Variational Bayes and beyond:

Foundations of scalable Bayesian inference (Part II)

Tamara Broderick
Associate Professor
MIT

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

Approximate Bayesian inference

Approximate Bayesian inference

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

Approximate Bayesian inference

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

Optimization

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

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

- Coordinate descent

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

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

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

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

Air pollution: Particulate matter

[Krongut 2020]

Air pollution: Particulate matter

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

- Model:

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

- Exercise: check

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

[Krongut 2020]

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

[Krongut 2020]

- Exercise: check

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

- MFVB approximation:

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

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

[Krongut 2020]

- Exercise: check

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

- MFVB approximation:

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

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

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

[Krongut 2020]

- Exercise: check

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

- MFVB approximation:

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

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

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

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

[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter

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

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

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

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

[Krongut 2020]

- Exercise: check

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

- MFVB approximation:

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

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

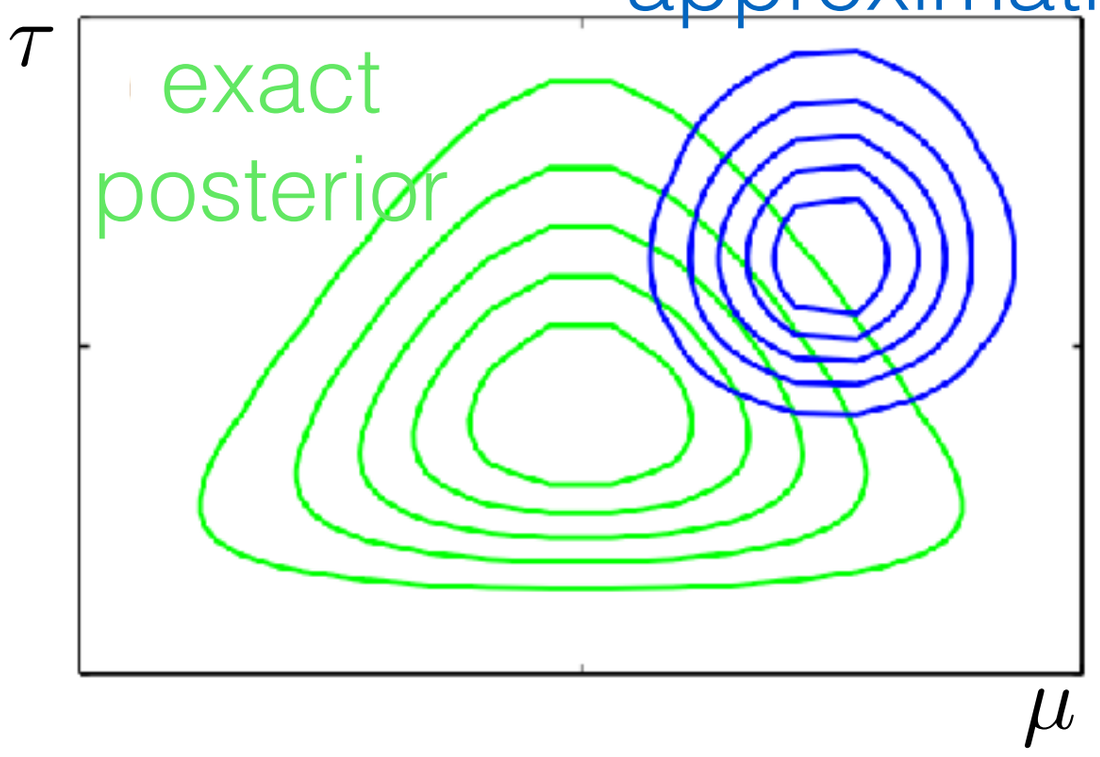
$$q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1})$$

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

"variational parameters"

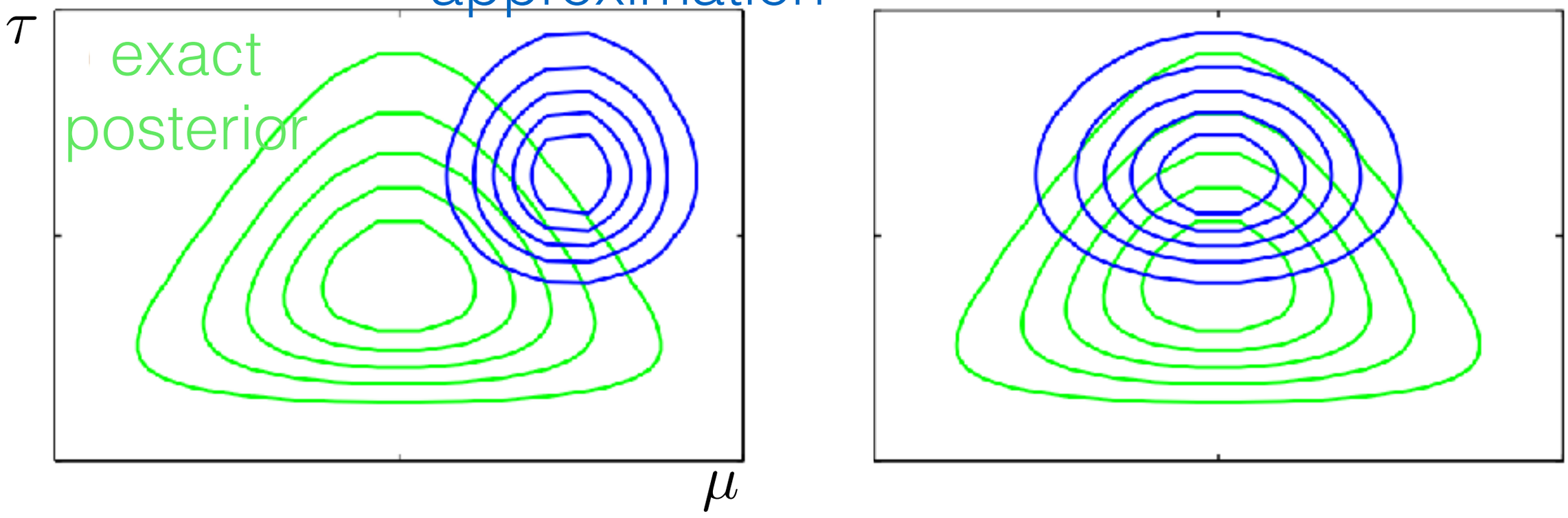
[MacKay 2003; Bishop 2006]

Air pollution: Particulate matter



[Bishop 2006]

Air pollution: Particulate matter



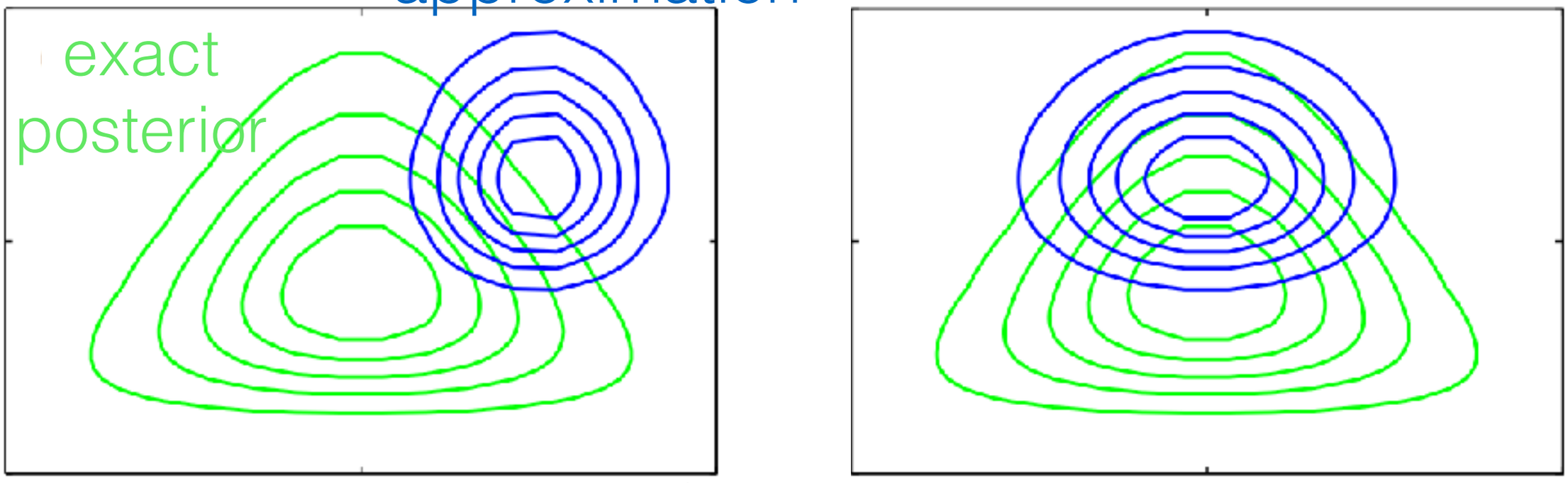
[Bishop 2006]

Air pollution: Particulate matter

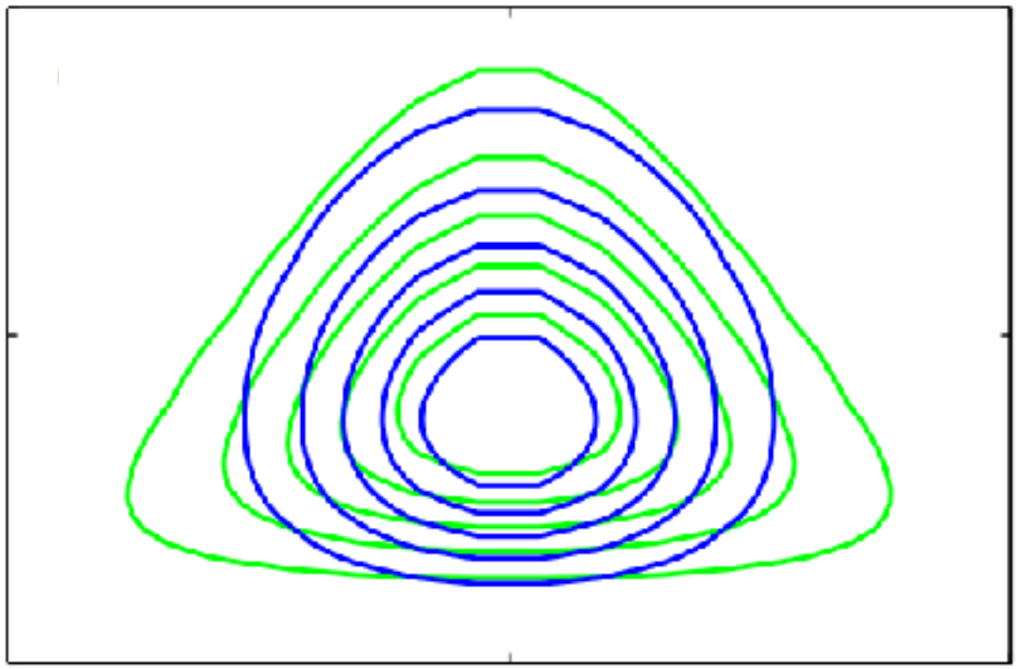
approximation

τ

exact posterior

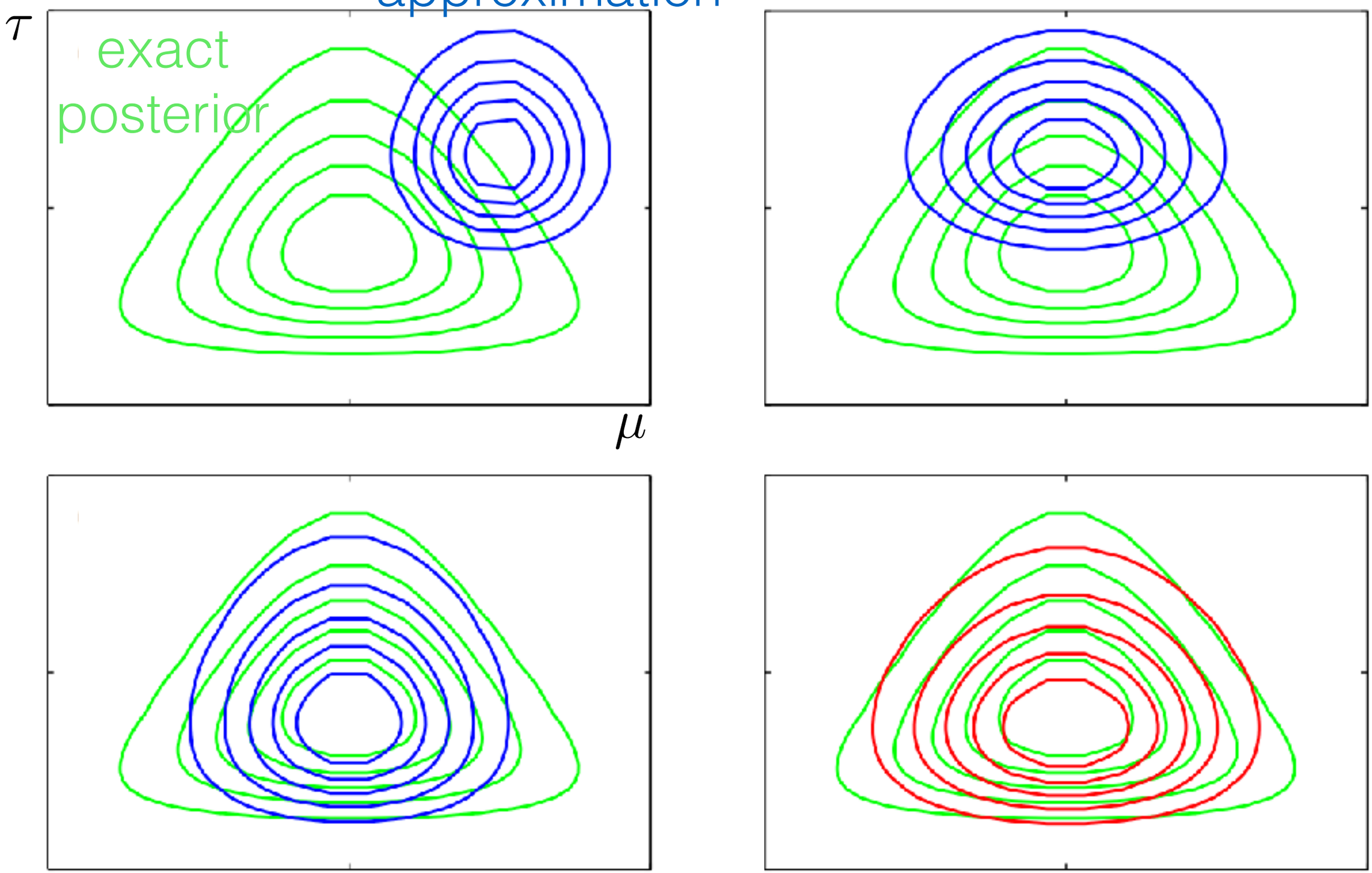


μ



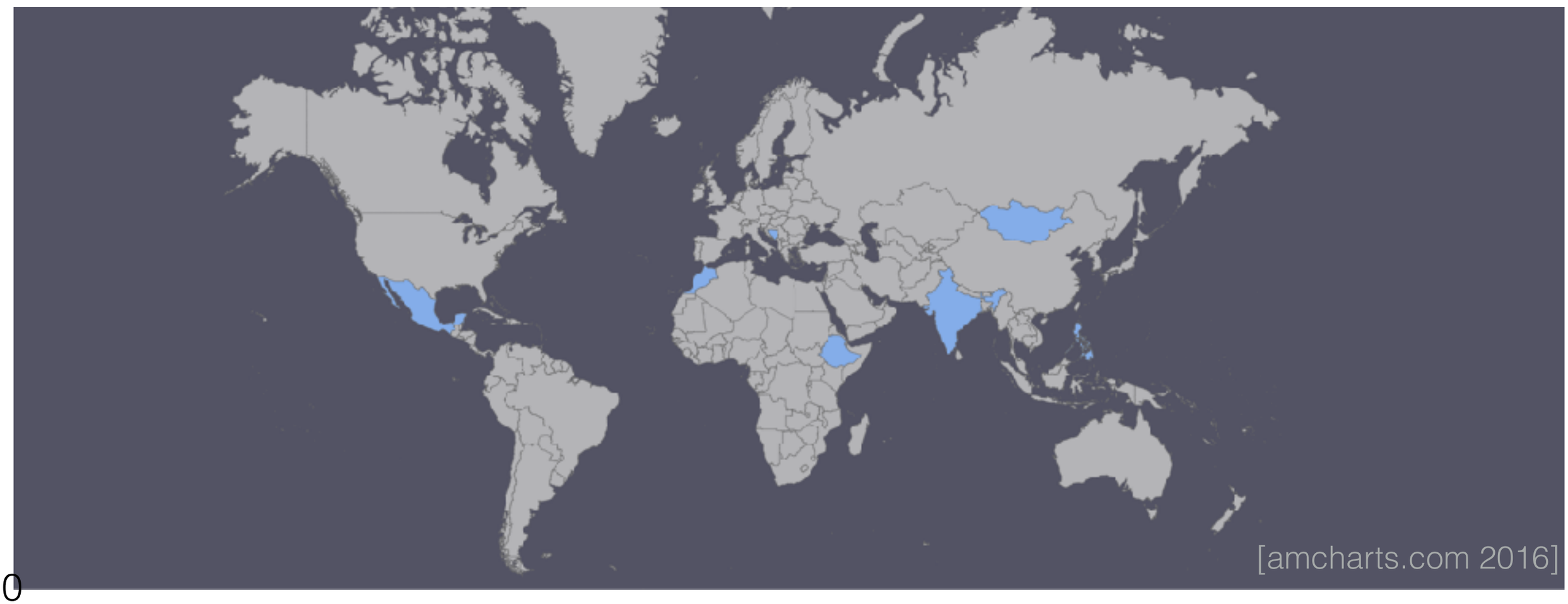
[Bishop 2006]

