

# Variational Bayes and beyond:

# Foundations of scalable Bayesian inference

Tamara Broderick

Associate Professor  
MIT

[http://tamarabroderick.com/tutorial\\_2021\\_ssc.html](http://tamarabroderick.com/tutorial_2021_ssc.html)

Rough schedule:

- Part I: 11 am Eastern Time
- Break: 12 noon
- Part II: 12:30 pm
- Break: 1:30 pm
- Part III after the Break
- Finish: by 3:00 pm ET

# Approximate Bayesian inference

# Approximate Bayesian inference

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

# Approximate Bayesian inference

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

Optimization

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

# Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

# Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

# Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

# Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

- Coordinate descent



# Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

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

# Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

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

# Approximate Bayesian inference

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

Optimization

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

**Variational Bayes**

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

Mean-field variational Bayes

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

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

# Approximate Bayesian inference

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

## Optimization

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

## Variational Bayes

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

## Mean-field variational Bayes

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

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

# Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

**Mean-field variational Bayes**

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

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

# 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?

# Air pollution: Particulate matter



[Krongut 2020]



# Air pollution: Particulate matter

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



[Krongut 2020]

# Air pollution: Particulate matter

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

- Model:

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



[Krongut 2020]

# Air pollution: Particulate matter

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

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



[Krongut 2020]

# Air pollution: Particulate matter

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

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

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

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



[Krongut 2020]

# Air pollution: Particulate matter

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

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

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

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



[Krongut 2020]

# Air pollution: Particulate matter

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

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

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

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



[Krongut 2020]

# Air pollution: Particulate matter

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

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

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

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



[Krongut 2020]

# Air pollution: Particulate matter

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

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

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

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



[Krongut 2020]



# Air pollution: Particulate matter

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

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

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

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



[Krongut 2020]

# Air pollution: Particulate matter

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

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

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

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



[Krongut 2020]

# Air pollution: Particulate matter

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

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

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

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



[Krongut 2020]

# Air pollution: Particulate matter

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

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

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

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

- Exercise: check

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



[Krongut 2020]

# Air pollution: Particulate matter

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

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

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

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

- Exercise: check

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

- MFVB approximation:

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



[Krongut 2020]

# Air pollution: Particulate matter

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

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

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

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

- Exercise: check

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

- MFVB approximation:

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

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



[Krongut 2020]



# Air pollution: Particulate matter

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

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

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

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

- Exercise: check

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

- MFVB approximation:

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

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

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

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



[Krongut 2020]

# Air pollution: Particulate matter

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

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

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

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

- Exercise: check

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

- MFVB approximation:

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

- 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”

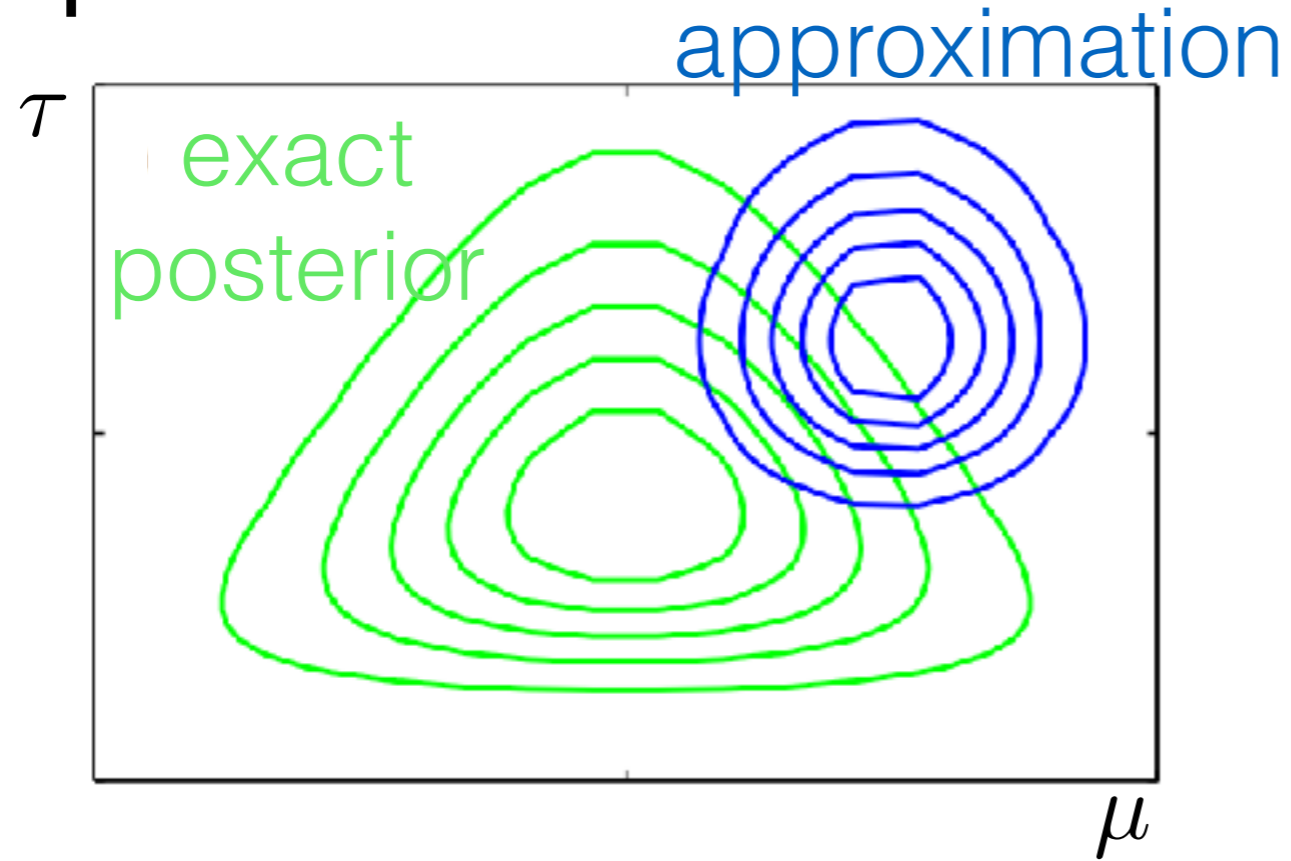
[MacKay 2003; Bishop 2006]



[Krongut 2020]

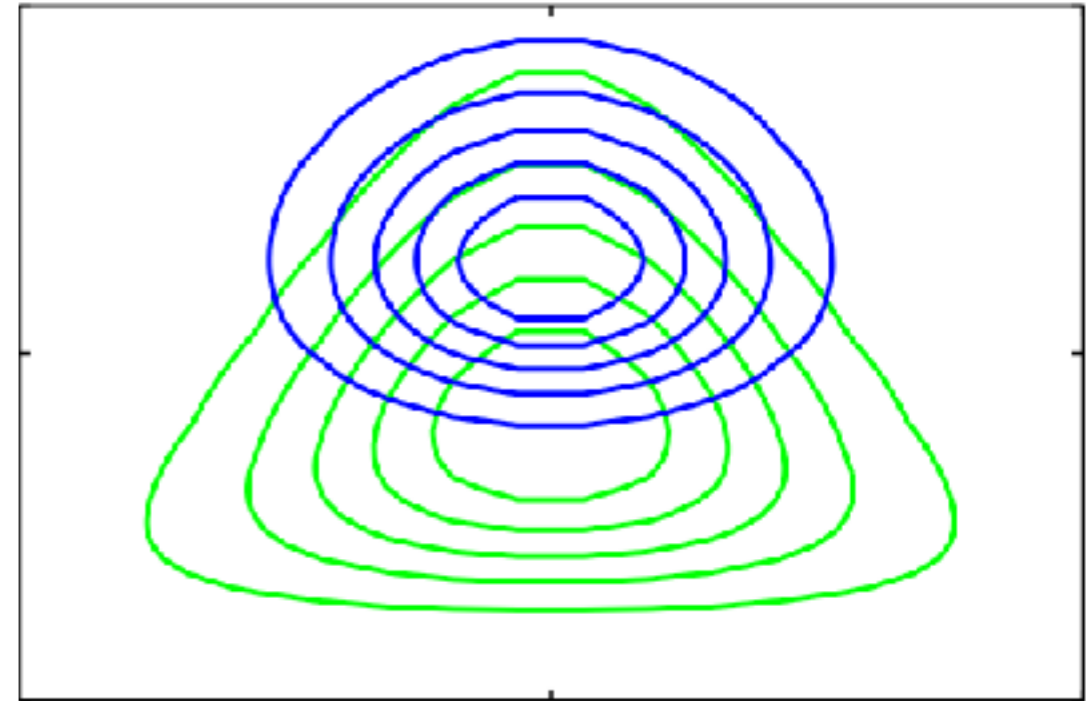
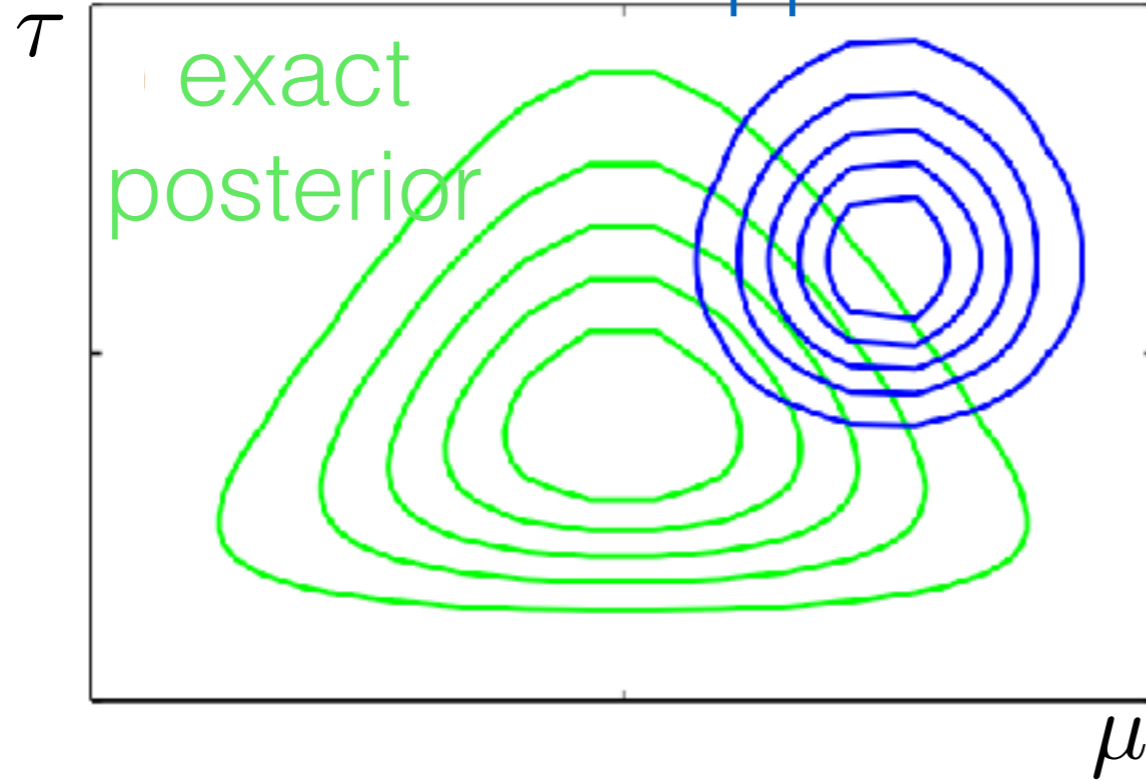


