Covariance matrices for mean field variational Bayes

Ryan Giordano, Tamara Broderick, Michael I. Jordan
Berkeley ITT Career Development Assistant Professor, MIT Berkeley
Statistical/computational trade-offs
Statistical/computational trade-offs

• Bayesian inference
Statistical/computational trade-offs

- Bayesian inference
  - modular, complex models

[Broderick, Kulis, Jordan 2013]
Statistical/computational trade-offs

- Bayesian inference
  - modular, complex models
  - all information about the parameter in the posterior
Statistical/computational trade-offs

- Bayesian inference
  - modular, complex models
  - all information about the parameter in the posterior
- Approximating the posterior can be computationally expensive
Statistical/computational trade-offs

- Bayesian inference
  - modular, complex models
  - all information about the parameter in the posterior
- Approximating the posterior can be computationally expensive
- Computational/statistical gains for trading off some posterior knowledge

[Broderick, Kulis, Jordan 2013]
Statistical/computational trade-offs

- Bayesian inference
  - modular, complex models
  - all information about the parameter in the posterior
- Approximating the posterior can be computationally expensive
- Computational/statistical gains for trading off some posterior knowledge
  - point estimates: e.g., MAD-Bayes

[Broderick, Kulis, Jordan 2013]
Statistical/computational trade-offs

- Bayesian inference
  - modular, complex models
  - all information about the parameter in the posterior
- Approximating the posterior can be computationally expensive
- Computational/statistical gains for trading off some posterior knowledge
  - point estimates: e.g., MAD-Bayes
  - covariances, coherent estimates of uncertainty

[Broderick, Kulis, Jordan 2013]
What about uncertainty?

• Variational Bayes (VB)

• Approximation for posterior

• Minimize Kullback-Liebler (KL) divergence:

\[
p(✓ | x) \approx q(✓) \quad \text{minimize } KL(p(✓ | x) || q(✓))
\]

• VB practical success

• point estimates and prediction

• fast
What about uncertainty?

• Variational Bayes (VB)
What about uncertainty?

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

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

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

• Variational Bayes (VB)
  • Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
What about uncertainty?

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

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

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

- VB practical success
  - point estimates and prediction
What about uncertainty?

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

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

• VB practical success
  • point estimates and prediction
  • fast

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

- Variational Bayes (VB)
- Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
- Minimize Kullback-Liebler (KL) divergence:
  \[
  KL(q\|p(\cdot|x))
  \]
- VB practical success
  - point estimates and prediction
  - fast, streaming, distributed

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

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

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

- Variational Bayes
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

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

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

- Variational Bayes

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

- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)

[Bishop 2006]
What about uncertainty?

- Variational Bayes

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

- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)

- No covariance estimates

[Bishop 2006]
What about uncertainty?

- Variational Bayes

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

- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)

- No covariance estimates

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

• Variational Bayes

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

• Mean-field variational Bayes (MFVB)

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

• Underestimates variance (sometimes severely)

• No covariance estimates

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

[Dunson 2014; Bardenet, Doucet, Holmes, 2015]
1. Derive *Linear Response Variational Bayes* (LRVB) variance/covariance correction

2. Accuracy experiments

3. Scalability experiments
Linear response
Linear response

• Cumulant-generating function
Linear response

- Cumulant-generating function

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

- Cumulant-generating function

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

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

- True posterior covariance
Linear response

- Cumulant-generating function

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

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

- True posterior covariance

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

[Bishop 2006]
Linear response

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

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

- Cumulant-generating function

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

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

- True posterior covariance vs MFVB covariance

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

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

[Bishop 2006]
Linear response

• Cumulant-generating function

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

• True 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

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

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

• True posterior covariance vs MFVB covariance

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

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

• “Linear response”

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

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

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

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

[Bishop 2006]
Linear response

• Cumulant-generating function

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

• True posterior covariance vs MFVB covariance

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

• “Linear response”

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

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

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

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

- Cumulant-generating function

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

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

- True 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”

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

• Cumulant-generating function

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

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

• True 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
Linear response

• Cumulant-generating function

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

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

• True 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

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

[Bishop 2006]
Linear response

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

- True posterior covariance 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
  \[ \Sigma = \left. \frac{d}{dt^T} \mathbb{E}_{p_t} \theta \right|_{t=0} \]
Linear response

