Fast Quantification of Uncertainty and Robustness with Variational Bayes

Tamara Broderick
ITT Career Development Assistant Professor, MIT

With: Ryan Giordano, Rachael Meager, Jonathan H. Huggins, Michael I. Jordan
• Bayesian inference
• Bayesian inference
  • Complex, modular models
• Bayesian inference
  • Complex, modular models; posterior distribution
• Bayesian inference
  • Complex, modular models; posterior distribution

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

• Complex, modular models; posterior distribution

• Challenge: Express prior beliefs in a distribution

• Time-consuming; subjective

Some reasonable priors
• Bayesian inference \[ p(\theta|x) \propto p(x|\theta)p(\theta) \]
• Complex, modular models; posterior distribution
• Challenge: Express prior beliefs in a distribution
  • Time-consuming; subjective

Some reasonable priors

Bayes Theorem
• Bayesian inference \[ p(\theta|x) \propto \theta \frac{p(x|\theta)p(\theta)}{\theta} \]
• Complex, modular models; posterior distribution
• Challenge: Express prior beliefs in a distribution
• Time-consuming; subjective

Some reasonable priors

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

Bayes Theorem

Some reasonable priors
robustness quantification

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

Some reasonable priors

Bayes Theorem
robustness quantification

- Bayesian inference
  \[ p(θ|x) \propto p(x|θ)p(θ) \]
- Complex, modular models; posterior distribution
- Challenge: Express prior beliefs in a distribution
  - Time-consuming; subjective; complex models
- Challenge: Approximating the posterior can be computationally expensive
robustness quantification

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

- Challenge: Express prior beliefs in a distribution
  - Time-consuming; subjective; complex models
- Challenge: Approximating the posterior can be computationally expensive
Uncertainty & robustness quantification

- Bayesian inference \[ p(\theta|x) \propto \theta \cdot p(x|\theta)p(\theta) \]
- Complex, modular models; posterior distribution

- Challenge: Express prior beliefs in a distribution
  - Time-consuming; subjective; complex models

- Challenge: Approximating the posterior can be computationally expensive

Bayes Theorem

Variational Bayes
Uncertainty & robustness quantification

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

- Challenge: Express prior beliefs in a distribution
  - Time-consuming; subjective; complex models

- Challenge: Approximating the posterior can be computationally expensive

- We propose: linear response variational Bayes
Uncertainty & robustness quantification

• Bayesian inference
  \[ p(\theta|x) \propto \theta \cdot p(x|\theta)p(\theta) \]
  
  • Complex, modular models; posterior distribution

• Challenge: Express prior beliefs in a distribution
  
  • Time-consuming; subjective; complex models

• Challenge: Approximating the posterior can be computationally expensive

• We propose: linear response variational Bayes

[see also Opper, Winther 2003]
Roadmap
Roadmap

• Variational Bayes as an alternative to MCMC
Roadmap

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

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

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

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

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

• Variational Bayes (VB)
• Approximation for posterior
• Minimize Kullback-Liebler (KL) divergence:
  \[ p(\theta | x) = \arg\min_{q(\theta)} KL(q(\theta) || p(\theta | x)) \]
• VB practical success
  • point estimates and prediction
  • fast
Variational Bayes

• VB approximation
Variational Bayes

• VB approximation
Variational Bayes

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

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

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

$p(\theta|x)$

$q(\theta)$
Variational Bayes

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

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

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

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

- VB approximation
- Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
- Minimize Kullback-Leibler (KL) divergence:
  \[ KL(q||p(\cdot|x)) \]
- VB practical success
Variational Bayes

- VB approximation
- Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
- Minimize Kullback-Leibler (KL) divergence:
  \[ KL(q\|p(\cdot|x)) \]

- VB practical success
  - point estimates and prediction
Variational Bayes

- VB approximation
- Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
- Minimize Kullback-Leibler (KL) divergence:
  $$KL(q||p(\cdot|x))$$

- VB practical success
  - point estimates and prediction
  - fast

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

- VB approximation
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - Minimize Kullback-Leibler (KL) divergence:
    $$KL(q\|p(\cdot|x))$$

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

[Broderick, Boyd, Wibisono, Wilson, Jordan 2013]
What about uncertainty?
What about uncertainty?

- Variational Bayes
What about uncertainty?

- Variational Bayes

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

\[
p(\theta|x)
\]

[Bishop 2006]
What about uncertainty?

- Variational Bayes

\[ KL(q||p(\cdot|x)) = \int_{\theta} q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta \]
What about uncertainty?

- Variational Bayes
  \[
  KL(q||p(\cdot|x)) = \int q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta
  \]

- Mean-field variational Bayes (MFVB)
  \[
  q(\theta) = \prod_{j=1}^{J} q(\theta_j)
  \]
What about uncertainty?

- Variational Bayes

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

- Mean-field variational Bayes (MFVB)

\[ q(\theta) = \prod_{j=1}^{J} q(\theta_j) \]
What about uncertainty?