Air pollution: Particulate matter



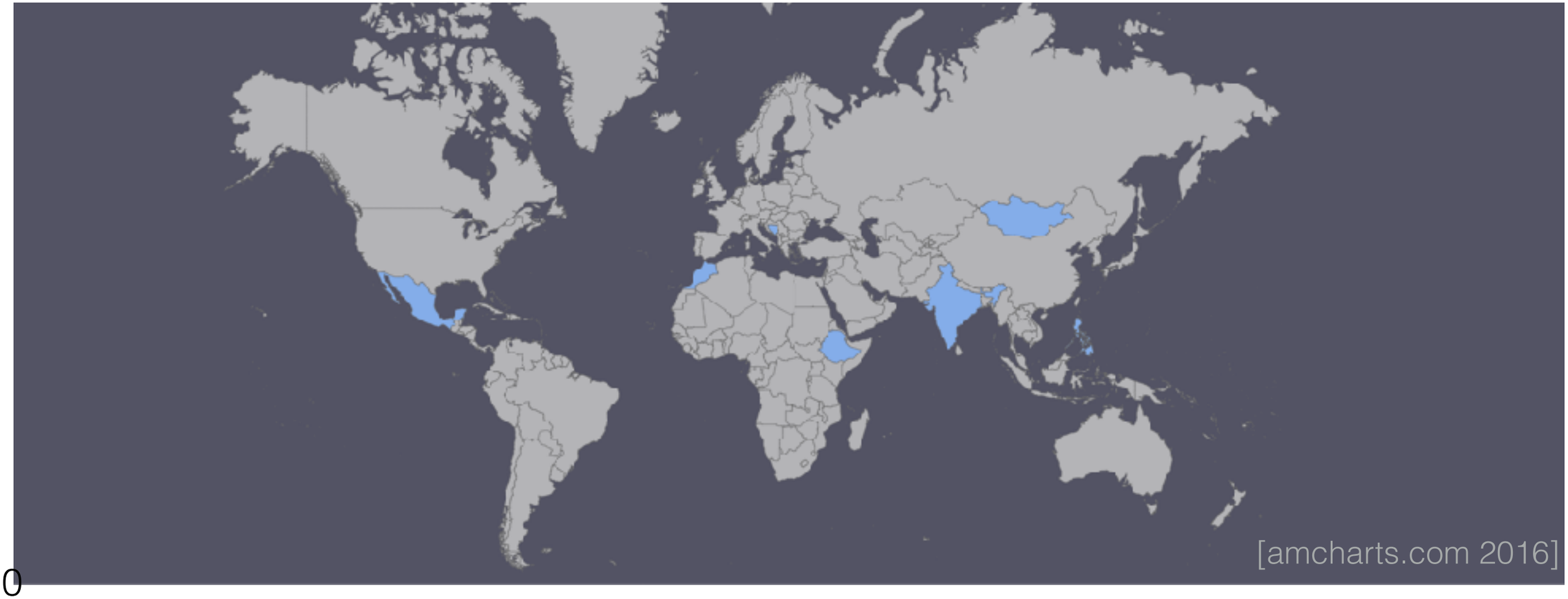
[Bishop 2006]

Microcredit Experiment



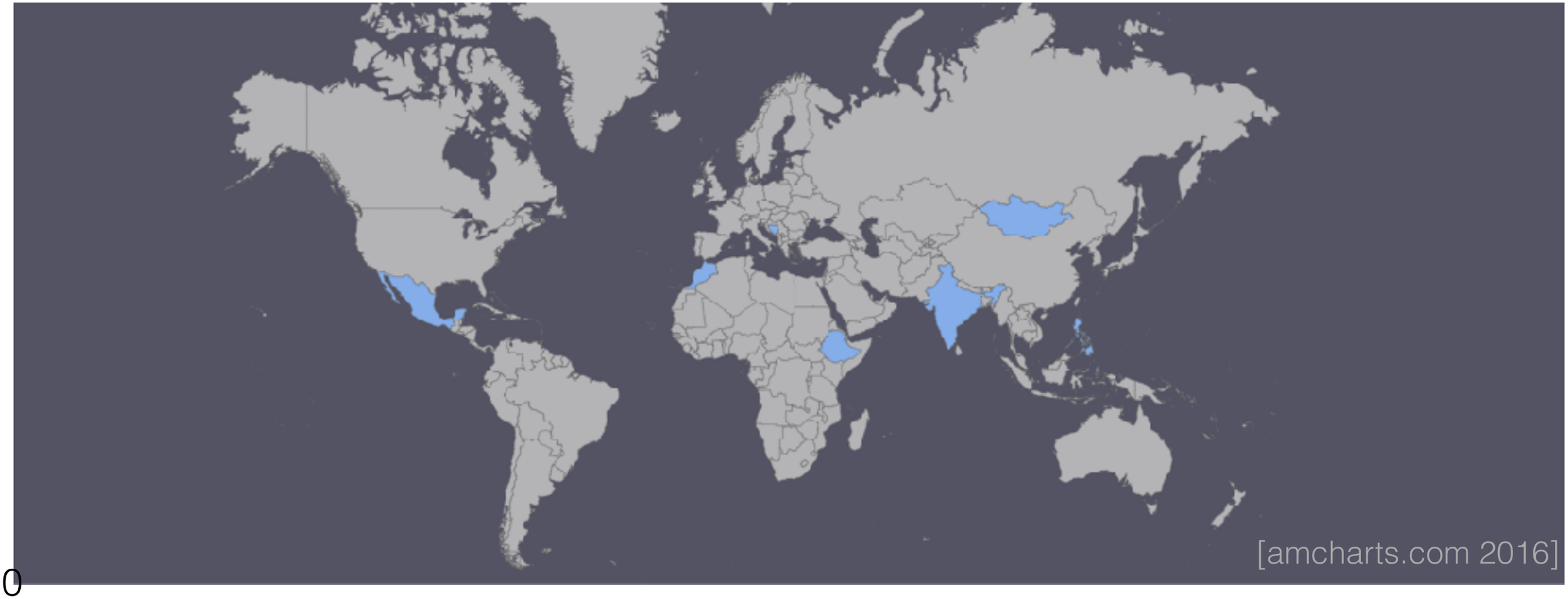
Microcredit Experiment

- Simplified from Meager (2019)



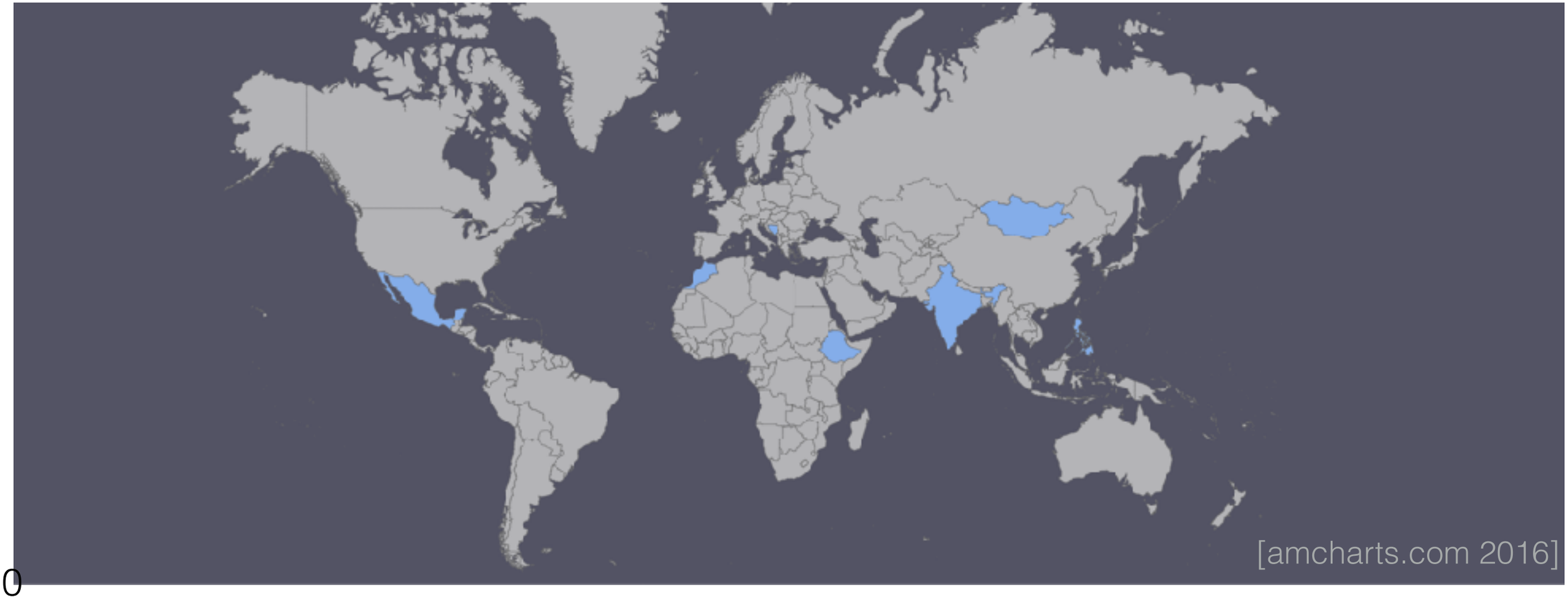
Microcredit Experiment

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



Microcredit Experiment

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



Microcredit Experiment

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

Microcredit Experiment

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

Microcredit Experiment

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

profit

$$y_{kn}$$

Microcredit Experiment

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

profit

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

Microcredit Experiment

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

profit

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

Microcredit Experiment

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

profit

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

Microcredit Experiment

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

1 if microcredit

profit

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

Microcredit Experiment

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

1 if microcredit

profit

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

Microcredit Experiment

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

1 if microcredit

profit

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

Microcredit Experiment

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

1 if microcredit

profit

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

Microcredit Experiment

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

1 if microcredit

profit

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

- Priors and hyperpriors:

Microcredit Experiment

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

1 if microcredit

profit

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

- Priors and hyperpriors:

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

Microcredit Experiment

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

1 if microcredit

profit

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

- Priors and hyperpriors:

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

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

Microcredit Experiment

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

1 if microcredit

profit

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

- Priors and hyperpriors:

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

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

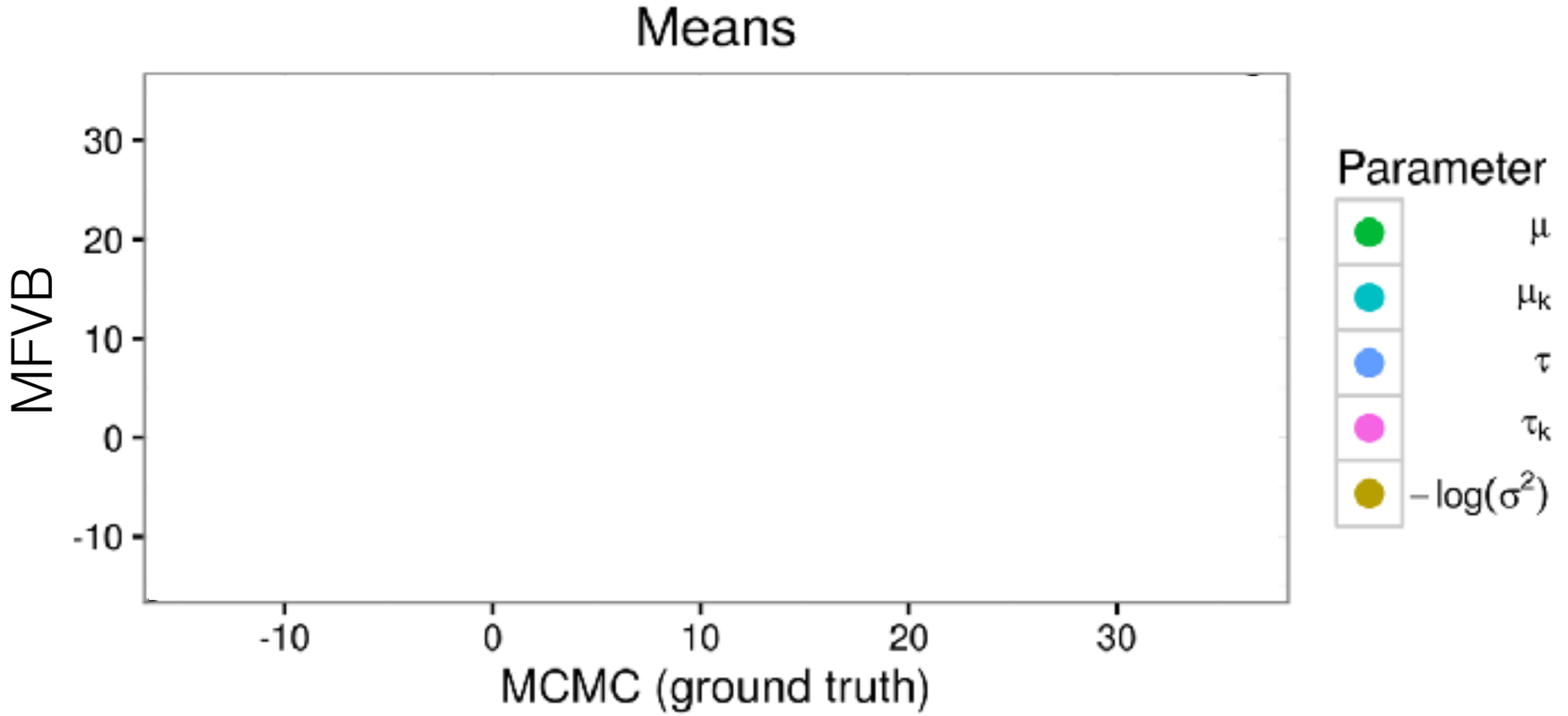
Microcredit

MFVB: Do we need to check the output?