# Air pollution: Particulate matter



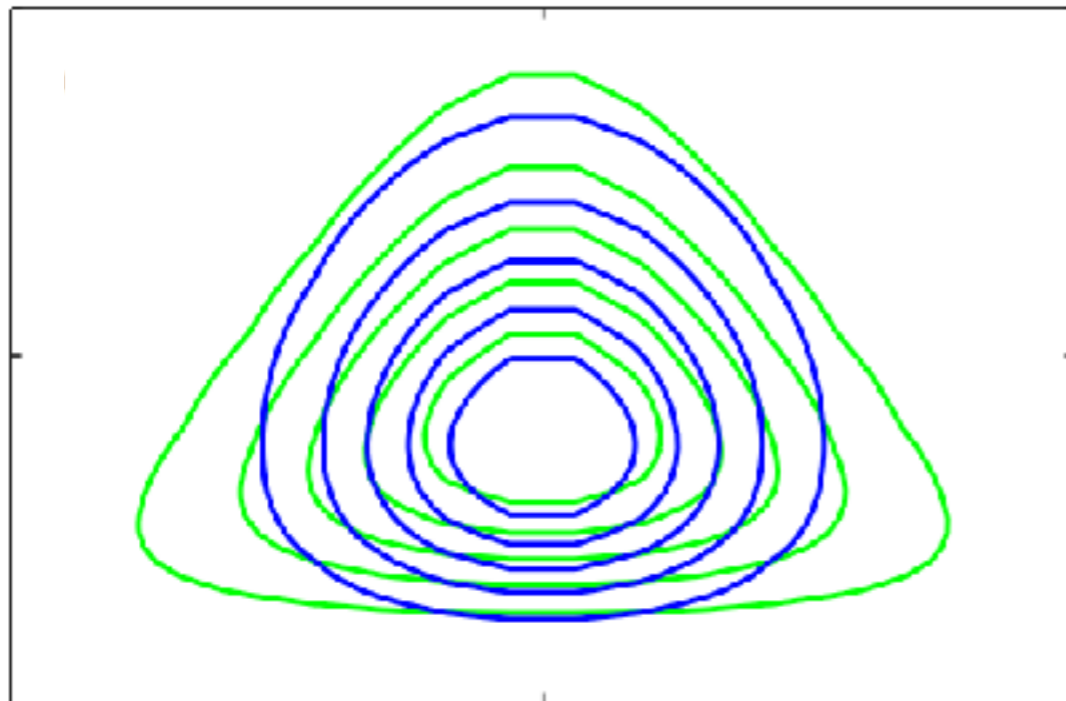
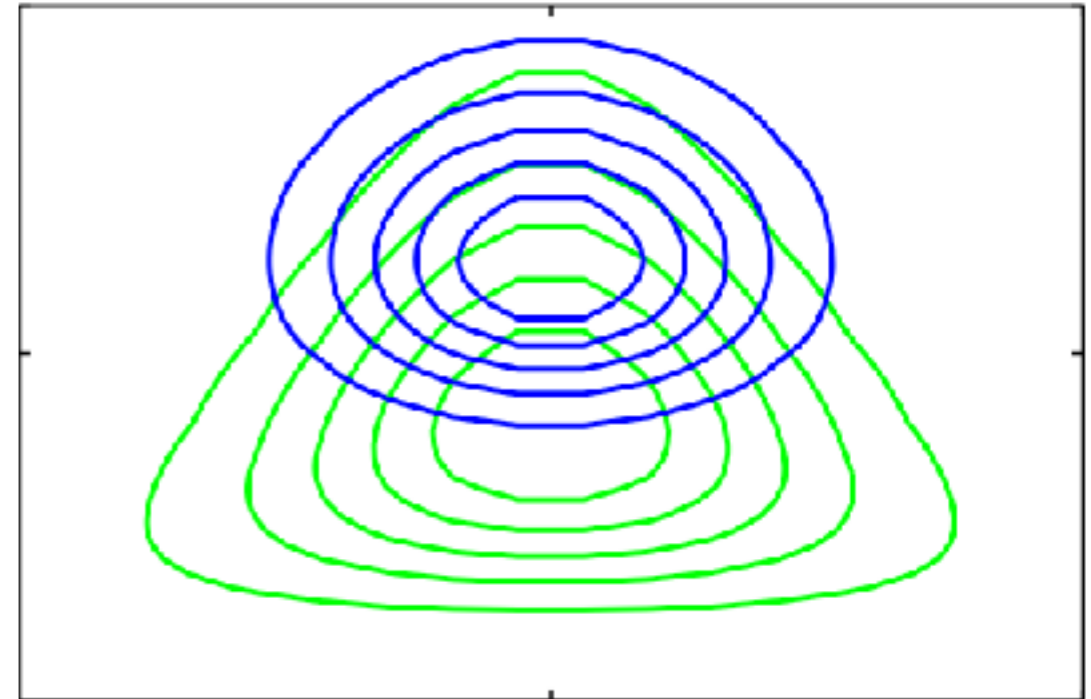
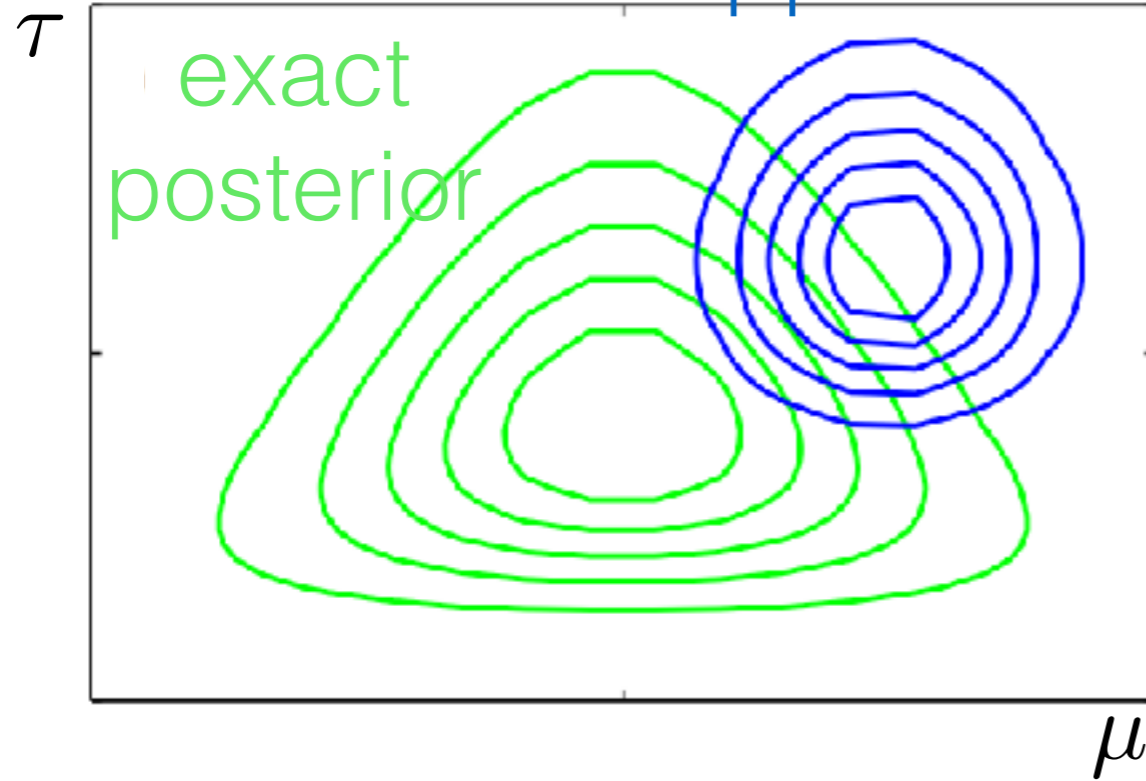
# Air pollution: Particulate matter

approximation



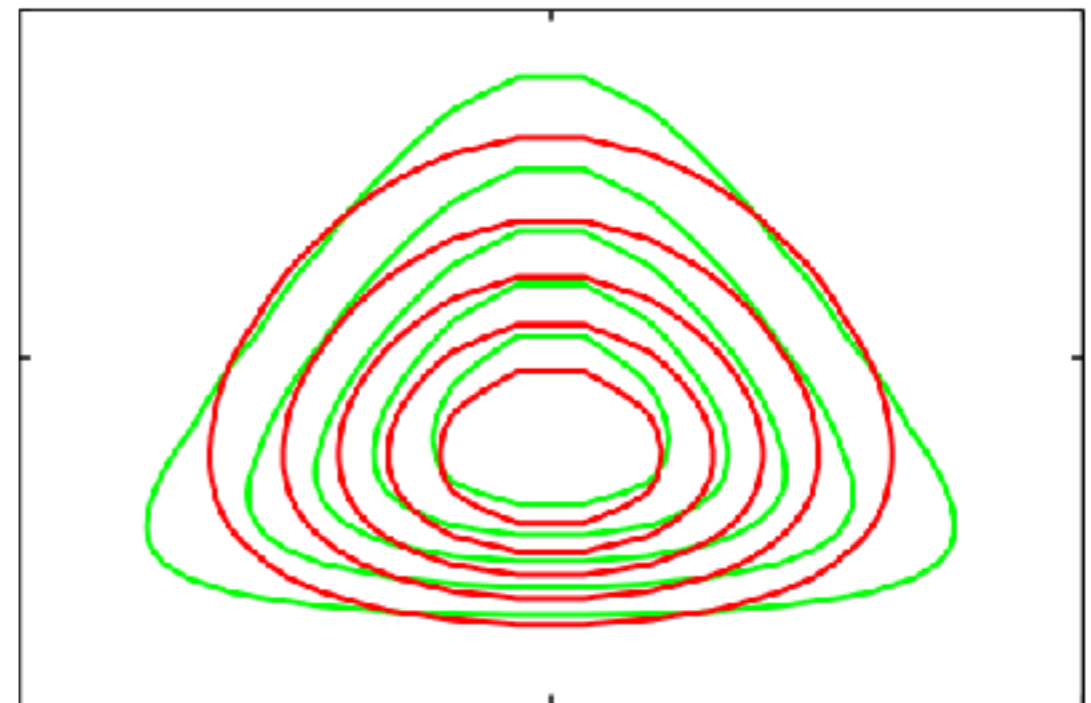
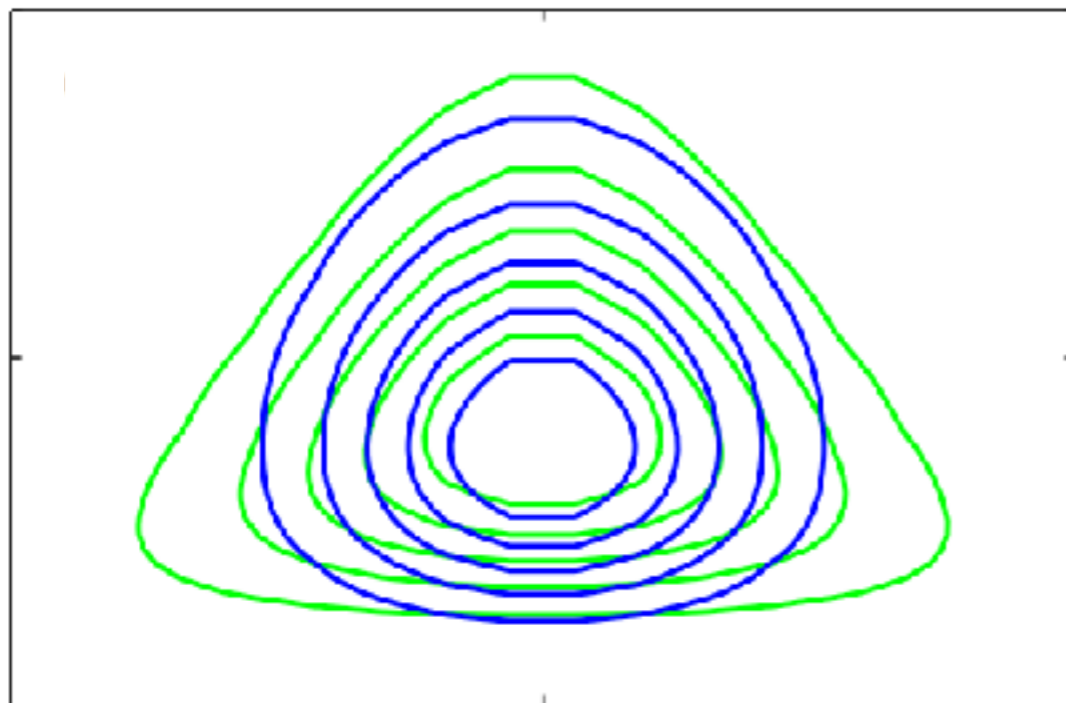
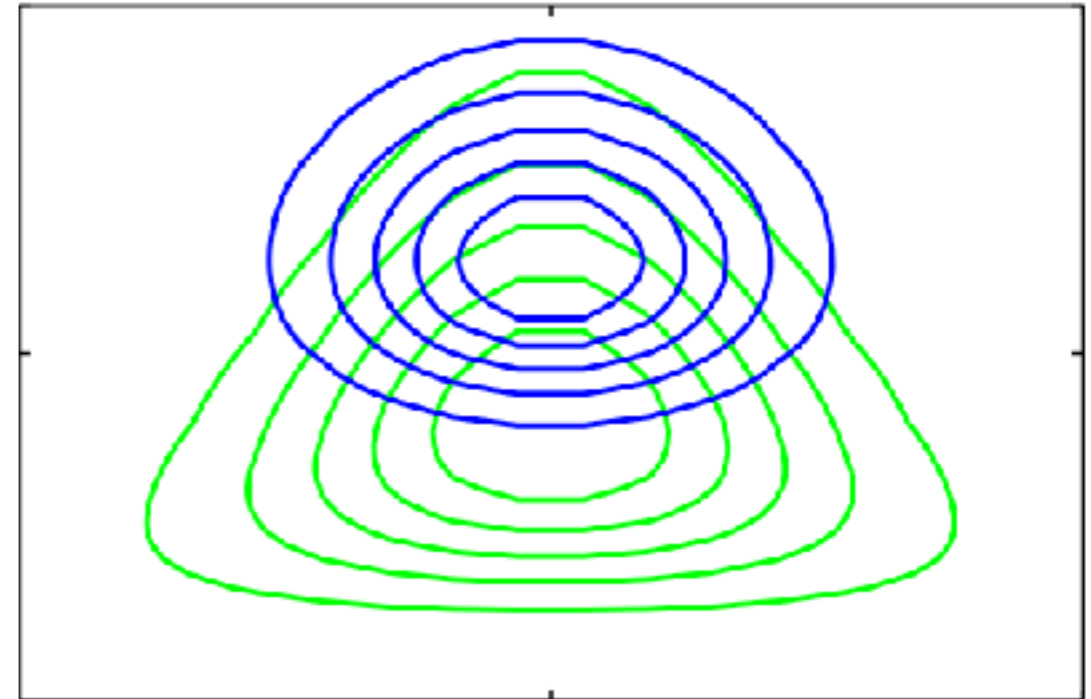
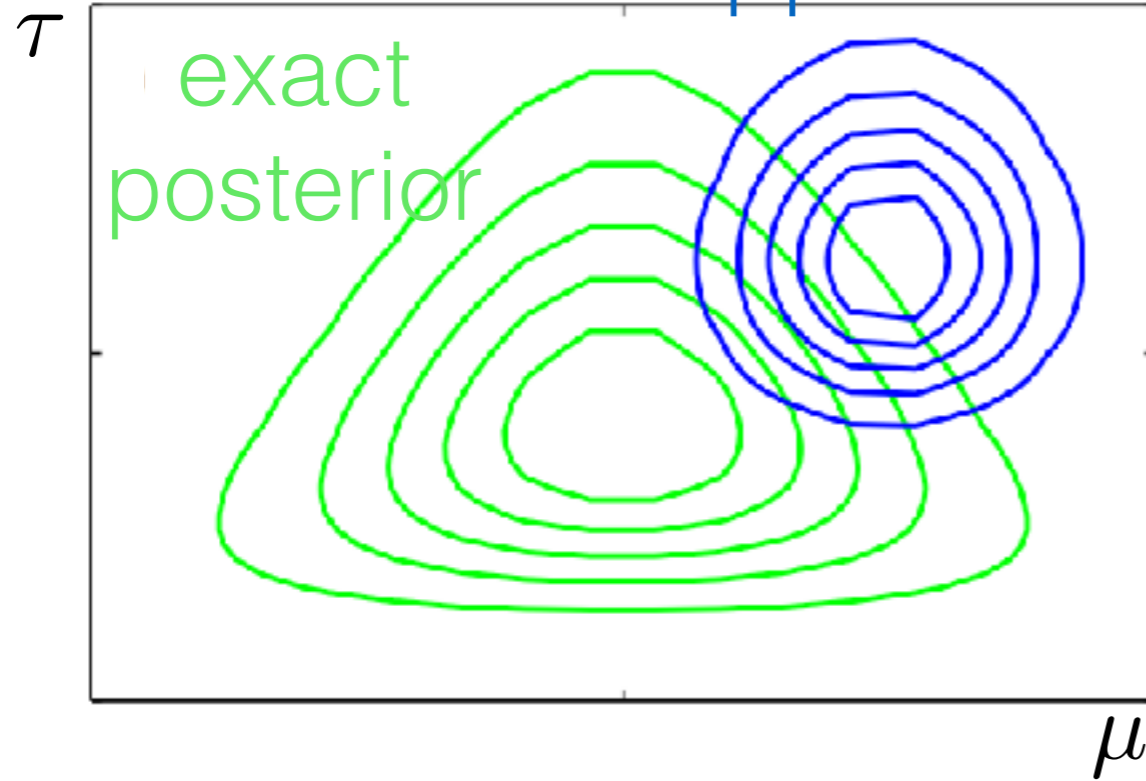
# Air pollution: Particulate matter