- Variational Bayes
  \[ KL(q||p(\cdot|x)) = \int q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta \]

- Mean-field variational Bayes (MFVB)
  \[ q(\theta) = \prod_{j=1}^{J} q(\theta_j) \]

- Underestimates variance (sometimes severely)

[Bishop 2006]
What about uncertainty?

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

- Mean-field variational Bayes (MFVB)
  
  $$q(\theta) = \prod_{j=1}^{J} q(\theta_j)$$

- Underestimates variance (sometimes severely)

- No covariance estimates
What about uncertainty?

• Variational Bayes

\[ KL(q||p(\cdot|x)) = \int \theta \log \frac{q(\theta)}{p(\theta|x)} d\theta \]

• Mean-field variational Bayes (MFVB)

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

• Underestimates variance (sometimes severely)

• No covariance estimates

[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011]
What about uncertainty?

- Variational Bayes
  \[ KL(q\|p(\cdot|x)) = \int_{\theta} q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta \]

- Mean-field variational Bayes (MFVB)
  \[ q(\theta) = \prod_{j=1}^{J} q(\theta_j) \]

- Underestimates variance (sometimes severely)

- No covariance estimates

[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011]
[Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2015]
Linear response
Linear response

- Cumulant-generating function

\[ C(t) = \log \mathbb{E}_t \mathbb{E} \mathbf{x}^T \mathbf{x} | \mathbf{x} \]

[see also Opper, Winther 2003]
Linear response

- Cumulant-generating function

\[ C(t) := \log \mathbb{E} e^{t^T \theta} \]
Linear response

• Cumulant-generating function

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

mean \[= \left. \frac{d}{dt} C(t) \right|_{t=0} \]
Linear response

- Cumulant-generating function
  \[ C(t) := \log \mathbb{E} e^{t^T \theta} \]
  \[ \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0} \]

- Exact posterior covariance

[see also Opper, Winther 2003]
Linear response

- Cumulant-generating function

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

- Exact posterior covariance

\[ \Sigma := \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \bigg|_{t=0} \]
Linear response

- Cumulant-generating function
  \[ C(t) := \log \mathbb{E} e^{t^T \theta} \]
  \[ \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0} \]

- Exact posterior covariance vs MFVB covariance
  \[ \Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \]

[see also Opper, Winther 2003]
Linear response

- Cumulant-generating function

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

- Exact posterior covariance vs MFVB covariance

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

[see also Opper, Winther 2003]
Linear response

• Cumulant-generating function

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

\[ \text{mean} = \frac{d}{dt} C(t) \bigg|_{t=0} \]

• Exact posterior covariance vs MFVB covariance

\[ \Sigma := \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \bigg|_{t=0} \]

\[ V := \frac{d^2}{dt^T dt} C_{q^*}(t) \bigg|_{t=0} \]

• “Linear response”
Linear response

• Cumulant-generating function

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

\[ \text{mean} = \frac{d}{dt} C(t) \bigg|_{t=0} \]

• Exact posterior covariance vs MFVB covariance

\[ \Sigma := \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \bigg|_{t=0} \]

\[ V := \frac{d^2}{dt^T dt} C_{q^*}(t) \bigg|_{t=0} \]

• “Linear response”

\[ \log p(\theta|x) \]

[see also Opper, Winther 2003]

[Bishop 2006]
Linear response

- Cumulant-generating function
  \[ C(t) := \log \mathbb{E} e^{t^T \theta} \]
  \[ \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0} \]

- Exact posterior covariance vs MFVB covariance
  \[ \Sigma := \left. \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \right|_{t=0} \]
  \[ V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0} \]

- “Linear response”
  \[ \log p(\theta|x) + t^T \theta \]

[see also Opper, Winther 2003]
Linear response

- Cumulant-generating function

\[ C(t) := \log \mathbb{E} e^{t^T \theta} \quad \text{mean} = \frac{d}{dt} C(t) \bigg|_{t=0} \]

- Exact posterior covariance vs MFVB covariance

\[ \Sigma := \frac{d^2}{dt^T dt} C_{\mathbb{P}(\cdot|x)}(t) \bigg|_{t=0} \quad V := \frac{d^2}{dt^T dt} C_{q^*}(t) \bigg|_{t=0} \]

- “Linear response”

\[ \log p(\theta|x) + t^T \theta \]
Linear response

- Cumulant-generating function

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

\[ \text{mean} = \frac{d}{dt} C(t) \bigg|_{t=0} \]

- Exact posterior covariance vs MFVB covariance

\[ \Sigma := \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \bigg|_{t=0} \]

\[ V := \frac{d^2}{dt^T dt} C_{q^*}(t) \bigg|_{t=0} \]

- “Linear response”

\[ \log p_t(\theta) := \log p(\theta|x) + t^T \theta \]
Linear response

- **Cumulant-generating function**

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

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

- **Exact posterior covariance vs MFVB covariance**

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

\[ V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0} \]