Microcredit

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

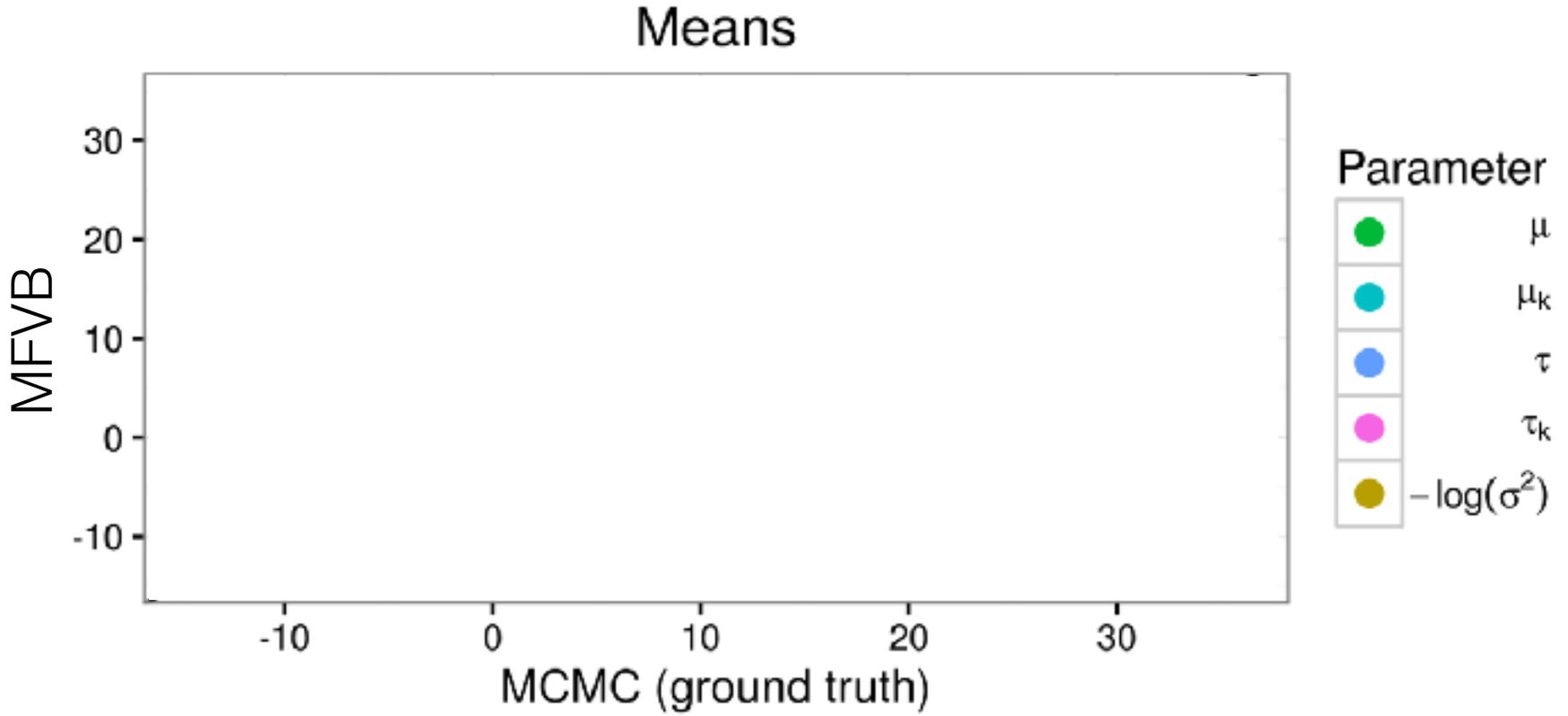
Microcredit



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

Microcredit

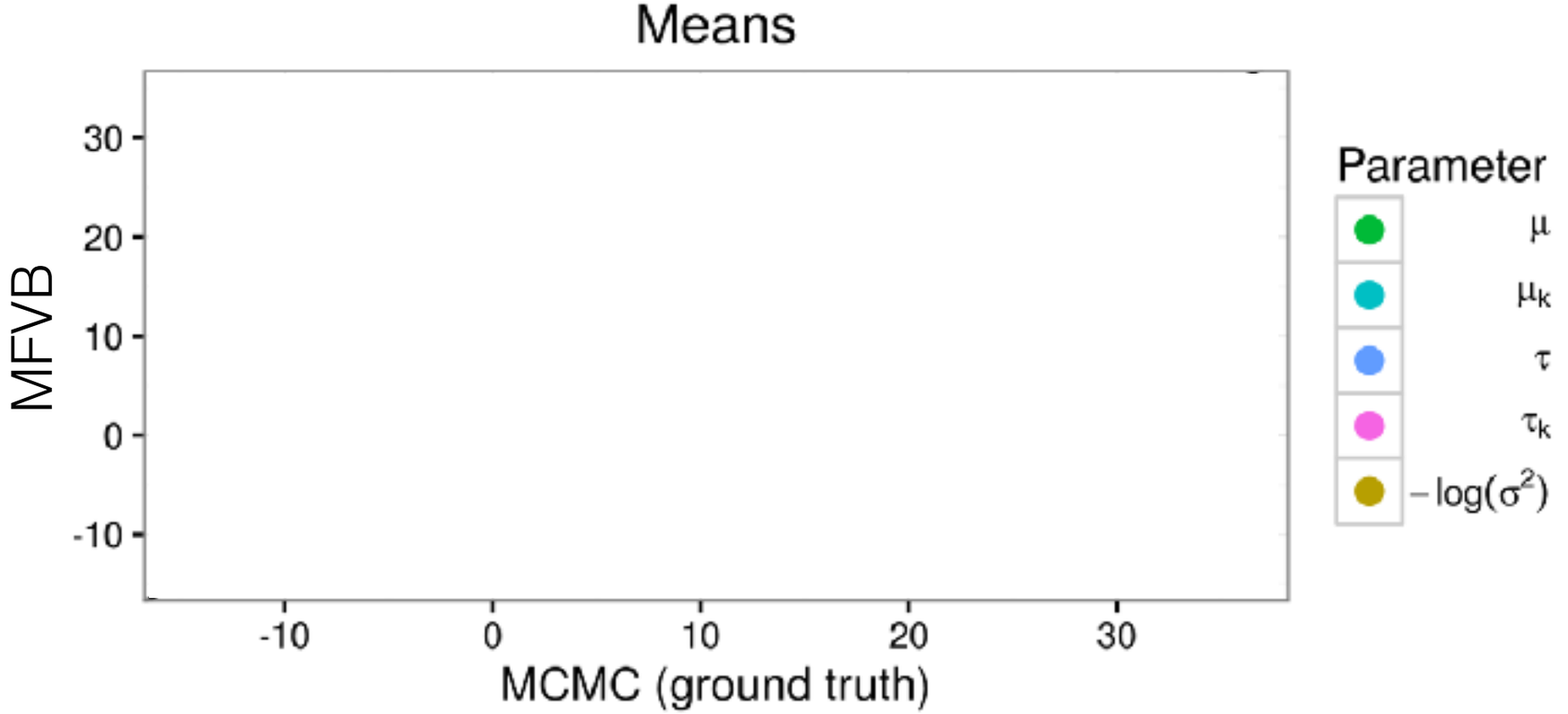
- *One set* of 2500 MCMC draws: **45 minutes**



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

Microcredit

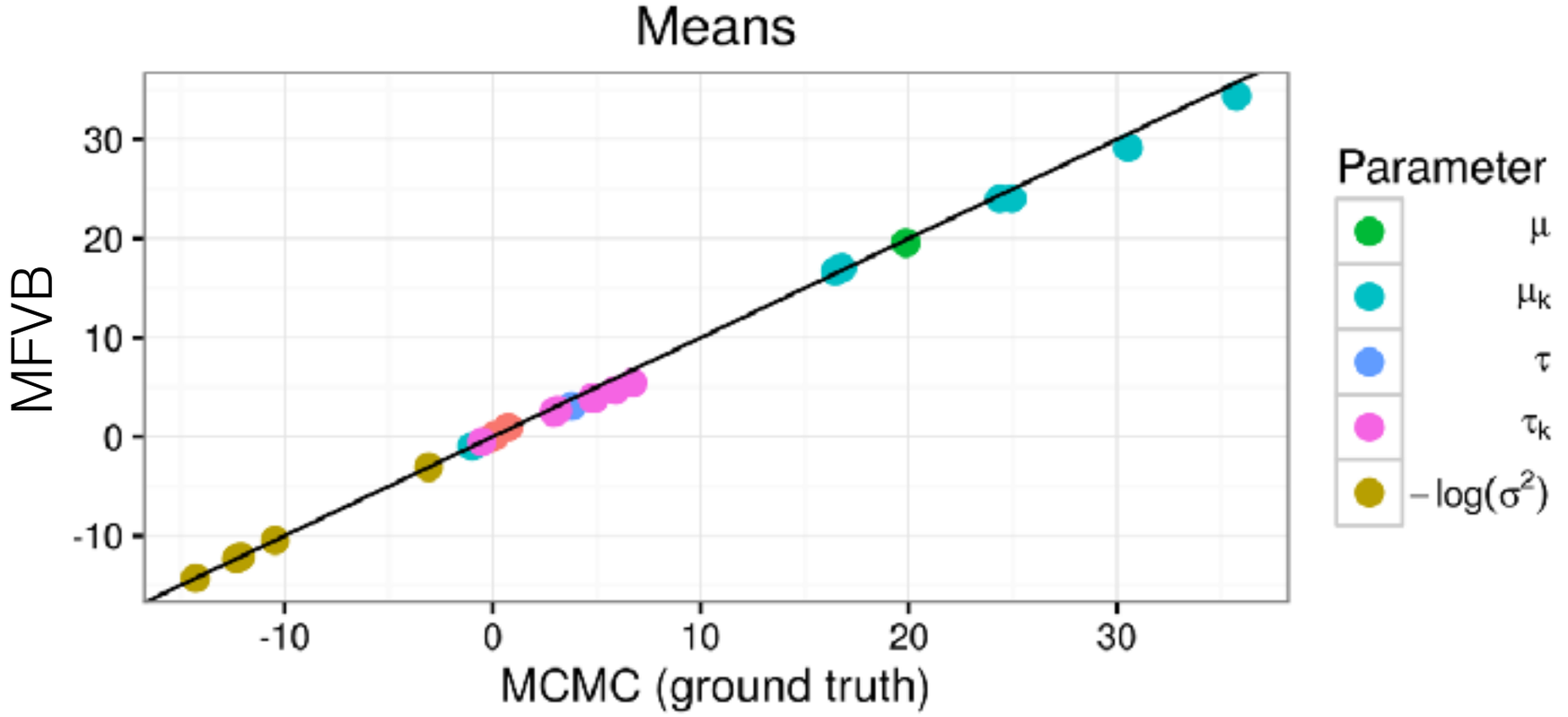
- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



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

Microcredit

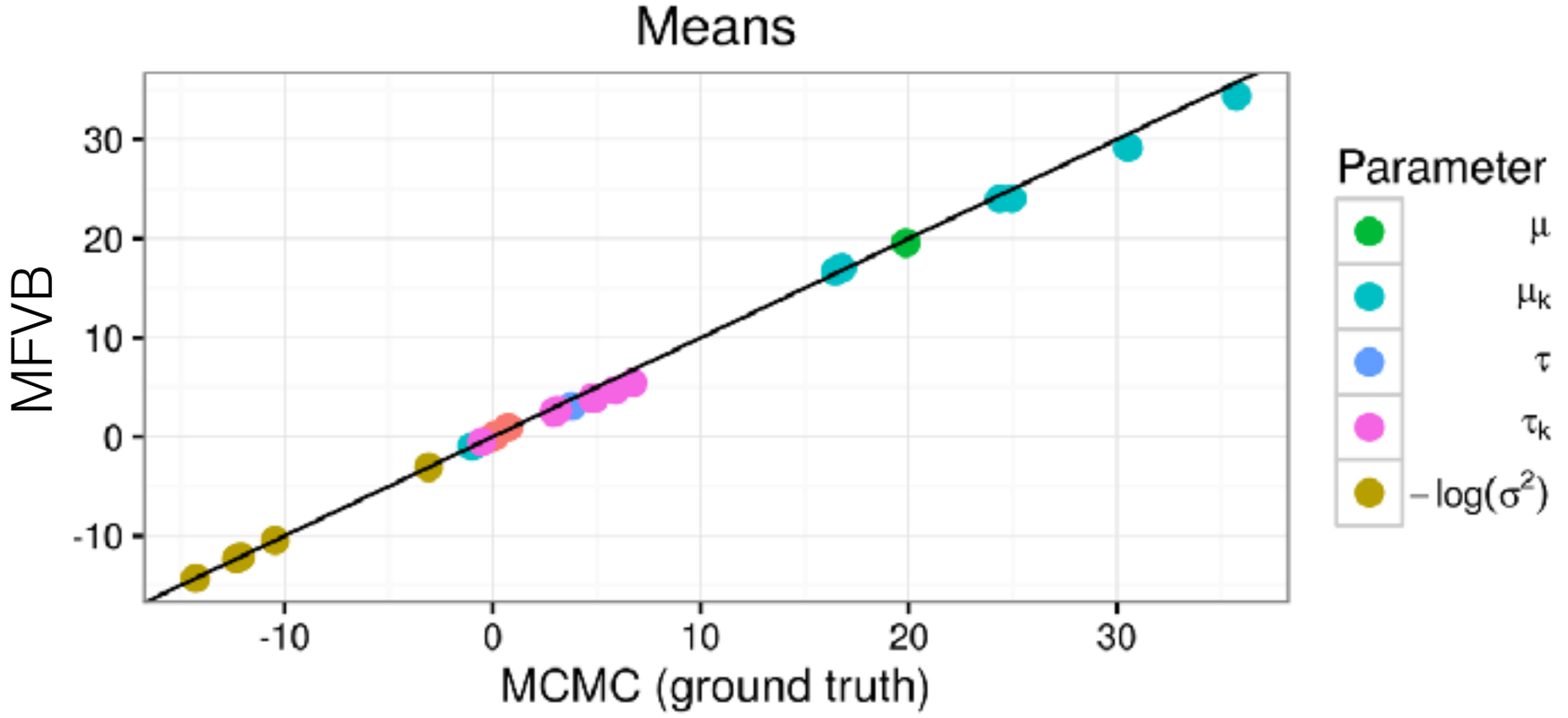
- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



