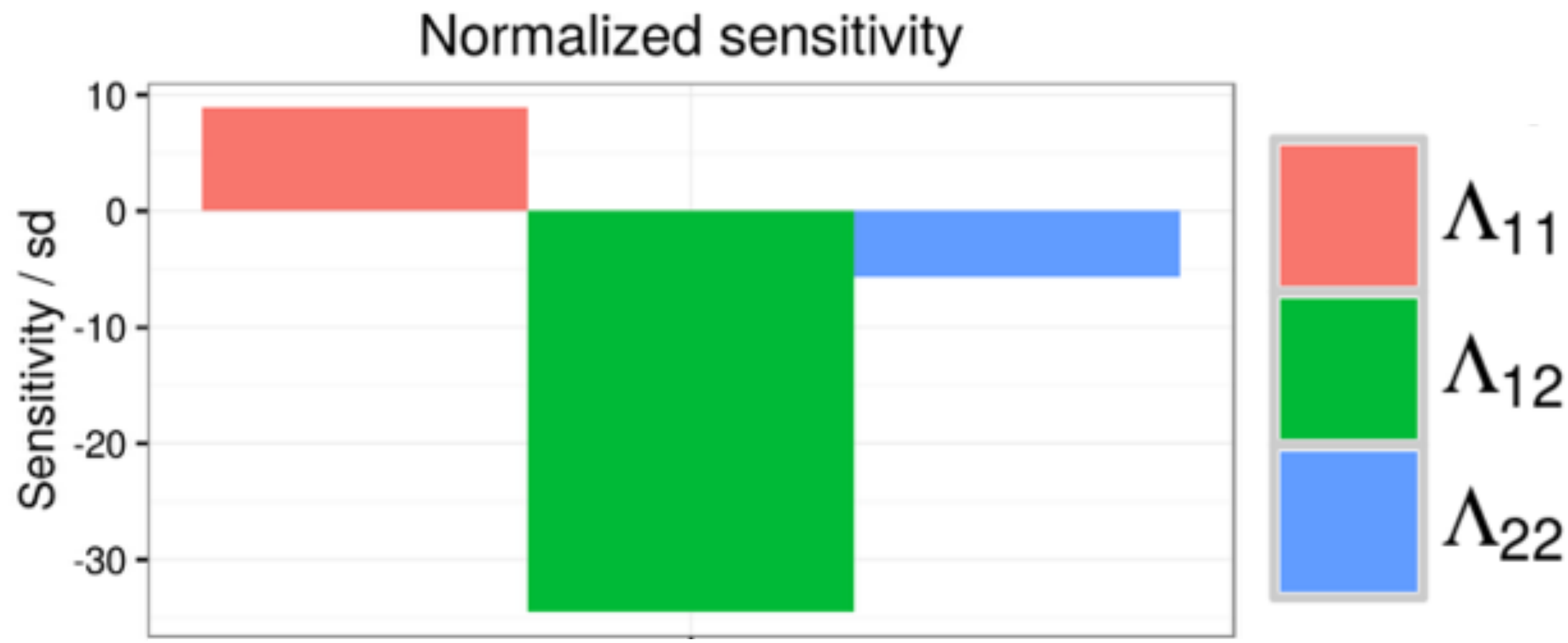
- Sensitivity of the expected microcredit effect (τ)