- **“Linear response”**

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

[see also Opper, Winther 2003]

[Bishop 2006]
Linear response

- Cumulant-generating function

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

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

- Exact posterior covariance vs MFVB covariance

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

\[ V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0} \]

- “Linear response”

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

[Bishop 2006]

[see also Opper, Winther 2003]
Linear response

- Cumulant-generating function

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

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

- Exact posterior covariance vs MFVB covariance

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

\[ V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0} \]

- "Linear response"

\[ \log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t), \text{ MFVB } q_t^* \]

[see also Opper, Winther 2003]
Linear response

- Cumulant-generating function
  \[ C(t) := \log \mathbb{E} e^{t^T \theta} \]
  \[ \text{mean} = \left. \frac{d}{dt} C(t) \right|_{t=0} \]

- Exact posterior covariance vs MFVB covariance
  \[ \Sigma := \left. \frac{d^2}{dt^T dt} C_p(\cdot|x)(t) \right|_{t=0} \]
  \[ V := \left. \frac{d^2}{dt^T dt} C_{q^*}(t) \right|_{t=0} \]

- “Linear response”
  \[ \log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t), \text{MFVB } q_t^* \]

- The LRVB approximation

[see also Opper, Winther 2003]
Linear response

- Cumulant-generating function

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

- Exact posterior covariance vs MFVB covariance

\[ \Sigma := \left. \frac{d^2}{dtT dt} C_p(\cdot|x)(t) \right|_{t=0} \quad V := \left. \frac{d^2}{dtT dt} C_{q^*}(t) \right|_{t=0} \]

- "Linear response"

\[ \log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t), \text{ MFVB } q_t^* \]

- The LRVB approximation

\[ \Sigma = \left. \frac{d}{dtT} \left[ \frac{d}{dt} C_p(\cdot|x)(t) \right] \right|_{t=0} \]

[see also Opper, Winther 2003]
Linear response

• Cumulant-generating function

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

\[ \text{mean} = \frac{d}{dt} C(t) \Bigg|_{t=0} \]

• Exact posterior covariance vs MFVB covariance

\[ \Sigma := \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \Bigg|_{t=0} \]

\[ V := \frac{d^2}{dt^T dt} C_{q^*}(t) \Bigg|_{t=0} \]

• “Linear response”

\[ \log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t), \quad \text{MFVB } q^*_t \]

• The LRVB approximation

\[ \Sigma = \frac{d}{dt^T} \mathbb{E}_{\theta} \theta \Bigg|_{t=0} \]

[Bishop 2006]

[see also Opper, Winther 2003]
Linear response

- Cumulant-generating function
  \[ C(t) := \log \mathbb{E} e^{t^T \theta} \quad \text{mean} = \frac{d}{dt} C(t) \bigg|_{t=0} \]

- Exact posterior covariance vs MFVB covariance
  \[ \Sigma := \frac{d^2}{dt^T dt} C_{p(\cdot | x)}(t) \bigg|_{t=0} \quad V := \frac{d^2}{dt^T dt} C_{q^*}(t) \bigg|_{t=0} \]

- “Linear response”
  \[ \log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t), \text{ MFVB } q^*_t \]

- The LRVB approximation
  \[ \Sigma = \frac{d}{dt^T} \mathbb{E}_{p_t} \theta \bigg|_{t=0} \]
Linear response

- Cumulant-generating function
  \[ C(t) := \log \mathbb{E} e^{t^T \theta} \]
  \[ \text{mean} = \frac{d}{dt} C(t) \bigg|_{t=0} \]

- Exact posterior covariance vs MFVB covariance
  \[ \Sigma := \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \bigg|_{t=0} \]
  \[ V := \frac{d^2}{dt^T dt} C_{q^*}(t) \bigg|_{t=0} \]

- “Linear response”
  \[ \log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t), \text{ MFVB } q_t^* \]

- The LRVB approximation
  \[ \Sigma = \frac{d}{dt^T} \mathbb{E}_{p_t} \theta \bigg|_{t=0} \]

[see also Opper, Winther 2003]
Linear response

- Cumulant-generating function

\[ C(t) := \log \mathbb{E} e^{t^T \theta} \quad \text{mean} = \frac{d}{dt} C(t) \bigg|_{t=0} \]

- Exact posterior covariance vs MFVB covariance

\[ \Sigma := \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \bigg|_{t=0} \quad \text{V} := \frac{d^2}{dt^T dt} C_{q^*}(t) \bigg|_{t=0} \]

- “Linear response”

\[ \log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t), \text{ MFVB } q^*_t \]

- The LRVB approximation

\[ \Sigma = \frac{d}{dt^T} \mathbb{E}_{p_t} \theta \bigg|_{t=0} \approx \frac{d}{dt^T} \mathbb{E}_{q^*_t} \theta \bigg|_{t=0} \]

[see also Opper, Winther 2003]