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

Microcredit

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



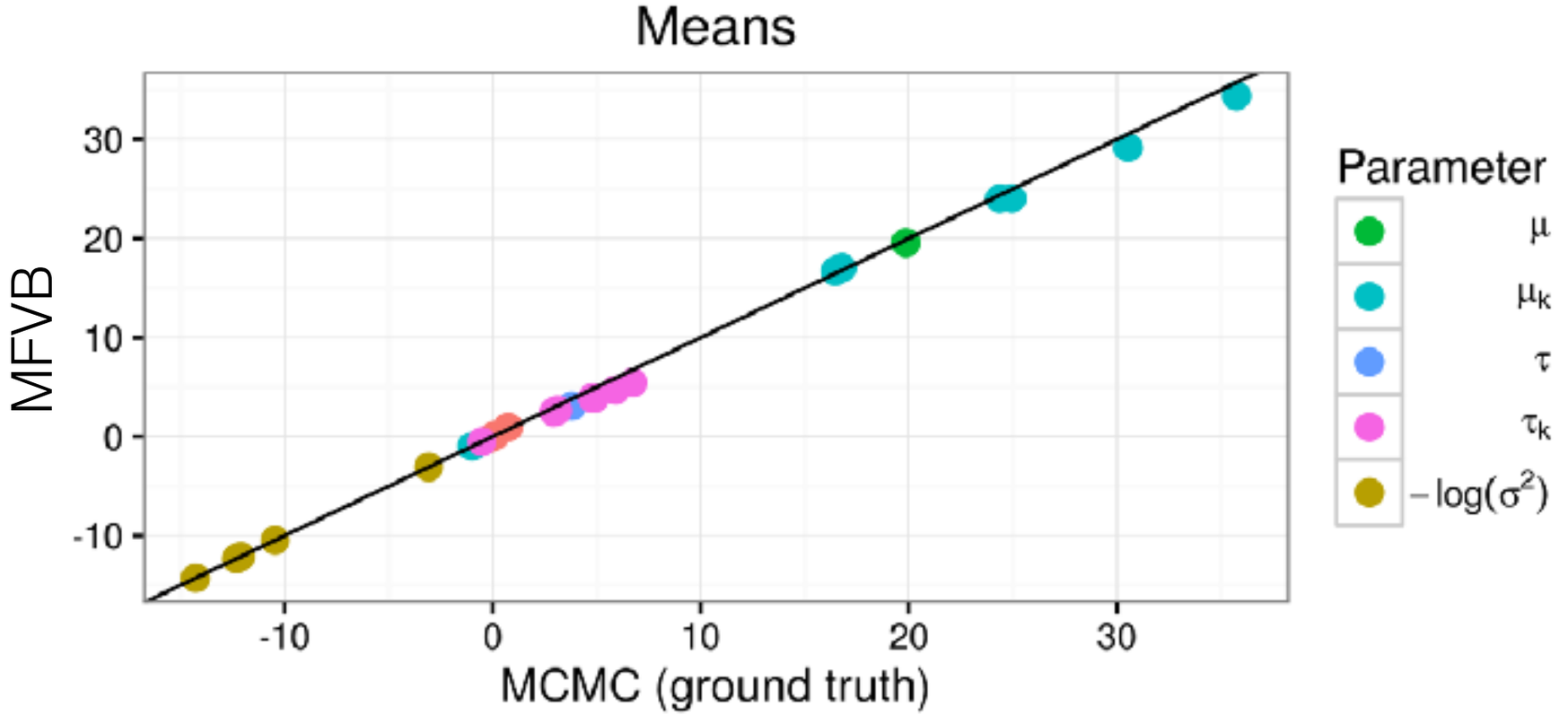
Criteo Online Ads Experiment

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

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

Microcredit

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



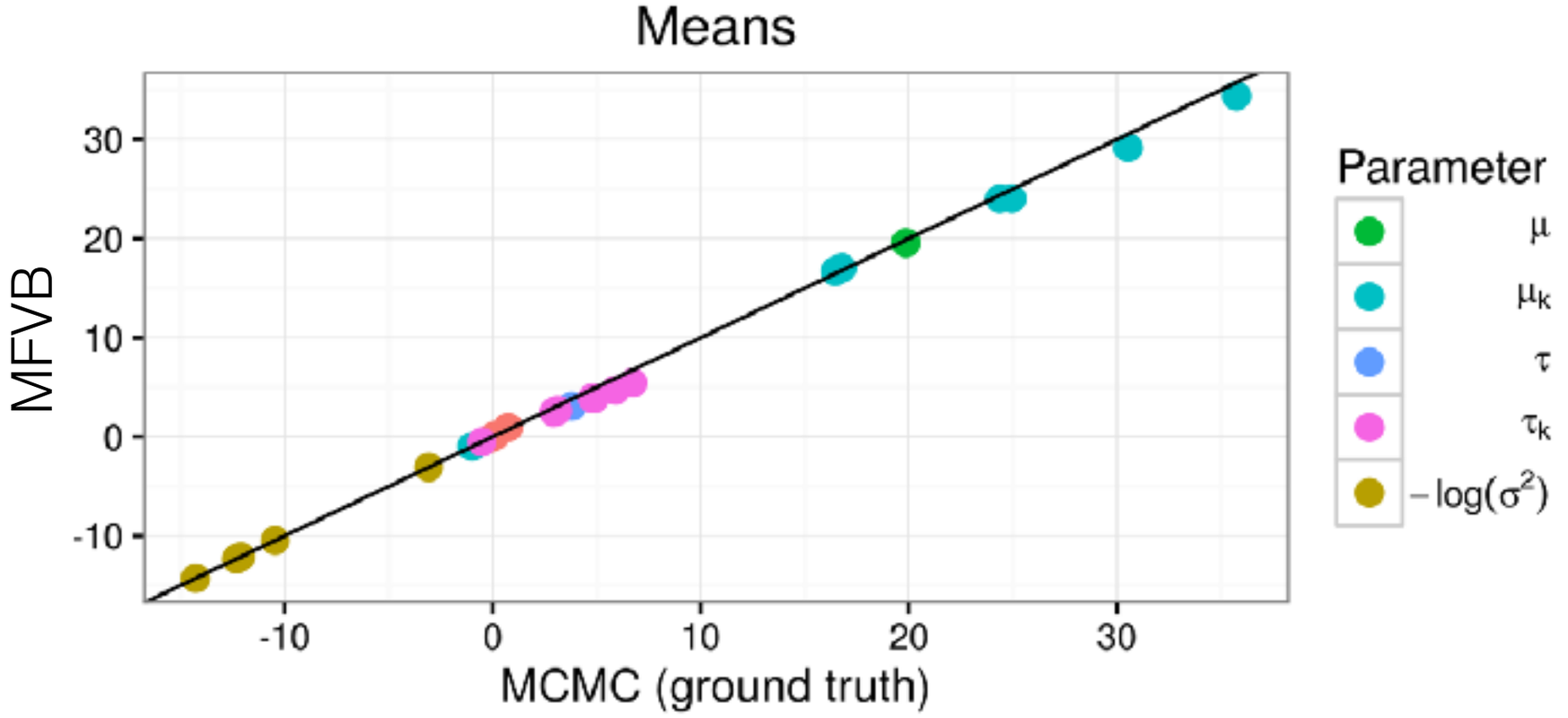
Criteo Online Ads Experiment

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

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

Microcredit

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



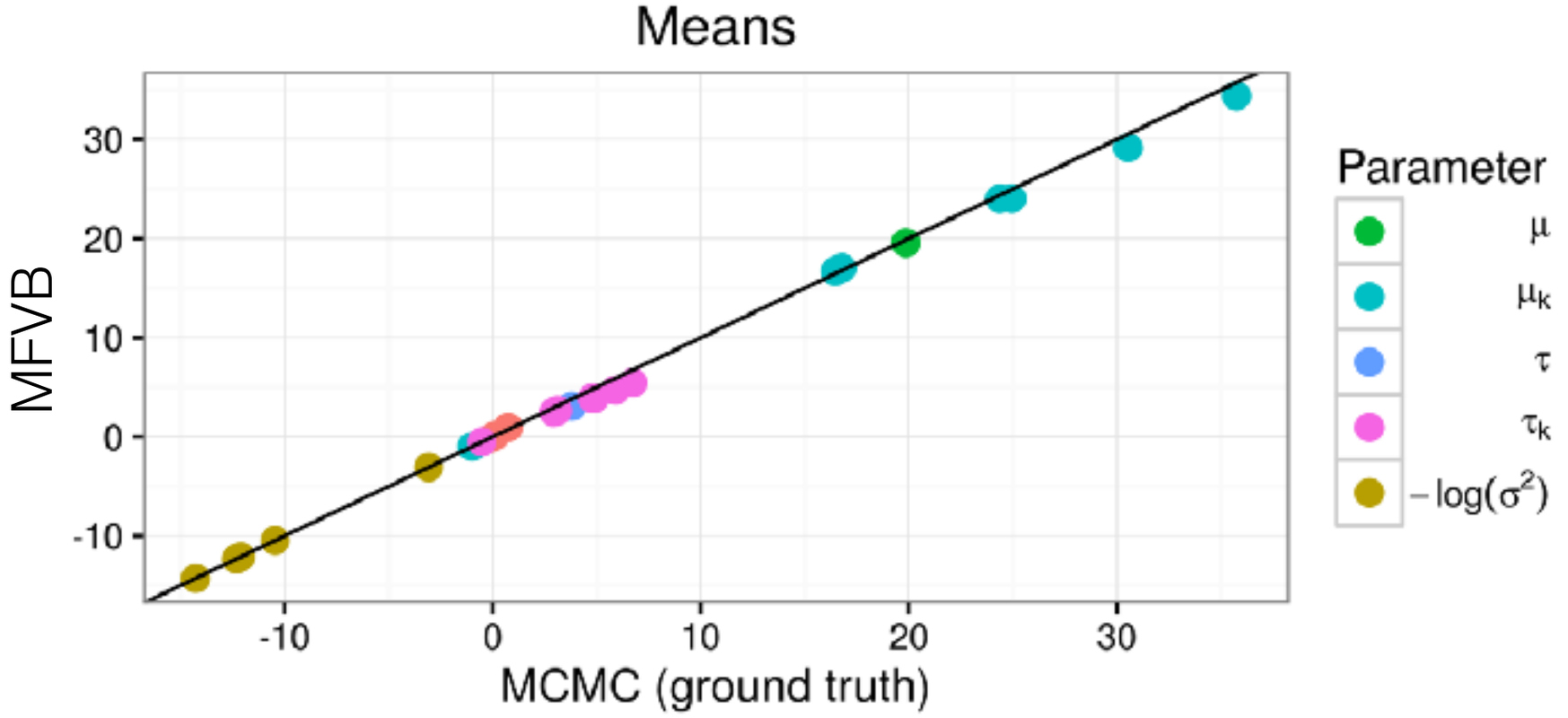
Criteo Online Ads Experiment

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

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

Microcredit

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



Criteo Online Ads Experiment

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

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

Criteo Online Ads Experiment

[Giordano, Broderick, Jordan 2018]

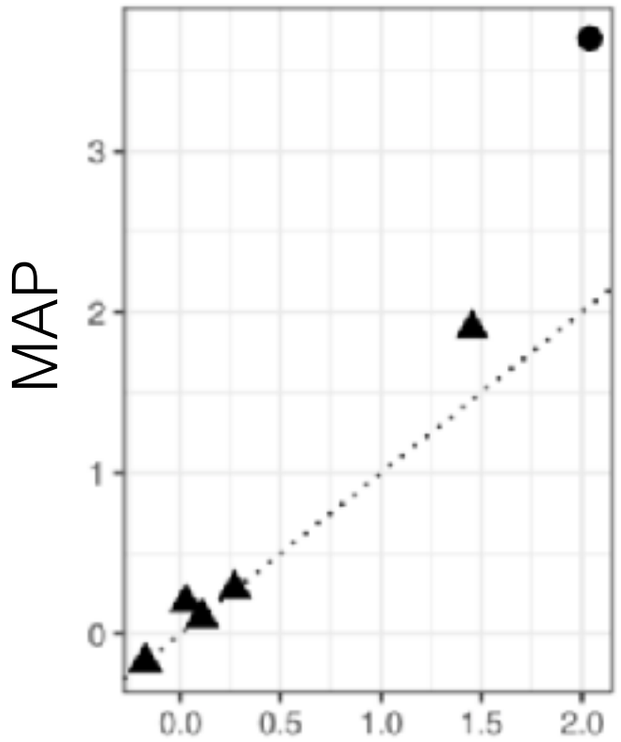
Criteo Online Ads Experiment

- MAP: **12 s**

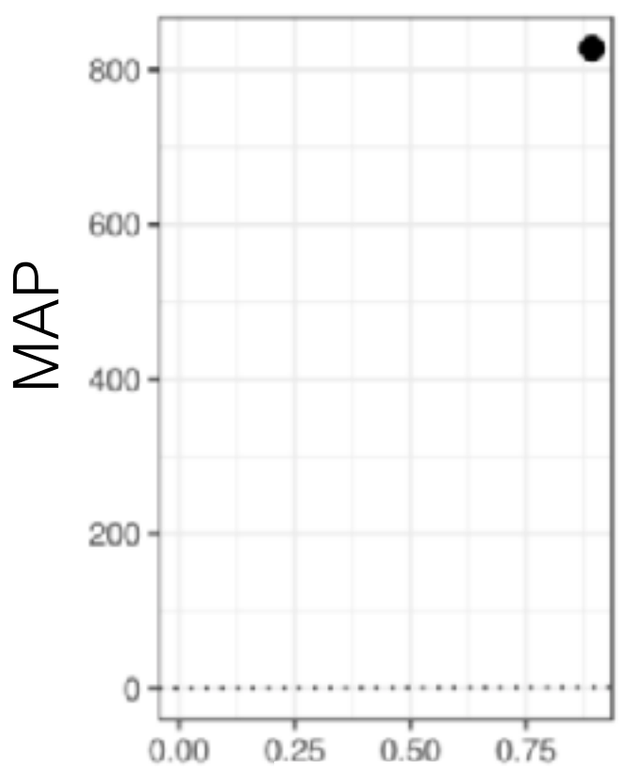
[Giordano, Broderick, Jordan 2018]

Criteo Online Ads Experiment

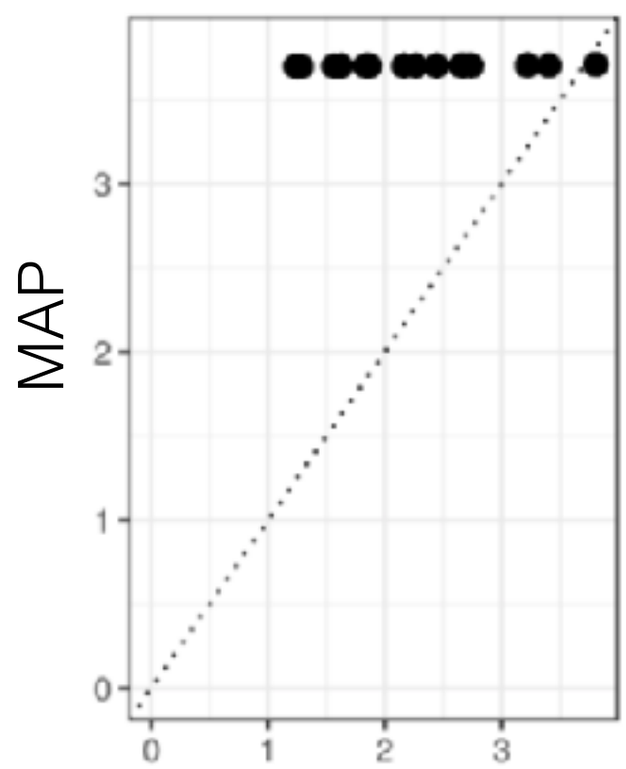
Global parameters (-τ)



Global parameter τ



Local parameters

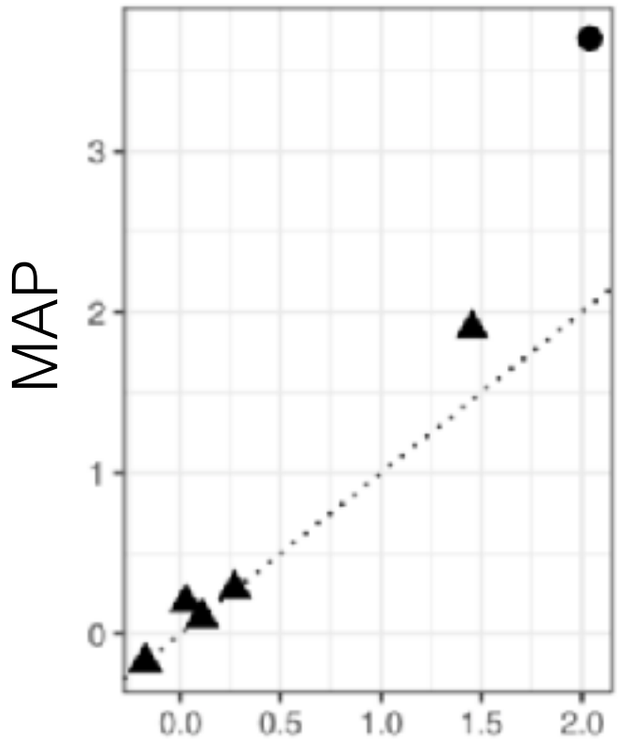


- MAP: **12 s**

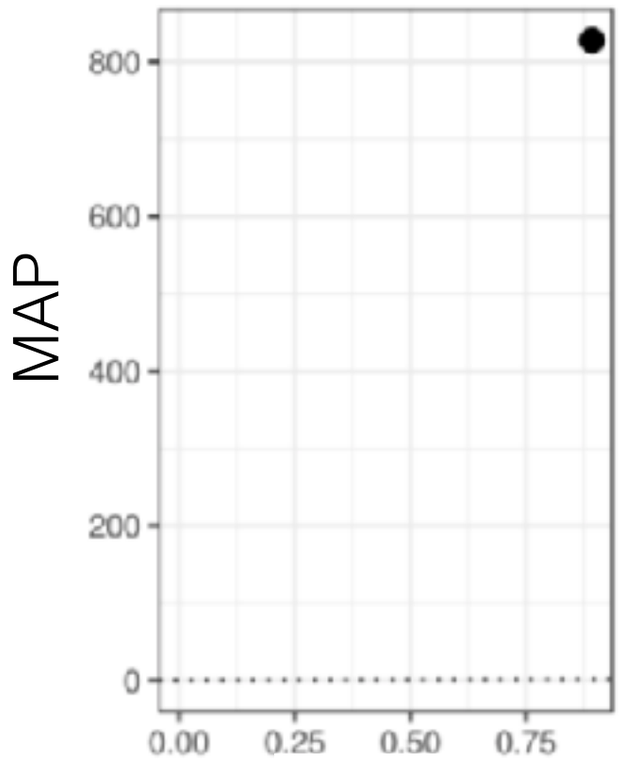
[Giordano, Broderick, Jordan 2018]

Criteo Online Ads Experiment

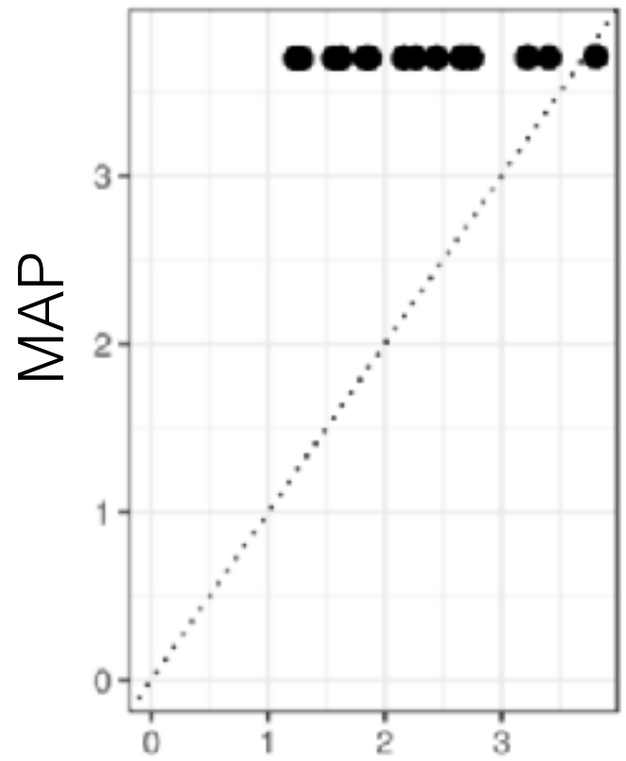
Global parameters (-τ)



Global parameter τ



Local parameters

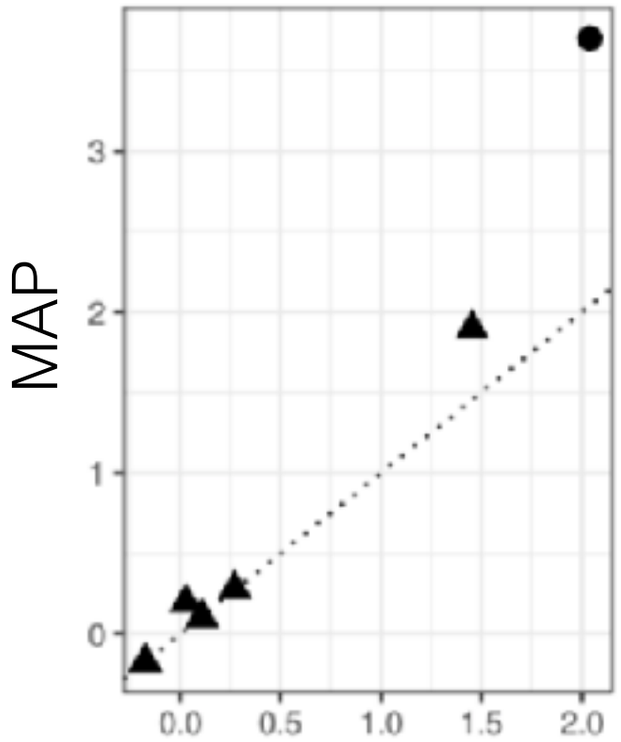


- MAP: **12 s**
- MFVB: **57 s**

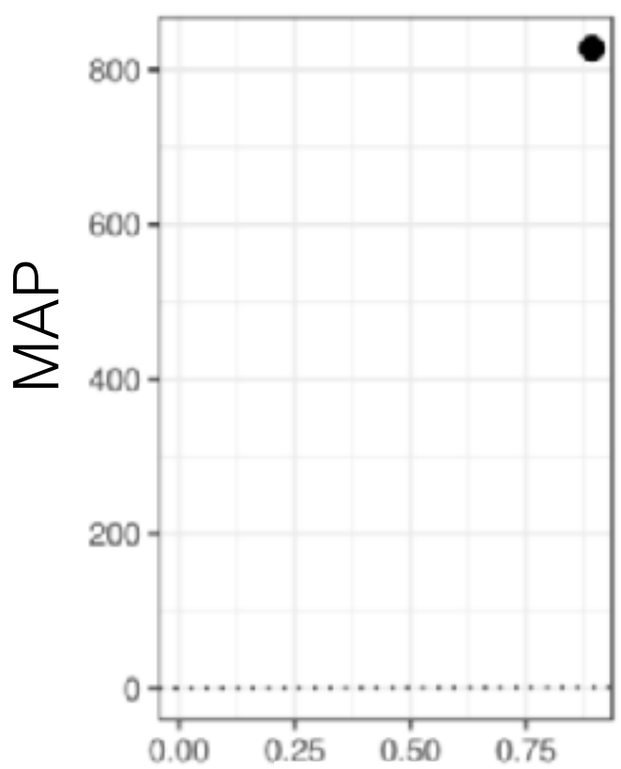
[Giordano, Broderick, Jordan 2018]

Criteo Online Ads Experiment

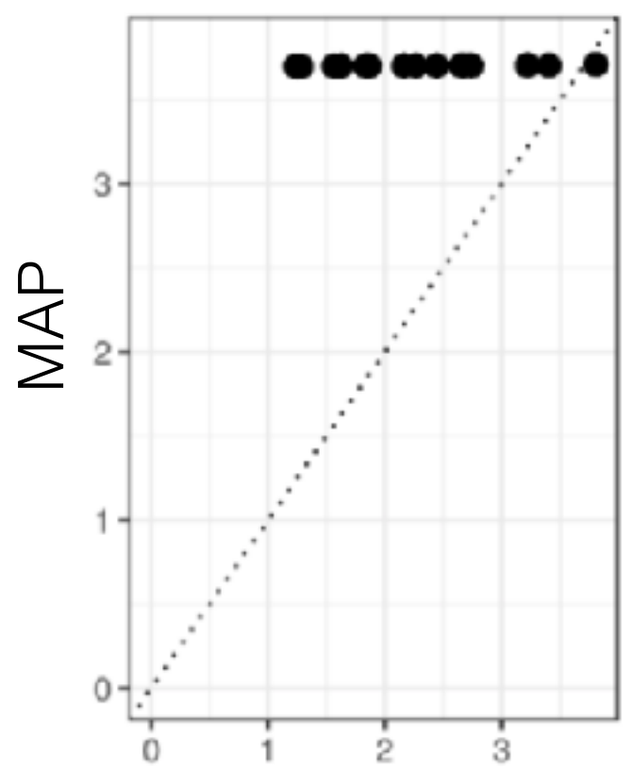
Global parameters (-τ)