Microcredit Experiment

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

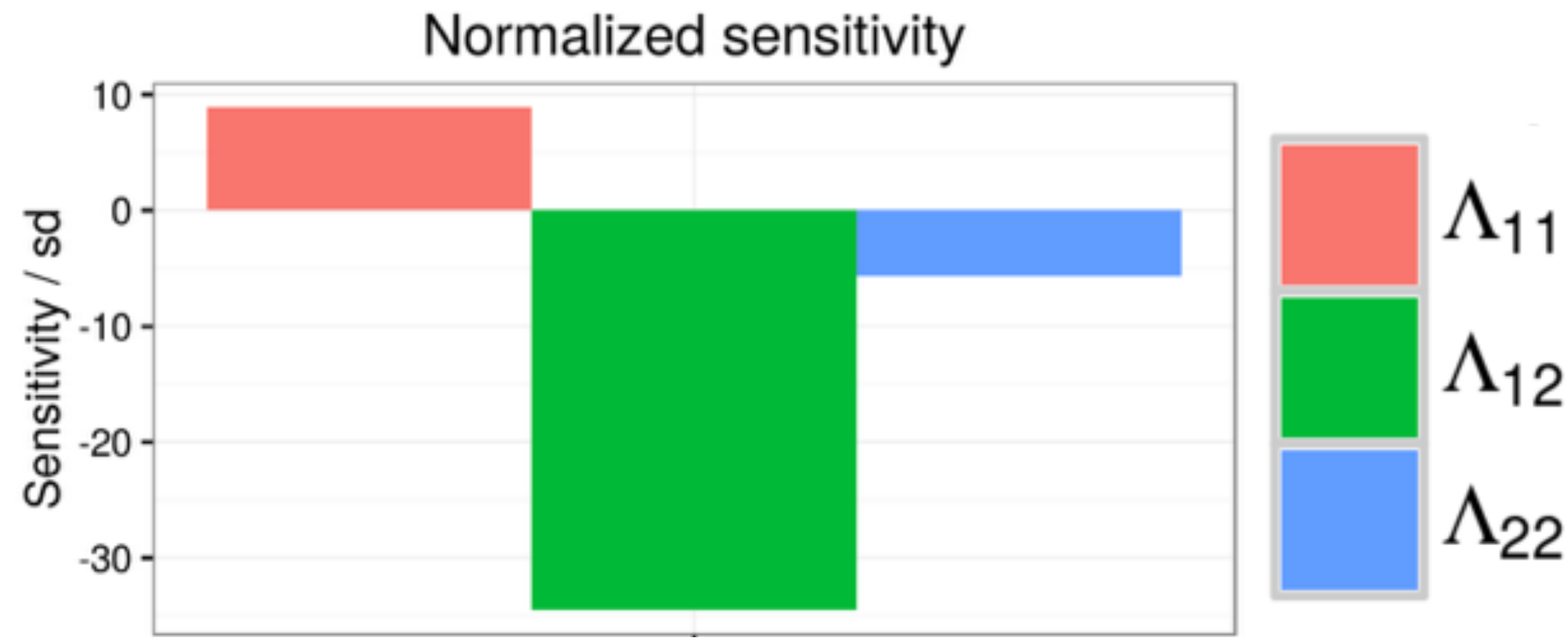
Microcredit Experiment

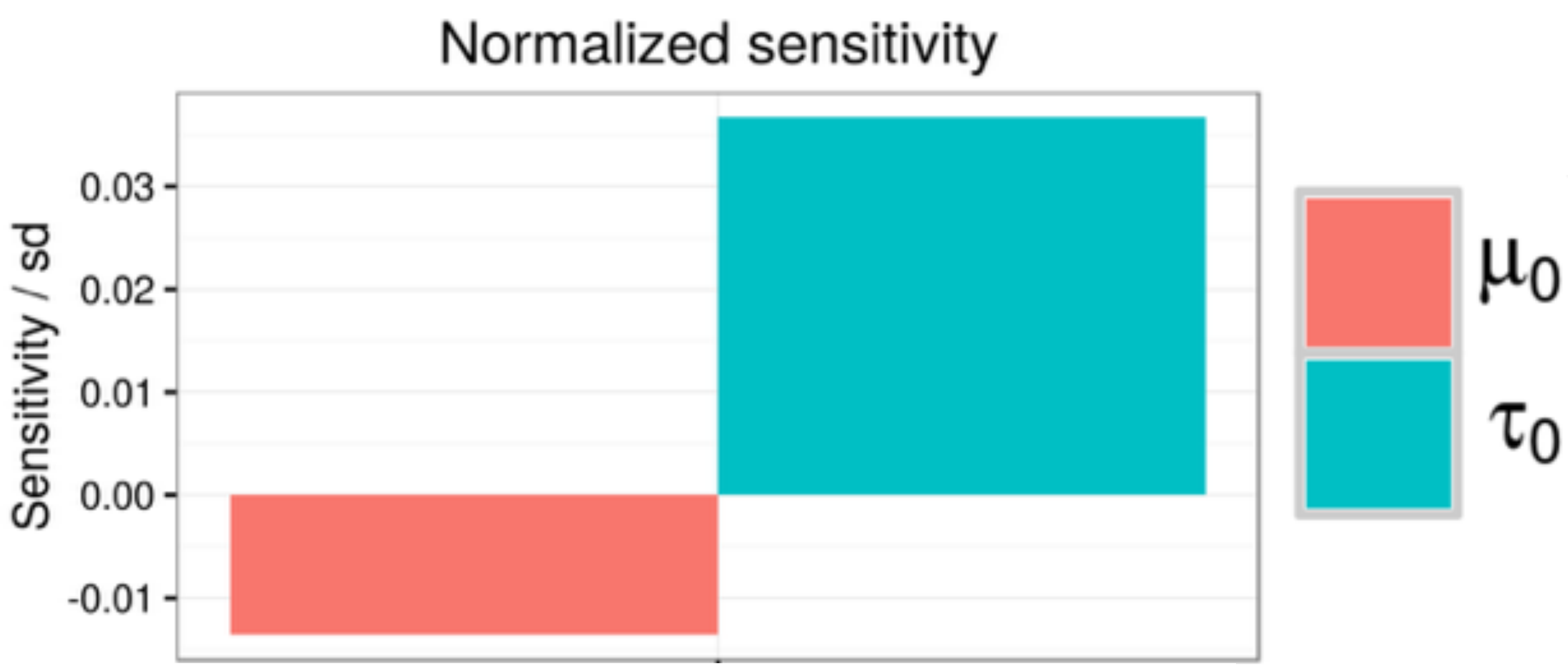
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of τ std devs