[Bishop 2006]
Linear response

- Cumulant-generating function

\[ C(t) := \log \mathbb{E} e^{t^T \theta} \quad \text{mean} = \frac{d}{dt} C(t) \bigg|_{t=0} \]

- Exact posterior covariance vs MFVB covariance

\[ \Sigma := \frac{d^2}{dt^T dt} C_{p(\cdot|x)}(t) \bigg|_{t=0} \quad V := \frac{d^2}{dt^T dt} C_{q^*(t)}(t) \bigg|_{t=0} \]

- “Linear response”

\[ \log p_t(\theta) := \log p(\theta|x) + t^T \theta - C(t), \text{ MFVB } q_t^* \]

- The LRVB approximation

\[ \Sigma = \frac{d}{dt^T} \mathbb{E}_{p_t} \theta \bigg|_{t=0} \approx \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \bigg|_{t=0} =: \hat{\Sigma} \]

[see also Opper, Winther 2003]
LRVB estimator

- LRVB covariance estimate $\hat{\Sigma} := \frac{dT}{dt} \mathbb{E}_{q_t^*} \theta \bigg|_{t=0}$
LRVB estimator

- LRVB covariance estimate: \( \hat{\Sigma} \equiv \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \bigg|_{t=0} \)

- Suppose \( q_t \) exponential family
LRVB estimator

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

- LRVB covariance estimate $\hat{\Sigma} := \frac{d}{dtT} \mathbb{E}_{q_t^*} \theta \bigg|_{t=0}$

- Suppose $q_t$ exponential family with mean parametrization $m_t$

\[ \hat{\Sigma} = \]
LRVB estimator

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

\[
\hat{\Sigma} = \left( \frac{\partial^2 KL}{\partial m \partial m^T} \bigg|_{m=m^*} \right)^{-1}
\]
LRVB estimator

• LRVB covariance estimate \( \hat{\Sigma} := \frac{d}{dtT} \mathbb{E}_{q_t^*} \theta \bigg|_{t=0} \)

• Suppose \( q_t \) exponential family with mean parametrization \( m_t \)

\[
\hat{\Sigma} = \left( \frac{\partial^2 KL}{\partial m \partial m^T} \bigg|_{m=m^*} \right)^{-1} = (I - VH)^{-1}V
\]
LRVB estimator

- LRVB covariance estimate:  
  \[ \hat{\Sigma} := \frac{d}{dt^T} \mathbb{E}_{q_t} \theta \bigg|_{t=0} \]

- Suppose \( q_t \) exponential family with mean parametrization \( m_t \)

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

- Symmetric and positive definite at local min of KL
LRVB estimator

• LRVB covariance estimate \( \hat{\Sigma} := \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \bigg|_{t=0} \)

• Suppose \( q_t \) exponential family with mean parametrization \( m_t \)

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

• Symmetric and positive definite at local min of KL

• The LRVB assumption: \( \mathbb{E}_{p_t} \theta \approx \mathbb{E}_{q_t^*} \theta \)
LRVB estimator

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

- Suppose \( q_t \) exponential family with mean parametrization \( m_t \)

\[
\hat{\Sigma} = \left( \frac{\partial^2 K L}{\partial m \partial m^T} \right)_{m=m^*}^{-1} = (I - VH)^{-1} V
\]

- Symmetric and positive definite at local min of KL

- The LRVB assumption: \( \mathbb{E}_{p_t} \theta \approx \mathbb{E}_{q_t^*} \theta \)

[Bishop 2006]
LRVB estimator

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

- Suppose $q_t$ exponential family with mean parametrization $m_t$

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

- Symmetric and positive definite at local min of KL

- The LRVB assumption: $\mathbb{E}_{p_t} \theta \approx \mathbb{E}_{q_t^*} \theta$

- LRVB estimate is exact when MFVB gives exact mean (e.g. multivariate normal)

[Bishop 2006]
Microcredit Experiment

• Simplified from Meager (2015)

• $k$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)

• $N_k$ businesses in $k$th site (~900 to ~17K)

• Profit of $n$th business at $k$th site:

• Priors and hyperpriors:
Microcredit Experiment

• Simplified from Meager (2016)
Microcredit Experiment

• Simplified from Meager (2016)
• $K$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
Microcredit Experiment

- Simplified from Meager (2016)
- $K$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site (~900 to ~17K)
Microcredit Experiment

- Simplified from Meager (2016)
- $K$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site (~900 to ~17K)
- Profit of $n$th business at $k$th site:
Microcredit Experiment

- Simplified from Meager (2016)
- $K$ 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:

$$y_{kn}$$
Microcredit Experiment

- Simplified from Meager (2016)
- $K$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site (~900 to ~17K)
- Profit of $n$th business at $k$th site:

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

• Simplified from Meager (2016)
• $K$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
• $N_k$ businesses in $k$th site (~900 to ~17K)
• Profit of $n$th business at $k$th site:

\[ y_{kn} \sim \text{indep} \mathcal{N}(\mu_k, \sigma_k^2) \]
Microcredit Experiment