approximation

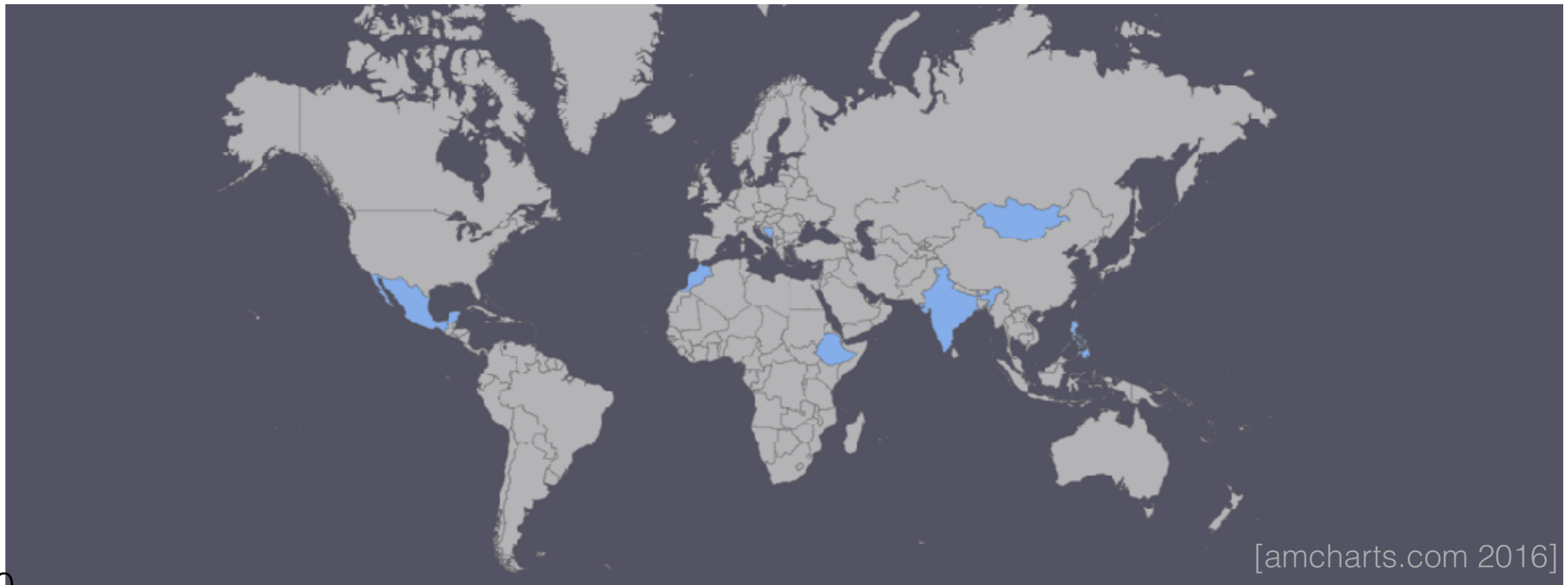


# Air pollution: Particulate matter

approximation



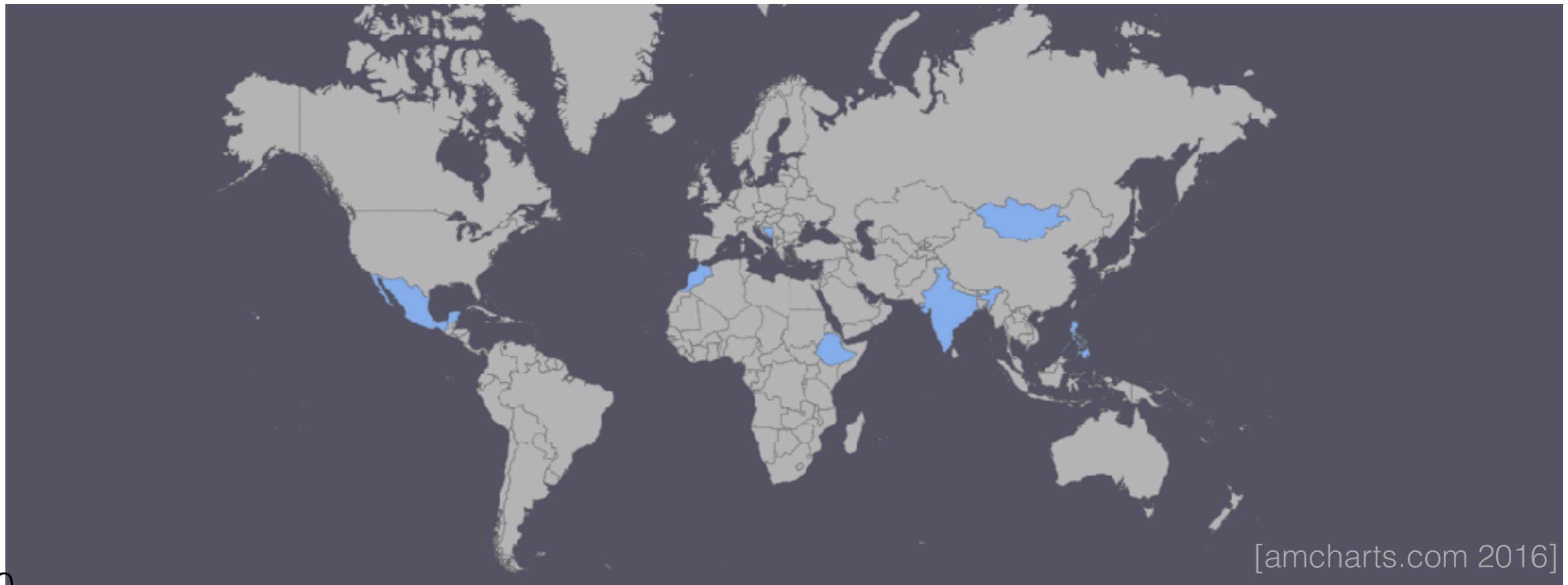
# Microcredit Experiment



[amcharts.com 2016]

# Microcredit Experiment

- Simplified from Meager (2019)

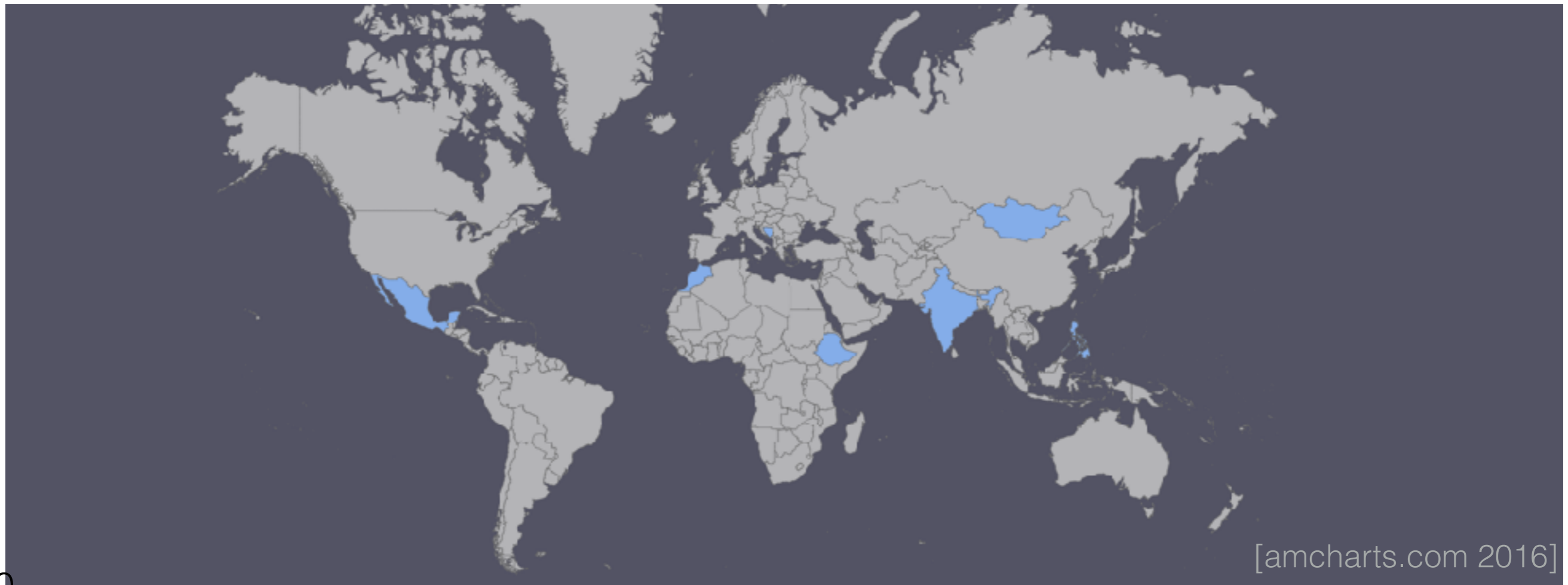


[amcharts.com 2016]