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

- True 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
  \[ \Sigma = \left. \frac{d}{dt^T} \mathbb{E}_{p_t} \theta \right|_{t=0} \]

[Bishop 2006]
Linear response

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

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

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

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

- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

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

- “Linear response”

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

- The LRVB approximation

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

[Bishop 2006]
Linear response

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

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

- The LRVB approximation
  \[ \Sigma = \frac{d}{dt^T} \mathbb{E}_{\rho_t} \theta \bigg|_{t=0} \approx \frac{d}{dt^T} \mathbb{E}_{q^*_t} \theta \bigg|_{t=0} =: \hat{\Sigma} \]
• LRVB covariance estimate \( \hat{\Sigma} := \left. \frac{d}{dt} \mathbf{E}_{q_t^*} \theta \right|_{t=0} \)
Getting rid of $t$

- LRVB covariance estimate  $\hat{\Sigma} := \left. \frac{d}{dt} \mathbb{E}_{q^*_t} \theta \right|_{t=0}$
Getting rid of $t$

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

- Suppose $q_t$ exponential family
Getting rid of $t$

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

- Suppose $q_t$ exponential family with mean parametrization $m_t$
Getting rid of $t$

- 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 $m_t$
Getting rid of $t$

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

- Suppose $q_t$ exponential family with mean parametrization $m_t$
Getting rid of $t$

- LRVB covariance estimate $\hat{\Sigma} := \frac{d}{dt^T} m_t^* \Bigg|_{t=0}$

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

- KL optimization: fixed point equation in the mean params

\[ 0 = \frac{\partial}{\partial m_t} KL_t \Bigg|_{m_t=m_t^*} \]
Getting rid of $t$

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

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

- KL optimization: fixed point equation in the mean params

$$m_t^* = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t = m_t^*} + m_t^*$$
Getting rid of $t$

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

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

- KL optimization: fixed point equation in the mean params

$$m_t^* = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t=m_t^*} + m_t^*$$

$$\hat{\Sigma} = \left( \left. \frac{\partial^2 KL}{\partial m \partial m^T} \right|_{m=m_t^*} \right)^{-1}$$
Getting rid of $t$

- LRVB covariance estimate $\hat{\Sigma} := \frac{d}{dt^T}m_t^* \bigg|_{t=0}$
- Suppose $q_t$ exponential family with mean parametrization $m_t$
- KL optimization: fixed point equation in the mean params
  $$m_t^* = \frac{\partial}{\partial m_t} KL_t \bigg|_{m_t=m_t^*} + m_t^*$$
  $$\hat{\Sigma} = \left( \frac{\partial^2 KL}{\partial m \partial m^T} \bigg|_{m=m^*} \right)^{-1}$$
- KL decomposition: $KL = \mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta|x) =: S - L$
Getting rid of $t$

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

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

- KL optimization: fixed point equation in the mean params

$$m_t^* = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t=m_t^*} + m_t^*$$

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

- KL decomposition: $KL = \mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta|x) =: S - L$

$$\hat{\Sigma} = (V^{-1} - H)^{-1}$$
Getting rid of $t$

- LRVB covariance estimate $\hat{\Sigma} := \frac{d}{dt^T} m_t^* \bigg|_{t=0}

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

- KL optimization: fixed point equation in the mean params

$$m_t^* = \frac{\partial}{\partial m_t} KL_t \bigg|_{m_t=m_t^*} + m_t^*$$

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

- KL decomposition: $KL = \mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta|x) =: S - L$

$$\hat{\Sigma} = (V^{-1} - H)^{-1} = (I - VH)^{-1} V$$
Getting rid of $t$

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

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

- KL optimization: fixed point equation in the mean params

$$m_t^* = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t=m_t^*} + m_t^*$$

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

- KL decomposition: $KL = \mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta|x) =: S - L$

$$\hat{\Sigma} = \left( V^{-1} - H \right)^{-1} = (I - VH)^{-1} V \quad \text{for} \quad \left. \frac{\partial^2 L}{\partial m \partial m^T} \right|_{m=m^*}$$
LRVB estimator

