Part II: Variational Bayes and beyond

Tamara Broderick
ITT Career Development
Assistant Professor,
MIT

http://www.tamarabroderick.com/tutorial_2018_lugano.html

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

- 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 MFVB?
- When can we trust MFVB?
- Where do we go from here?

Roadmap

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

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$

[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Model:

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta): \quad y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

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

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

[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

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

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

[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

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

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

[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

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

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

[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

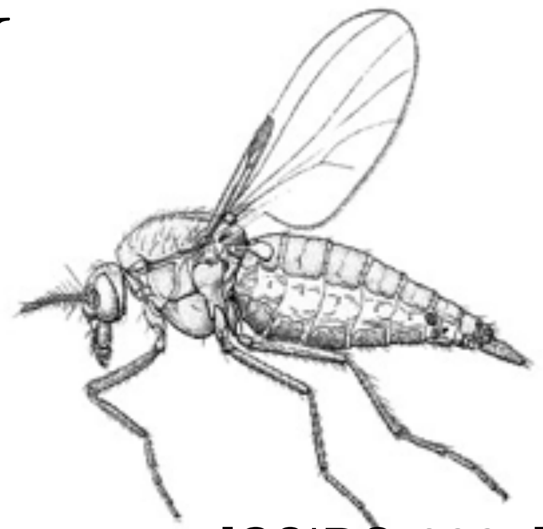
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:

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

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

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



[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

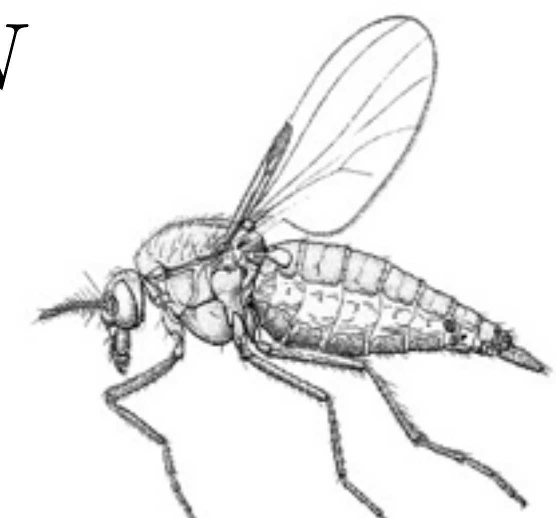
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:

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

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

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



[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

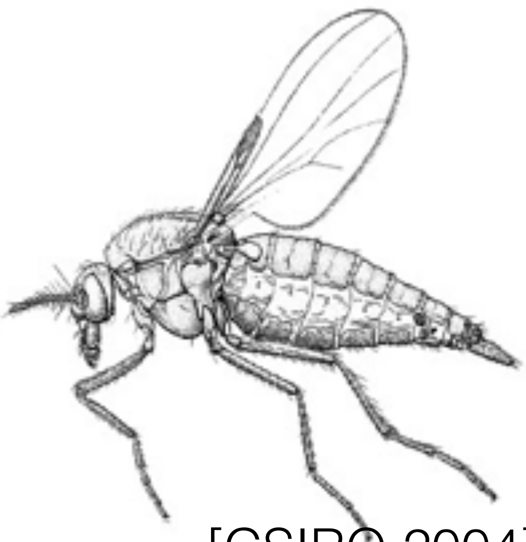
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:

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

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

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



[CSIRO 2004]

- Exercise: check $p(\mu, \tau | y) \neq f_1(\mu, y) f_2(\tau, y)$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

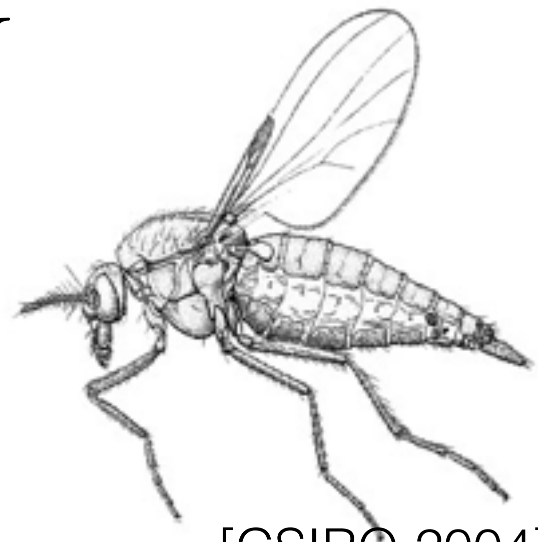
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:

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

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

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



[CSIRO 2004]

- Exercise: check $p(\mu, \tau | y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

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

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

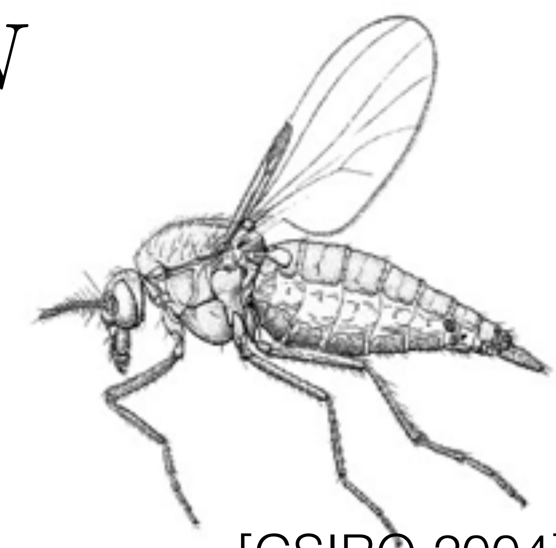
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:

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

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

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



[CSIRO 2004]

- Exercise: check $p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

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

- Coordinate descent (derivation shortly) [Bishop 2006, Sec 10.1.3]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

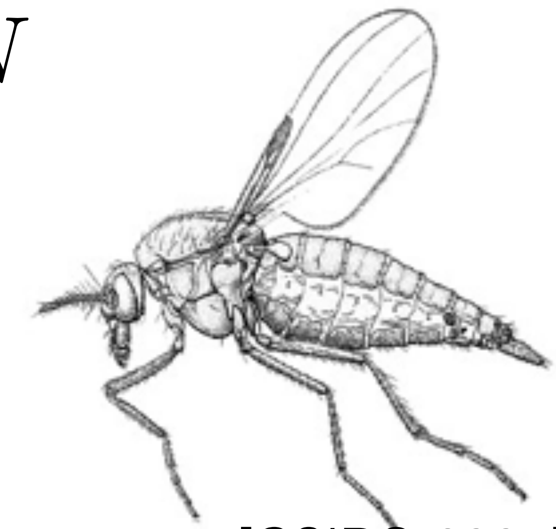
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:

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

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

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



[CSIRO 2004]

- Exercise: check $p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

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

- Coordinate descent (derivation shortly) [Bishop 2006, Sec 10.1.3]

$$q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \qquad q^*_\tau(\tau) = \mathrm{Gamma}(\tau|a_N, b_N)$$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

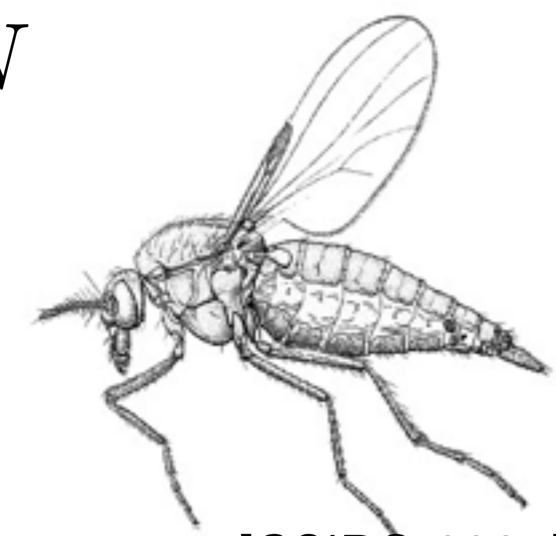
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:

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

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

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



[CSIRO 2004]

- Exercise: check $p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

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

- Coordinate descent (derivation shortly) [Bishop 2006, Sec 10.1.3]

$$q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \qquad q^*_\tau(\tau) = \mathrm{Gamma}(\tau|a_N, b_N)$$

"variational parameters"

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

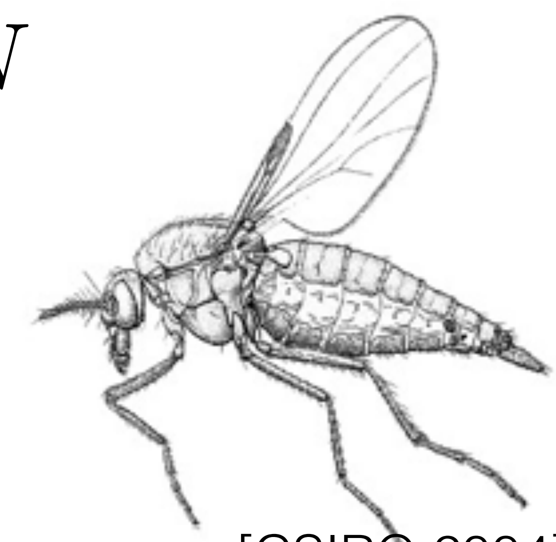
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:

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

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

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



[CSIRO 2004]

- Exercise: check $p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

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

- Coordinate descent (derivation shortly) [Bishop 2006, Sec 10.1.3]

$$q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \qquad q^*_\tau(\tau) = \operatorname{Gamma}(\tau|a_N, b_N)$$

  - Iterate: $(\mu_N, \rho_N) = f(a_N, b_N)$

    $(a_N, b_N) = g(\mu_N, \rho_N)$

“variational parameters”

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

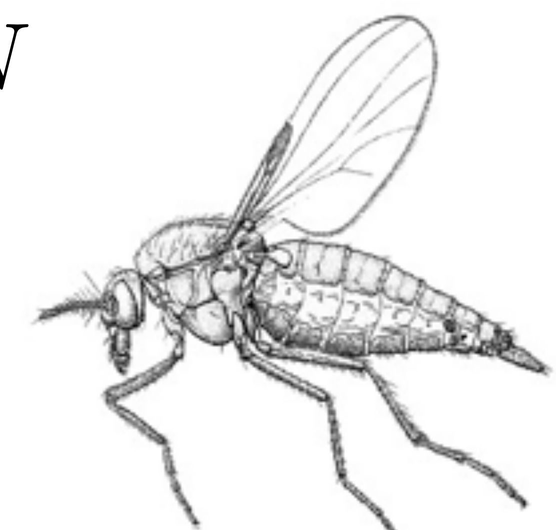
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior):

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

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

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



[CSIRO 2004]

- Exercise: check $p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

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

- Coordinate descent (derivation shortly) [Bishop 2006, Sec 10.1.3]

$$q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \qquad q^*_\tau(\tau) = \mathrm{Gamma}(\tau|a_N, b_N)$$

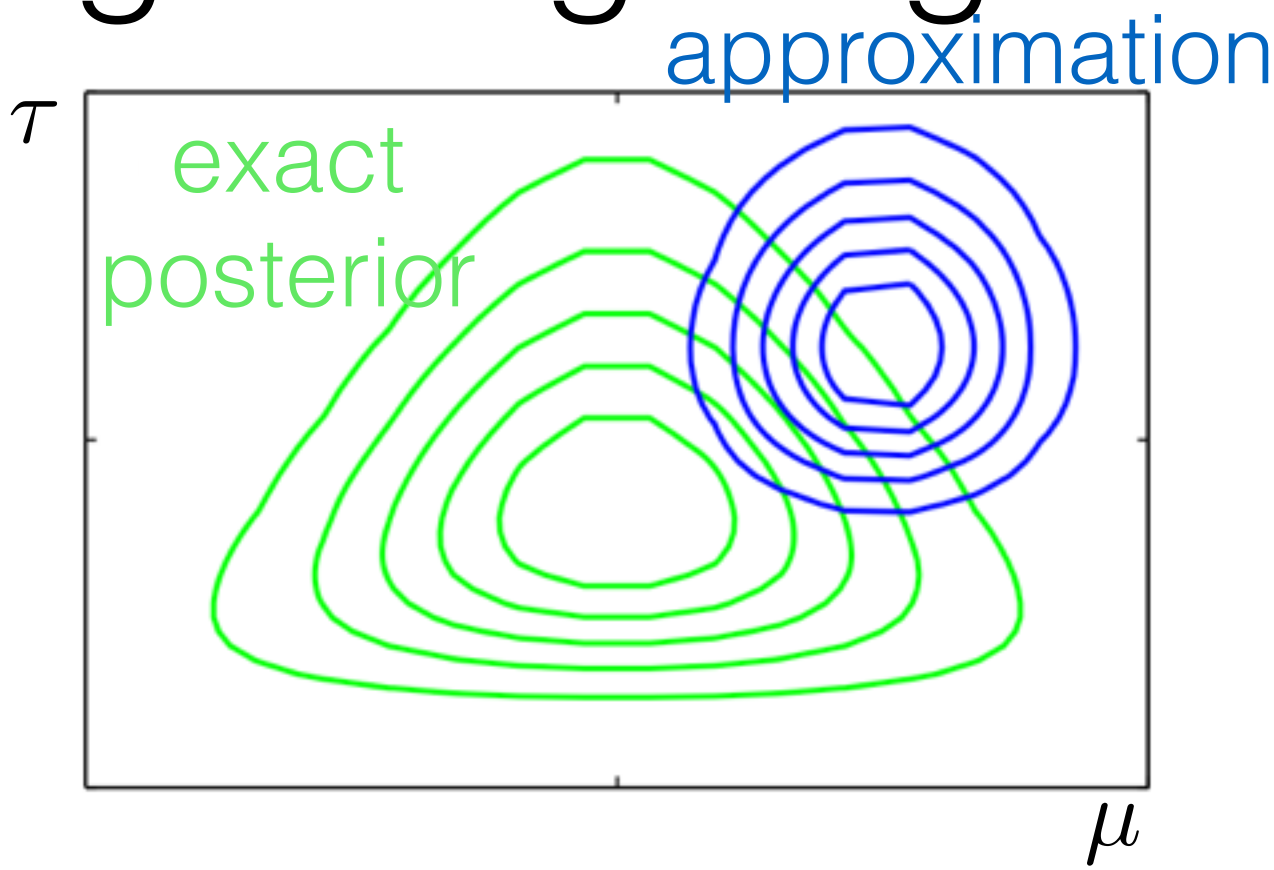
  - Iterate: $(\mu_N, \rho_N) = f(a_N, b_N)$