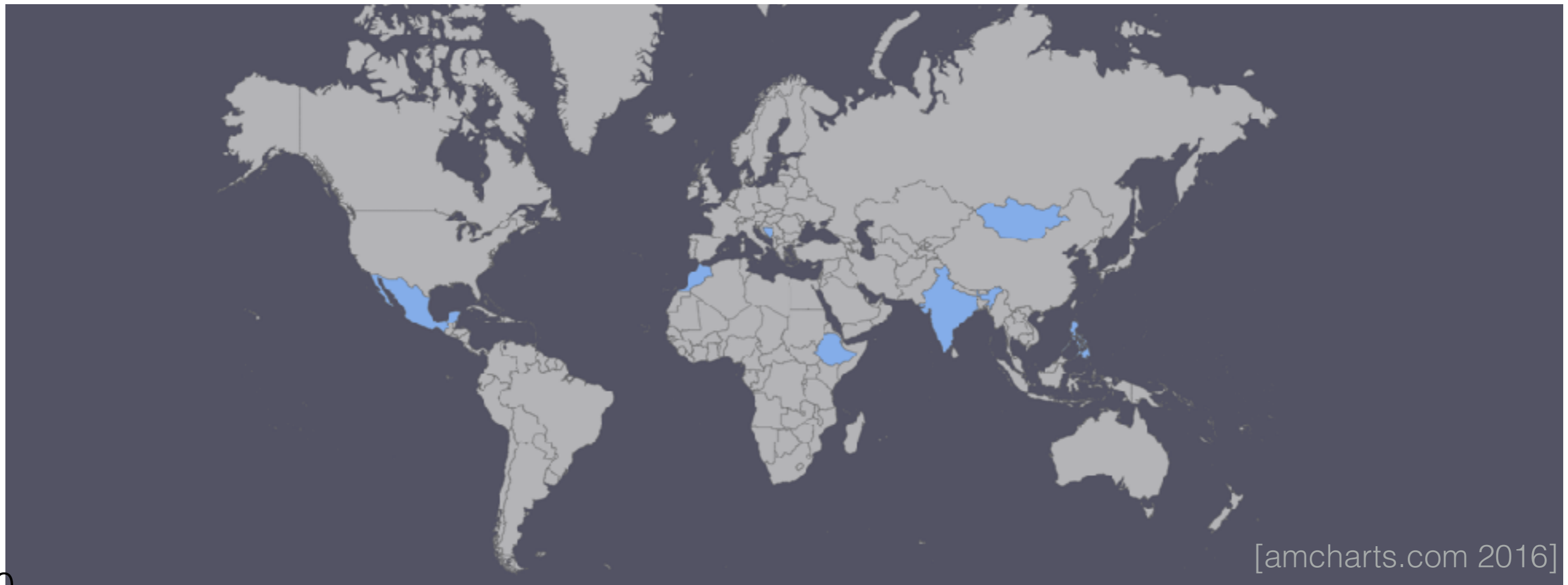
# Microcredit Experiment

- 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  $k$ th site ( $\sim 900$  to  $\sim 17K$ )





# Microcredit Experiment


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

# Microcredit Experiment

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

# Microcredit Experiment

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

profit  $\rightarrow y_{kn} \stackrel{\text{indep}}{\sim} \mathcal{N}(\quad, \quad)$

# Microcredit Experiment

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

profit  $\rightarrow y_{kn} \stackrel{\text{indep}}{\sim} \mathcal{N}(\mu_k, \sigma_k^2)$

# Microcredit Experiment

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

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

# Microcredit Experiment

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

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

1 if microcredit

# Microcredit Experiment

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

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

$\rightarrow$  1 if microcredit



# Microcredit Experiment

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

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

1 if microcredit

# Microcredit Experiment

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

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

$\rightarrow$  1 if microcredit

# Microcredit Experiment

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

profit  $\rightarrow$   $y_{kn} \stackrel{\text{indep}}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$   $\leftarrow$  1 if microcredit

- Priors and hyperpriors:

# Microcredit Experiment

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

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

$\rightarrow$  1 if microcredit

- 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 (2019)
- $K = 7$  microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$  businesses in  $k$ th site ( $\sim 900$  to  $\sim 17K$ )
- Profit of  $n$ th business at  $k$ th site:

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

$\rightarrow$  1 if microcredit

- 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 (2019)
- $K = 7$  microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$  businesses in  $k$ th site ( $\sim 900$  to  $\sim 17K$ )
- Profit of  $n$ th business at  $k$ th site:

profit  $\rightarrow$   $y_{kn} \stackrel{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$   $\leftarrow$  1 if microcredit

- Priors and hyperpriors:

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

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

$$C \sim \text{Sep\&LKJ}(\eta, c, d)$$

# Microcredit

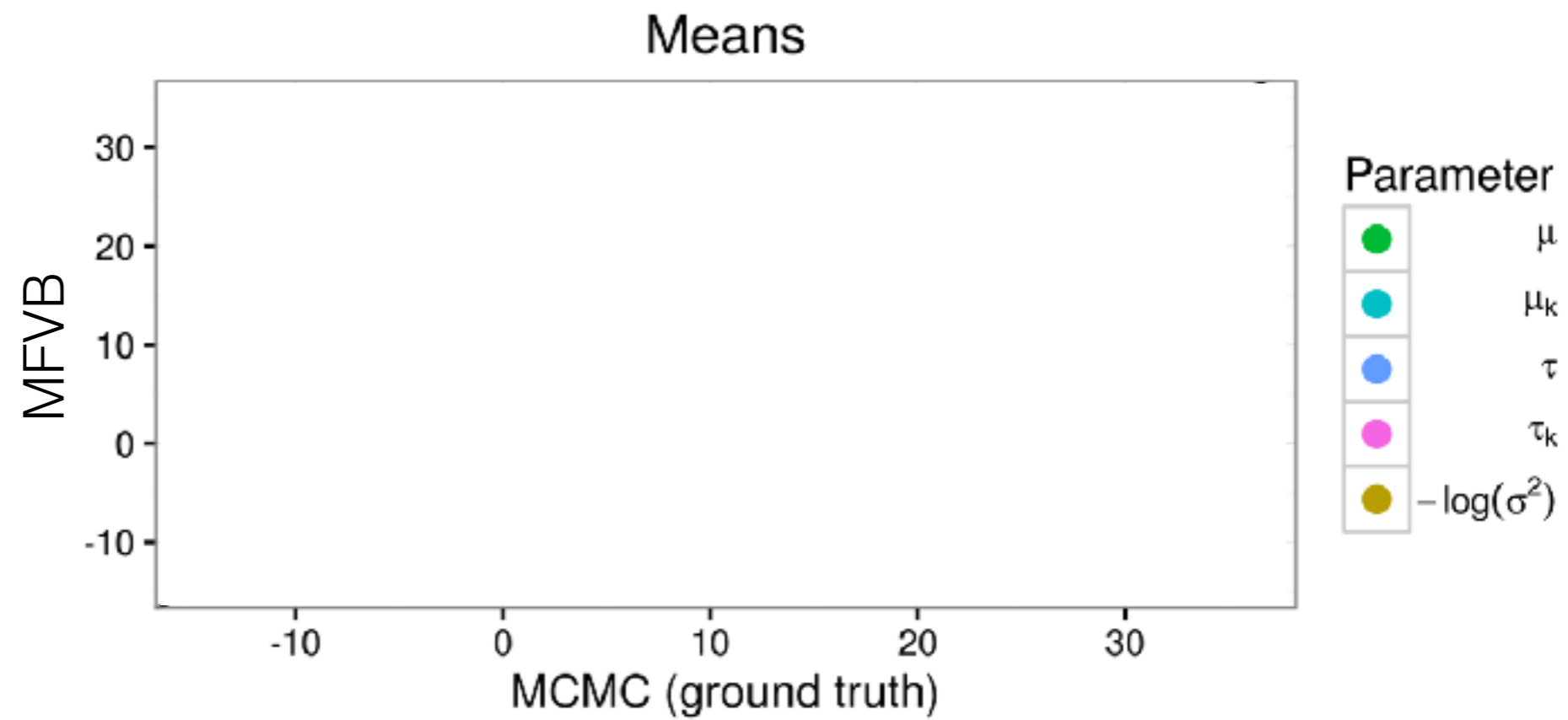
MFVB: Do we need to check the output?