- Simplified from Meager (2016)
- $K$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site (~900 to ~17K)
- Profit of $n$th business at $k$th site:

profit

\[
y_{kn} \overset{\text{indep}}{\sim} \mathcal{N}(\mu_k + T_{kn} \tau_k, )
\]
Microcredit Experiment

- Simplified from Meager (2016)
- \( K \) microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- \( N_k \) businesses in \( k \)th site (~900 to ~17K)
- Profit of \( n \)th business at \( k \)th site:

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

- Simplified from Meager (2016)
- $K$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site (~900 to ~17K)
- Profit of $n$th business at $k$th site:

\[
y_{kn} \sim \mathcal{N} \left( \mu_k + T_{kn} \tau_k, \sigma^2 \right)
\]
Microcredit Experiment

- Simplified from Meager (2016)
- $K$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site (~900 to ~17K)
- Profit of $n$th business at $k$th site:

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

profit \quad 1 \text{ if microcredit}
Microcredit Experiment

• Simplified from Meager (2016)
• $K$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
• $N_k$ businesses in $k$th site (~900 to ~17K)
• Profit of $n$th business at $k$th site:

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

1 if microcredit
Microcredit Experiment

- Simplified from Meager (2016)
- $K$ 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:
  
  $$ y_{kn} \sim \text{iid} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2) $$

- Priors and hyperpriors:

  $1$ if microcredit

profit

$y_{kn}$
Microcredit Experiment

- Simplified from Meager (2016)
- \( K \) microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- \( N_k \) businesses in \( k \)th site (~900 to ~17K)
- Profit of \( n \)th business at \( k \)th site:

\[
y_{kn} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)
\]

- Priors and hyperpriors:

\[
\begin{pmatrix}
\mu_k \\
\tau_k
\end{pmatrix} \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{pmatrix}
\mu \\
\tau
\end{pmatrix}, C\right)
\]
Microcredit Experiment

- Simplified from Meager (2016)
- $K$ 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:
  \[
  y_{kn} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_k + T_{kn} \tau_k, \sigma_k^2)
  \]
- Priors and hyperpriors:
  \[
  \begin{pmatrix}
  \mu_k \\
  \tau_k
  \end{pmatrix}
  \overset{\text{iid}}{\sim} \mathcal{N}
  \left(
  \begin{pmatrix}
  \mu \\
  \tau
  \end{pmatrix}, C
  \right)
  \]
  \[
  \sigma_k^{-2} \overset{iid}{\sim} \Gamma(a, b)
  \]
Microcredit Experiment

- Simplified from Meager (2016)
- $K$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site (~900 to ~17K)
- Profit of $n$th business at $k$th site:

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

- Priors and hyperpriors:

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

$\sigma_k^{-2} \overset{iid}{\sim} \Gamma(a, b) \quad C \sim \text{SepLKJ}(\eta, c, d)$
Microcredit Experiment
Microcredit Experiment

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

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

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

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

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

![Means](image1)

![Standard deviations](image2)
Microcredit Experiment

- *One set* of 2500 MCMC draws: **45 minutes**
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures: **58 seconds**
- $\tau$ mean (MFVB): 3.08 USD PPP
- $\tau$ std dev (LRVB): 1.83 USD PPP
Microcredit Experiment

- **One set of 2500 MCMC draws:**
  - 45 minutes
- All of MFVB optimization, LRVB uncertainties, all sensitivity measures:
  - 58 seconds
- $\tau$ mean (MFVB): 3.08 USD PPP
- $\tau$ std dev (LRVB): 1.83 USD PPP
- Mean is 1.68 std dev from 0
Experiments
Experiments

- Gaussian mixture model

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

with conjugate priors on \( \pi, \mu, \Lambda \)
Experiments

• Gaussian mixture model

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

with conjugate priors on \( \pi, \mu, \Lambda \)

• 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R \texttt{bayesm} package
Experiments

- Gaussian mixture model
  \[ P(z_{nk} = 1) = \pi_k, \quad p(x | \pi, \mu, \Lambda, z) = \prod_{n=1:N} \prod_{k=1:K} N(x_n | \mu_k, \Lambda_k^{-1})^{z_{nk}} \]
  with conjugate priors on \( \pi, \mu, \Lambda \)

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R \texttt{bayesm} package

LRVB, MFVB
Experiments

- Gaussian mixture model
  \[ P(z_{nk} = 1) = \pi_k, \quad p(x|\pi, \mu, \Lambda, z) = \prod_{n=1:N} \prod_{k=1:K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{nk}} \]
  with conjugate priors on \( \pi, \mu, \Lambda \)

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R \texttt{bayesm} package

Experiments

- LRVB, MFVB
Experiments

- Gaussian mixture model

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

with conjugate priors on \( \pi, \mu, \Lambda \)