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

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

- KL optimization: fixed point equation in the mean params
  \[ m_t^* = \frac{\partial}{\partial m_t} KL_t \bigg|_{m_t=m_t^*} + m_t^* \]
  \[ \hat{\Sigma} = \left( \frac{\partial^2 KL}{\partial m \partial m^T} \bigg|_{m=m^*} \right)^{-1} \]

- KL decomposition: \( KL = E_q \log q(\theta) - E_q \log p(\theta|x) =: S - L \)
  \[ \hat{\Sigma} = (V^{-1} - H)^{-1} = (I - VH)^{-1} V \quad \text{for} \quad H := \frac{\partial^2 L}{\partial m \partial m^T} \bigg|_{m=m^*} \]
LRVB estimator

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

- KL decomposition: \( KL = \mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta|x) =: S - L \)
  \[
  \hat{\Sigma} = \left( V^{-1} - H \right)^{-1} = (I - VH)^{-1} V \quad \text{for} \quad H := \frac{\partial^2 L}{\partial m \partial m^T} \bigg|_{m=m^*}
  \]
LRVB estimator

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

- KL decomposition:
  \[ KL = E_q \log q(\theta) - E_q \log p(\theta|x) =: S - L \]
  \[ \hat{\Sigma} = (V^{-1} - H)^{-1} = (I - VH)^{-1}V \]
  for \[ H := \left. \frac{\partial^2 L}{\partial m \partial m^T} \right|_{m=m^*} \]
LRVB estimator

- LRVB covariance estimate

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

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

\[
\hat{\Sigma} = (I - VH)^{-1} V
\]
LRVB estimator

- LRVB covariance estimate

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

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

\[ \hat{\Sigma} = (I - VH)^{-1}V \]
LRVB estimator

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

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

\[
\hat{\Sigma} = (I - VH)^{-1} V
\]
LRVB estimator

• LRVB covariance estimate
  \[ \hat{\Sigma} := \left. \frac{d}{dt^T} \mathbb{E}_{q_t^*} \theta \right|_{t=0} \]
  \[ \hat{\Sigma} = \left( \frac{\partial^2 KL}{\partial m \partial m^T} \right)_{m=m^*}^{-1} \]
  \[ \hat{\Sigma} = (I - VH)^{-1}V \]

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

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

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

\[ \hat{\Sigma} = (I - VH)^{-1} V \]

• 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 \( \hat{\Sigma} := \left. \frac{d}{d t^T} \mathbb{E}_{q^*_t} \theta \right|_{t=0} \)

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

\[
\hat{\Sigma} = (I - VH)^{-1} V
\]

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

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

\[
\hat{\Sigma} = (I - VH)^{-1} V
\]

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

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

\[
\hat{\Sigma} = (I - VH)^{-1} V
\]

• 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 VB gives exact mean (e.g. multivariate normal)

[Bishop 2006]
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:
  \[
  \theta = (\alpha^T, z^T)^T
  \]

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

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

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

- Schur complement

\[
\begin{bmatrix}
  H_{\alpha} & H_{\alpha z} \\
  H_{z\alpha} & H_z
\end{bmatrix}
\]
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 \( \hat{\Sigma} = (I - VH)^{-1}V \)

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

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

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

\[
\begin{bmatrix}
H_\alpha & H_{\alpha z} \\
H_{z\alpha} & H_z
\end{bmatrix}
\]

• 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
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
  \[
  V \quad H \quad I - VH
  \]
1. Derive *Linear Response Variational Bayes* (LRVB) variance/covariance correction

2. Accuracy experiments

3. Scalability experiments
Experiments
Experiments

• Non-conjugate normal-Poisson generalized linear mixed model
Experiments

• Non-conjugate normal-Poisson generalized linear mixed model

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

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

• Non-conjugate normal-Poisson generalized linear mixed model

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

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

• MFVB assumption:

\[ q(\beta, \tau, z) = q(\beta)q(\tau) \prod_{n=1}^{N} q(z_n) \]
Experiments

• Non-conjugate normal-Poisson generalized linear mixed model

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

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

• MFVB assumption:

\[ q(\beta, \tau, z) = q(\beta) q(\tau) \prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathcal{N}(z_n) \]
Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