# Microcredit

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

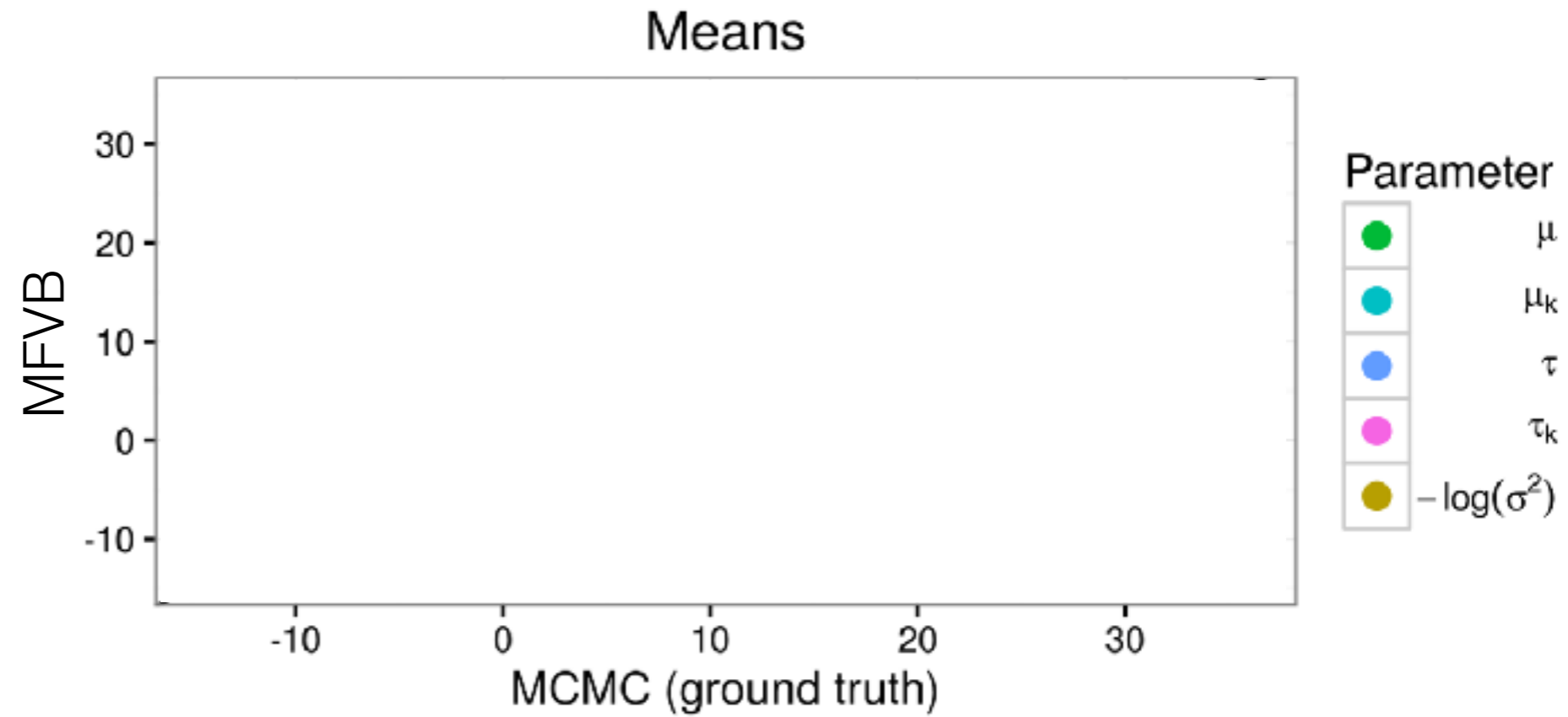


# Microcredit



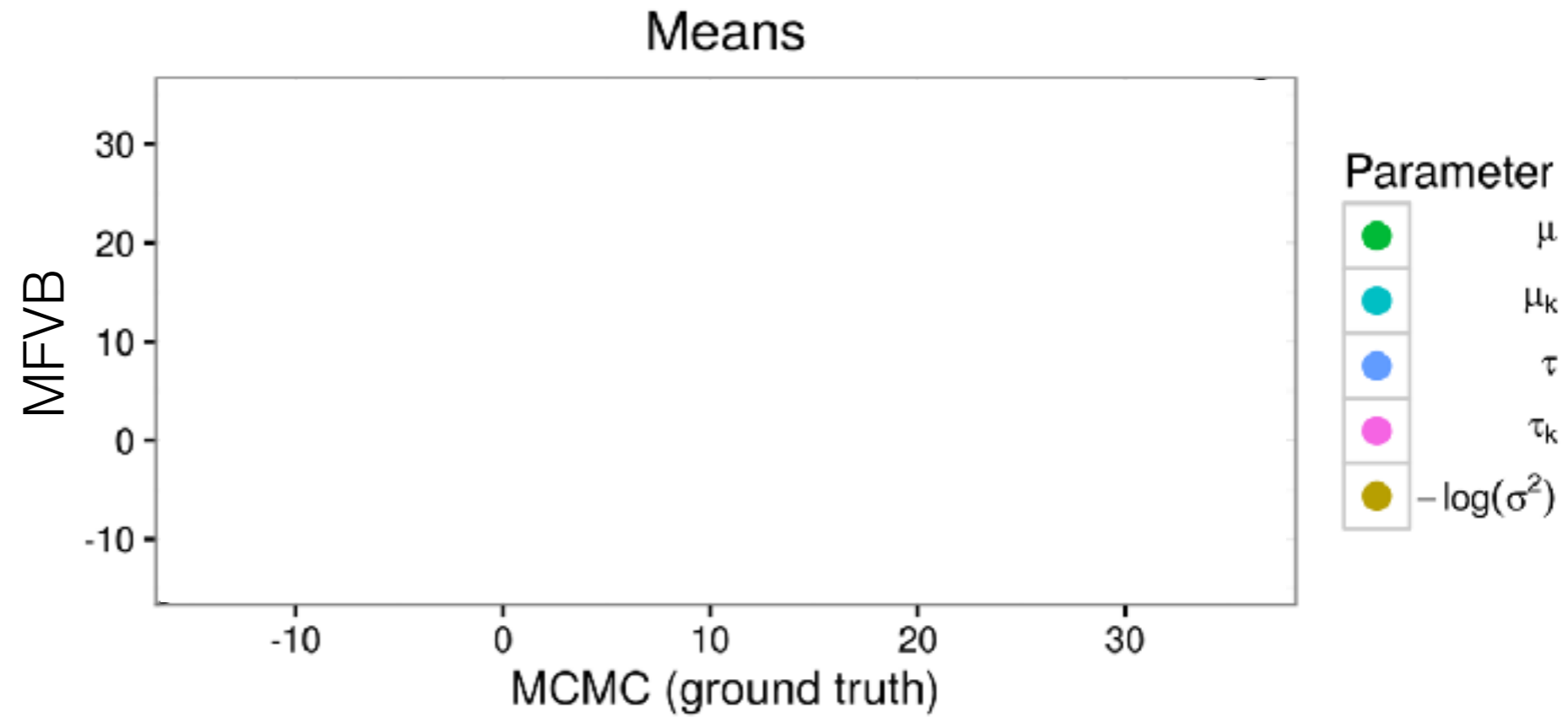
# Microcredit

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



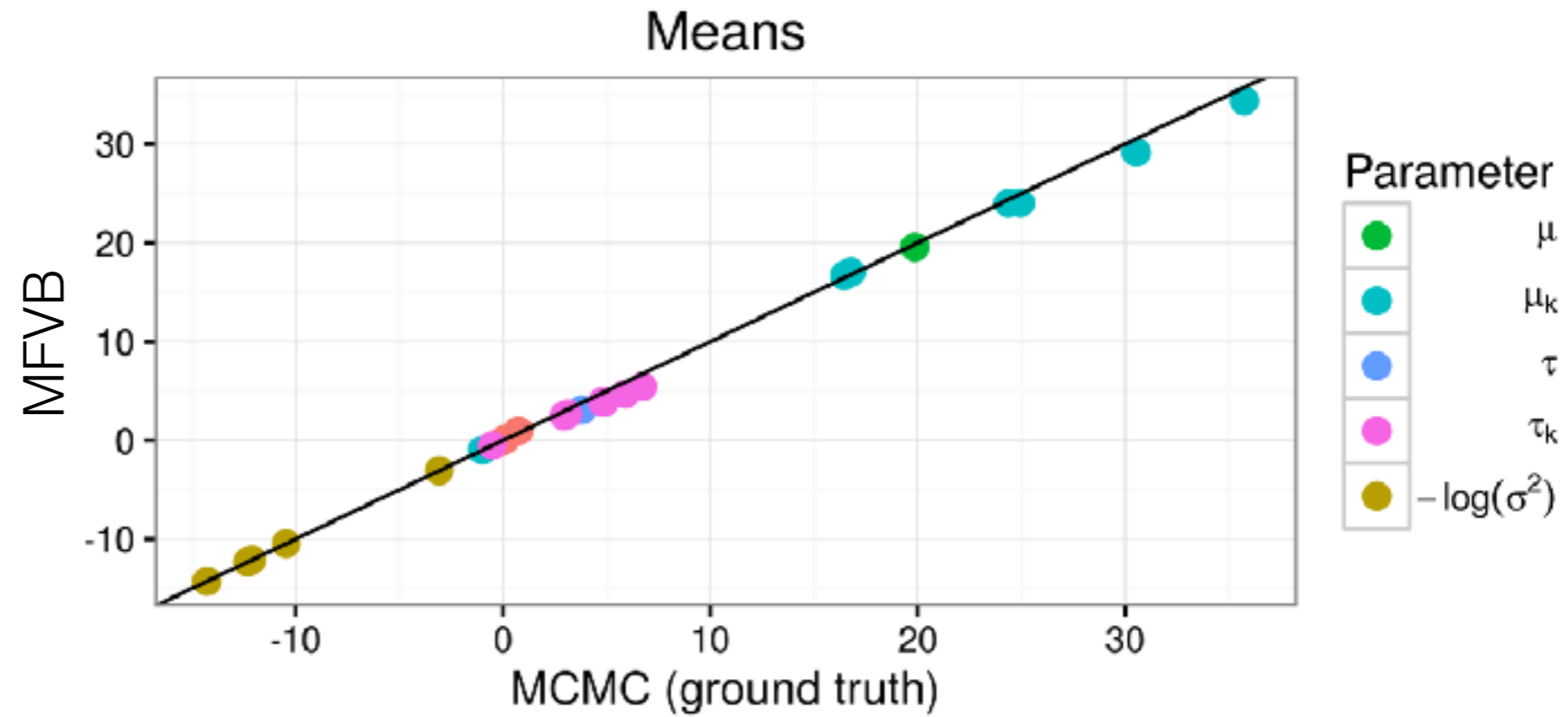
# Microcredit

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



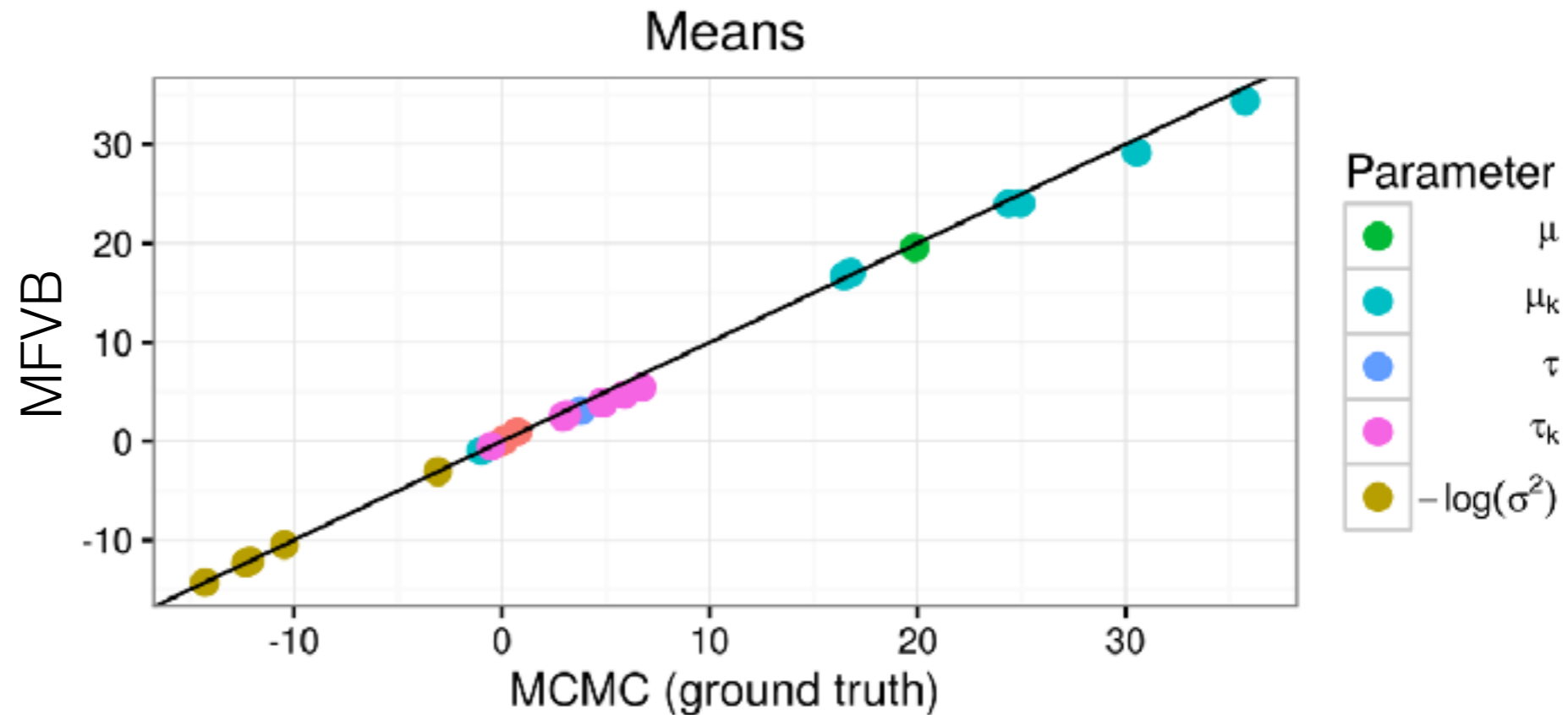
# Microcredit

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



# Microcredit

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

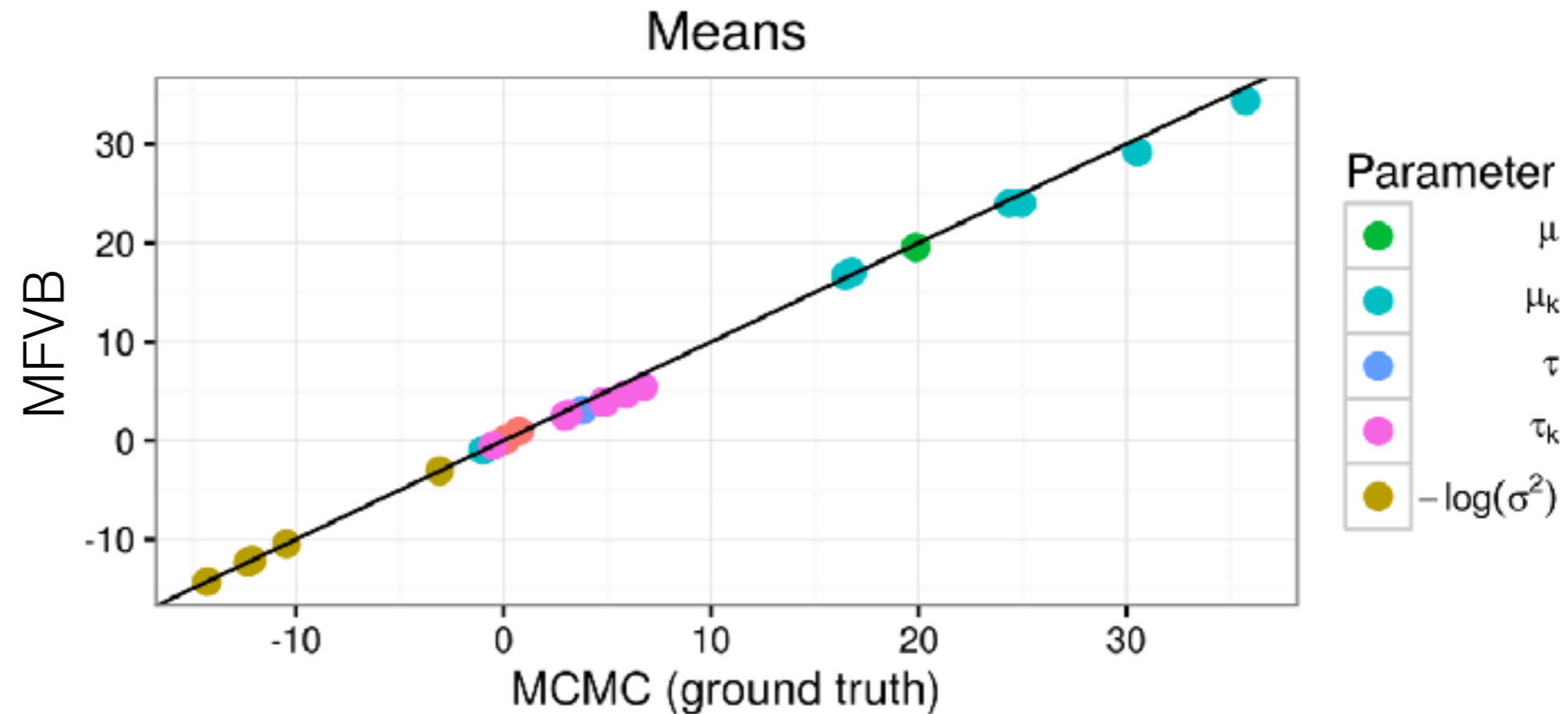


# Criteo Online Ads Experiment

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