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

- Gaussian mixture model
  \[ P(z_{nk} = 1) = \pi_k, \quad p(x|\pi, \mu, \Lambda, z) = \prod_{n=1:N} \prod_{k=1:K} \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1})^{z_{nk}} \]
  with conjugate priors on \( \pi, \mu, \Lambda \)

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, \( R \) \texttt{bayesm} package

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

\[ \text{LRVB, MFVB} \]
Experiments

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

  with conjugate priors on \( \pi, \mu, \Lambda \)

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R \texttt{bayesm} package

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

LRVB, MFVB
Experiments
Experiments

• Non-conjugate normal-Poisson generalized linear mixed model

\[ z_n | \beta, \tau \sim \text{iid } \mathcal{N}(z_n | \beta x_n, \tau^{-1}) , \quad y_n | z_n \sim \text{iid } \text{Poisson}(y_n | \exp(z_n)) , \]

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

- Non-conjugate normal-Poisson generalized linear mixed model
  \[ z_n | \beta, \tau \overset{\text{indep}}{\sim} \mathcal{N} ( z_n | \beta x_n, \tau^{-1} ) , \quad y_n | z_n \overset{\text{indep}}{\sim} \text{Poisson} \left( y_n | \exp(z_n) \right) , \]
  \[ \beta \sim \mathcal{N} ( \beta | 0, \sigma_\beta^2 ) , \quad \tau \sim \text{Gamma} ( \tau | \alpha_\tau, \beta_\tau ) \]

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

- Non-conjugate normal-Poisson generalized linear mixed model
  \[ z_n | \beta, \tau \overset{\text{indep}}{\sim} \mathcal{N}(z_n | \beta x_n, \tau^{-1}), \quad y_n | z_n \overset{\text{indep}}{\sim} \text{Poisson}(y_n | \exp(z_n)), \]
  \[ \beta \sim \mathcal{N}(\beta | 0, \sigma_\beta^2), \quad \tau \sim \text{Gamma}(\tau | \alpha_\tau, \beta_\tau) \]

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

• Non-conjugate normal-Poisson generalized linear mixed model
  \[ z_n \mid \beta, \tau \overset{\text{indep}}{\sim} \mathcal{N}(z_n \mid \beta x_n, \tau^{-1}), \quad y_n \mid z_n \overset{\text{indep}}{\sim} \text{Poisson}(y_n \mid \exp(z_n)), \quad \beta \sim \mathcal{N}(\beta \mid 0, \sigma_\beta^2), \quad \tau \sim \text{Gamma}(\tau \mid \alpha_\tau, \beta_\tau) \]

• 100 simulated data sets, 500 data points each, R MCMCglmm package

LRVB, MFVB
Experiments

• Non-conjugate normal-Poisson generalized linear mixed model
  \[ z_n \mid \beta, \tau \, \text{indep} \sim \mathcal{N}(z_n \mid \beta x_n, \tau^{-1}), \quad y_n \mid z_n \, \text{indep} \sim \text{Poisson}(y_n \mid \exp(z_n)), \]
  \[ \beta \sim \mathcal{N}(\beta \mid 0, \sigma^2_\beta), \quad \tau \sim \text{Gamma}(\tau \mid \alpha_{\tau}, \beta_{\tau}) \]

• 100 simulated data sets, 500 data points each, R MCMCglmm package

LRVB, MFVB
Experiments

• Non-conjugate normal-Poisson generalized linear mixed model

\[ z_n \mid \beta, \tau \sim_{\text{indep}} \mathcal{N} \left( z_n \mid \beta x_n, \tau^{-1} \right), \quad y_n \mid z_n \sim_{\text{indep}} \text{Poisson} \left( y_n \mid \exp(z_n) \right), \]

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

• 100 simulated data sets, 500 data points each, R \textit{MCMCglmm} package

LRVB, MFVB
Experiments

- Non-conjugate normal-Poisson generalized linear mixed model
  \[ z_n | \beta, \tau \overset{\text{indep}}{\sim} \mathcal{N}(z_n | \beta x_n, \tau^{-1}), \quad y_n | z_n \overset{\text{indep}}{\sim} \text{Poisson}(y_n | \exp(z_n)), \]
  \[ \beta \sim \mathcal{N}(\beta | 0, \sigma^2_\beta), \quad \tau \sim \text{Gamma}(\tau | \alpha_\tau, \beta_\tau) \]

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

LRVB, MFVB
Scaling the matrix inverse
Scaling the matrix inverse

- LRVB estimate \( \hat{\Sigma} = (I - VH)^{-1}V \)
Scaling the matrix inverse

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

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

• Decomposition of parameter vector

\[ \theta = (\alpha^T, z^T)^T \]
Scaling the matrix inverse

- LRVB estimate \( \hat{\Sigma} = (I - VH)^{-1}V \)
- Decomposition of parameter vector
  \[ \theta = (\alpha^T, z^T)^T \]

\[
\begin{bmatrix}
H_\alpha & H_{\alpha z} \\
H_{z \alpha} & H_z
\end{bmatrix}
\]
Scaling the matrix inverse