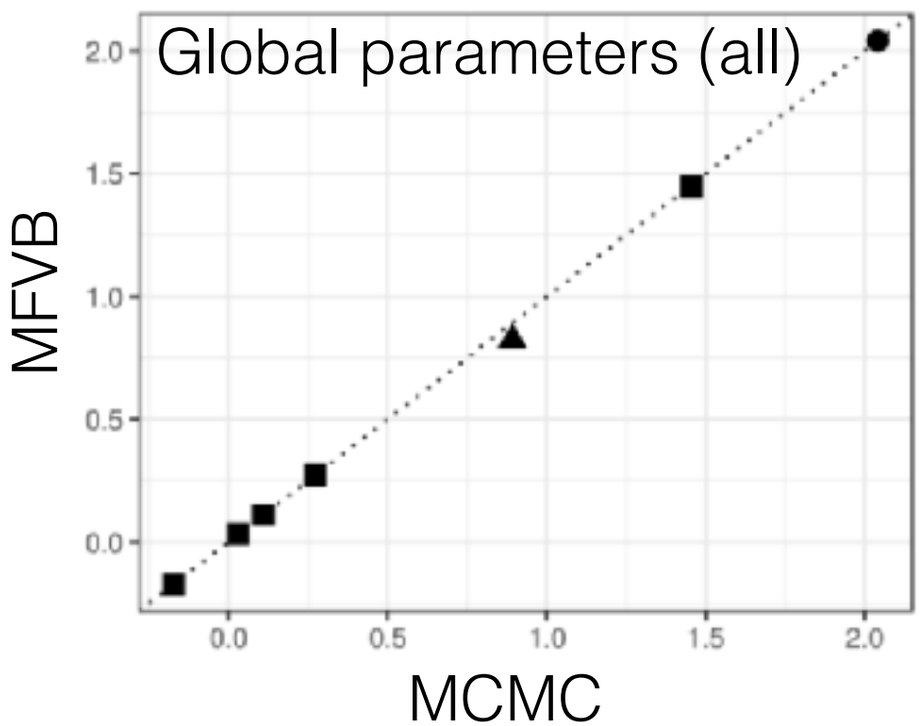
Global parameter τ

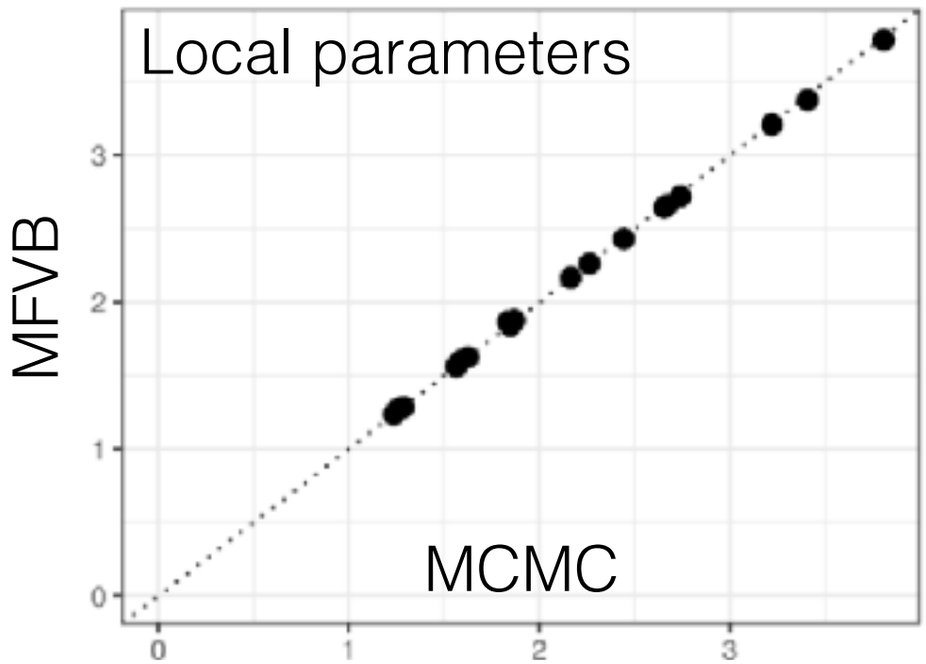


Local parameters



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

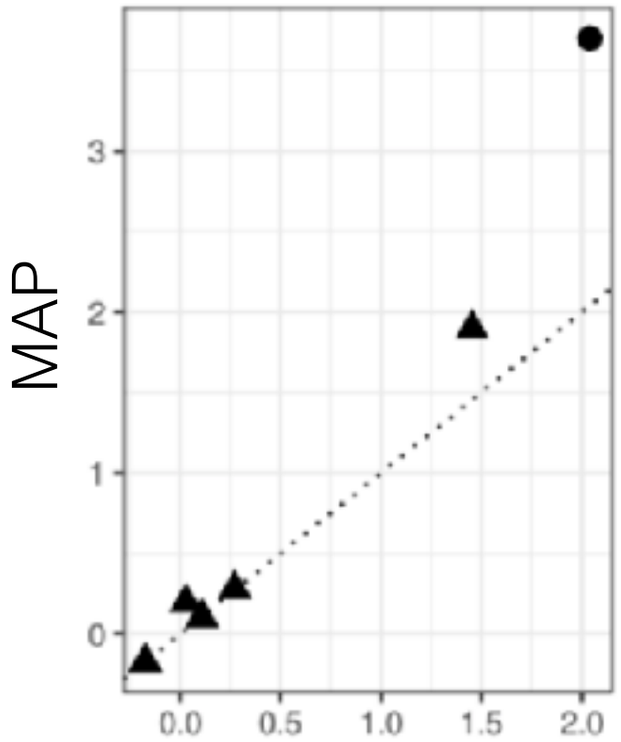




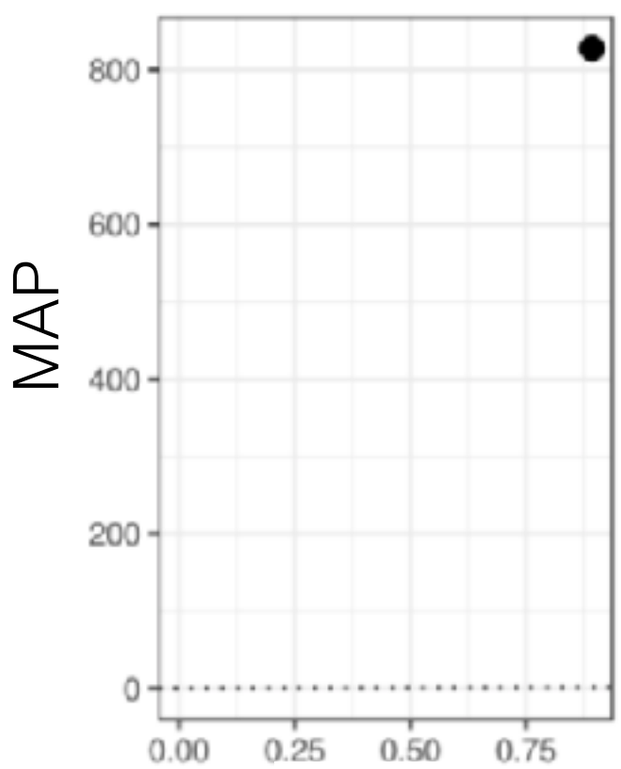
[Giordano, Broderick, Jordan 2018]

Criteo Online Ads Experiment

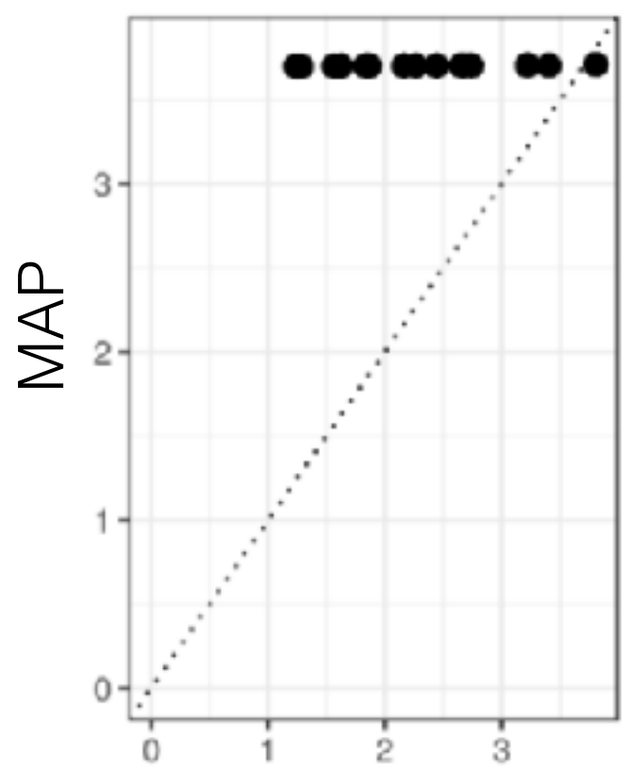
Global parameters (-τ)



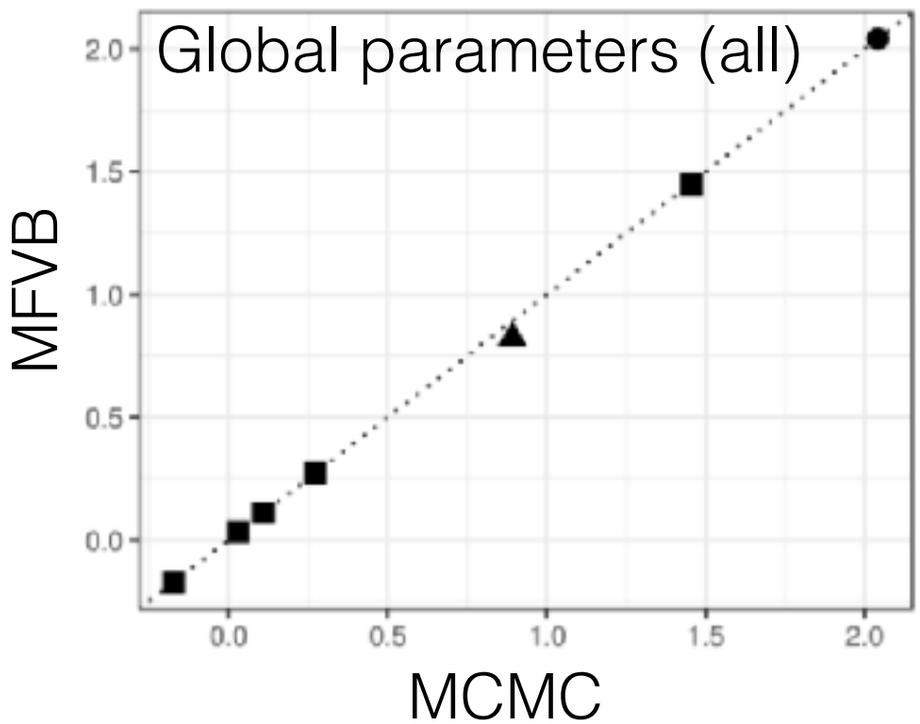
Global parameter τ

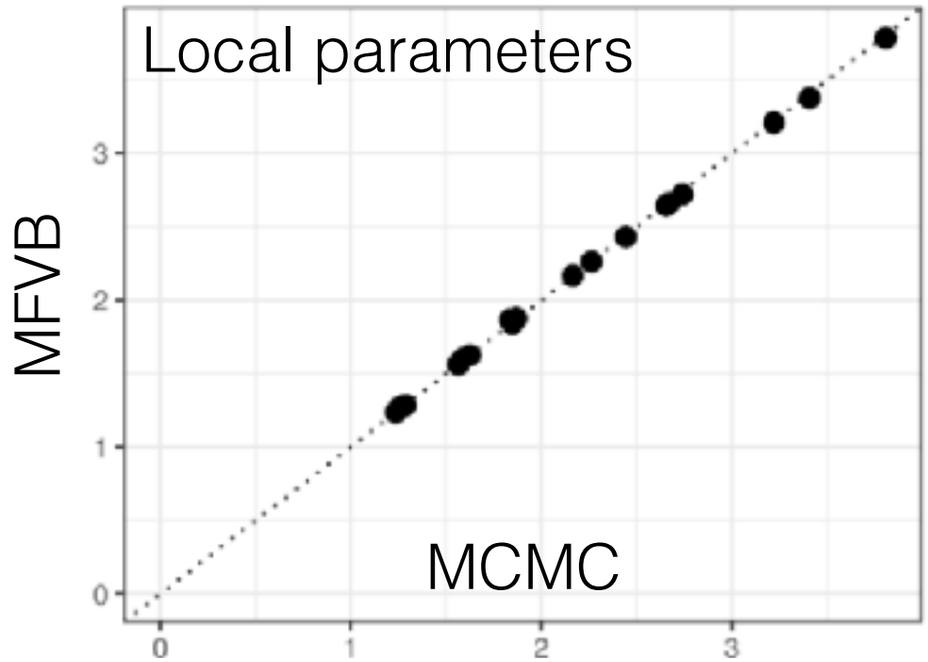


Local parameters



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





[Giordano, Broderick, Jordan 2018]

Why use MFVB?

- Topic discovery

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

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

[Blei et al 2003]

Why use MFVB?

- Topic discovery
  - Latent Dirichlet allocation (LDA)

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

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

[Blei et al 2003]

Why use MFVB?

- Topic discovery
  - Latent Dirichlet allocation (LDA): 33,000+ citations

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

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

[Blei et al 2003]

Roadmap

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

Roadmap

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

What about uncertainty?

What about uncertainty?

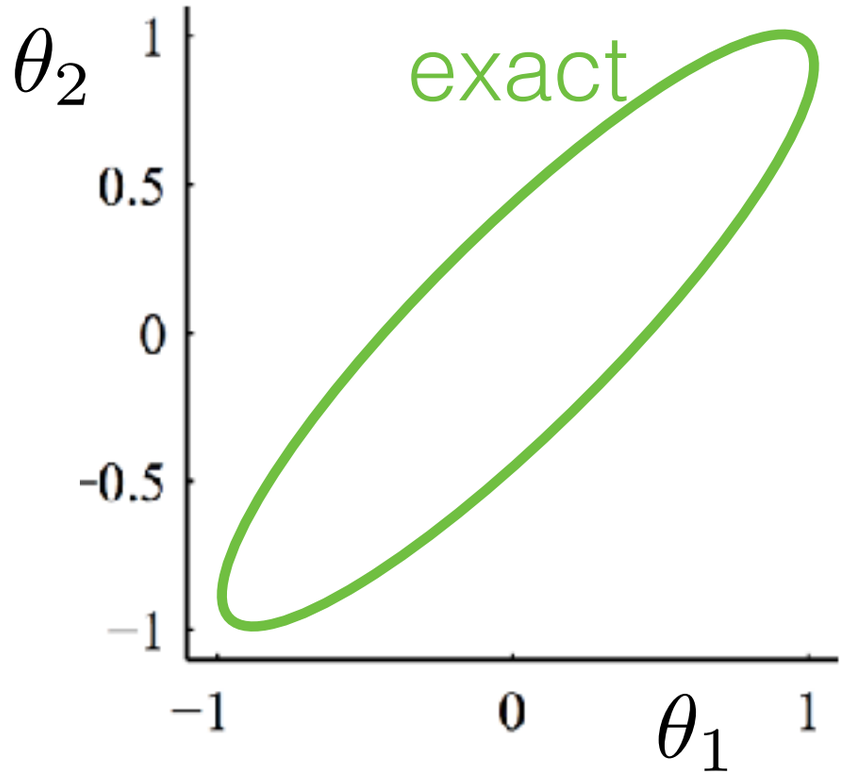
$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

What about uncertainty?

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

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

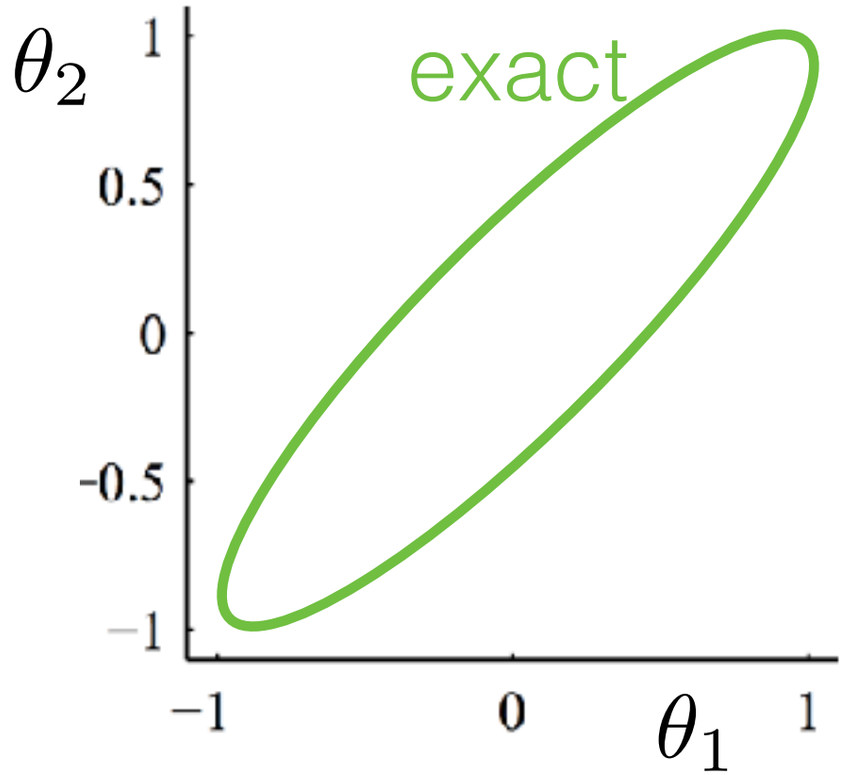


[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

What about uncertainty?