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

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

- MFVB assumption:

\[ q(\beta, \tau, z) = q(\beta)q(\tau) \prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathcal{N} (z_n) \]

- 100 simulated data sets, 500 data points each, R MCMCglmm package (20,000 samples)
Experiments

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

• MFVB assumption:
\[ q(\beta, \tau, z) = q(\beta)q(\tau) \prod_{n=1}^{N} q(z_n), \quad q(z_n) = N(z_n) \]

• 100 simulated data sets, 500 data points each, R MCMCglmm package (20,000 samples)
Experiments

• Non-conjugate normal-Poisson generalized linear mixed model

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

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

• MFVB assumption:

\[ q(\beta, \tau, z) = q(\beta)q(\tau) \prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathcal{N}(z_n) \]

• 100 simulated data sets, 500 data points each, R MCMCglm package (20,000 samples)

LRVB, MFVB
Experiments

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

- MFVB assumption:
  \[ q(\beta, \tau, z) = q(\beta)q(\tau) \prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathcal{N}(z_n) \]

- 100 simulated data sets, 500 data points each, R MCMCglmm package (20,000 samples)
Experiments

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

- MFVB assumption:
  \[ q(\beta, \tau, z) = q(\beta)q(\tau) \prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathcal{N}(z_n) \]

- 100 simulated data sets, 500 data points each, R \textit{MCMCglmm} package (20,000 samples)
Experiments

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

- MFVB assumption:
  \[ q(\beta, \tau, z) = q(\beta)q(\tau) \prod_{n=1}^{N} q(z_n) , \quad q(z_n) = \mathcal{N}(z_n) \]

- 100 simulated data sets, 500 data points each, R `MCMCglmm` package (20,000 samples)
Experiments
Experiments

- Linear model with random effects
Experiments

- Linear model with random effects

\[ y_n | \beta, z, \tau \overset{\text{indep}}{\sim} \mathcal{N} \left( y_n | \beta^T x_n + r_n z_k(n), \tau^{-1} \right), \quad z_k | \nu \overset{\text{iid}}{\sim} \mathcal{N} \left( z_k | 0, \nu^{-1} \right) \]
\[ \beta \sim \mathcal{N}(\beta | 0, \Sigma_\beta), \quad \nu \sim \Gamma(\nu | \alpha_\nu, \beta_\nu), \quad \tau \sim \Gamma(\tau | \alpha_\tau, \beta_\tau) \]
Experiments

• Linear model with random effects

\[ y_n | \beta, z, \tau \sim \text{indep } \mathcal{N} \left( y_n | \beta^T x_n + r_n z_k(n), \tau^{-1} \right), \quad z_k | \nu \sim \text{iid } \mathcal{N} \left( z_k | 0, \nu^{-1} \right) \]

\[ \beta \sim \mathcal{N}(\beta | 0, \Sigma_\beta), \quad \nu \sim \Gamma(\nu | \alpha_\nu, \beta_\nu), \quad \tau \sim \Gamma(\tau | \alpha_\tau, \beta_\tau) \]

• MFVB assumption:

\[ q(\beta, \nu, \tau, z) = q(\beta)q(\tau)q(\nu) \prod_{k=1}^{K} q(z_n) \]
Experiments

- Linear model with random effects
  \( y_n | \beta, z, \tau \overset{\text{indep}}{\sim} \mathcal{N} (y_n | \beta^T x_n + r_n z_{k(n)}, \tau^{-1}) \), \( z_k | \nu \overset{iid}{\sim} \mathcal{N} (z_k | 0, \nu^{-1}) \)
  \( \beta \sim \mathcal{N}(\beta | 0, \Sigma_\beta), \quad \nu \sim \Gamma(\nu | \alpha_\nu, \beta_\nu), \quad \tau \sim \Gamma(\tau | \alpha_\tau, \beta_\tau) \)

- MFVB assumption: \( q(\beta, \nu, \tau, z) = q(\beta)q(\tau)q(\nu) \prod_{k=1}^{K} q(z_n) \)