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

- Decomposition of parameter vector:
  $$\theta = (\alpha^T, z^T)^T$$

- Schur complement

$$H = \begin{bmatrix}
H_\alpha & H_{\alpha z} \\
H_{z\alpha} & H_z
\end{bmatrix}$$
Scaling the matrix inverse

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

- Decomposition of parameter vector
  \[ \theta = (\alpha^T, z^T)^T \]

- Schur complement
  \[ \hat{\Sigma}_\alpha = (I_\alpha - V_\alpha H_\alpha - V_\alpha H_{\alpha z} (I_z - V_z H_z)^{-1} V_z H_{z \alpha})^{-1} V_\alpha \]
Scaling the matrix inverse

• LRVB estimate \( \hat{\Sigma} = (I - V H)^{-1} V \)

• Decomposition of parameter vector

\[
\theta = (\alpha^T, z^T)^T
\]

\[
H = \begin{pmatrix}
H_{\alpha} & H_{\alpha z} \\
H_{z\alpha} & H_z
\end{pmatrix}
\]

• Schur complement

\[
\hat{\Sigma}_\alpha = (I_\alpha - V_\alpha H_\alpha - V_\alpha H_{\alpha z} (I_z - V_z H_z)^{-1} V_z H_{z\alpha})^{-1} V_\alpha
\]
Scaling the matrix inverse

- LRVB estimate \( \hat{\Sigma} = (I - VH)^{-1}V \)
- Decomposition of parameter vector
  \[ \theta = (\alpha^T, z^T)^T \]
- Schur complement
  \[ \hat{\Sigma}_\alpha = (I_\alpha - V_\alpha H_\alpha - V_\alpha H_{\alpha z} \left( I_z - V_z H_z \right)^{-1} V_z H_{z\alpha})^{-1} V_\alpha \]
Scaling the matrix inverse

- LRVB estimate \( \hat{\Sigma} = (I - VH)^{-1}V \)
- Decomposition of parameter vector
  \[ \theta = (\alpha^T, z^T)^T \]

- Schur complement
  \[ \hat{\Sigma}_\alpha = (I_\alpha - V_\alpha H_\alpha - V_\alpha H_{\alpha z} (I_z - V_z H_z)^{-1}V_z H_{z \alpha})^{-1} V_\alpha \]
- Sparsity patterns
Scaling the matrix inverse

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

- Decomposition of parameter vector
  \[ \theta = (\alpha^T, z^T)^T \]

- Schur complement
  \[ \hat{\Sigma}_\alpha = (I_\alpha - V_\alpha H_\alpha - V_\alpha H_\alpha z (I_z - V_z H_z)^{-1}V_z H_z \alpha)^{-1} V_\alpha \]

- Sparsity patterns
Scaling the matrix inverse

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

- Decomposition of parameter vector

  \[ \theta = (\alpha^T, z^T)^T \]

- Schur complement

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

- Sparsity patterns

\[ H = \begin{bmatrix} H_\alpha & H_{\alpha z} \\ H_{z \alpha} & H_z \end{bmatrix} \]
Scaling the matrix inverse

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

- Decomposition of parameter vector \( \theta = (\alpha^T, z^T)^T \)

- Schur complement

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

- Sparsity patterns

\[
V \quad H \quad I - VH
\]
Roadmap

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

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

• Bayes Theorem

\begin{align*}
p(\theta|x) & \\
\propto & \; p(x|\theta)p(\theta)
\end{align*}
Robustness quantification

- Bayes Theorem

\[ p(\theta|x, \alpha) \propto \theta \cdot p(x|\theta)p(\theta|\alpha) \]
Robustness quantification

• Bayes Theorem

\[ p_\alpha(\theta) := p(\theta|x, \alpha) \]
\[ \propto \theta \cdot p(x|\theta)p(\theta|\alpha) \]
Robustness quantification

• Bayes Theorem

\[ p_{\alpha}(\theta) := p(\theta|x, \alpha) \]

\[ \propto_{\theta} p(x|\theta)p(\theta|\alpha) \]

• Sensitivity
Robustness quantification

• Bayes Theorem

\[ p_\alpha(\theta) := p(\theta|x, \alpha) \propto \theta \cdot p(x|\theta)p(\theta|\alpha) \]

• Sensitivity

Some reasonable priors
Robustness quantification

- **Bayes Theorem**
  
  \[
  p_\alpha(\theta) := p(\theta|x, \alpha) 
  \propto \theta \ p(x|\theta)p(\theta|\alpha)
  \]

- **Sensitivity**
  
  \[
  \mathbb{E}_{p_\alpha}[g(\theta)]
  \]
Robustness quantification

- Bayes Theorem
  \[ p_\alpha(\theta) := p(\theta|x, \alpha) \propto_\theta p(x|\theta)p(\theta|\alpha) \]