    $(a_N, b_N) = g(\mu_N, \rho_N)$

"variational parameters"

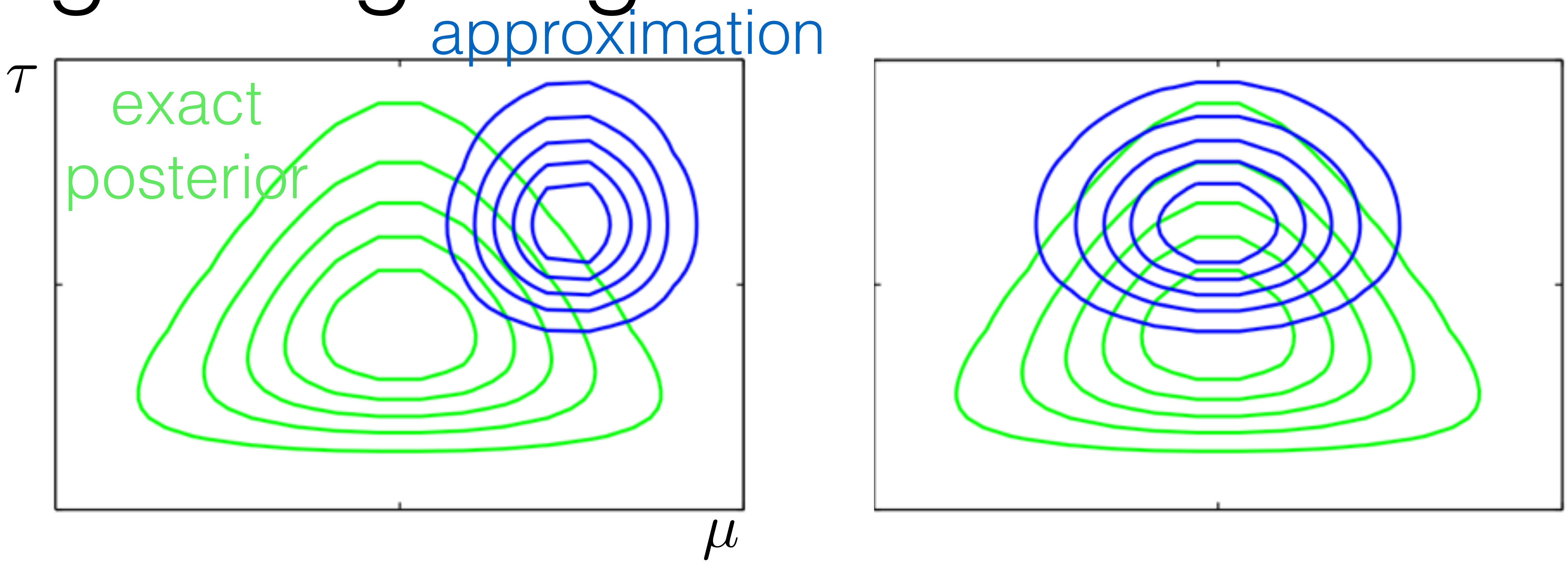
[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length



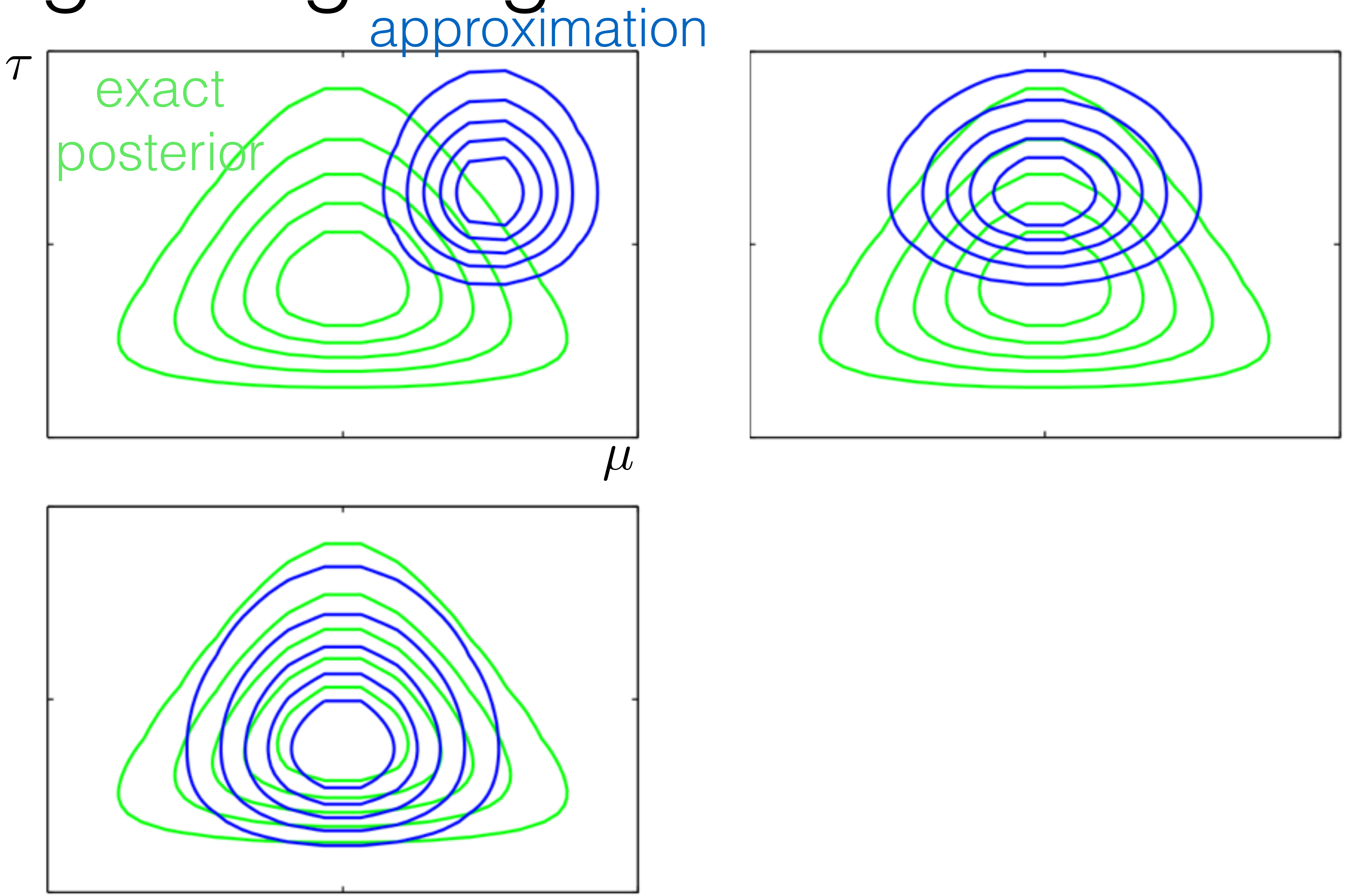
[Bishop 2006]

Midge wing length



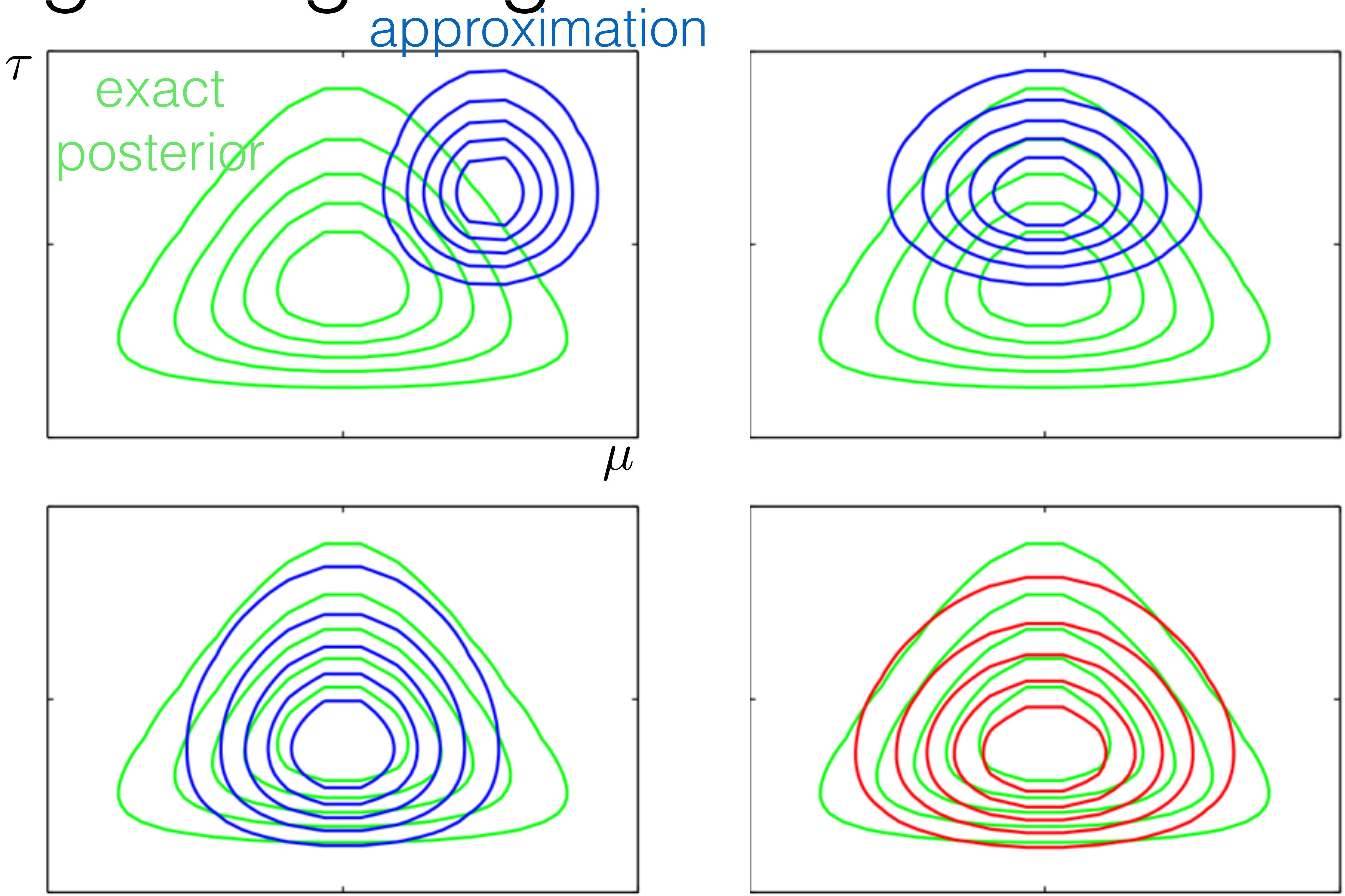
[Bishop 2006]

Midge wing length



[Bishop 2006]

Midge wing length



[Bishop 2006]

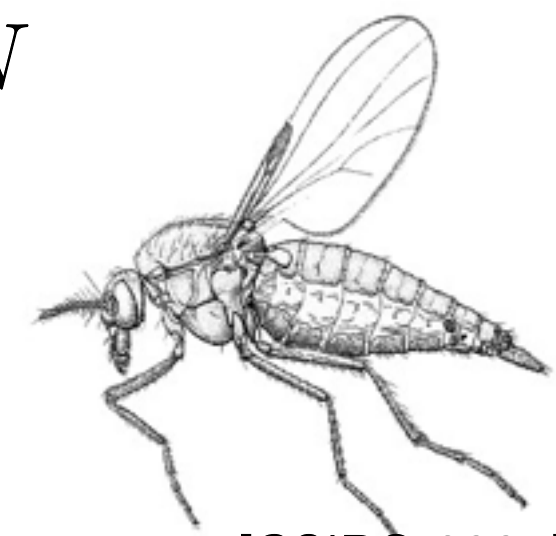
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:

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

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

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



[CSIRO 2004]

- Exercise: check $p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

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

- Coordinate descent (derivation shortly) [Bishop 2006, Sec 10.1.3]

$$q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \qquad q^*_\tau(\tau) = \mathrm{Gamma}(\tau|a_N, b_N)$$

  - Iterate: $(\mu_N, \rho_N) = f(a_N, b_N)$
    $(a_N, b_N) = g(\mu_N, \rho_N)$

"variational parameters"

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

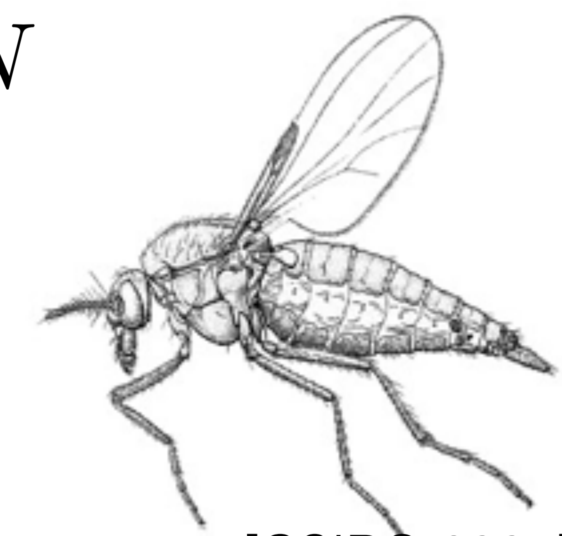
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:

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

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

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



[CSIRO 2004]

- 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 (derivation shortly) [Bishop 2006, Sec 10.1.3]

$$q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \qquad q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$$