Microcredit Experiment

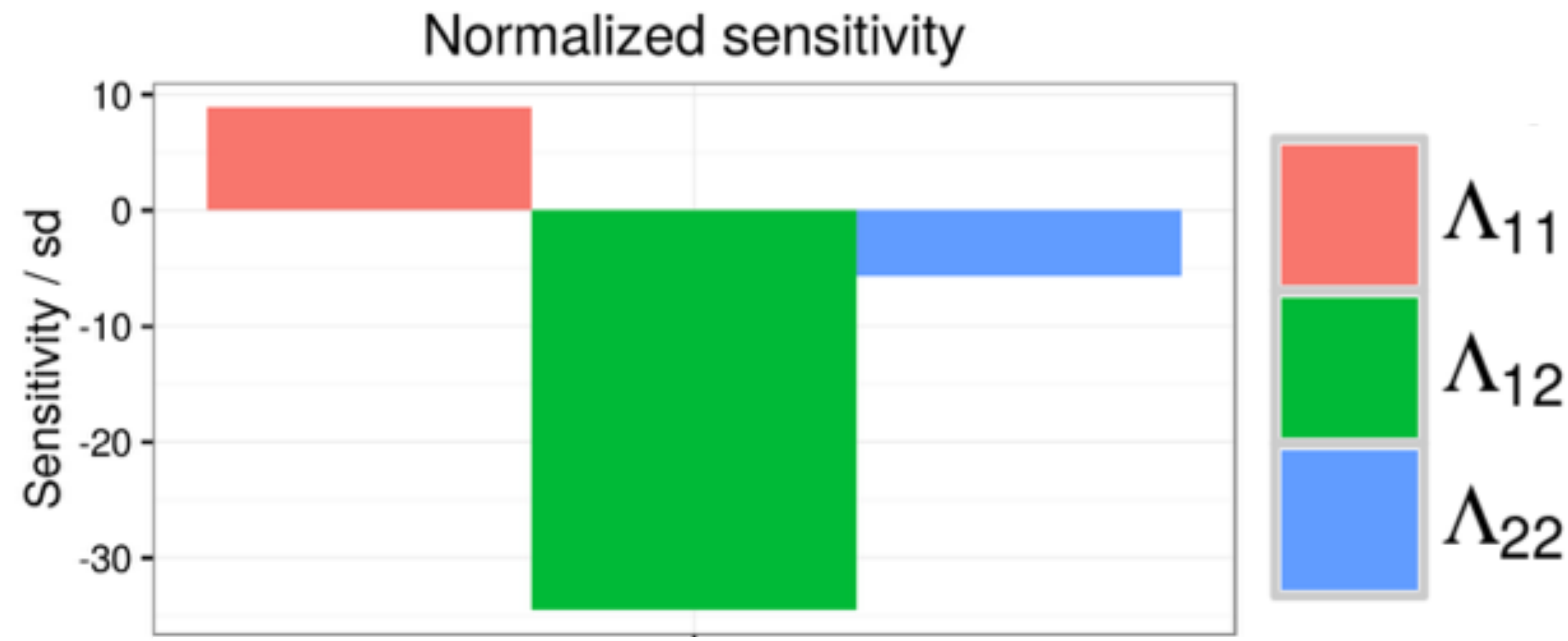
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of τ std devs