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

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



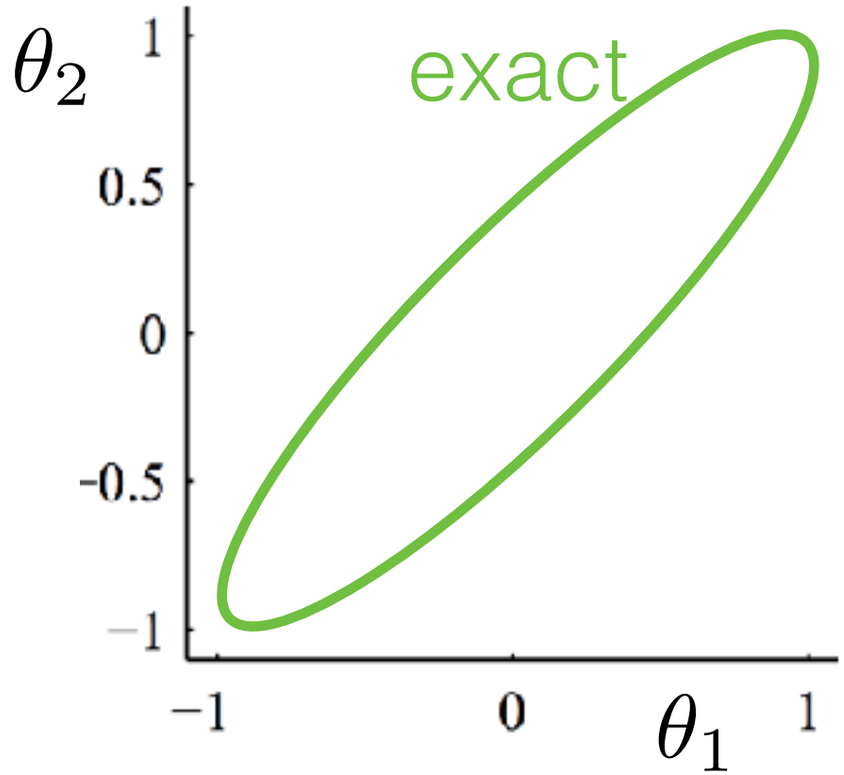
- Conjugate linear regression

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

What about uncertainty?

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

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



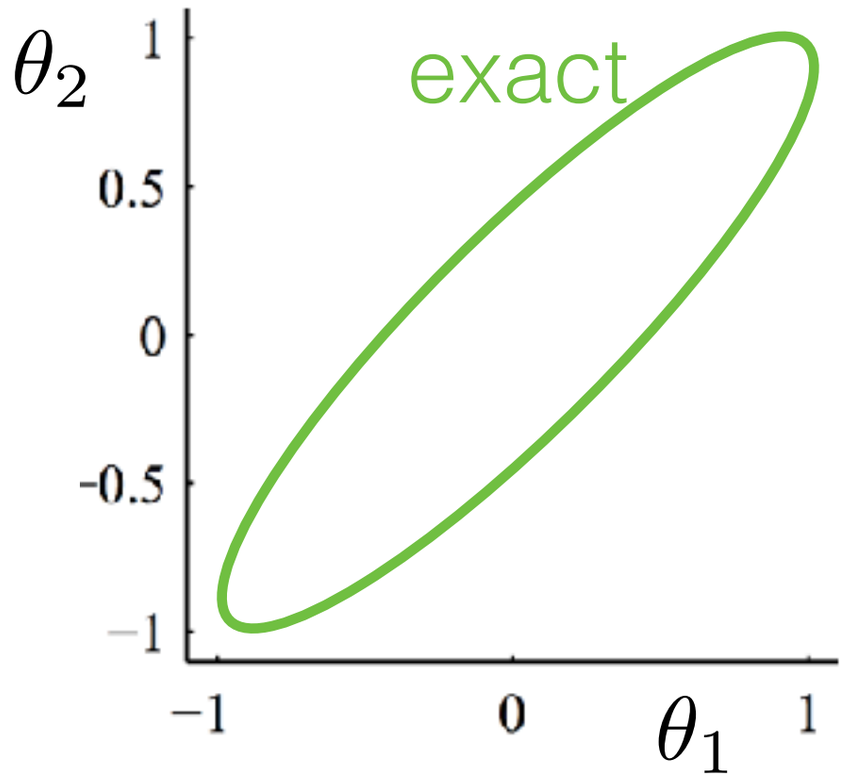
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

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

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



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

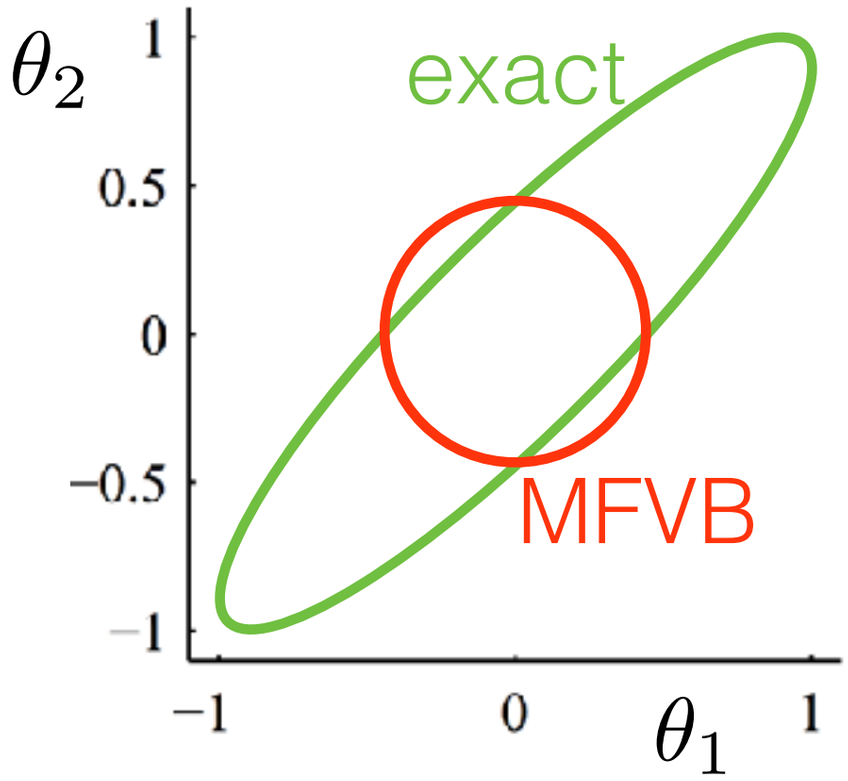
- Conjugate linear regression
- Bayesian central limit theorem

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

What about uncertainty?

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

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



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

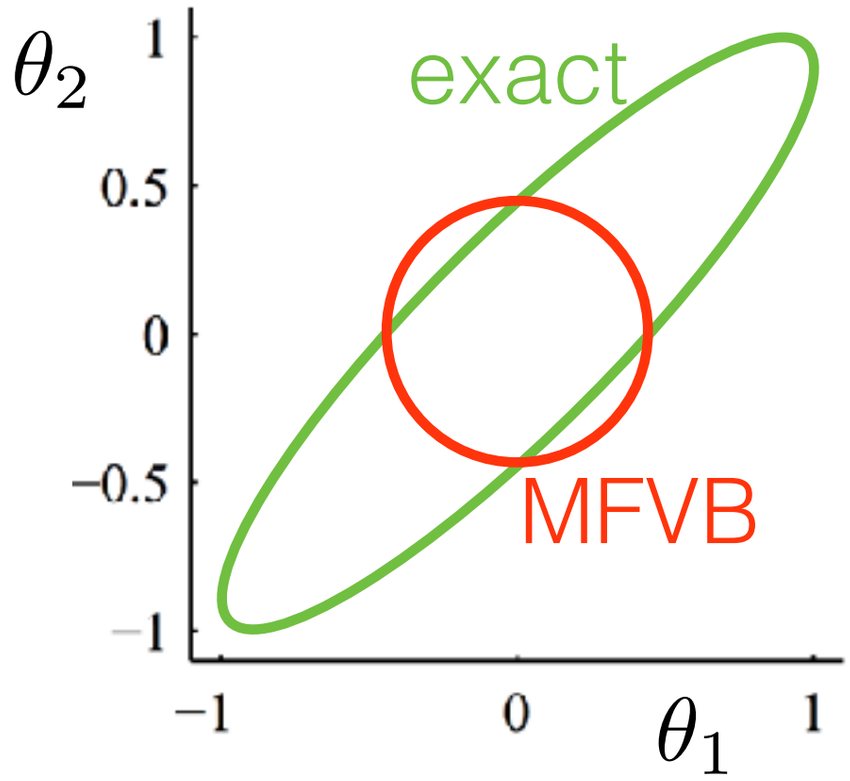
- Conjugate linear regression
- Bayesian central limit theorem

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

What about uncertainty?

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

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



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

- Conjugate linear regression
- Bayesian central limit theorem

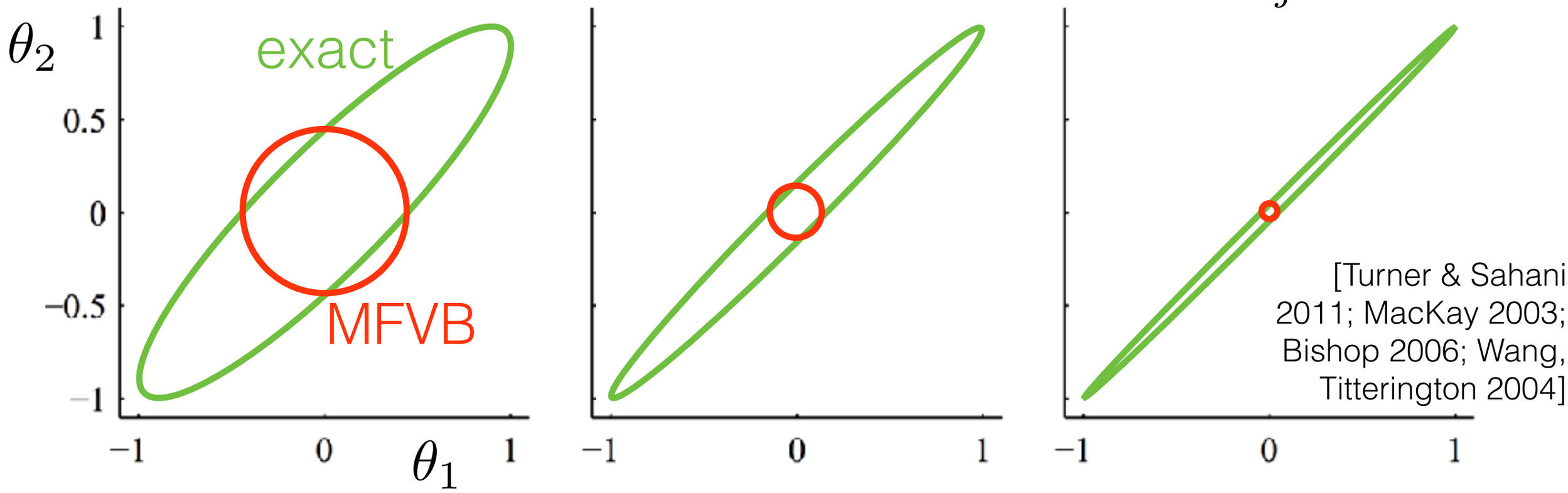
[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

- Underestimates variance (sometimes severely)

What about uncertainty?

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

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



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

- Conjugate linear regression
- Bayesian central limit theorem

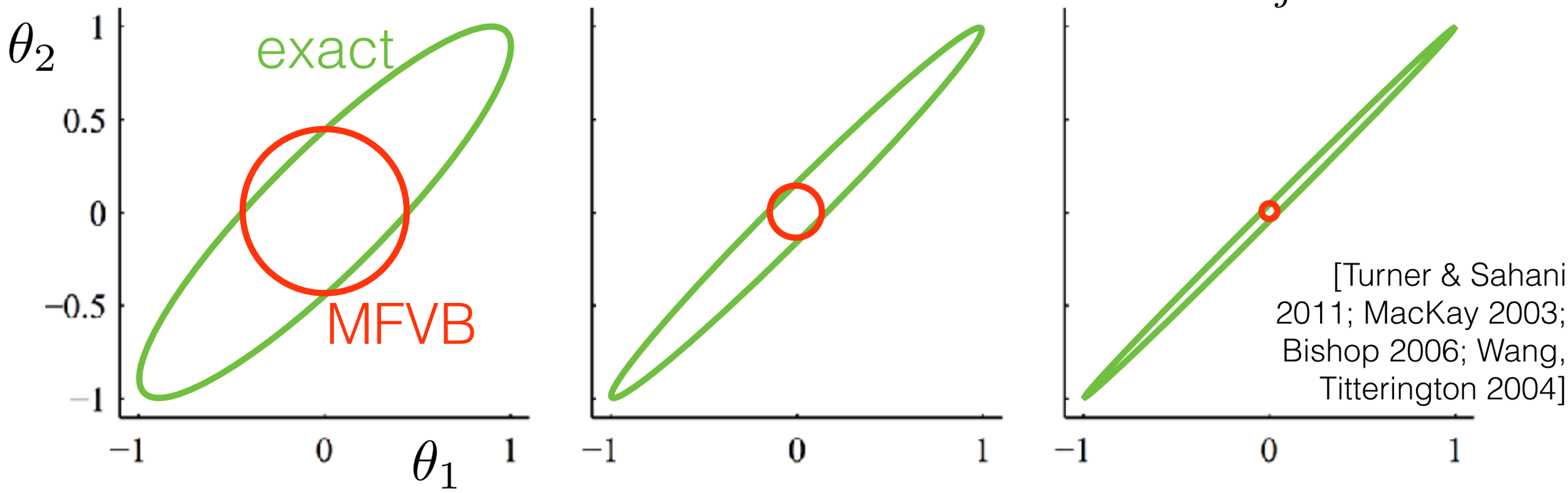
[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

- Underestimates variance (sometimes severely)

What about uncertainty?

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

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



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

- Conjugate linear regression
- Bayesian central limit theorem

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

- Underestimates variance (sometimes severely)
- No covariance estimates

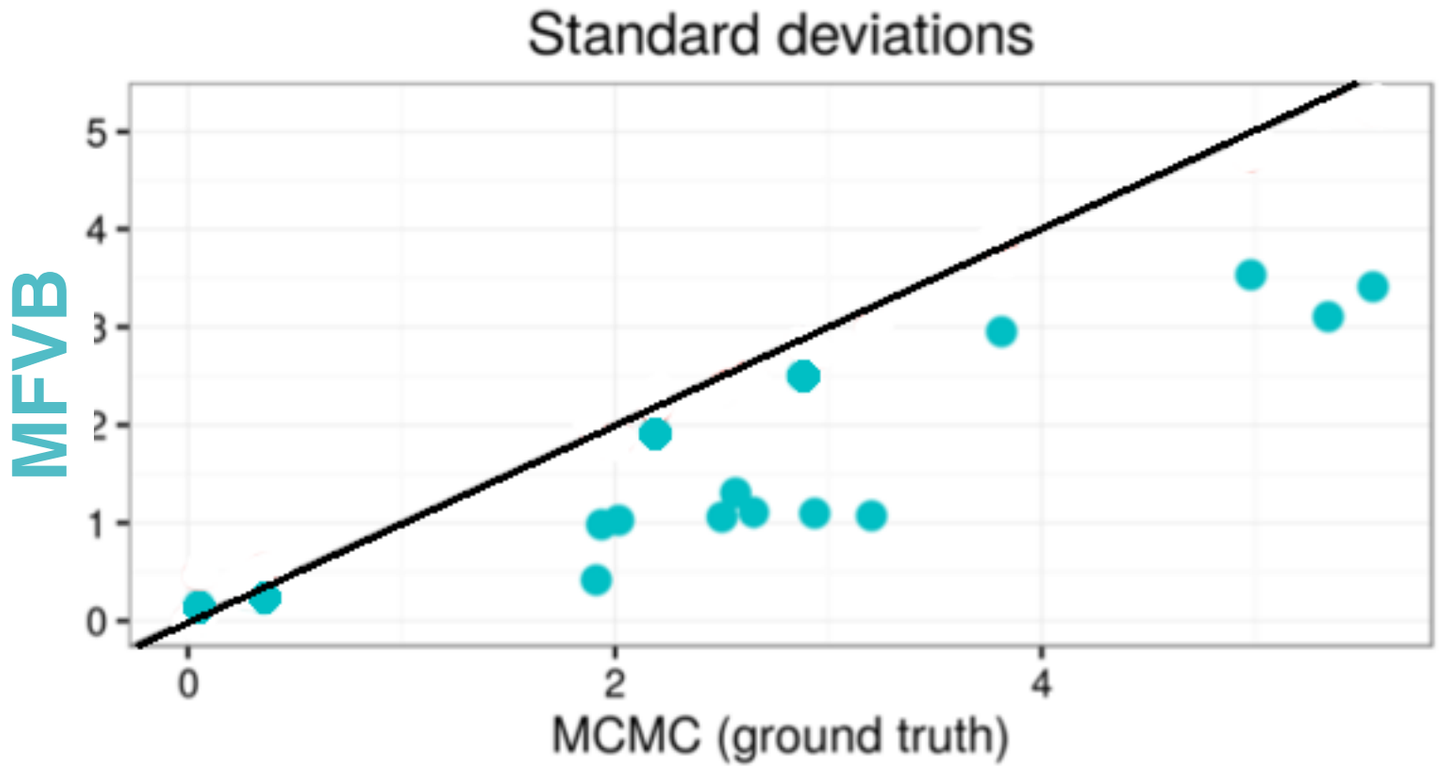
What about uncertainty?

- Microcredit

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

What about uncertainty?