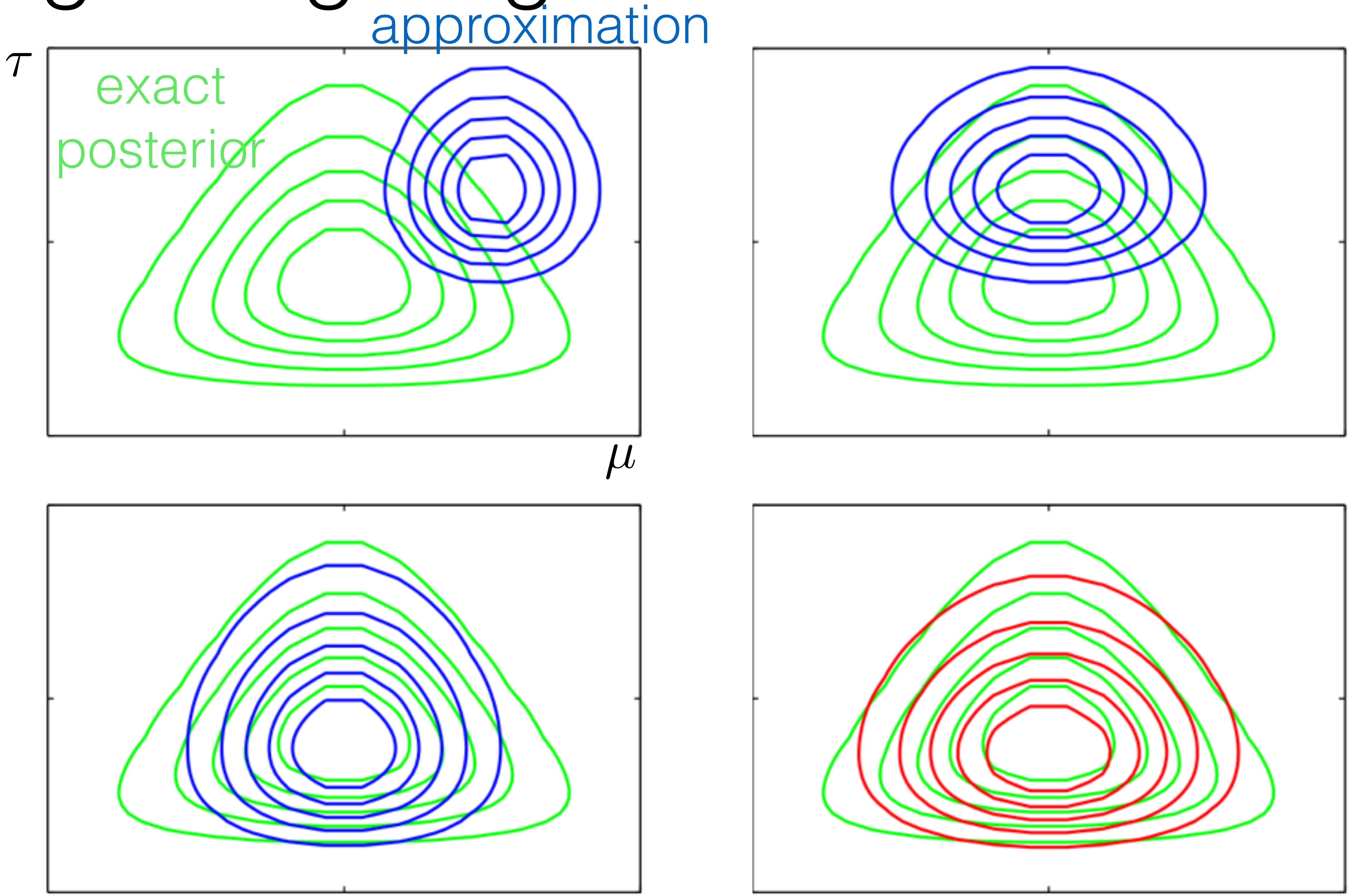
  - Iterate: $(\mu_N, \rho_N) = f(a_N, b_N)$

[board] $(a_N, b_N) = g(\mu_N, \rho_N)$

"variational parameters"

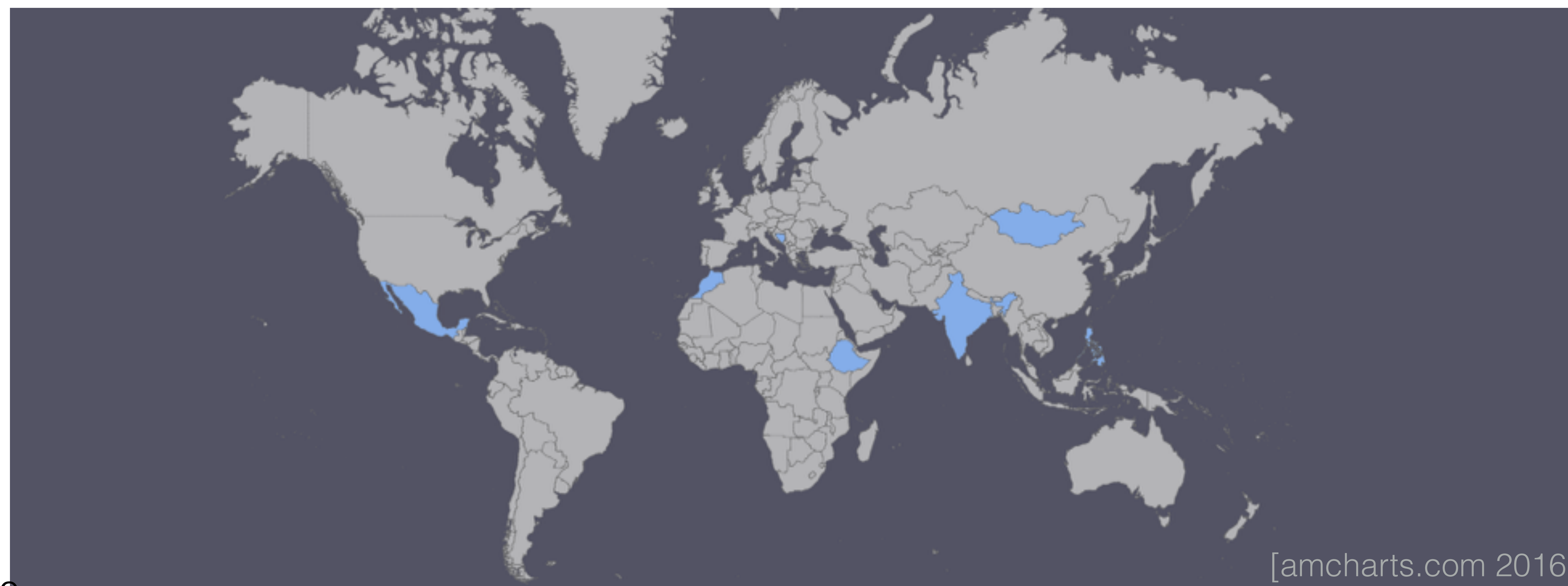
[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length



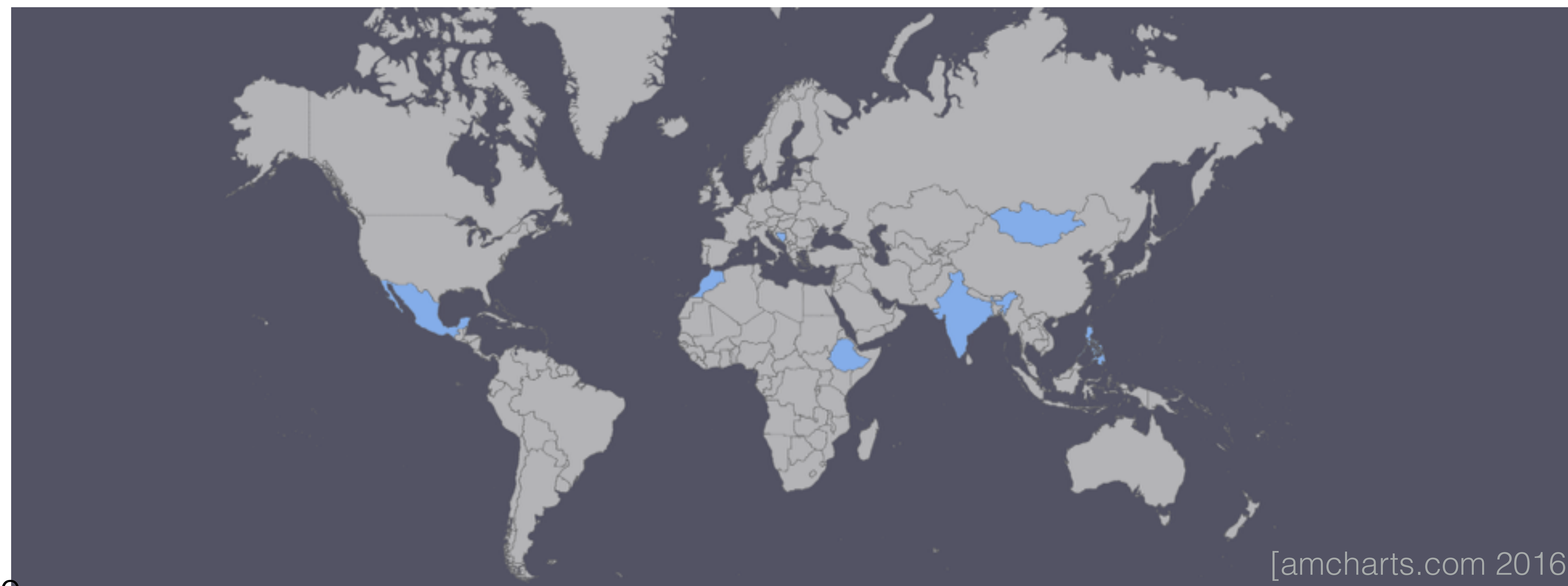
[Bishop 2006]

Microcredit Experiment



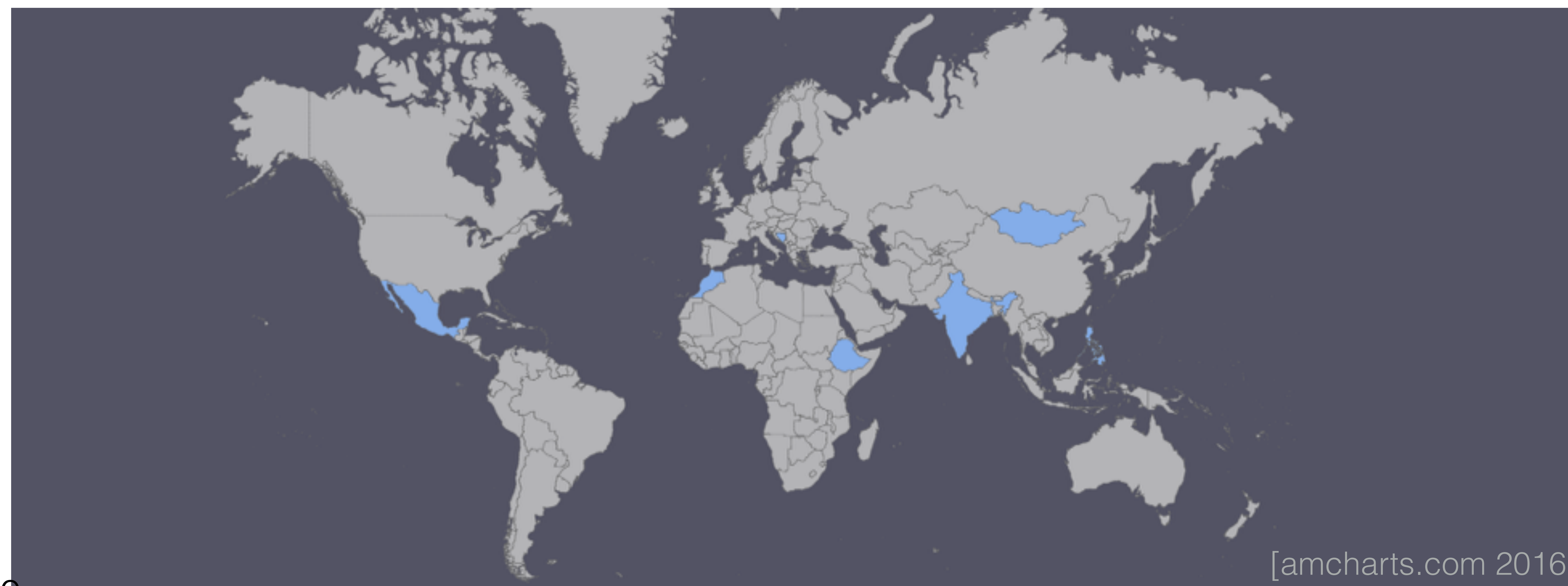
Microcredit Experiment

- Simplified from Meager (2018a)