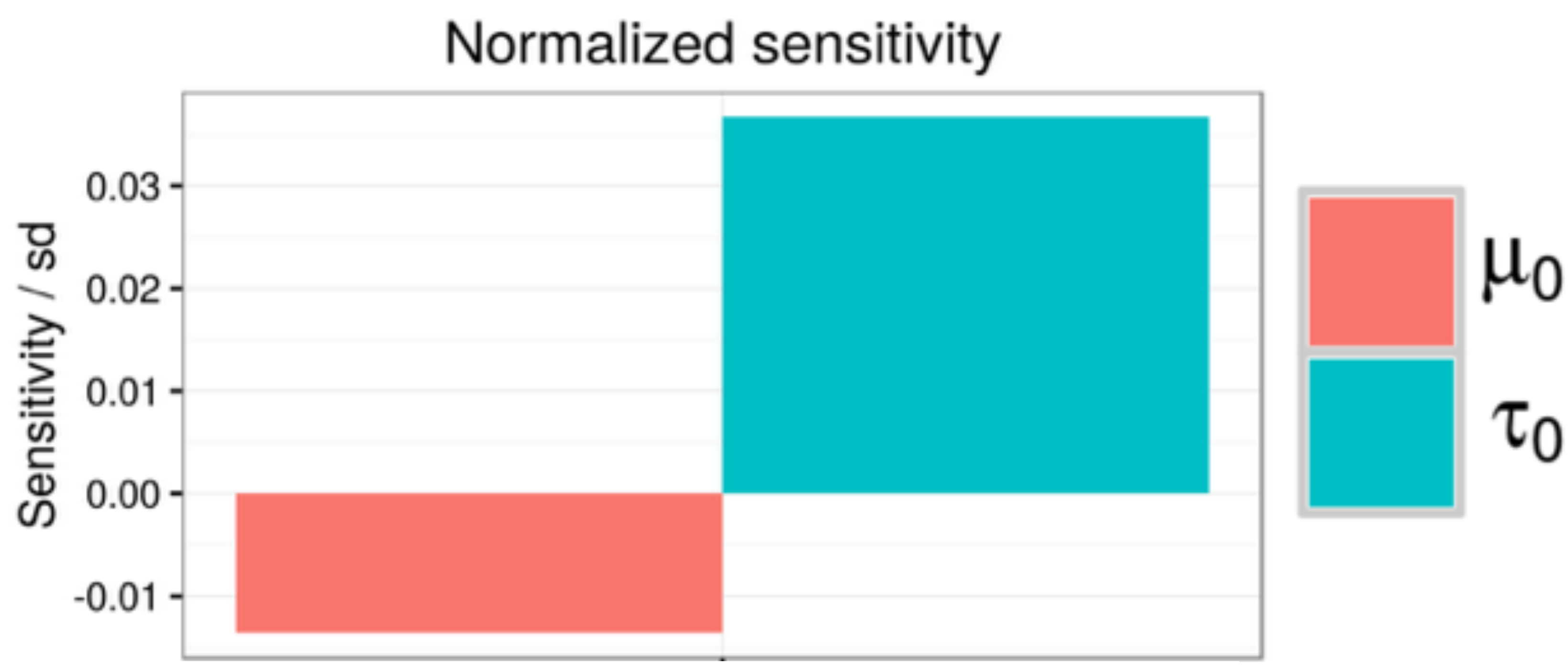




Microcredit Experiment

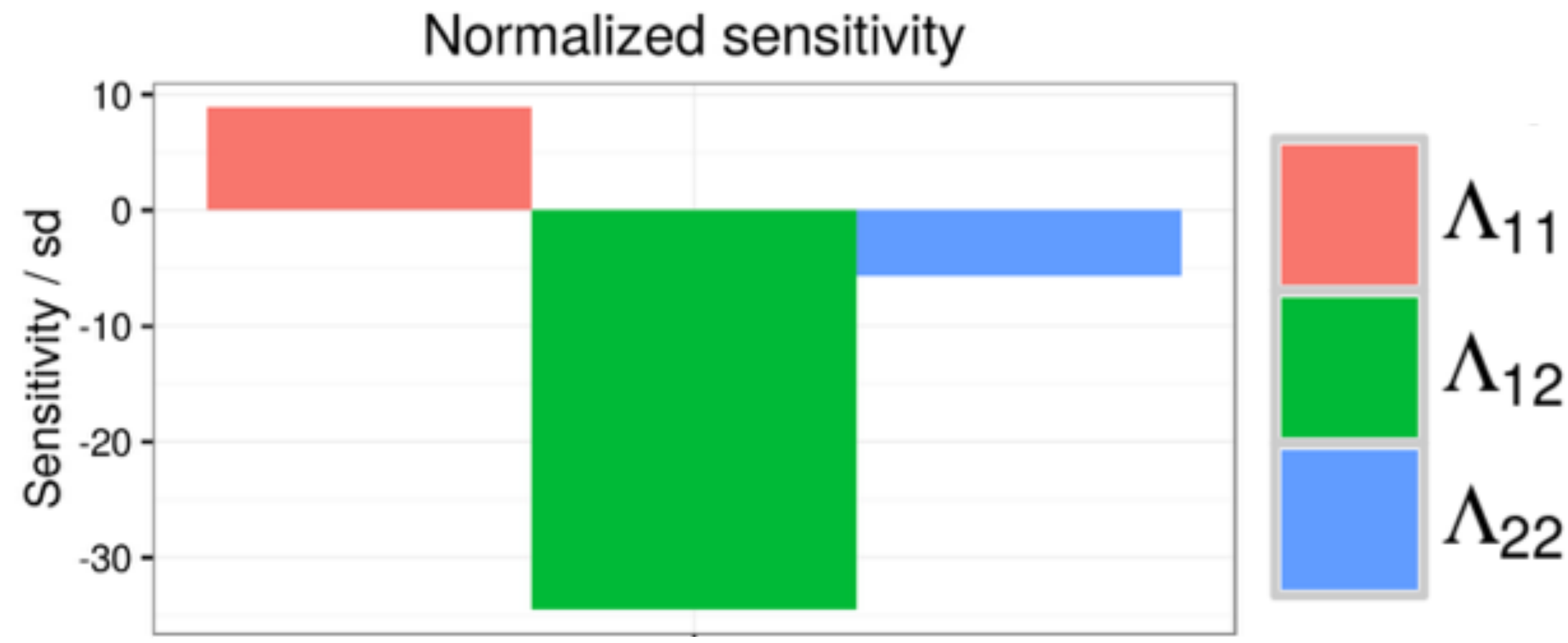
- Sensitivity of the expected microcredit effect (τ)
- Normalized to be on scale of τ std devs
- τ mean (MFVB): 3.08 USD PPP
- τ std dev (LRVB): 1.83 USD PPP
- Mean is 1.68 std dev from 0