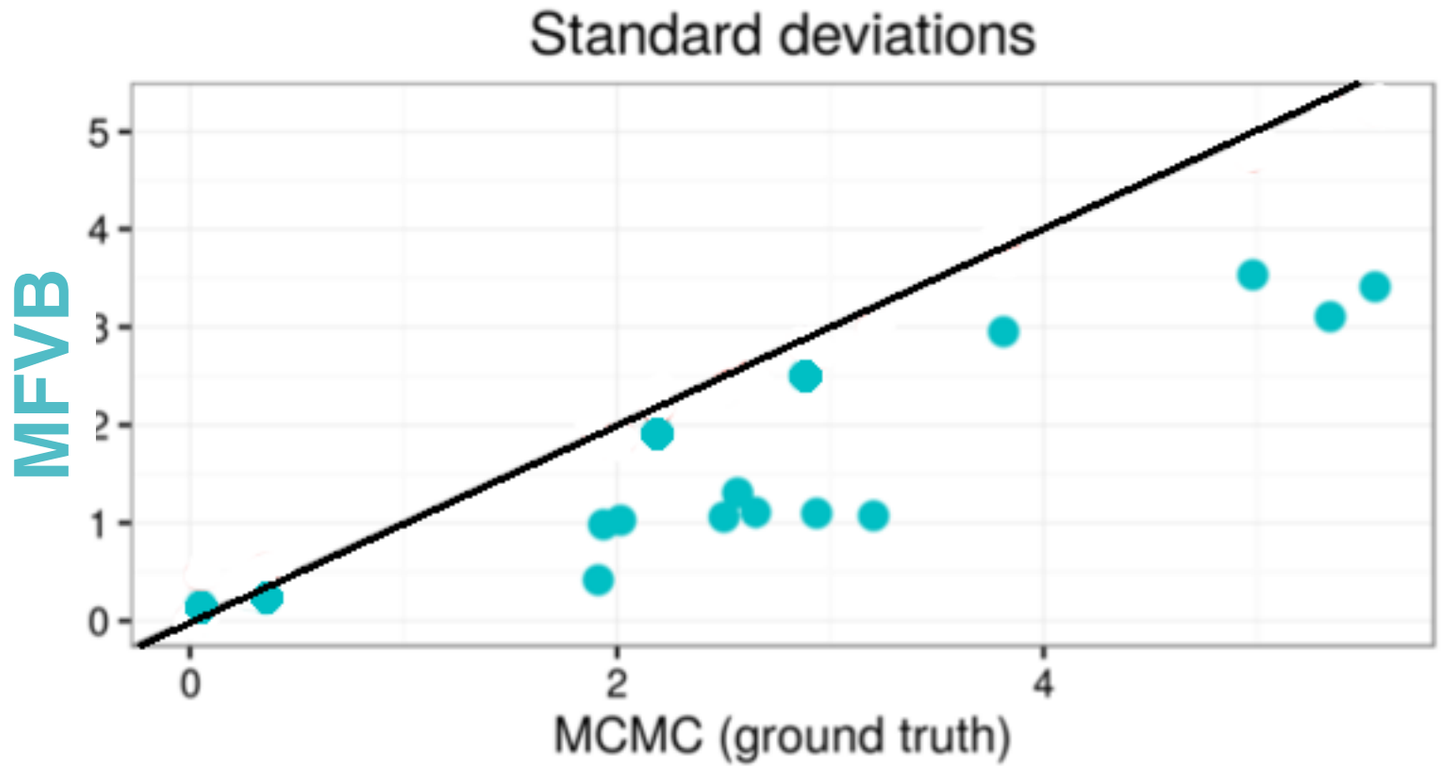
- Microcredit



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

What about uncertainty?

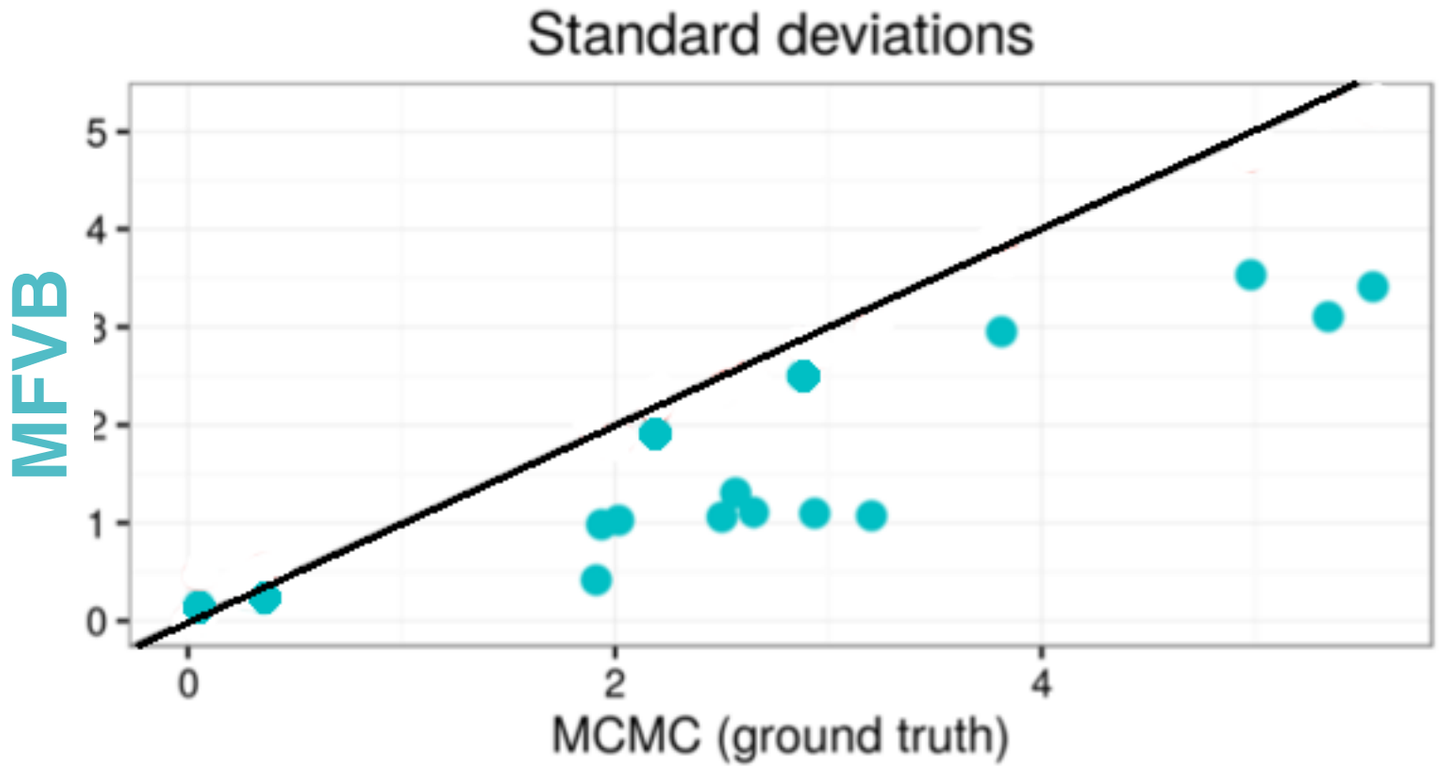
- Microcredit effect
- τ mean:
  3.08 USD PPP



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

What about uncertainty?

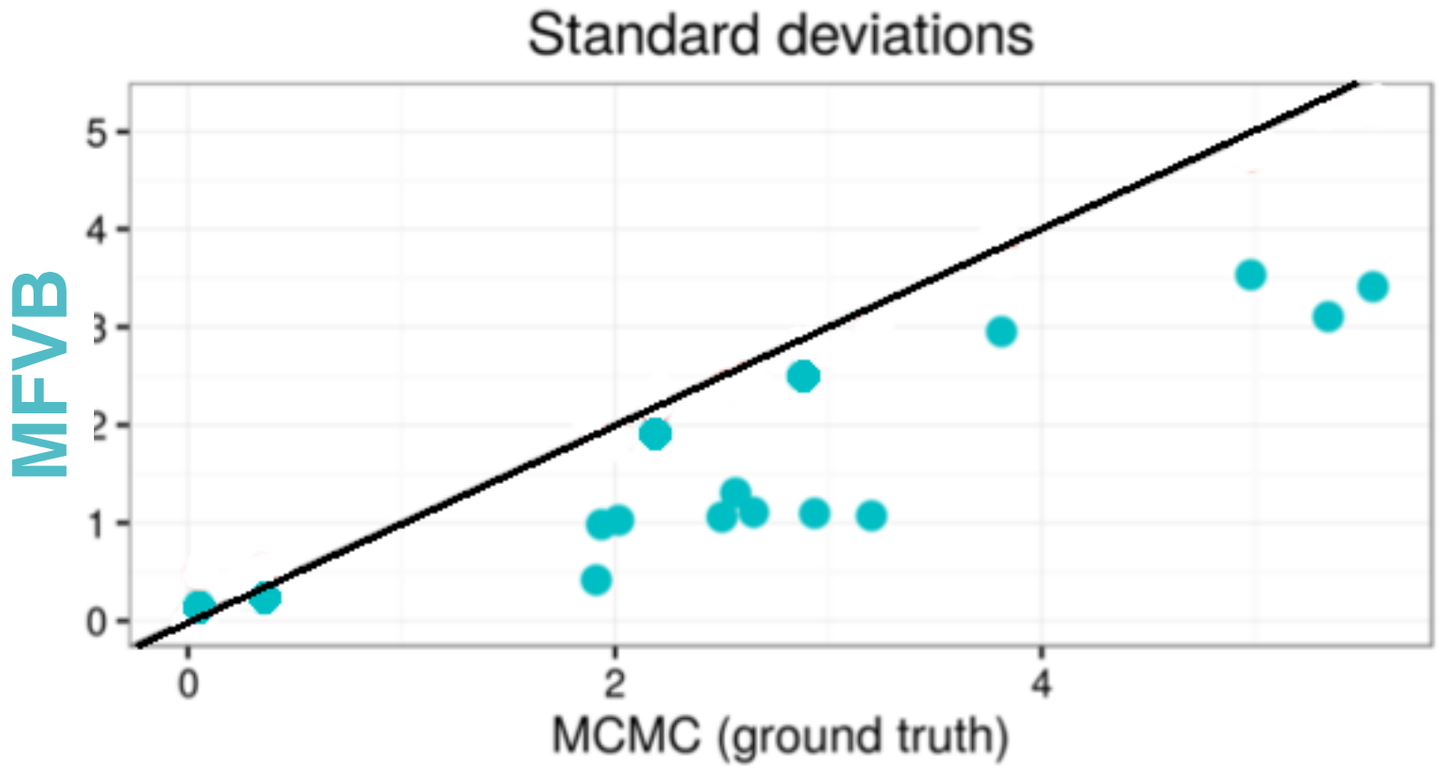
- Microcredit effect
- τ mean:
  3.08 USD PPP
- τ std dev:
  1.83 USD PPP



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

What about uncertainty?

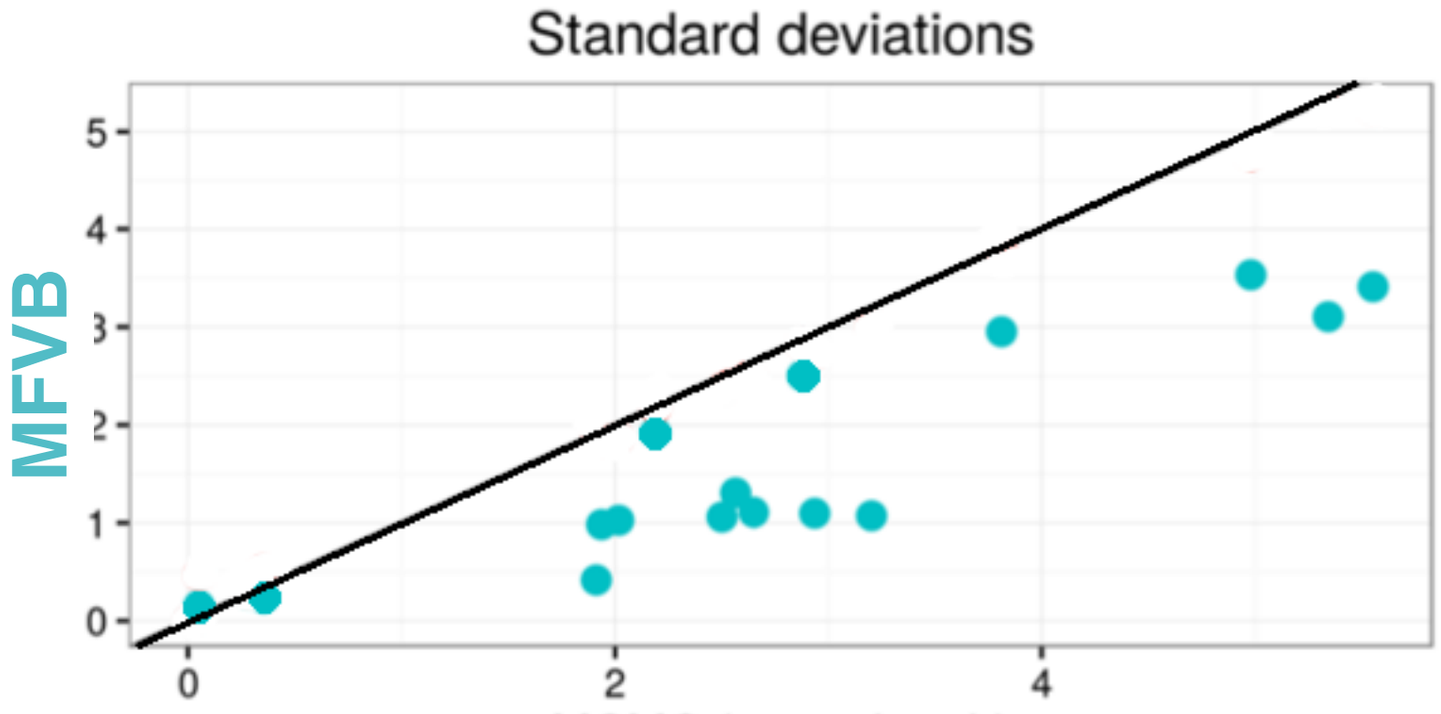
- Microcredit effect
- τ mean:
  3.08 USD PPP
- τ std dev:
  1.83 USD PPP
- Mean is 1.68 std dev from 0



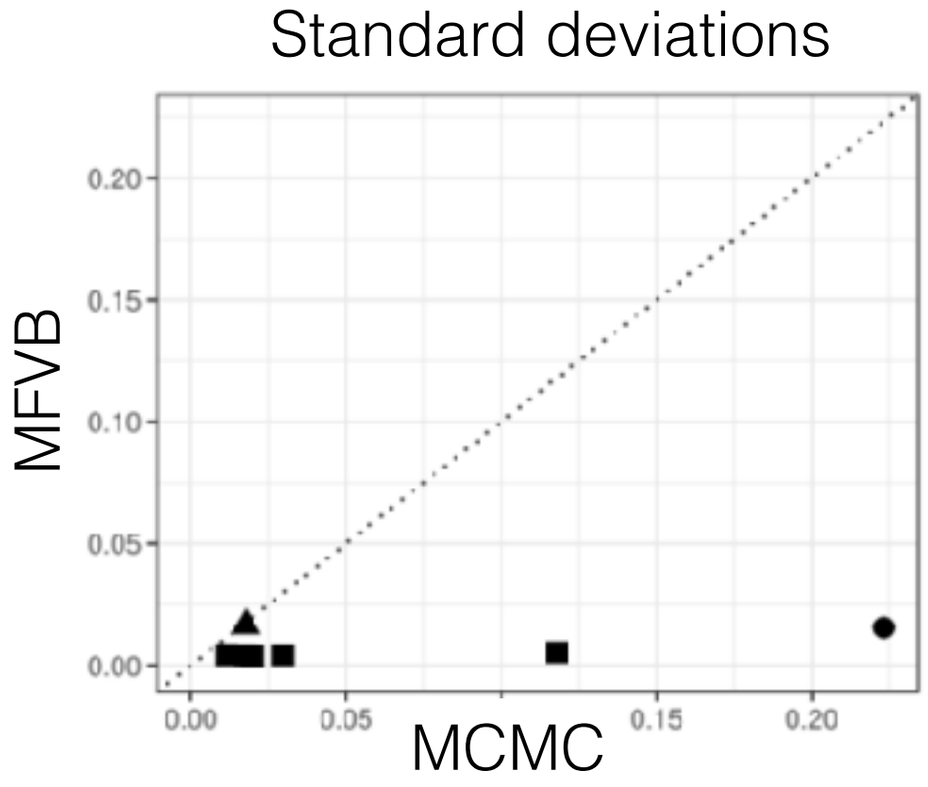
[Giordano, Broderick, Meager, Huggins, Jordan 2016]

What about uncertainty?