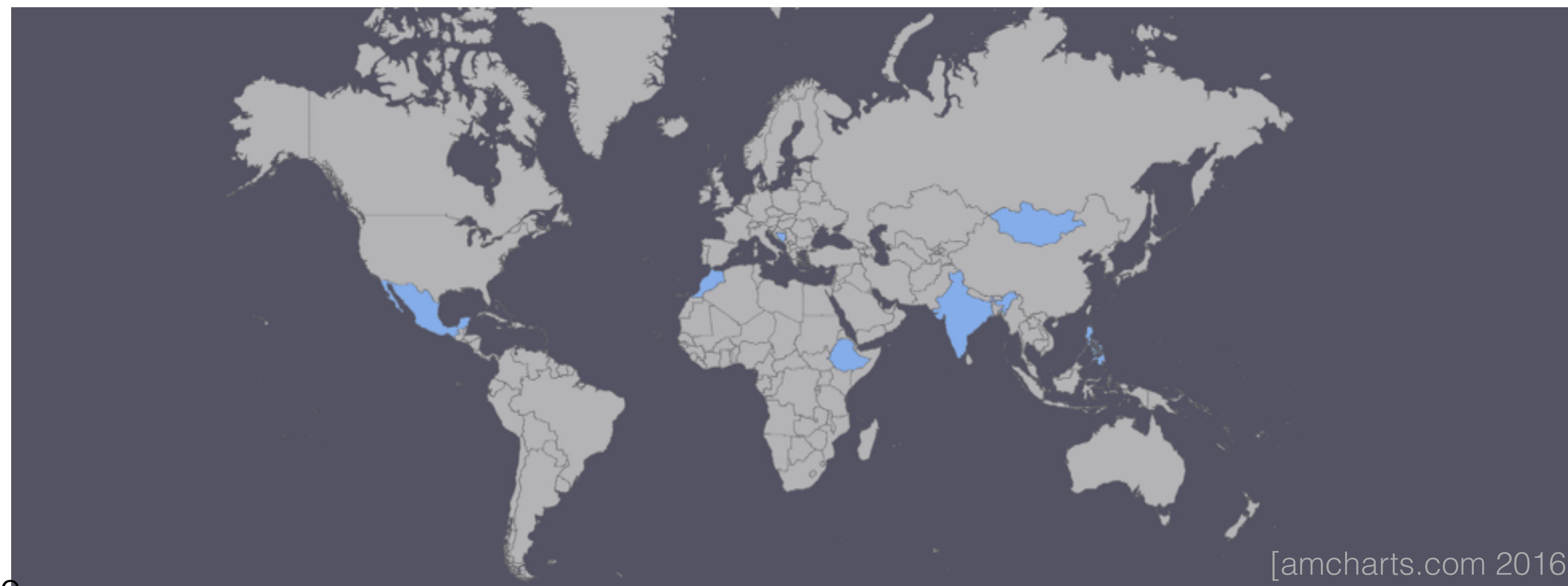
Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)



Microcredit Experiment

- Simplified from Meager (2018a)
- 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 (2018a)
- 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 (2018a)
- 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 (2018a)
- 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 (2018a)
- 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

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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 (2018a)
- 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

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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

$$\underbrace{y_{kn}}_{\text{profit}} \stackrel{indep}{\sim} \mathcal{N}(\mu_k + \overbrace{T_{kn}}^{\text{1 if microcredit}}\boxed{\tau_k}, \quad)$$

Microcredit Experiment

- Simplified from Meager (2018a)
- 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

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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

1 if microcredit

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

- Priors and hyperpriors:

Microcredit Experiment

- Simplified from Meager (2018a)
- 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

1 if microcredit

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

- Priors and hyperpriors:

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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

1 if microcredit

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

- Priors and hyperpriors:

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

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

Microcredit Experiment

- Simplified from Meager (2018a)
- 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

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

- Priors and hyperpriors:

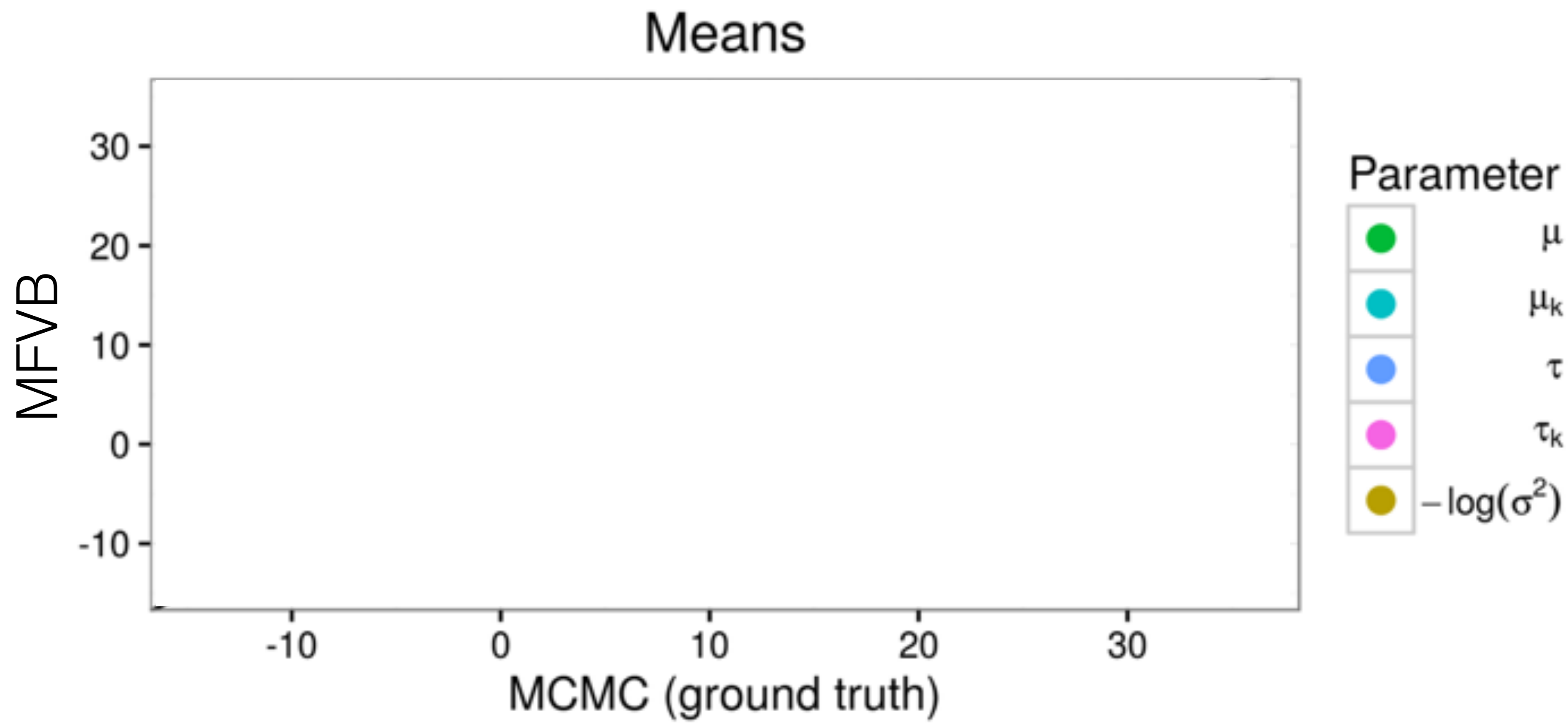
$$\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \stackrel{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, C\right) \qquad \begin{pmatrix} \mu \\ \tau \end{pmatrix} \stackrel{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu_0 \\ \tau_0 \end{pmatrix}, \Lambda^{-1}\right)$$

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

Microcredit

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

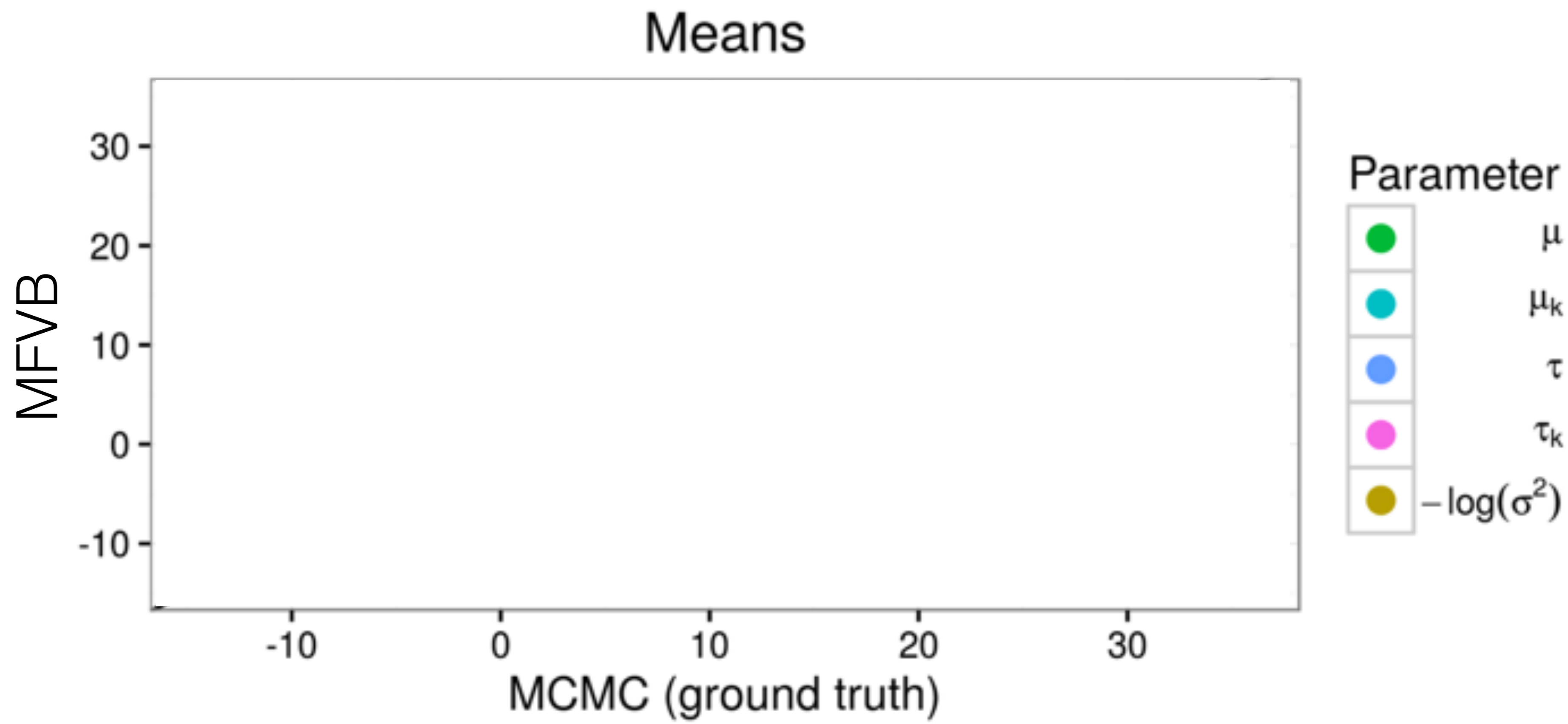
Microcredit



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

Microcredit

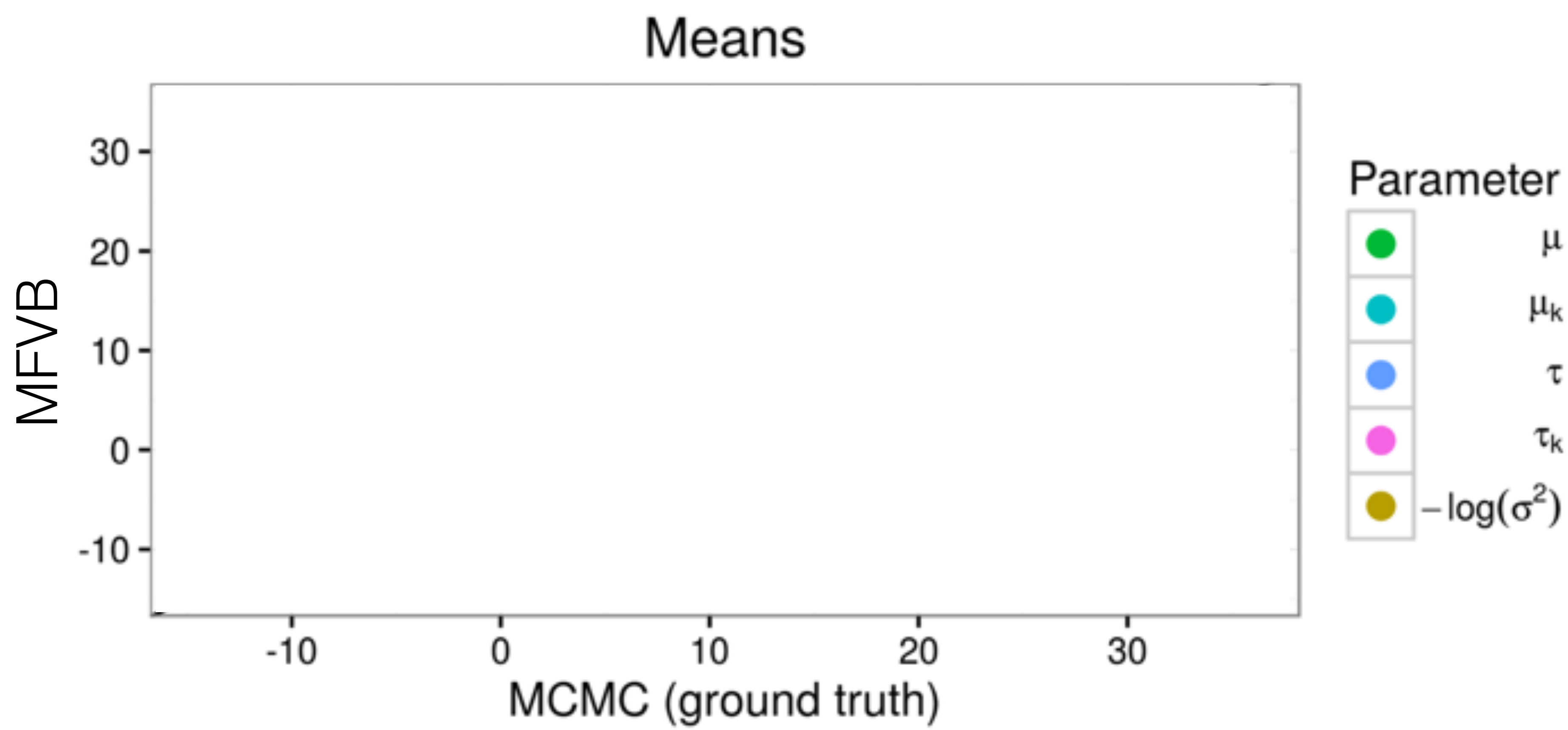
- *One set* of 2500 MCMC draws: **45 minutes**