- Sensitivity
  \[ \mathbb{E}_{p_\alpha}[g(\theta)] \]
Robustness quantification

- Bayes Theorem

\[ p_\alpha(\theta) := p(\theta|x, \alpha) \]
\[ \propto \theta p(x|\theta)p(\theta|\alpha) \]

- Sensitivity

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

- Bayes Theorem

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

- Sensitivity

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

• Bayes Theorem

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

• Sensitivity

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

- Bayes Theorem

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

\[ \propto_\theta p(x | \theta) p(\theta | \alpha) \]

- Sensitivity

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

\[ \approx \frac{d\mathbb{E}_{q_{\alpha}^*} [g(\theta)]}{d\alpha} \bigg|_{\alpha} \Delta \alpha =: \hat{S} \]
Robustness quantification

- **Bayes Theorem**
  
  \[ p_{\alpha}(\theta) := p(\theta|x, \alpha) \]
  
  \[ \propto_{\theta} p(x|\theta)p(\theta|\alpha) \]

- **Sensitivity**

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

Some reasonable priors
Robustness quantification

- Bayes Theorem

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

- Sensitivity

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

- When \( q^*_\alpha \) in exponential family

Some reasonable priors
Robustness quantification

- Bayes Theorem

\[ p_\alpha(\theta) := p(\theta|x, \alpha) \]
\[ \propto \theta p(x|\theta)p(\theta|\alpha) \]

- Sensitivity

\[ S := \frac{d\mathbb{E}_{p_\alpha}[g(\theta)]}{d\alpha} \bigg|_{\alpha} \Delta \alpha \]
\[ \approx \frac{d\mathbb{E}_{q_\alpha^*}[g(\theta)]}{d\alpha} \bigg|_{\alpha} \Delta \alpha =: \hat{S} \]

- When \( q_\alpha^* \) in exponential family

\[ \hat{S} = A \left( \frac{\partial^2 KL}{\partial m \partial m^T} \bigg|_{m=m^*} \right)^{-1} B \]
Microcredit Experiment

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

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

- Priors and hyperpriors:

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

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

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

- Perturb $\Lambda_{11}$: $0.03 \rightarrow 0.04$
Microcredit Experiment

• Perturb $\Lambda_{11}$: 0.03 $\rightarrow$ 0.04
Microcredit Experiment

- Perturb $\Lambda_{11}$: $0.03 \rightarrow 0.04$
Microcredit Experiment
Microcredit Experiment

- Sensitivity of the expected microcredit effect ($\tau$)

- Normalized to be on scale of $\tau$ std devs

- $\tau$ mean (MFVB): 3.08 USD PPP

- $\tau$ std dev (LRVB): 1.83 USD PPP

- Mean is 1.68 std dev from 0

- $\Lambda_{11} = 0.04$

  - Mean $> 2$ std dev
Microcredit Experiment

• Sensitivity of the expected microcredit effect ($\tau$)
• Normalized to be on scale of $\tau$ std devs
Microcredit Experiment

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

- $\tau$ mean (MFVB): 3.08 USD PPP
- $\tau$ std dev (LRVB): 1.83 USD PPP
- Mean is 1.68 std dev from 0

\[ \Lambda_{11} = 0.04 \]
Mean > 2 std dev Microcredit Experiment
Microcredit Experiment

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

- $\tau$ mean (MFVB): 3.08 USD PPP
- $\tau$ std dev (LRVB): 1.83 USD PPP
- Mean is 1.68 std dev from 0

[Graph showing normalized sensitivity with bars for $\lambda_{11}$, $\lambda_{12}$, and $\lambda_{22}$]

[Graph showing normalized sensitivity with bars for $\mu_0$ and $\tau_0$]
Microcredit Experiment

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

- Sensitivity of the expected microcredit effect ($\tau$)
- Normalized to be on scale of $\tau$ std devs
- $\tau$ mean (MFVB): 3.08 USD PPP
- $\tau$ std dev (LRVB): 1.83 USD PPP
- Mean is 1.68 std dev from 0
- $\Lambda_{11} = \pm 0.04$
Microcredit Experiment

- Sensitivity of the expected microcredit effect ($\tau$)
- Normalized to be on scale of $\tau$ std devs
- $\tau$ mean (MFVB): 3.08 USD PPP
- $\tau$ std dev (LRVB): 1.83 USD PPP
- Mean is 1.68 std dev from 0
- $\Lambda_{11} += 0.04$
  $\Rightarrow$ Mean > 2 std dev
Conclusions

• We provide linear response variational Bayes: supplements MFVB for fast & accurate covariance estimate

• More from LRVB: fast & accurate robustness quantification

• Interested in your data and models:
  • Sensitivity to prior perturbations
  • Sensitivity to likelihood, data perturbations

• Computational statistical trade-offs
  • New data summaries: coresets, approx. sufficient stats
  • Criteo data set: 40 million data points, 3 million features, our runtime: ~20 seconds on 24 cores
  • Theoretical guarantees on finite-sample quality

[Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick, submitted]
References


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

References