# Microcredit

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

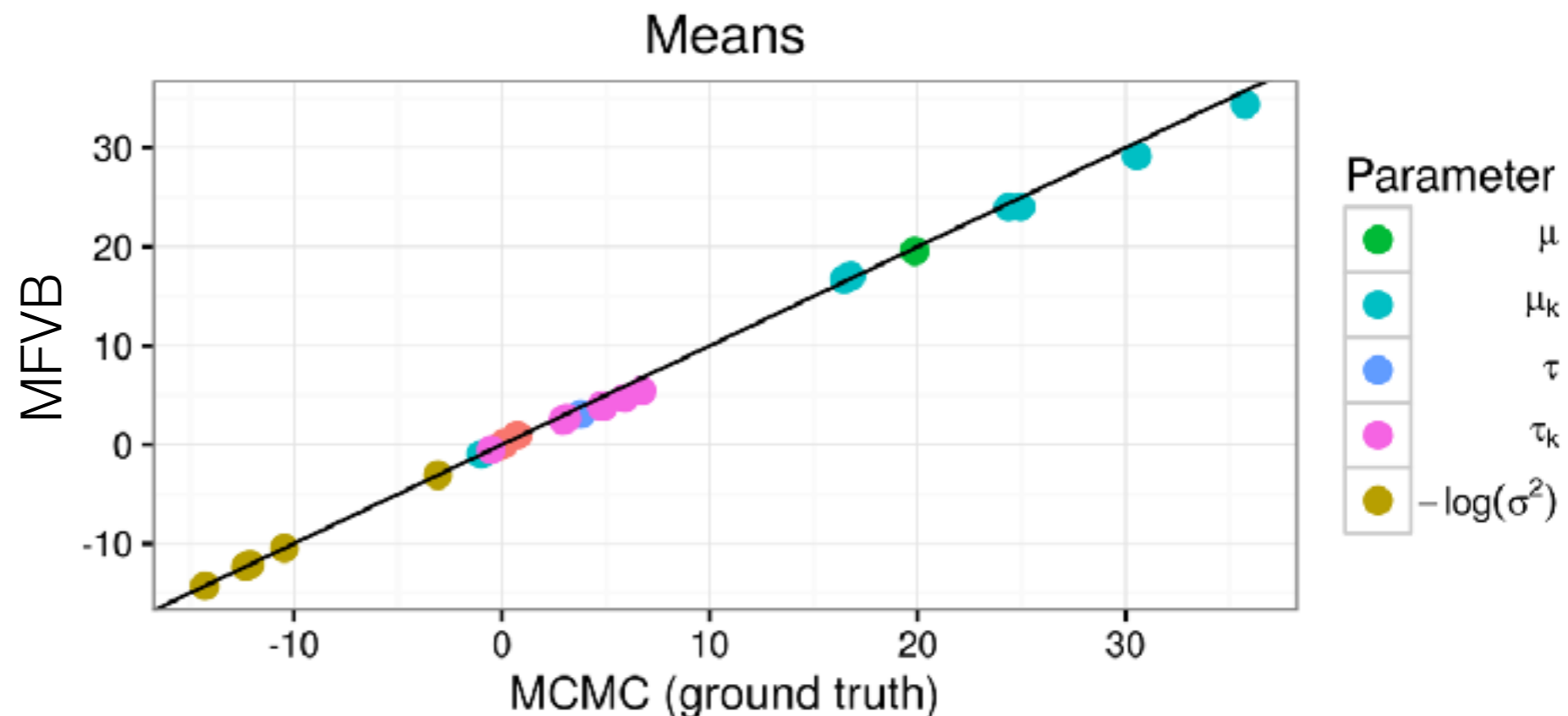


# Criteo Online Ads Experiment

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

# Microcredit

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

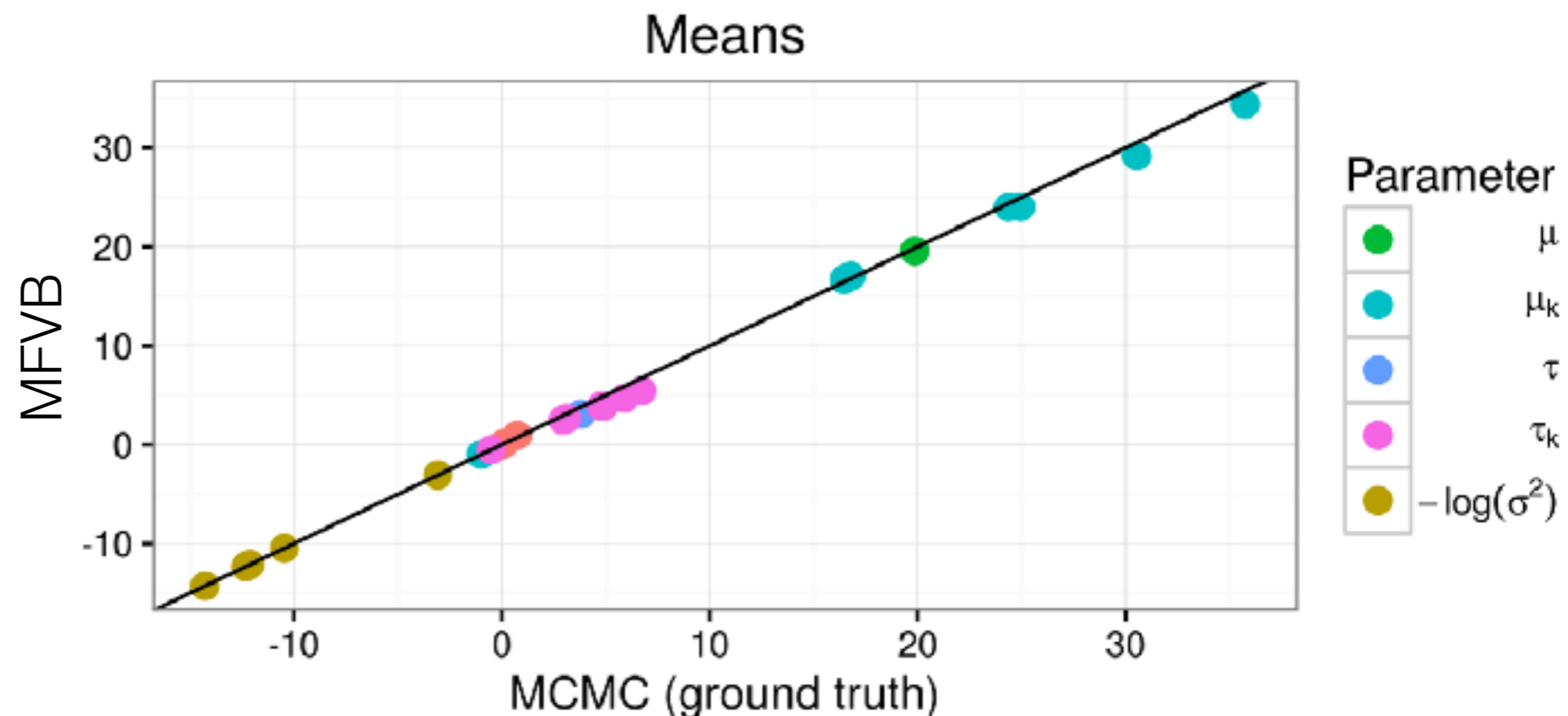


# Criteo Online Ads Experiment

- 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

# 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



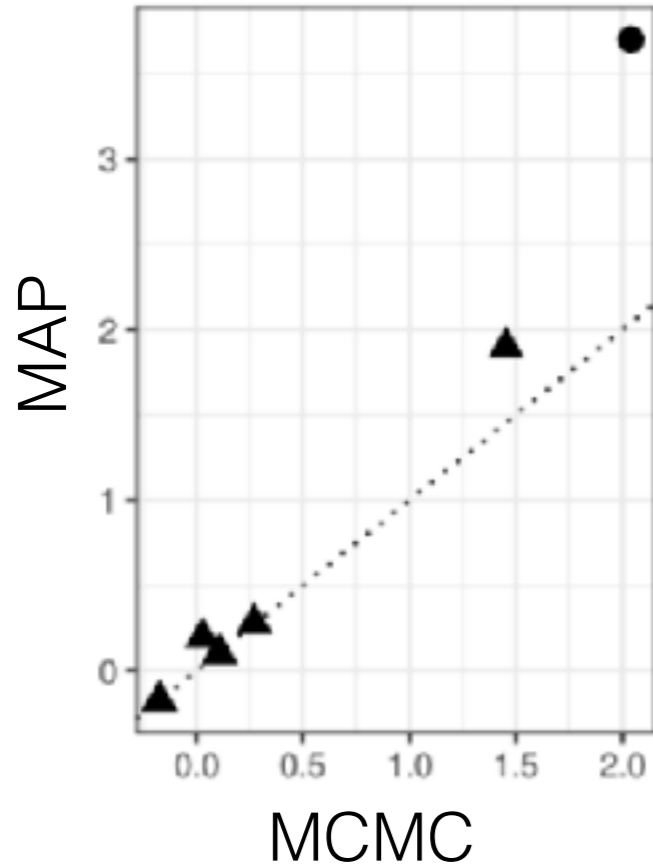
# Criteo Online Ads Experiment

# Criteo Online Ads Experiment

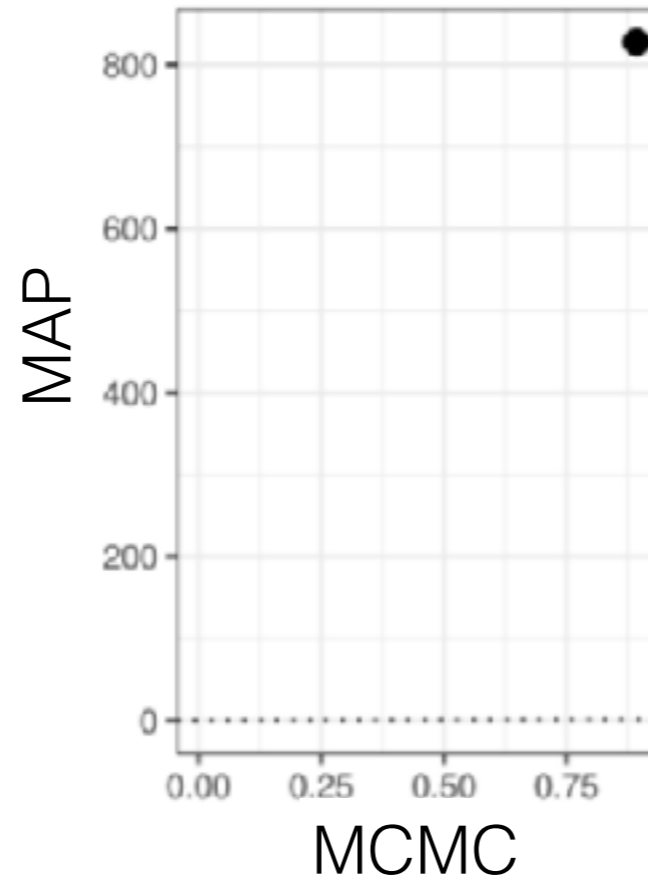
- MAP: **12 s**

# Criteo Online Ads Experiment

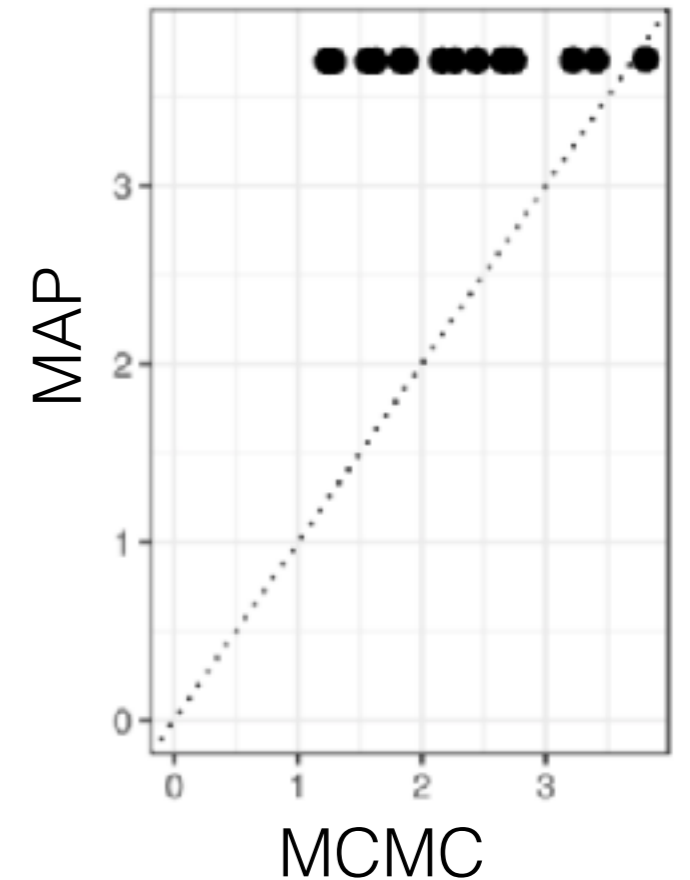
Global parameters ( $-\tau$ )