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

Microcredit

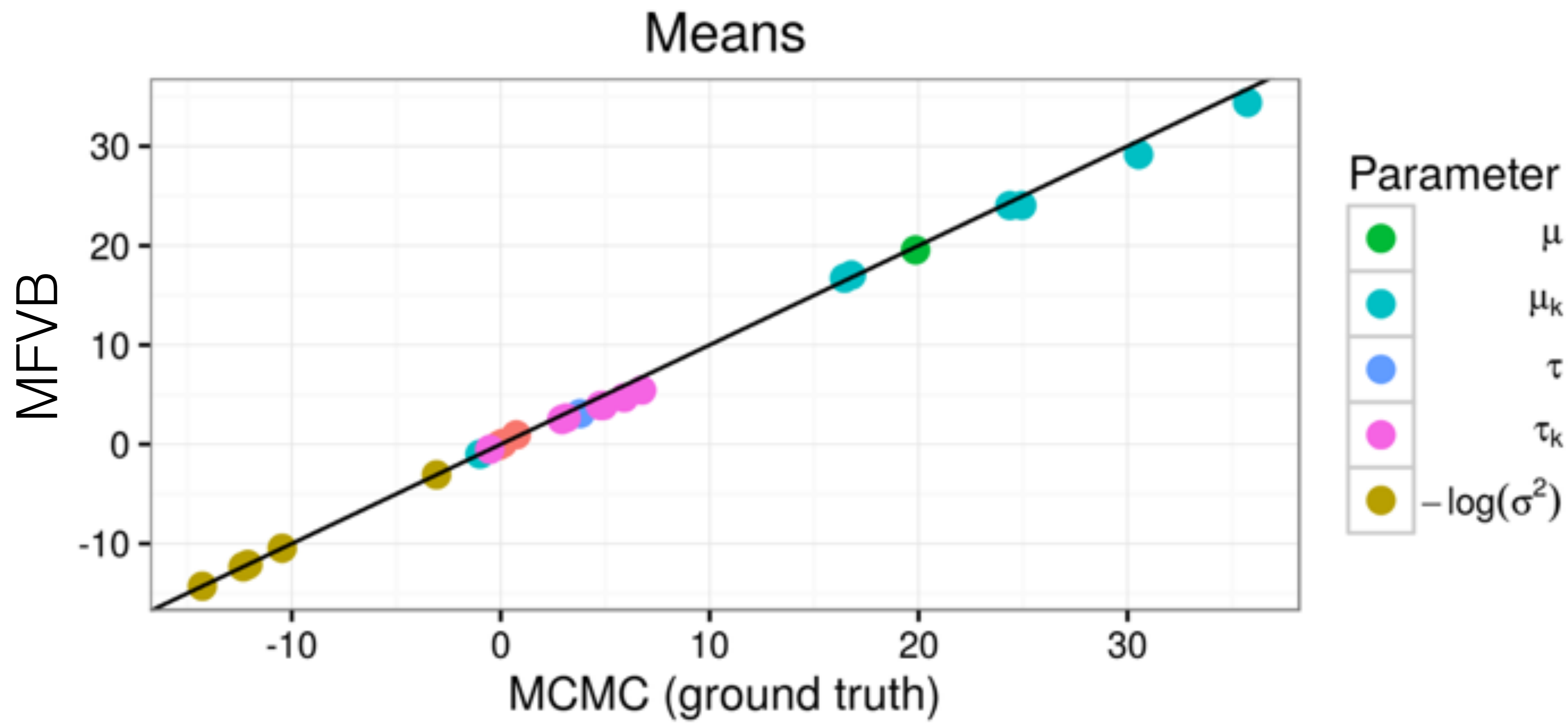
- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



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

Microcredit

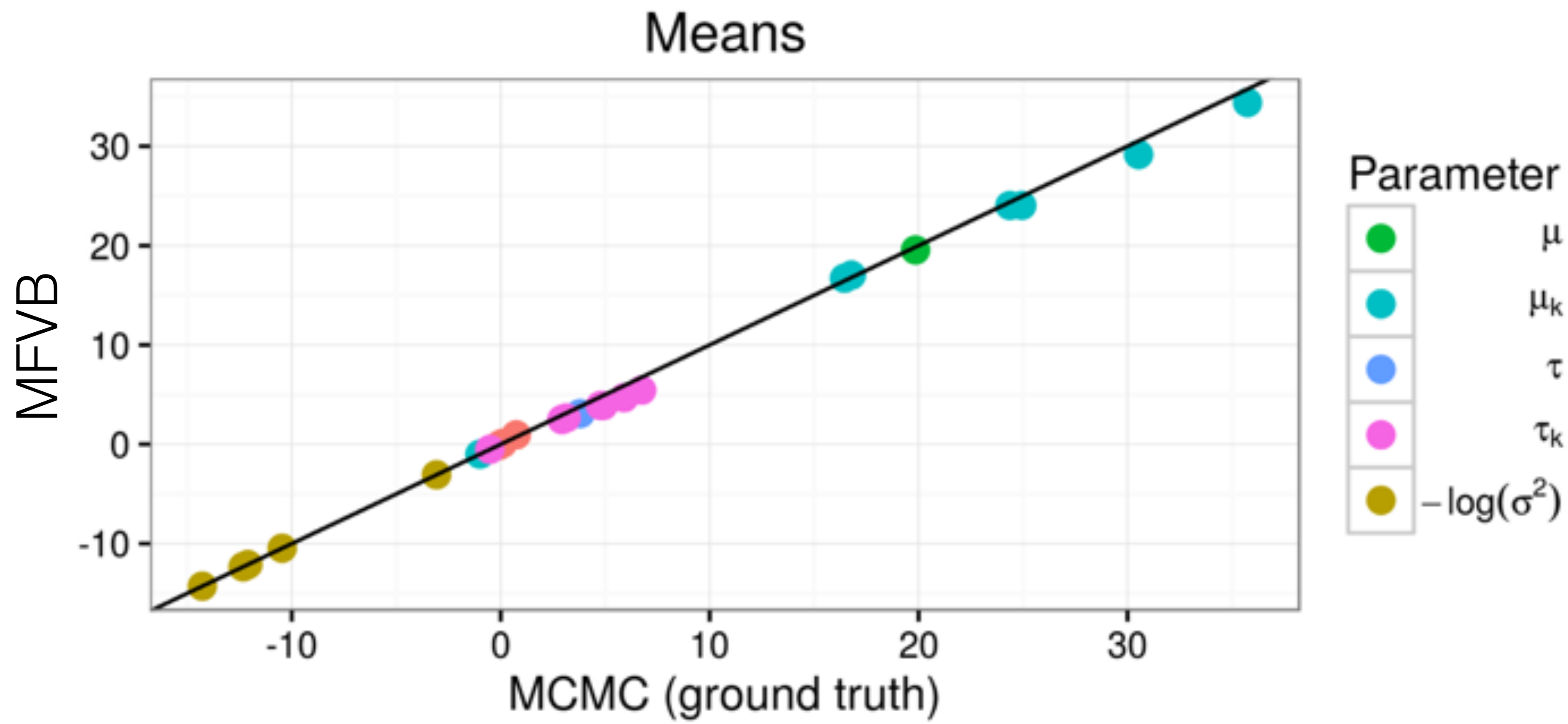
- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