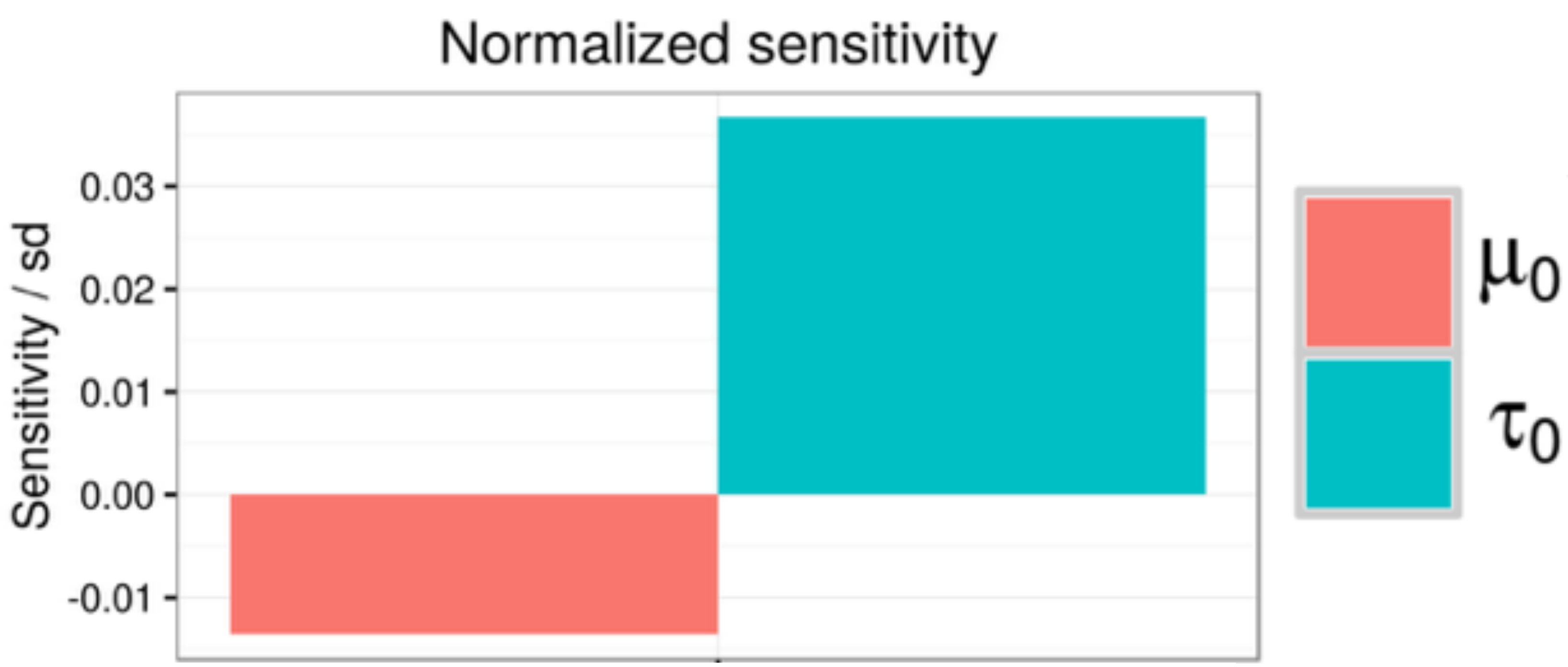




Microcredit Experiment

- 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

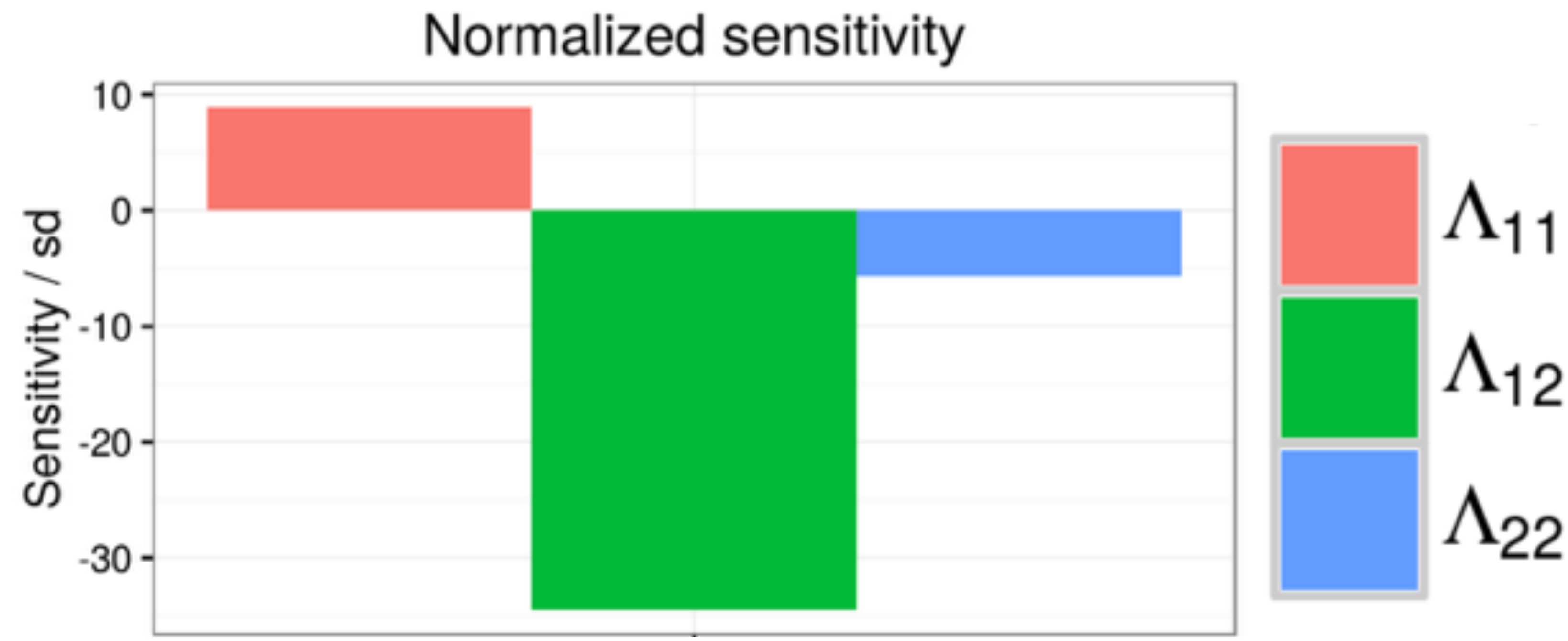


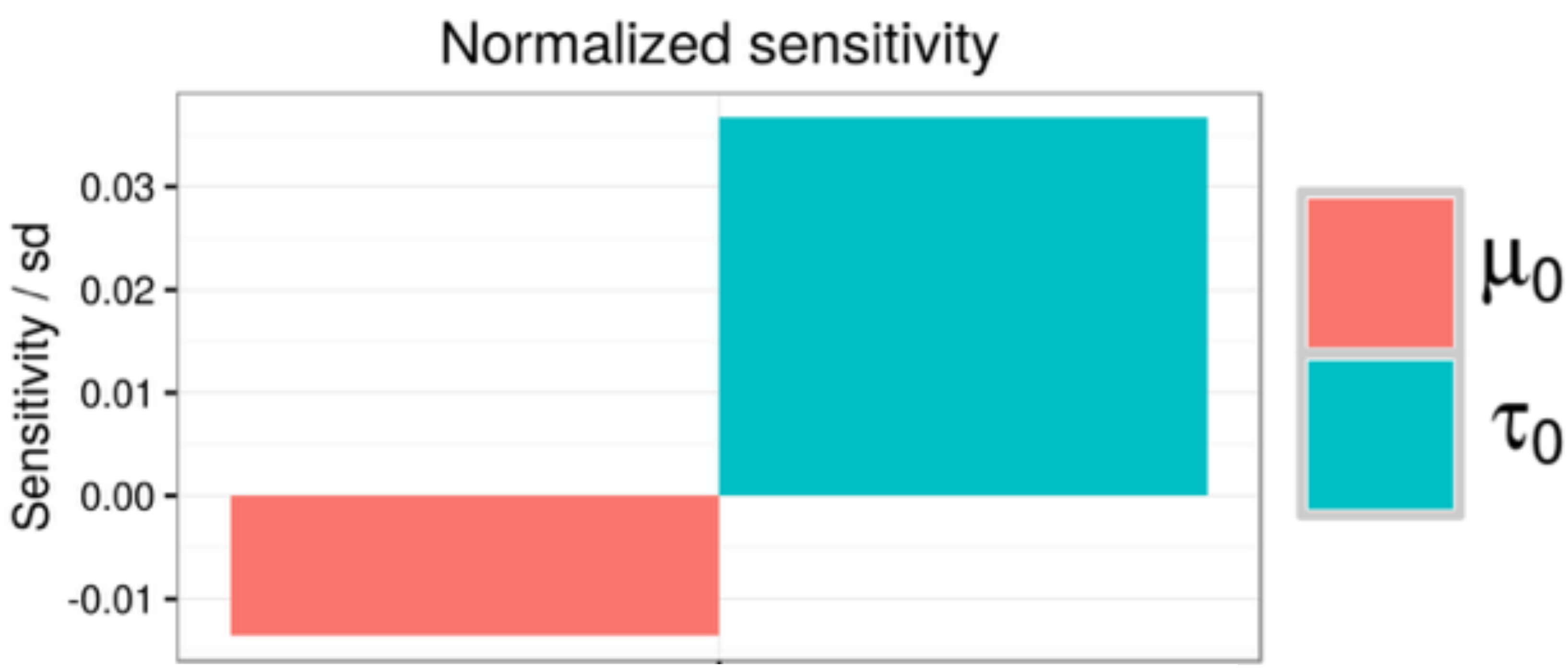


Microcredit Experiment

- 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

$\Rightarrow$ Mean > 2 std dev





Conclusions

- 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 likelihood, data perturbations

References

T Broderick, N Boyd, A Wibisono, AC Wilson, and MI 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, 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.

JH Huggins, T Campbell, and T Broderick. Core sets for scalable Bayesian logistic regression. *NIPS 2016, to appear*. 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*, 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 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.