Global parameter  $\tau$



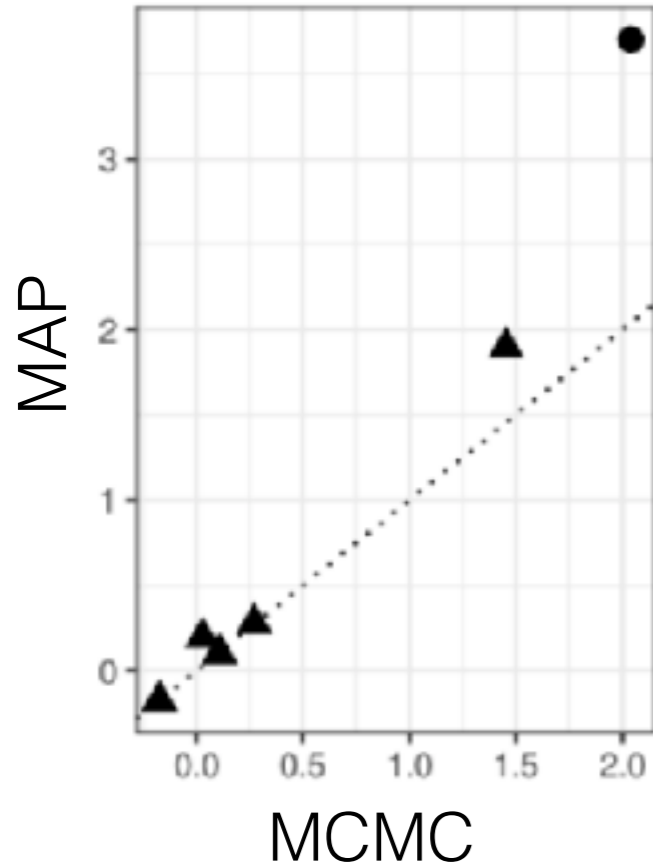
Local parameters



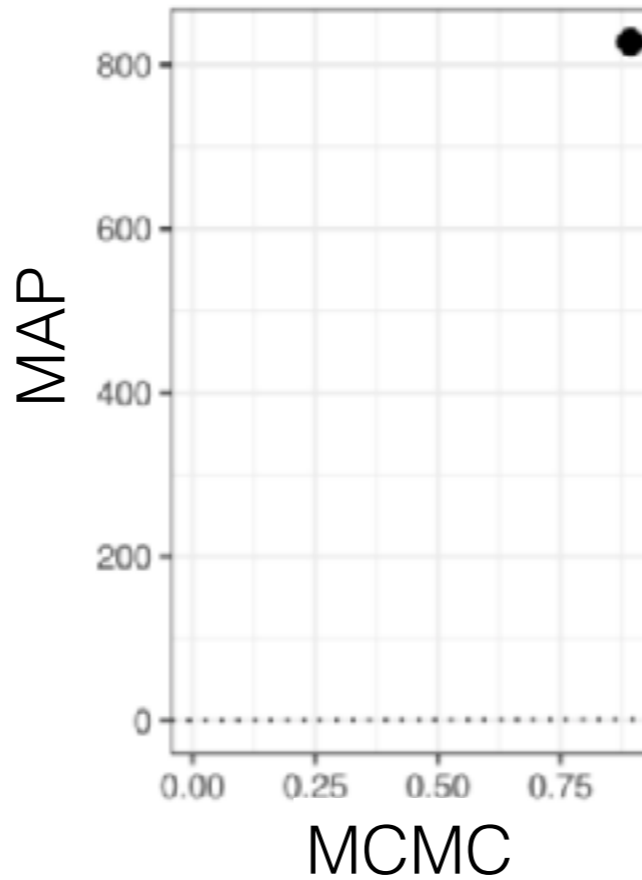
- MAP: **12 s**

# Criteo Online Ads Experiment

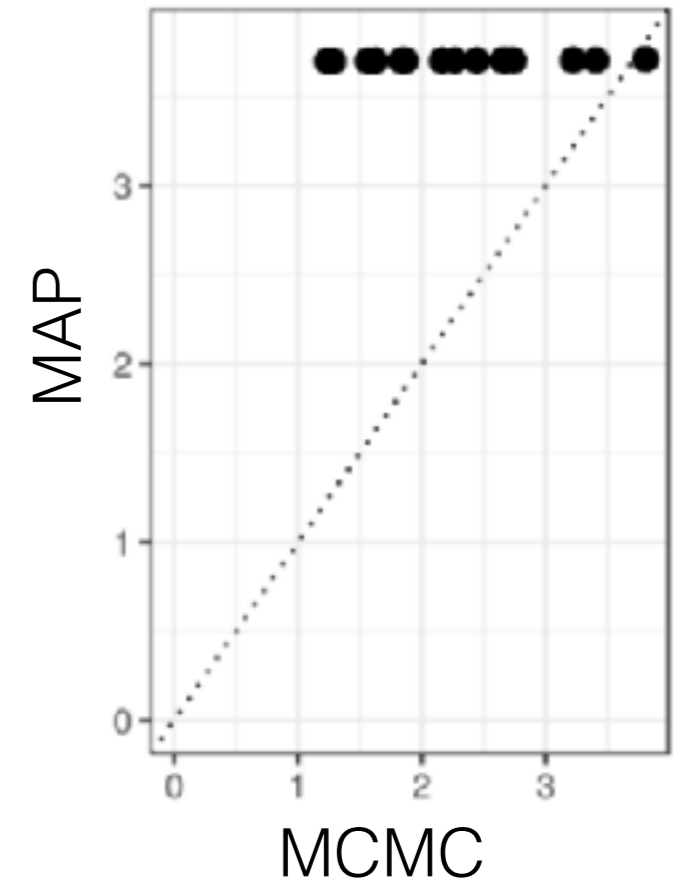
Global parameters ( $-\tau$ )



Global parameter  $\tau$



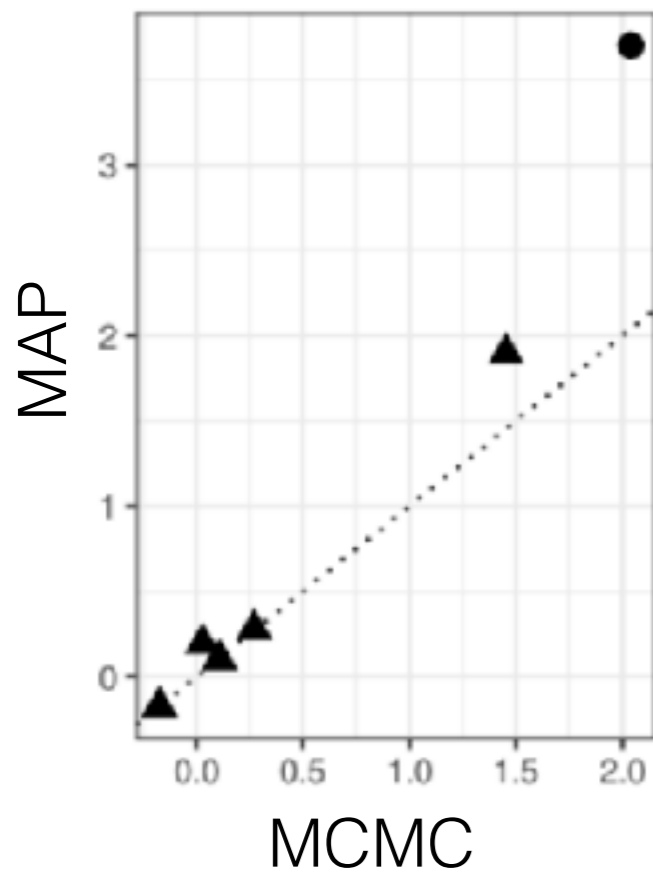
Local parameters



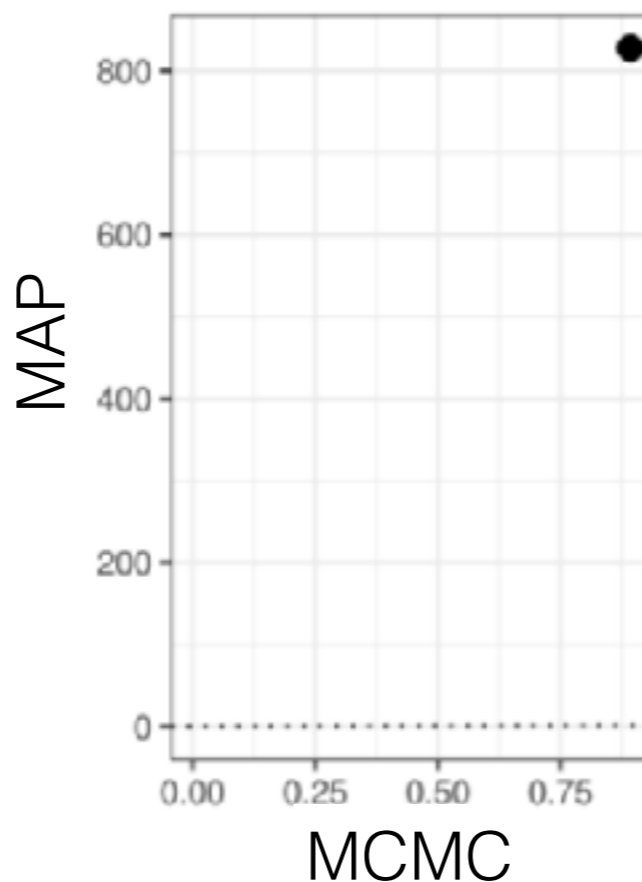
- MAP: **12 s**
- MFVB: **57 s**

# Criteo Online Ads Experiment

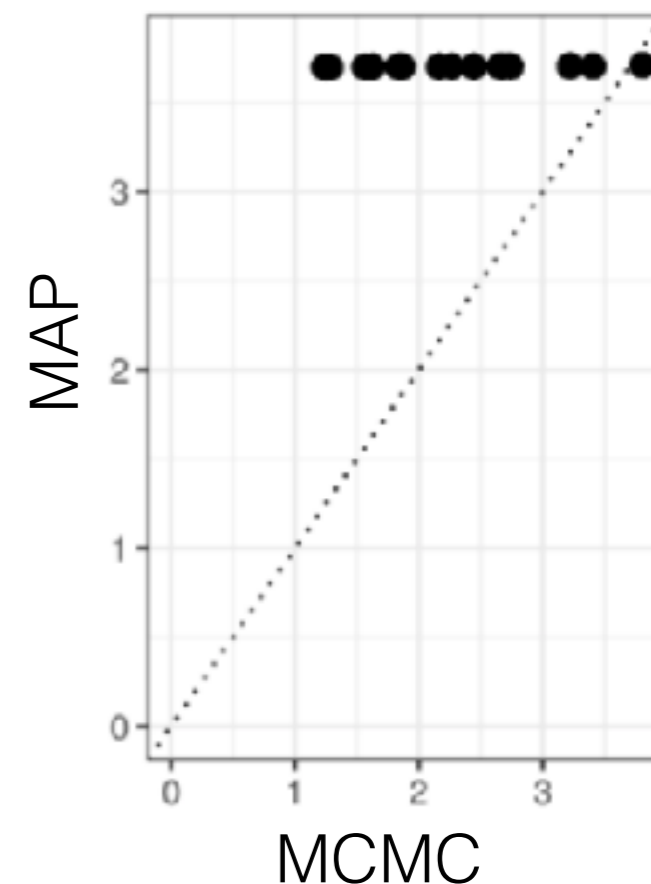
Global parameters ( $-\tau$ )