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

Microcredit

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



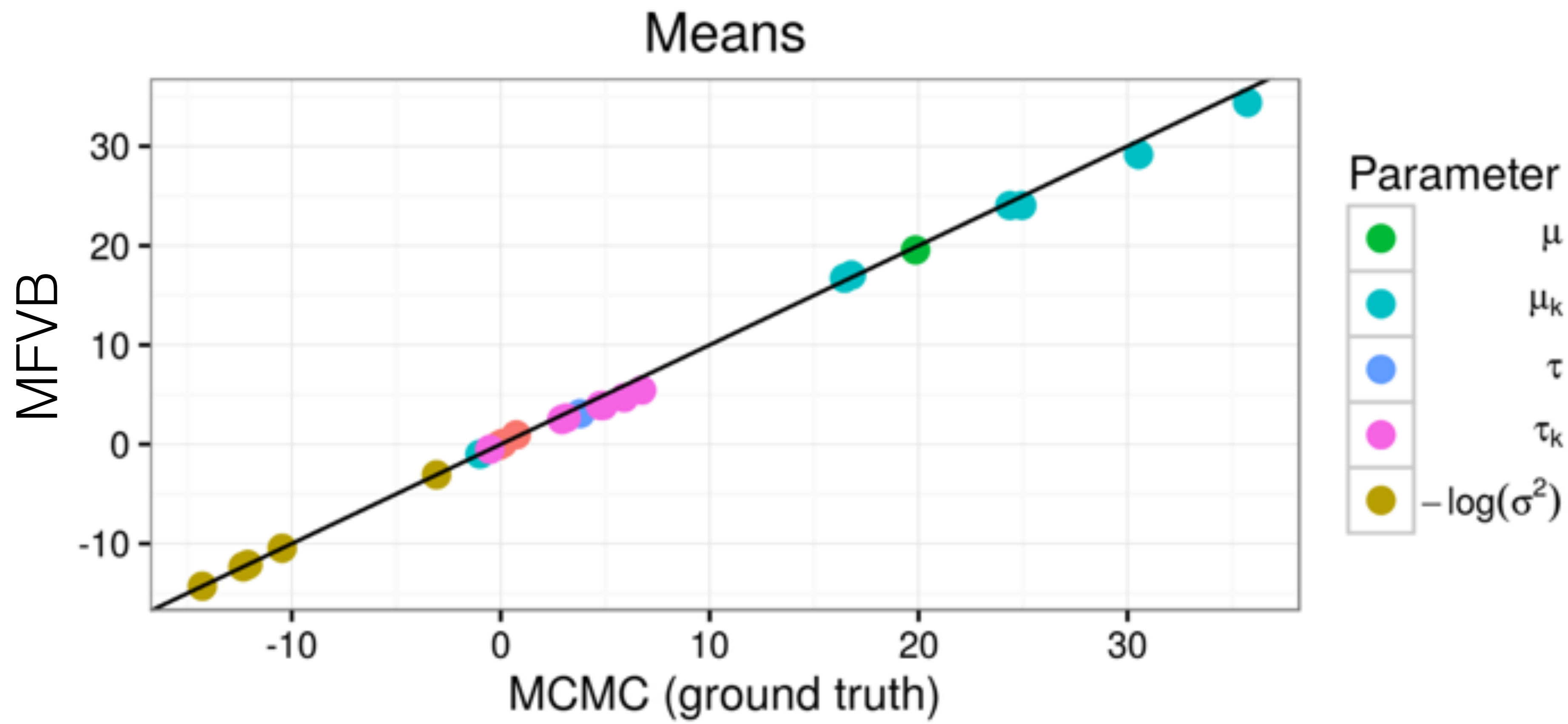
Criteo Online Ads Experiment

- 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**



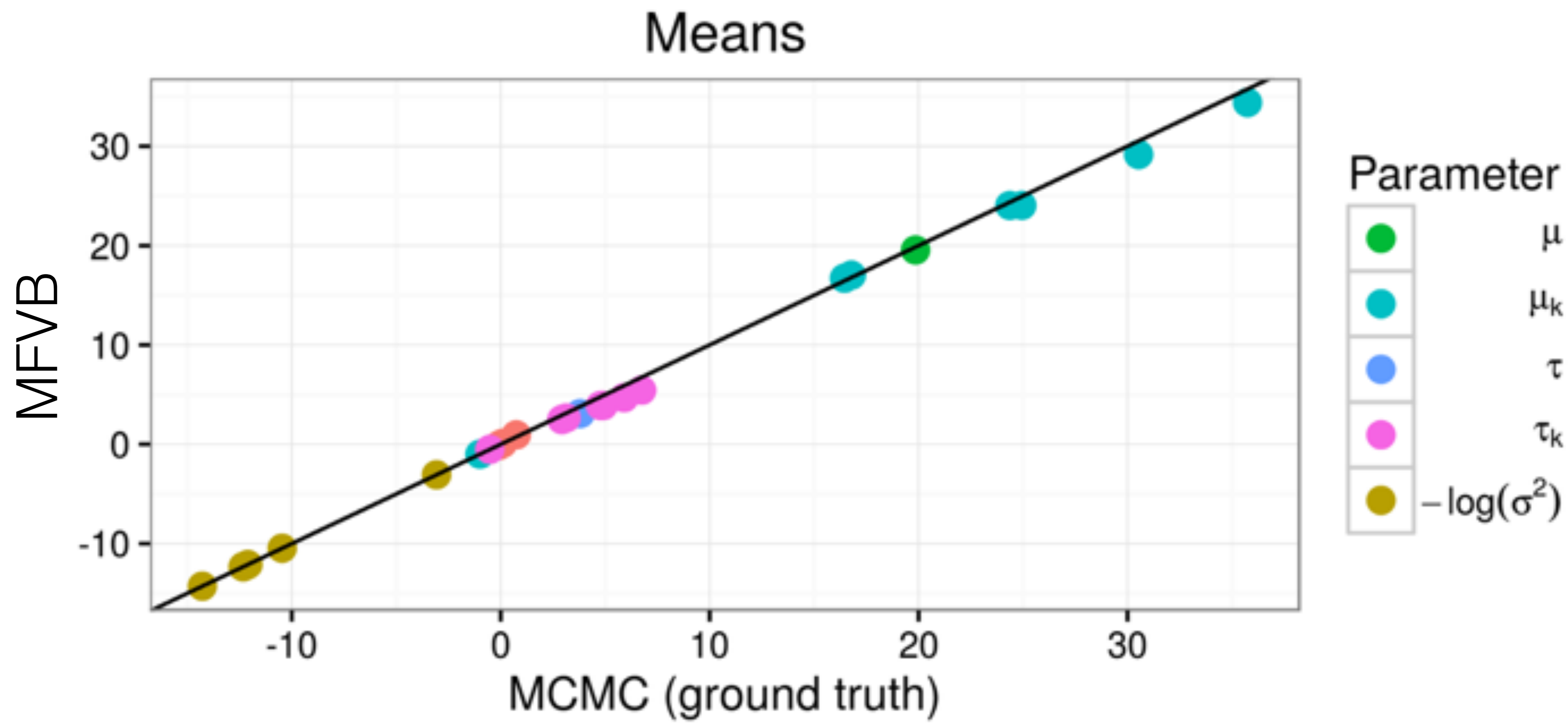
Criteo Online Ads Experiment

- 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**



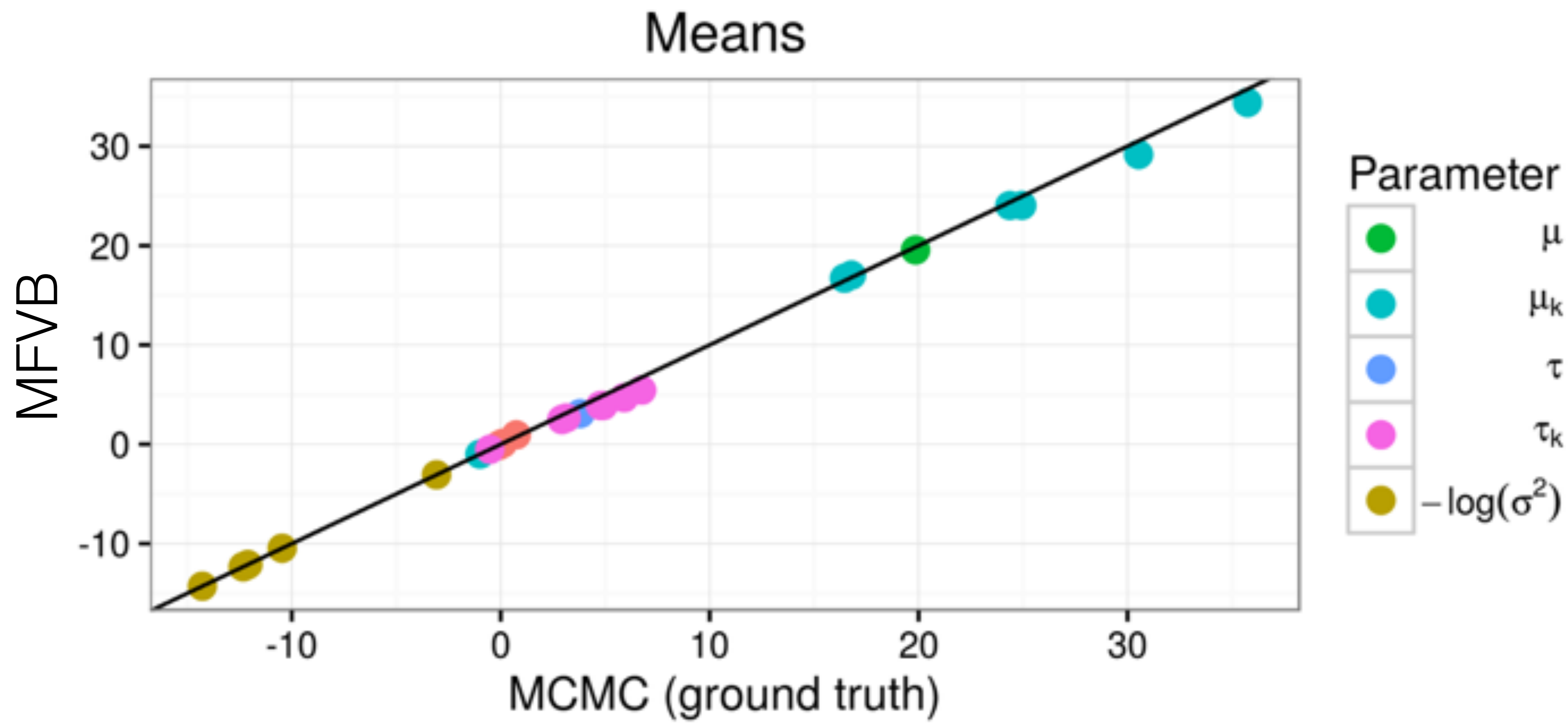
Criteo Online Ads Experiment

- 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 [board]

[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2017]

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 2017]

Criteo Online Ads Experiment

[Giordano, Broderick, Jordan 2017]

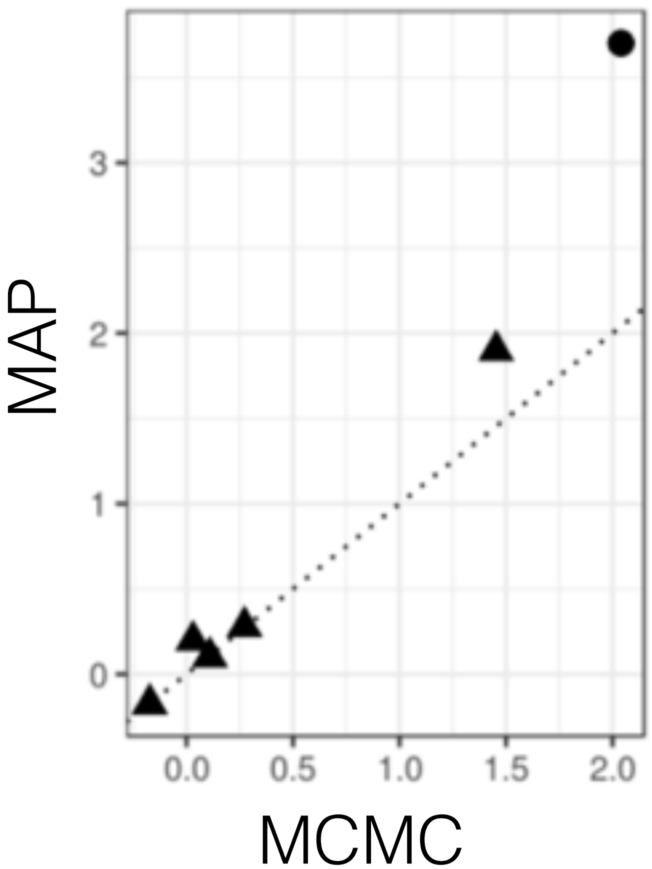
Criteo Online Ads Experiment

- MAP: **12 s**

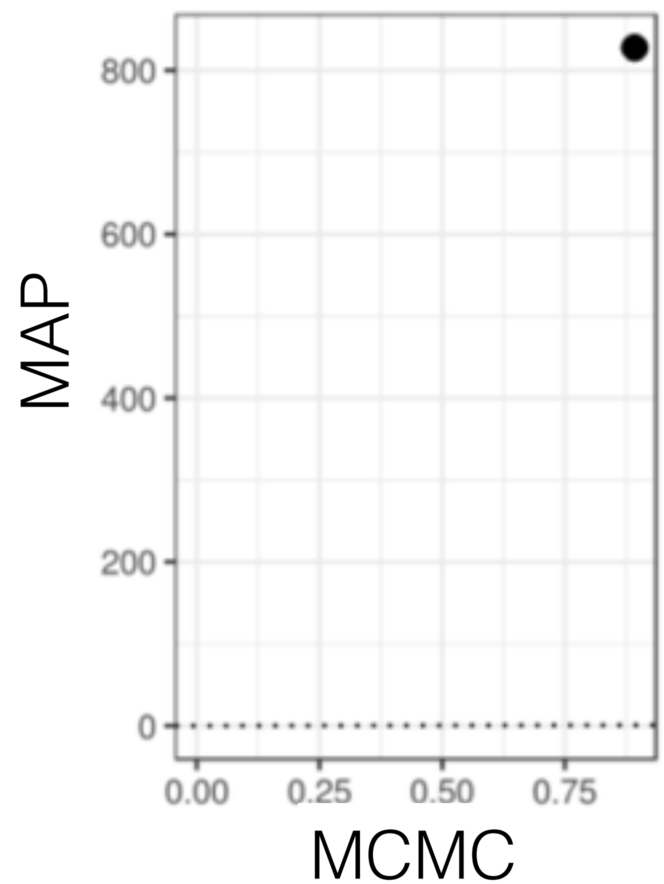
[Giordano, Broderick, Jordan 2017]

Criteo Online Ads Experiment

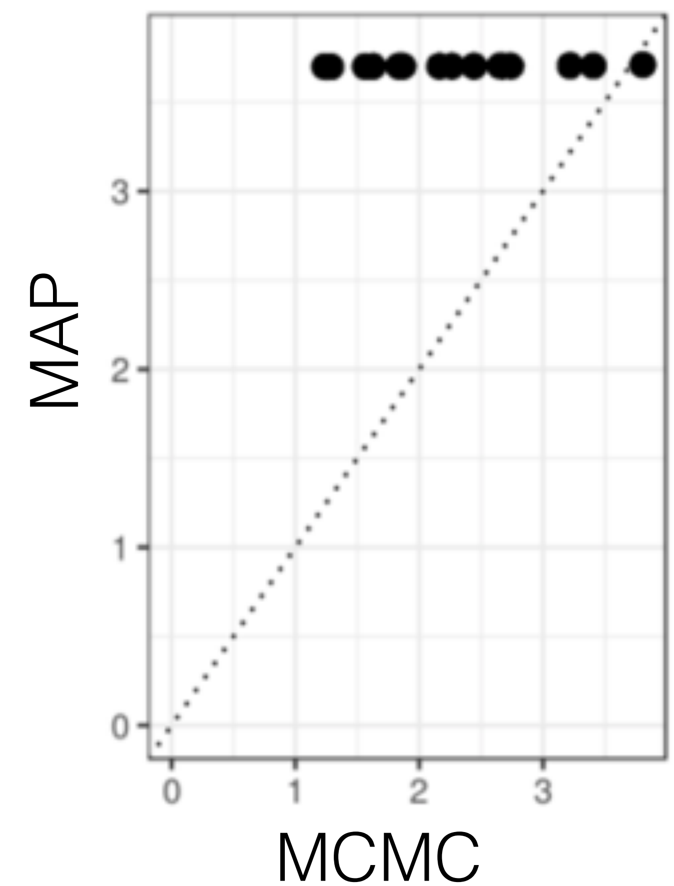
Global parameters (-τ)



Global parameter τ



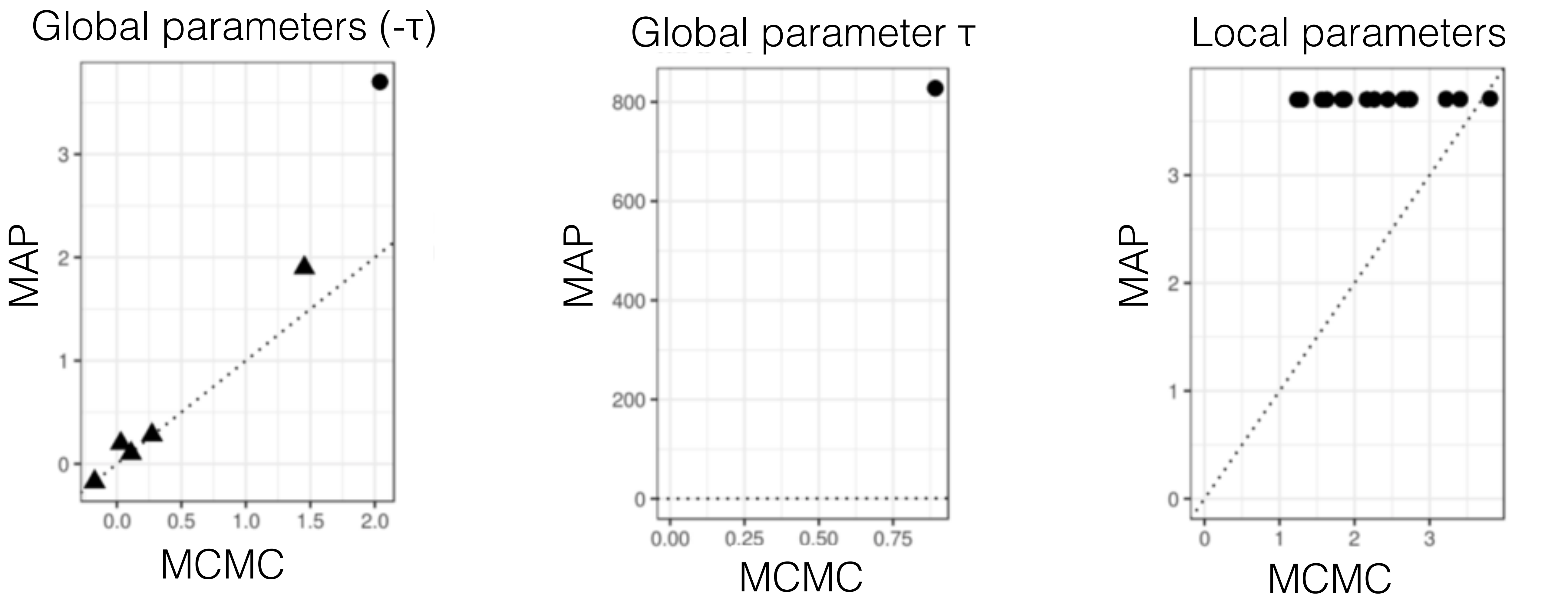
Local parameters



- MAP: **12 s**

[Giordano, Broderick, Jordan 2017]