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



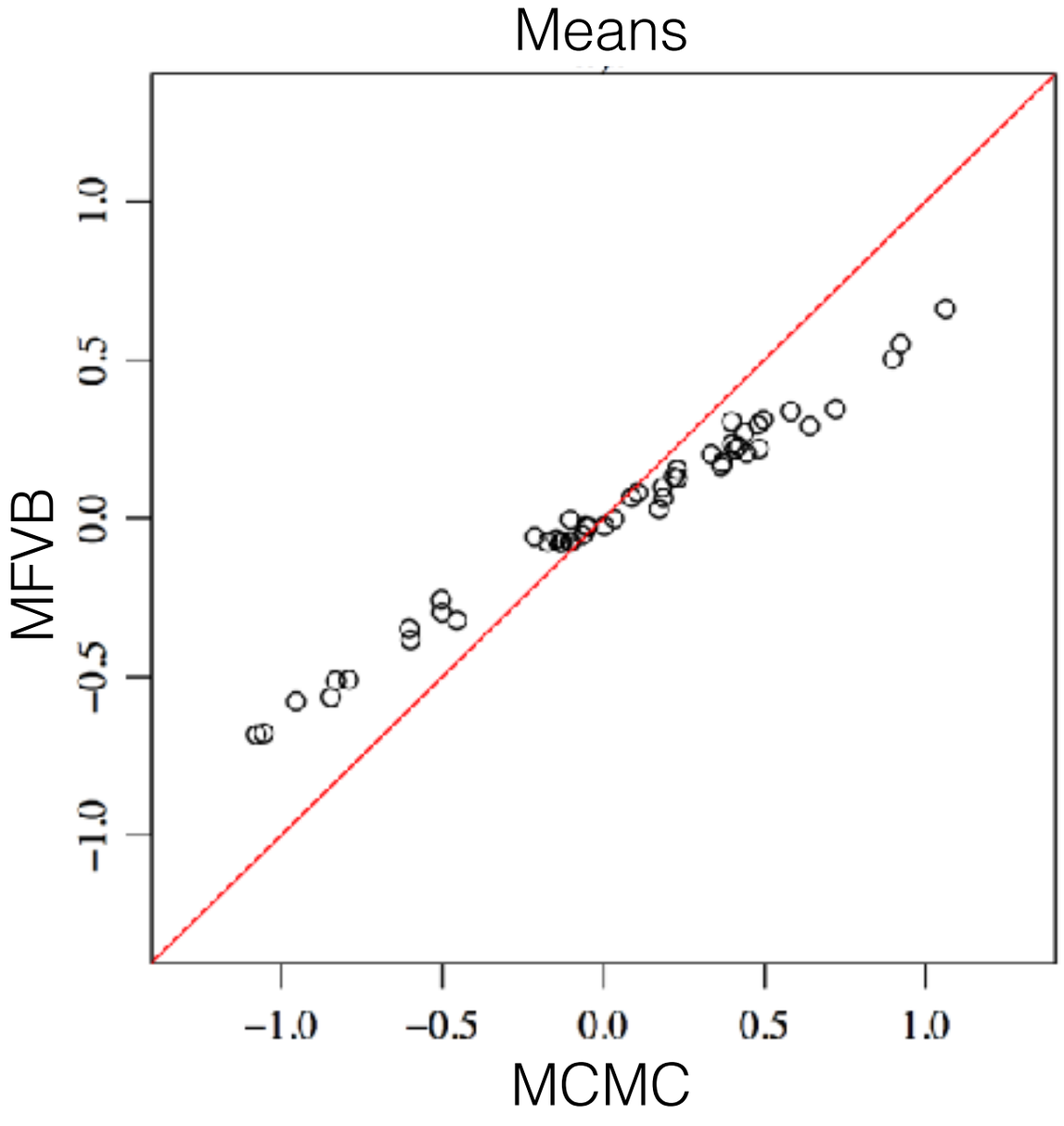
- Criteo online ads experiment



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

What about means?

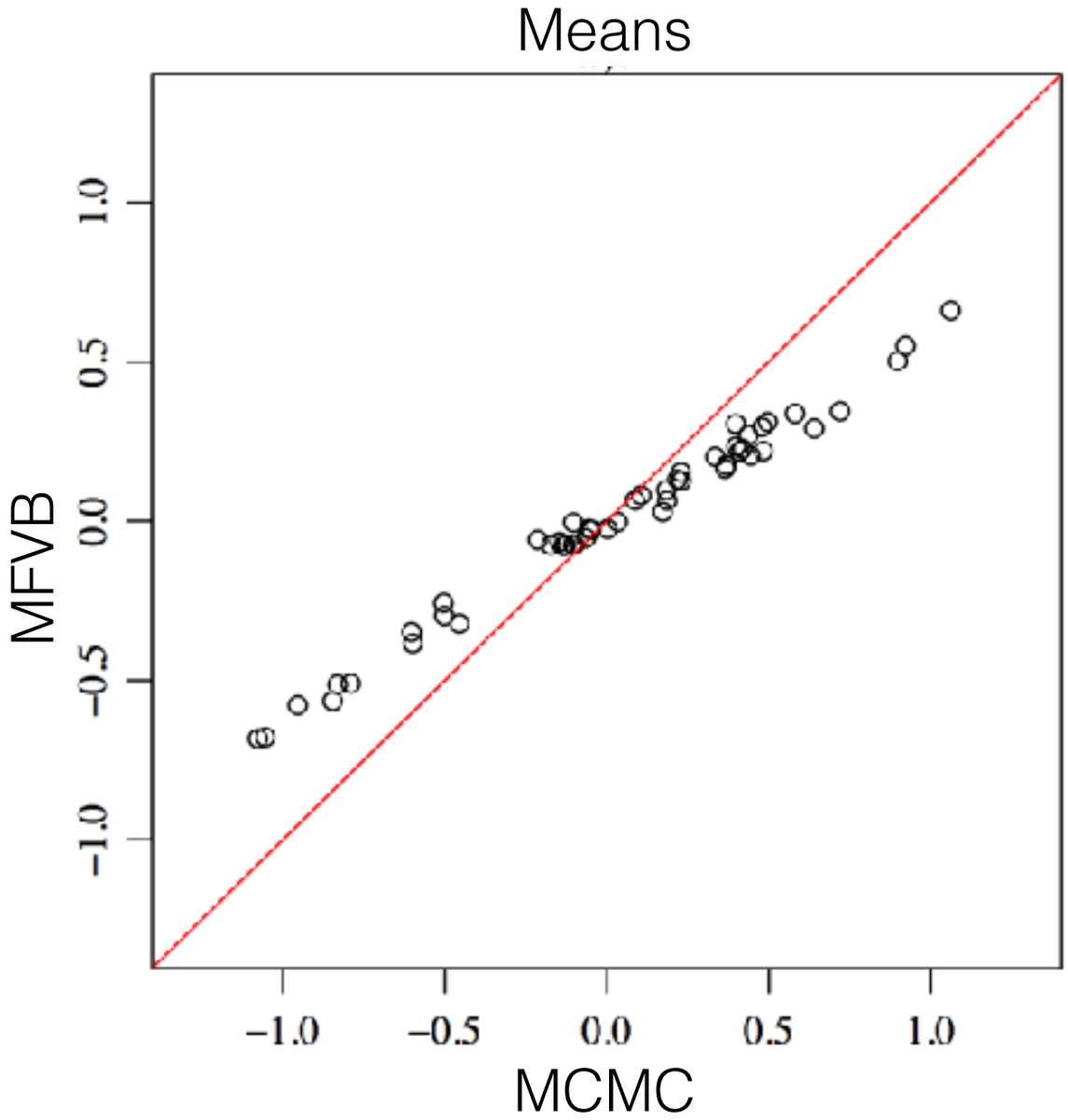
- Model for relational data with covariates
- When 1000+ nodes, MCMC > 1 day [Fosdick 2013, Ch 4]

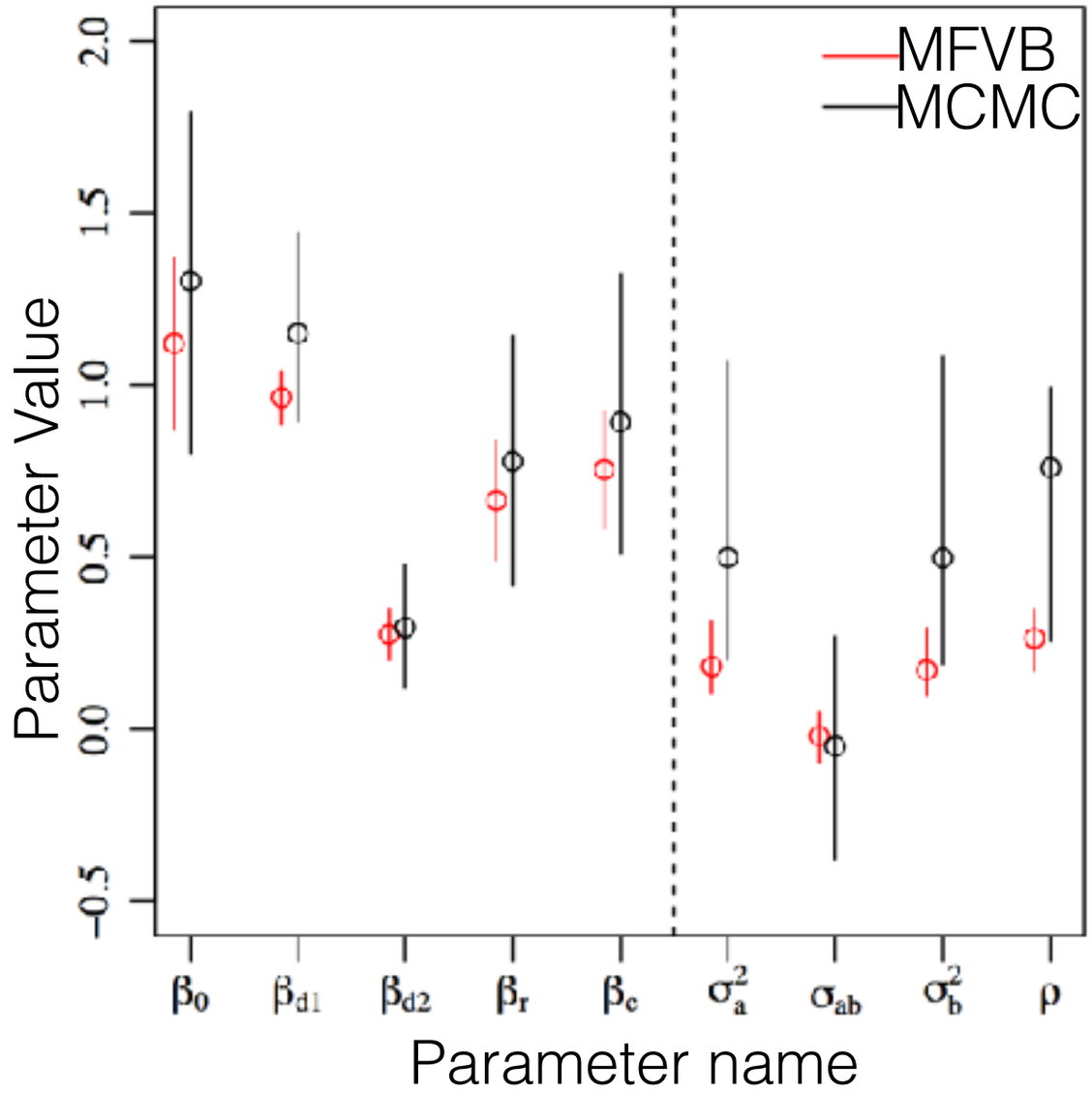


[Fosdick 2013, Ch 4, Fig 4.3]

What about means?

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





[Fosdick 2013, Ch 4, Fig 4.3]

Posterior means: revisited

- Want to predict college GPA y_n

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

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

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

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

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

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

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

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

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

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

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

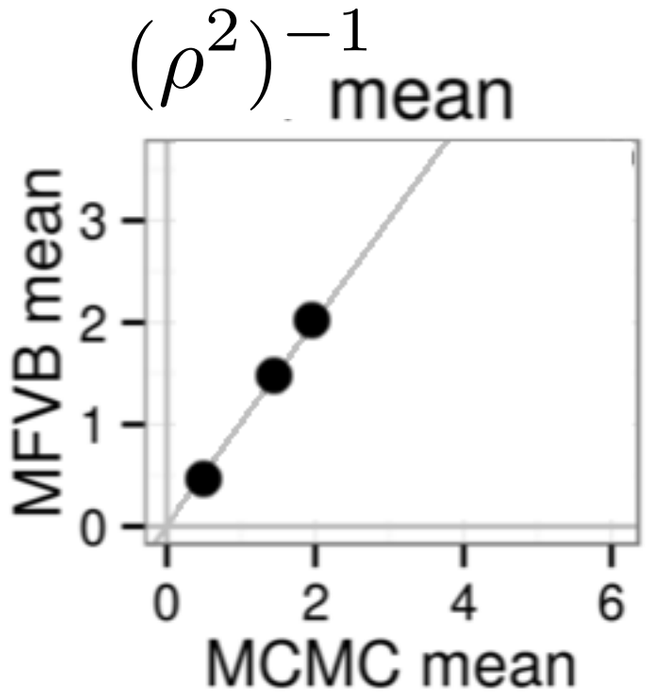
- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model:

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

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

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

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



[Giordano, Broderick, Jordan 2015]

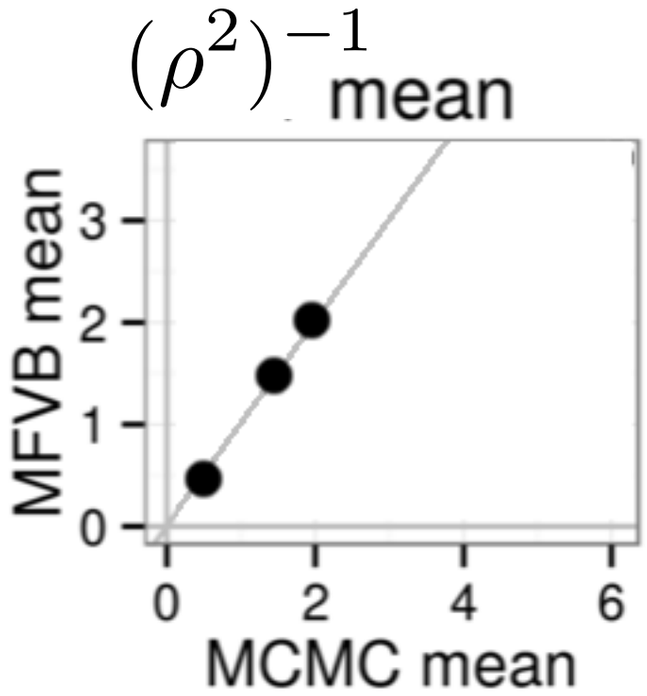
Posterior means: revisited

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

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

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

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



[Giordano, Broderick, Jordan 2015]

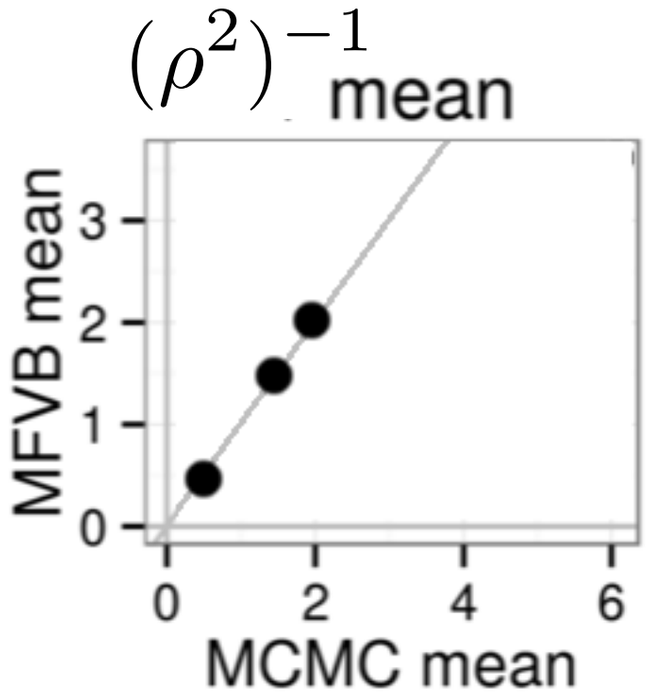
Posterior means: revisited

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

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

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

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



[Giordano, Broderick, Jordan 2015]

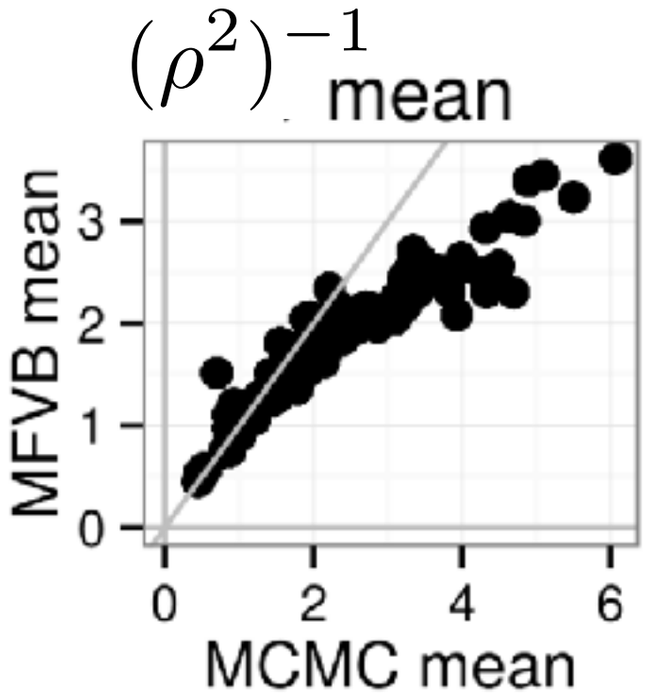
Posterior means: revisited

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

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

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

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



[Giordano, Broderick, Jordan 2015]

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

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

How deep is the issue?

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

Algorithm

How deep is the issue?

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

Algorithm

Implementation

How deep is the issue?

What to read next

Textbooks and Reviews

- Bishop. *Pattern Recognition and Machine Learning*, Ch 10. 2006.
- Blei, Kucukelbir, McAuliffe. Variational inference: A review for statisticians, *JASA* 2016.
- MacKay. *Information Theory, Inference, and Learning Algorithms*, Ch 33. 2003.
- Murphy. *Machine Learning: A Probabilistic Perspective*, Ch 21. 2012.
- Ormerod, Wand. Explaining variational approximations. *Amer Stat* 2010.
- Turner, Sahani. Two problems with variational expectation maximisation for time-series models. In *Bayesian Time Series Models*, 2011.
- Wainwright, Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 2008.

Our Experiments

- RJ Giordano, T Broderick, and MI Jordan. Linear response methods for accurate covariance estimates from mean field variational Bayes. *NeurIPS* 2015.
- RJ Giordano, T Broderick, R Meager, JH Huggins, and MI Jordan. Fast robustness quantification with variational Bayes. *ICML Data4Good Workshop* 2016.
- RJ Giordano, T Broderick, and MI Jordan. Covariances, robustness, and variational Bayes. *JMLR* 2018.
- J Huggins, M Kasprzak, T Campbell, T Broderick. Validated Variational Inference via Practical Posterior Error Bounds. ArXiv: 1910.04102. *AISTATS* 2020, to appear.
- T Campbell and T Broderick. Automated scalable Bayesian inference via Hilbert coresets. *JMLR* 2019.
- T Campbell and T Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. *ICML* 2018.

References

Full references at end of final slides