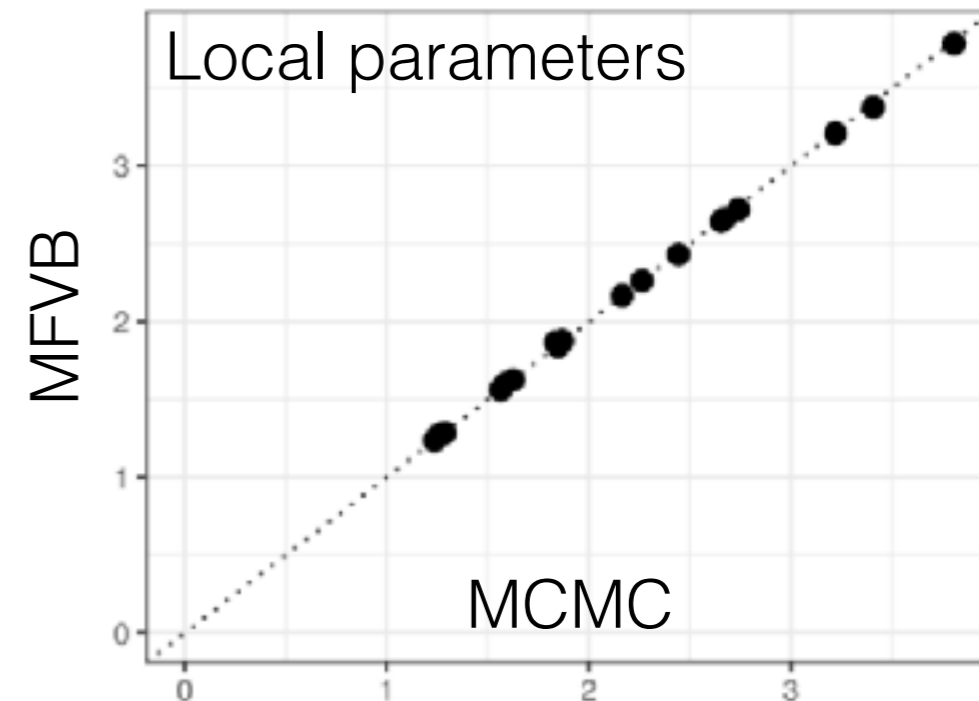
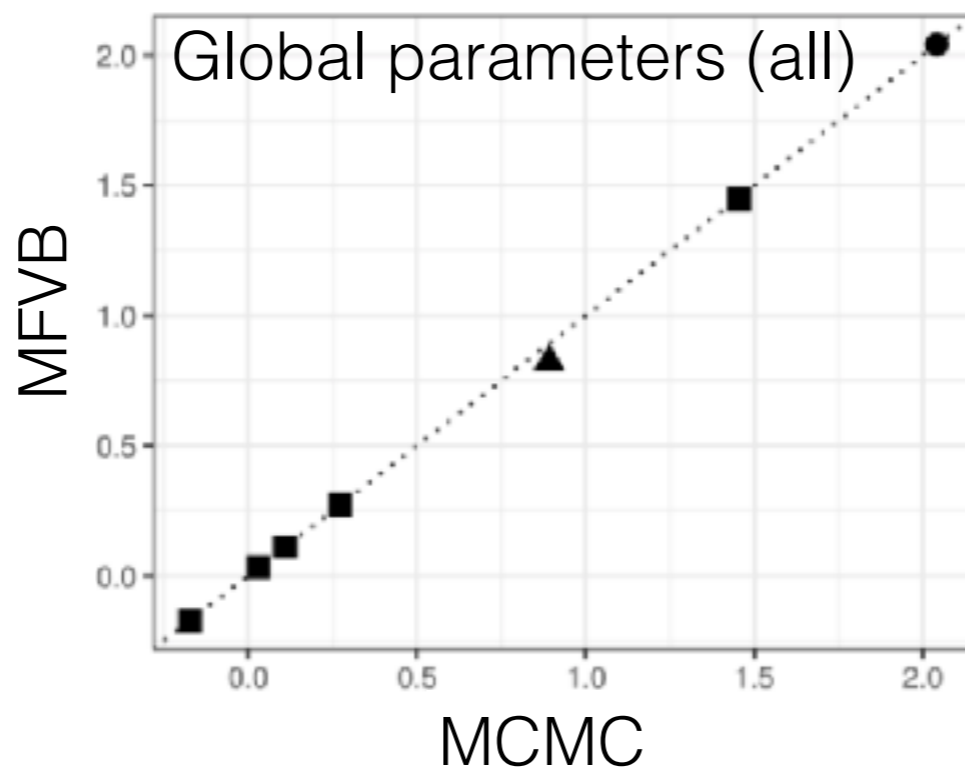
Global parameter  $\tau$



Local parameters

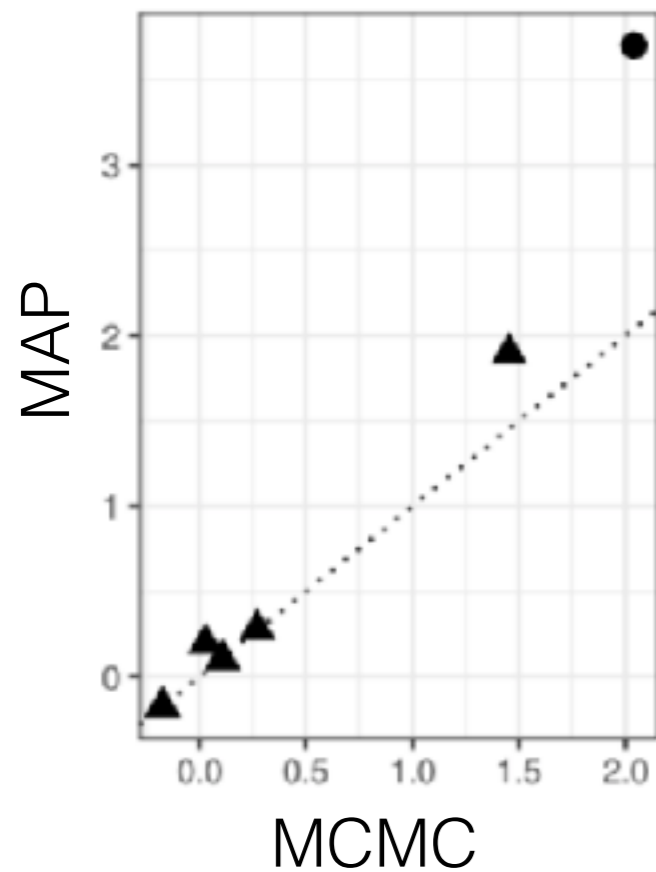


- MAP: **12 s**
- MFVB: **57 s**

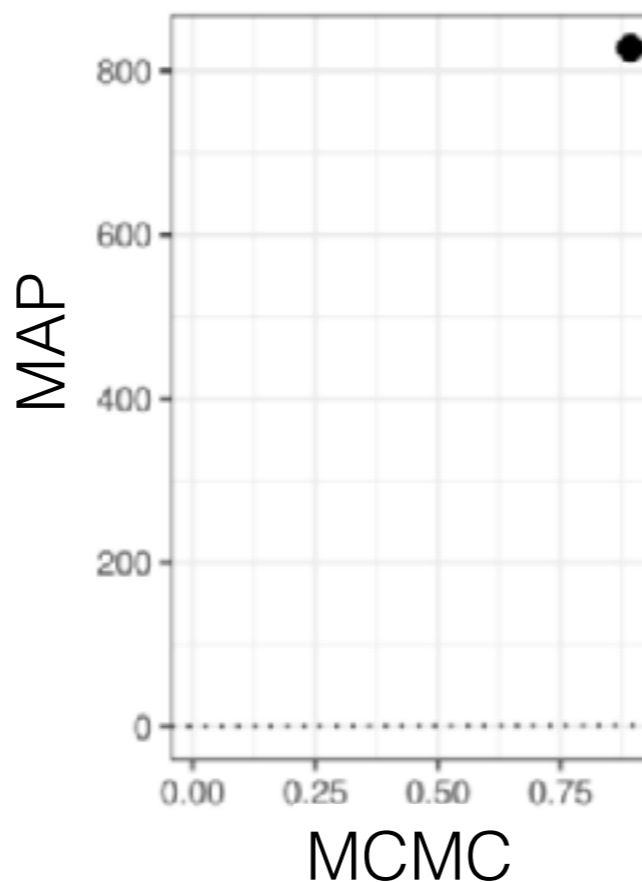


# Criteo Online Ads Experiment

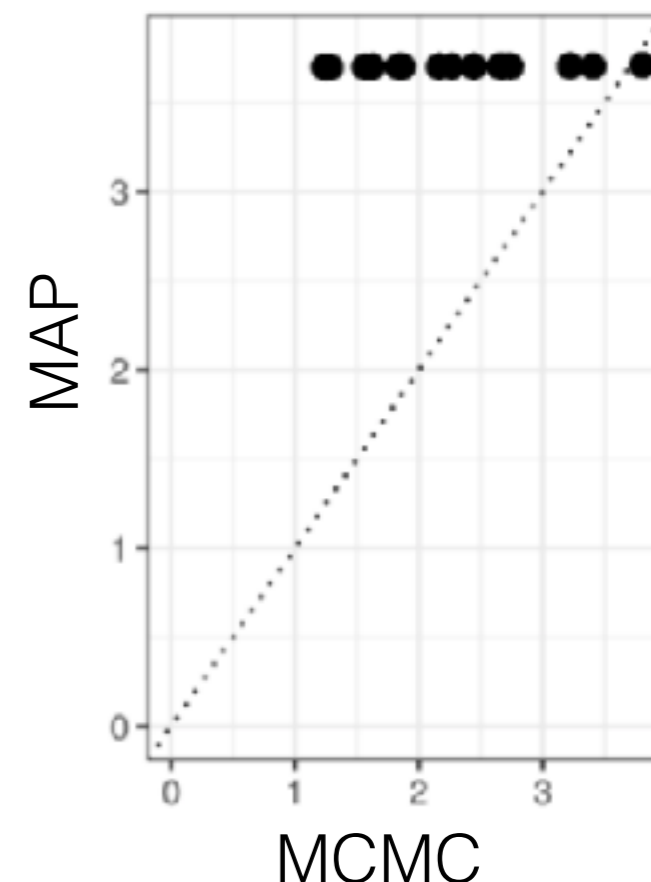
Global parameters ( $-\tau$ )



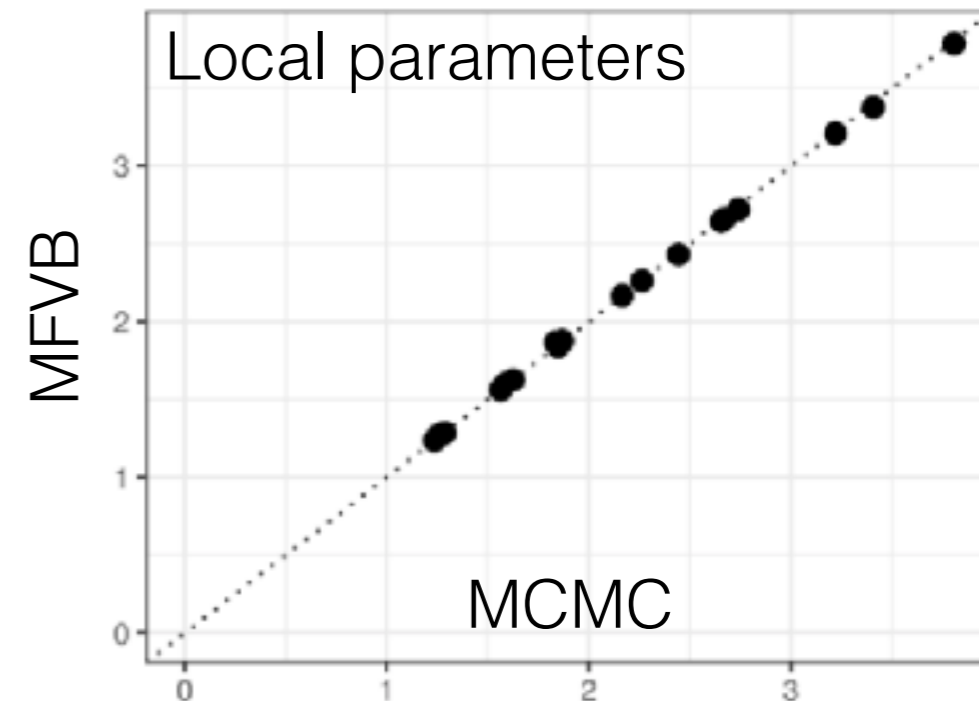
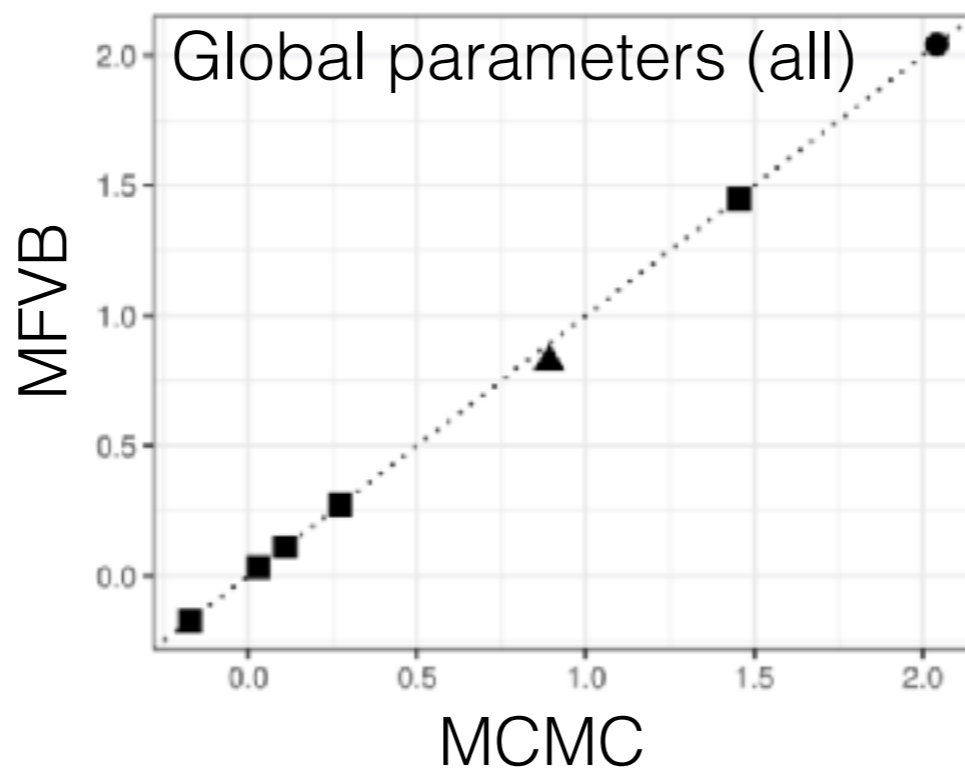
Global parameter  $\tau$



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.



# 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.



# Why use MFVB?

- Topic discovery
- Latent Dirichlet allocation (LDA): 38,000+ 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.

# 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?

# References

Full references at end of final slides