Criteo Online Ads Experiment

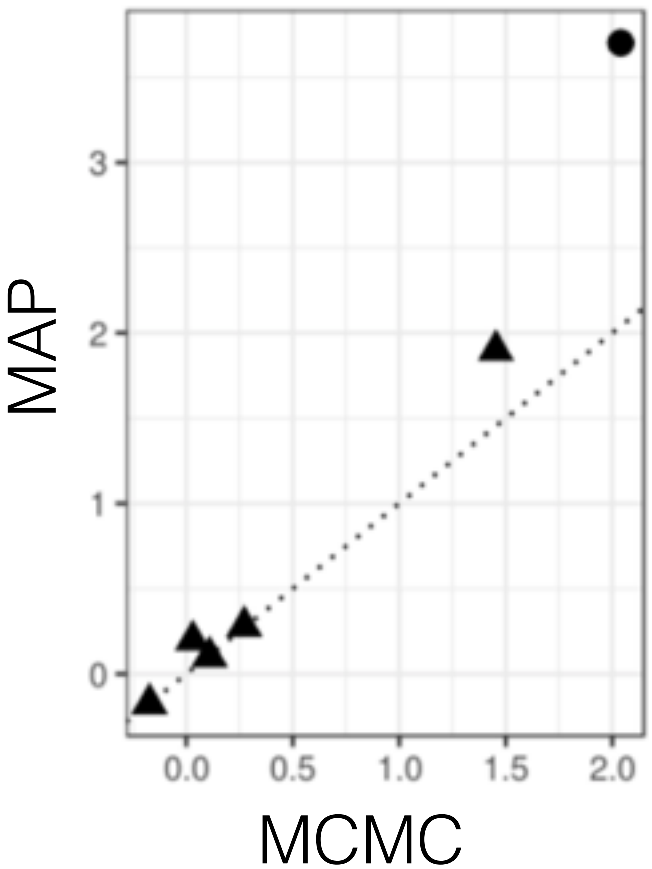


- MAP: **12 s**
- MFVB: **57 s**

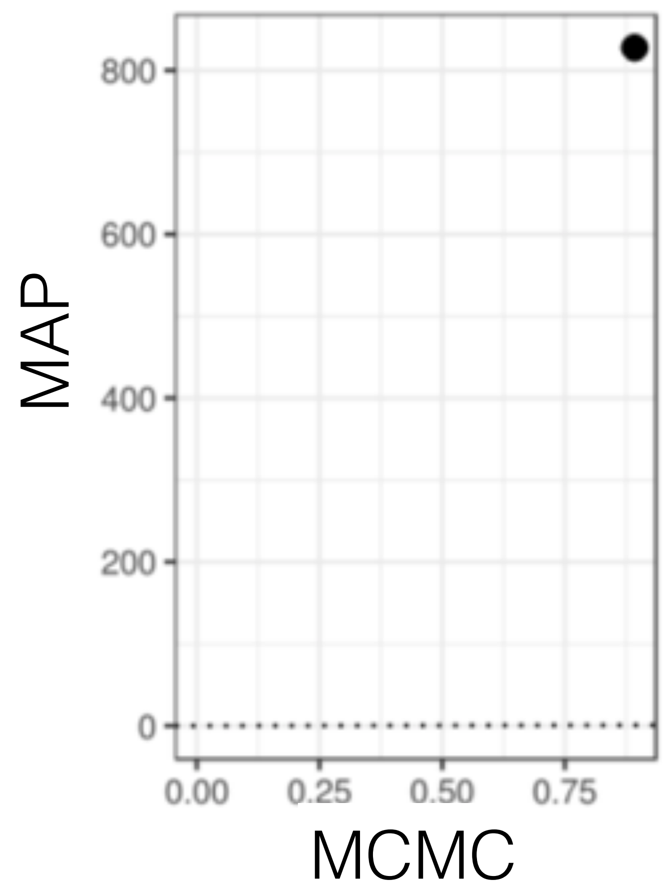
[Giordano, Broderick, Jordan 2017]

Criteo Online Ads Experiment

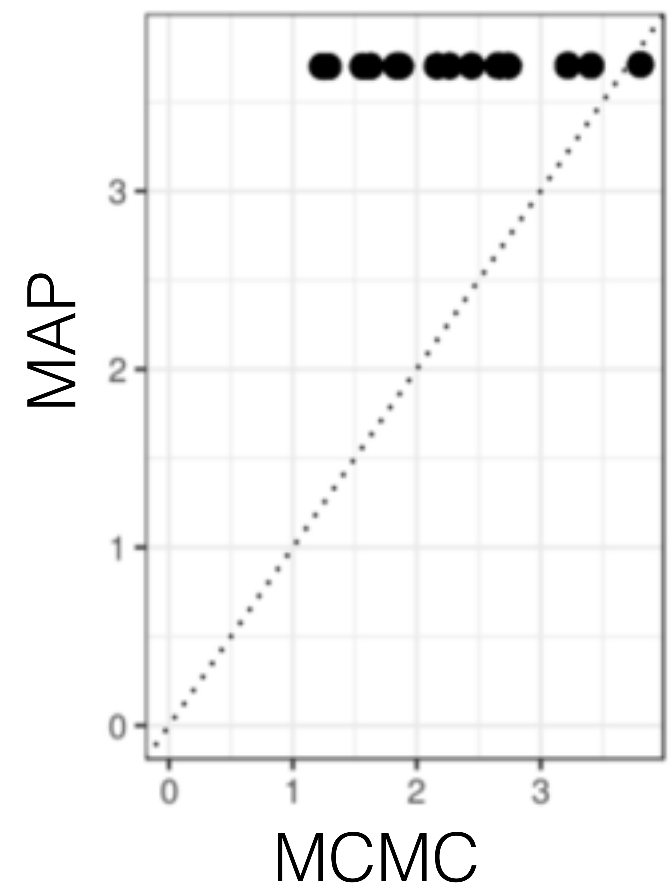
Global parameters (-τ)



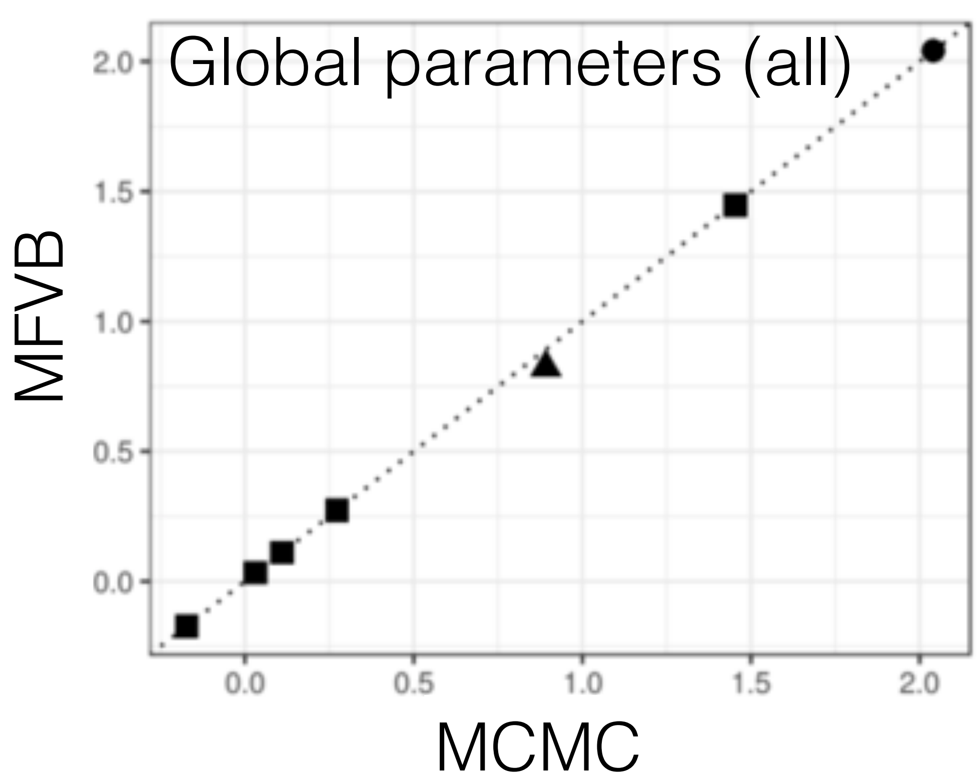
Global parameter τ

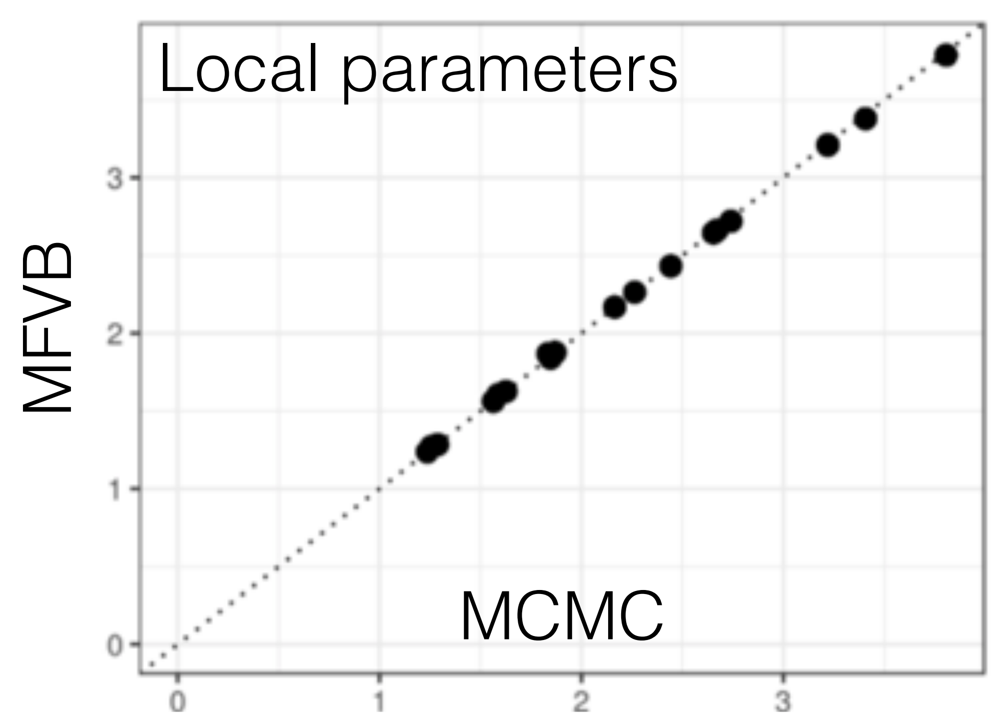


Local parameters



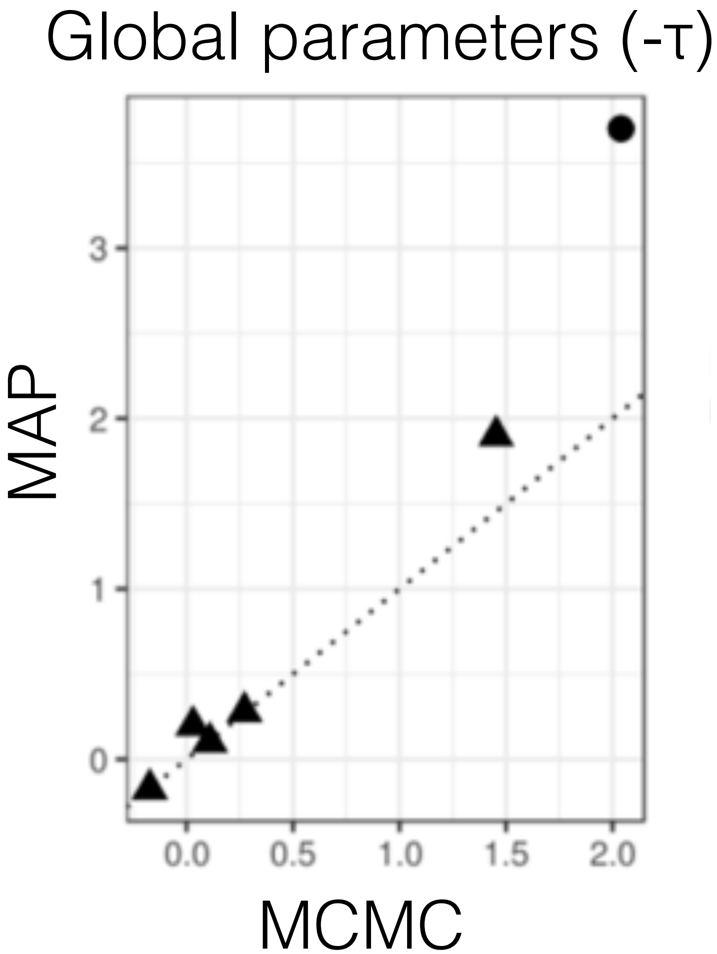
- MAP: **12 s**
- MFVB: **57 s**

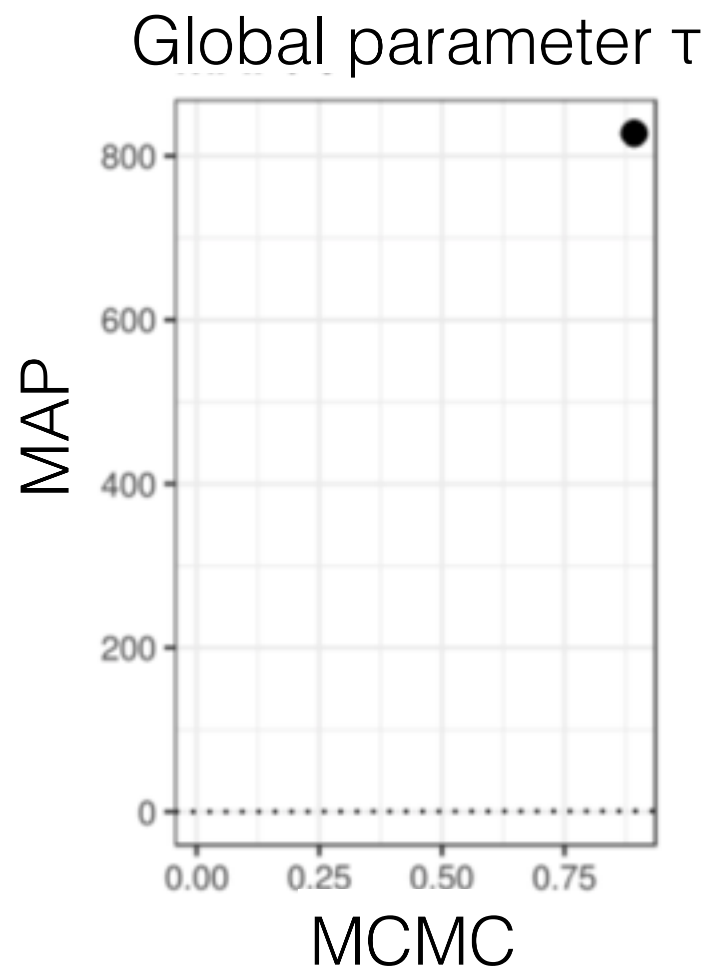


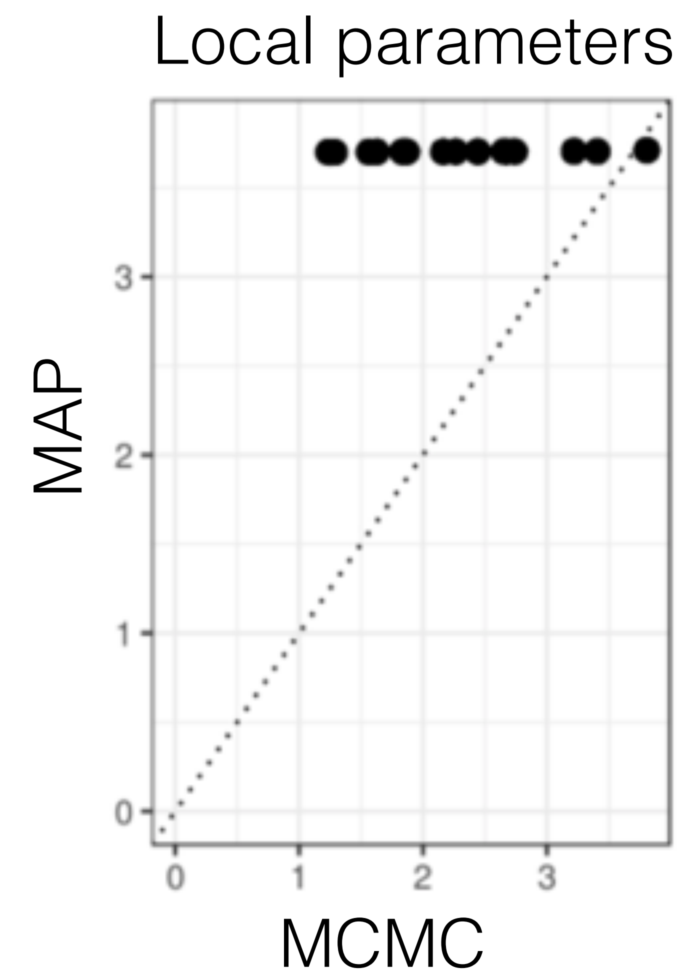


[Giordano, Broderick, Jordan 2017]

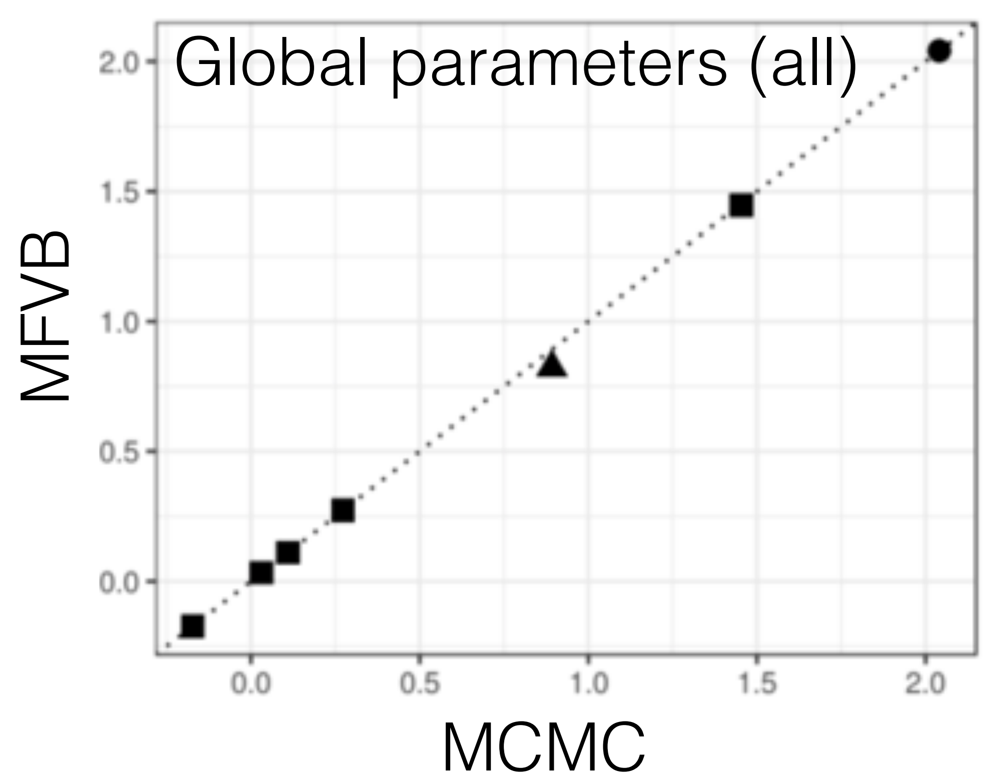
Criteo Online Ads Experiment

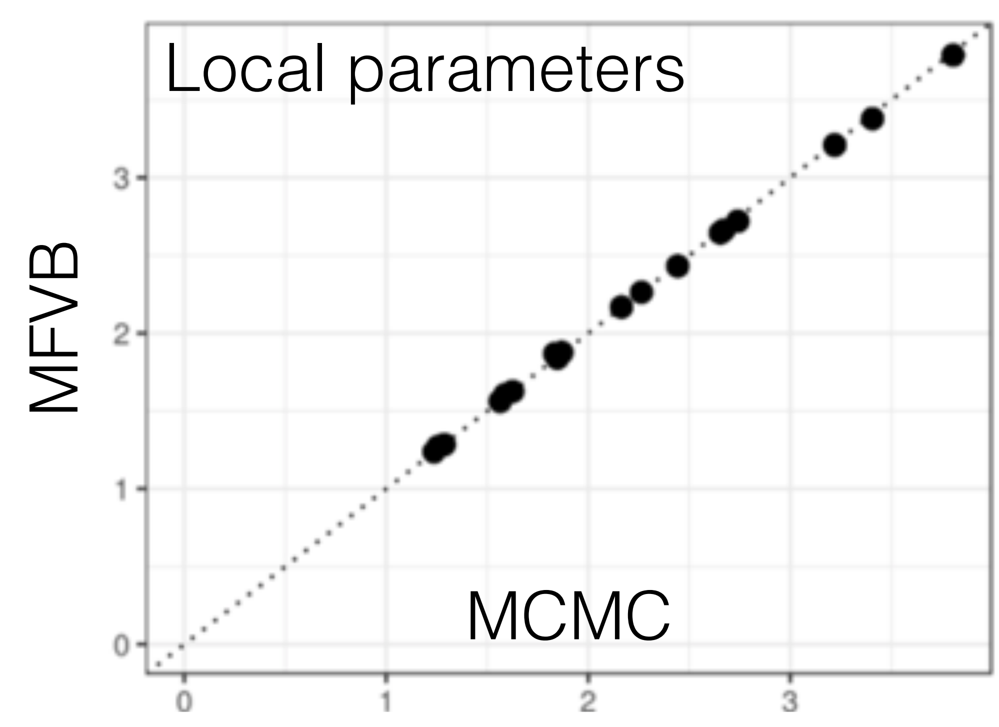






- MAP: **12 s**
- MFVB: **57 s**
- MCMC (5K samples): 21,066 s (**5.85 h**)





[Giordano, Broderick, Jordan 2017]

Roadmap

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

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA)

- Topics

Latent Dirichlet Allocation (LDA)

- Topics: two perspectives
  - Each document can belong to multiple groups
  - Cluster words in documents

Latent Dirichlet Allocation (LDA)

- Topics: two perspectives
  - Each document can belong to multiple groups
  - Cluster words in documents

"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

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, Ng, Jordan 2003]

Latent Dirichlet Allocation (LDA)

- Topics: two perspectives
  - Each document can belong to multiple groups
  - Cluster words in documents

"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

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, Ng, Jordan 2003; Pritchard, Stephens, Donnelly 2000]

Roadmap

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

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

- 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 MFVB?
- When can we trust MFVB?
- Where do we go from here?

Roadmap

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

References (1/6)

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

AG Baydin, BA Pearlmutter, AA Radul, and JM Siskind. "Automatic differentiation in machine learning: a survey." ArXiv:1502.05767v4 (2018).

DM Blei, A Kucukelbir, and JD McAuliffe. "Variational inference: A review for statisticians." *Journal of the American Statistical Association* 112.518 (2017): 859-877.

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

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

T Campbell and T Broderick. Automated scalable Bayesian inference via Hilbert coresets. Under review. ArXiv:1710.05053.

T Campbell and T Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. *ICML* 2018, to appear.

RJ 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, J Huggins, and MI Jordan. "Fast robustness quantification with variational Bayes." *ICML 2016 Workshop on #Data4Good: Machine Learning in Social Good Applications*, 2016.

R Giordano, T Broderick, and MI Jordan. Covariances, robustness, and variational Bayes, 2017. Under review. ArXiv:1709.02536.

References (2/6)

J Gorham and L Mackey. "Measuring sample quality with Stein's method." *NIPS* 2015.

J Gorham, and L Mackey. "Measuring sample quality with kernels." ArXiv:1703.01717 (2017).

PD Hoff. *A first course in Bayesian statistical methods*. Springer Science & Business Media, 2009.

MD Hoffman, DM Blei, C Wang, and J Paisley. "Stochastic variational inference." *The Journal of Machine Learning Research* 14.1 (2013): 1303-1347.

JH Huggins, T Campbell, and T Broderick. Coresets for scalable Bayesian logistic regression. *NIPS* 2016.

JH Huggins, RP Adams, and T Broderick. PASS-GLM: Polynomial approximate sufficient statistics for scalable Bayesian GLM inference. *NIPS* 2017.

JH Huggins, M Kasprzak, T Campbell, and T Broderick. Bayesian posterior mean and uncertainty estimates: a non-asymptotic approach. Forthcoming.

A Kucukelbir, R Ranganath, A Gelman, and D Blei. "Automatic variational inference in Stan." *NIPS* 2015.

A Kucukelbir, D Tran, R Ranganath, A Gelman, and DM Blei. "Automatic differentiation variational inference." *The Journal of Machine Learning Research* 18.1 (2017): 430-474.

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

Stan (open source software). http://mc-stan.org/ Accessed: 2018.

References (3/6)

S Talts, M Betancourt, D Simpson, A Vehtari, and A Gelman. "Validating Bayesian Inference Algorithms with Simulation-Based Calibration." aArXiv:1804.06788 (2018).

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.

Y Yao, A Vehtari, D Simpson, and A Gelman. "Yes, but Did It Work?: Evaluating Variational Inference." ArXiv:1802.02538 (2018).

Application References (4/6)

Abbott, Benjamin P., et al. "Observation of gravitational waves from a binary black hole merger." *Physical Review Letters* 116.6 (2016): 061102.

Abbott, Benjamin P., et al. "The rate of binary black hole mergers inferred from advanced LIGO observations surrounding GW150914." *The Astrophysical Journal Letters* 833.1 (2016): L1.

Airoldi, Edoardo M., David M. Blei, Stephen E. Fienberg, and Eric P. Xing. "Mixed membership stochastic blockmodels." *Journal of Machine Learning Research* 9.Sep (2008): 1981-2014.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation." *Journal of Machine Learning Research* 3.Jan (2003): 993-1022.

Chati, Yashovardhan Sushil, and Hamsa Balakrishnan. "A Gaussian process regression approach to model aircraft engine fuel flow rate." *Cyber-Physical Systems (ICCPS), 2017 ACM/IEEE 8th International Conference on*. IEEE, 2017.

Gershman, Samuel J., David M. Blei, Kenneth A. Norman, and Per B. Sederberg. "Decomposing spatiotemporal brain patterns into topographic latent sources." NeuroImage 98 (2014): 91-102.

Gillon, Michaël, et al. "Seven temperate terrestrial planets around the nearby ultracool dwarf star TRAPPIST-1." *Nature* 542.7642 (2017): 456.

Grimm, Simon L., et al. "The nature of the TRAPPIST-1 exoplanets." *Astronomy & Astrophysics* 613 (2018): A68.

Application References (5/6)

Grogan Jr, William L., and Willis W. Wirth. "A new American genus of predaceous midges related to Palpomyia and Bezzia (Diptera: Ceratopogonidae). Un nuevo género Americano de purrujas depredadoras relacionadas con Palpomyia y Bezzia (Diptera: Ceratopogonidae)." *Proceedings of the Biological Society of Washington*. 94.4 (1981): 1279-1305.

Kuikka, Sakari, Jarno Vanhatalo, Henni Pulkkinen, Samu Mäntyniemi, and Jukka Corander. "Experiences in Bayesian inference in Baltic salmon management." *Statistical Science* 29.1 (2014): 42-49.

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

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

Stegle, Oliver, Leopold Parts, Richard Durbin, and John Winn. "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies." PLoS computational biology 6.5 (2010): e1000770.

Stone, Lawrence D., Colleen M. Keller, Thomas M. Kratzke, and Johan P. Strumpfer. "Search for the wreckage of Air France Flight AF 447." *Statistical Science* (2014): 69-80.

Woodard, Dawn, Galina Nogin, Paul Koch, David Racz, Moises Goldszmidt, and Eric Horvitz. "Predicting travel time reliability using mobile phone GPS data." *Transportation Research Part C: Emerging Technologies* 75 (2017): 30-44.

Xing, Eric P., Wei Wu, Michael I. Jordan, and Richard M. Karp. "LOGOS: a modular Bayesian model for de novo motif detection." Journal of Bioinformatics and Computational Biology 2.01 (2004): 127-154.

Additional image references (6/6)

amCharts. Visited Countries Map. https://www.amcharts.com/visited_countries/ Accessed: 2016.

Baltic Salmon Fund. https://www.en.balticsalmonfund.org/about_us Accessed: 2018.

ESO/L. Calçada/M. Kornmesser. 16 October 2017, 16:00:00. Obtained from: https://commons.wikimedia.org/wiki/File:Artist%E2%80%99s_impression_of_merging_neutron_stars.jpg || Source: https://www.eso.org/public/images/eso1733a/ (Creative Commons Attribution 4.0 International License)

J. Herzog. 3 June 2016, 17:17:30. Obtained from: https://commons.wikimedia.org/wiki/File:Airbus_A350-941_F-WWCF_MSN002_ILA_Berlin_2016_17.jpg (Creative Commons Attribution 4.0 International License)

E. Xing. 2003. Slides "LOGOS: a modular Bayesian model for de novo motif detection." Obtained from: https://www.cs.cmu.edu/~epxing/papers/Old_papers/slide_CSB03/CSB1.pdf Accessed: 2018.