- 100 simulated data sets, 300 data points each, R MCMCglmm package (20,000 samples)
Experiments

• Linear model with random effects

\[ y_n | \beta, z, \tau \overset{\text{indep}}{\sim} \mathcal{N} ( y_n | \beta^T x_n + r_n z_k(n), \tau^{-1} ) , \quad z_k | \nu \overset{iid}{\sim} \mathcal{N} ( z_k | 0, \nu^{-1} ) \]

\[ \beta \sim \mathcal{N}(\beta|0, \Sigma_{\beta}), \quad \nu \sim \Gamma(\nu|\alpha_{\nu}, \beta_{\nu}), \quad \tau \sim \Gamma(\tau|\alpha_{\tau}, \beta_{\tau}) \]

• MFVB assumption:

\[ q(\beta, \nu, \tau, z) = q(\beta)q(\tau)q(\nu) \prod_{k=1}^{K} q(z_n) \]

• 100 simulated data sets, 300 data points each, R MCMC\texttt{glm}m package (20,000 samples)

LRVB, MFVB
Experiments

• Linear model with random effects
  \[ y_n | \beta, z, \tau \overset{\text{indep}}{\sim} \mathcal{N}(y_n | \beta^T x_n + r_n z_k(n), \tau^{-1}) \, , \, z_k | \nu \overset{iid}{\sim} \mathcal{N}(z_k | 0, \nu^{-1}) \]
  \[ \beta \sim \mathcal{N}((\beta | 0, \Sigma_\beta) \, , \, \nu \sim \Gamma(\nu | \alpha_\nu, \beta_\nu) \, , \, \tau \sim \Gamma(\tau | \alpha_\tau, \beta_\tau) \]

• MFVB assumption:
  \[ q(\beta, \nu, \tau, z) = q(\beta)q(\tau)q(\nu) \prod_{k=1}^{K} q(z_n) \]

• 100 simulated data sets, 300 data points each, R MCMCglmm package (20,000 samples)
Experiments

- Linear model with random effects
  \[ y_n | \beta, z, \tau \overset{\text{indep}}{\sim} \mathcal{N} \left( y_n | \beta^T x_n + r_n z_k(n), \tau^{-1} \right), \quad z_k | \nu \overset{iid}{\sim} \mathcal{N} \left( z_k | 0, \nu^{-1} \right) \]
  \[ \beta \sim \mathcal{N}(\beta | 0, \Sigma_\beta), \quad \nu \sim \Gamma(\nu | \alpha_\nu, \beta_\nu), \quad \tau \sim \Gamma(\tau | \alpha_\tau, \beta_\tau) \]

- MFVB assumption:
  \[ q(\beta, \nu, \tau, z) = q(\beta)q(\tau)q(\nu) \prod_{k=1}^{K} q(z_n) \]

- 100 simulated data sets, 300 data points each, R MCMCglmm package (20,000 samples)
Experiments

- Linear model with random effects

$$y_n | \beta, z, \tau \overset{\text{indep}}{\sim} N \left( y_n | \beta^T x_n + r_n z_k(n), \tau^{-1} \right), \quad z_k | \nu \overset{iid}{\sim} N \left( z_k | 0, \nu^{-1} \right)$$

$$\beta \sim N(\beta | 0, \Sigma_{\beta}), \quad \nu \sim \Gamma(\nu | \alpha_{\nu}, \beta_{\nu}), \quad \tau \sim \Gamma(\tau | \alpha_{\tau}, \beta_{\tau})$$

- MFVB assumption:

$$q(\beta, \nu, \tau, z) = q(\beta)q(\tau)q(\nu) \prod_{k=1}^{K} q(z_n)$$

- 100 simulated data sets, 300 data points each, R MCMCglmm package (20,000 samples)
Experiments
Experiments

- Gaussian mixture model
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 $\pi, \mu, \Lambda$

• MFVB assumption:

\[
\left[ \prod_{k=1}^{K} q(\mu_k)q(\Lambda_k)q(\pi_k) \right] \prod_{n=1}^{N} q(z_n)
\]
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 \( \pi, \mu, \Lambda \)

- MFVB assumption:
  \[
  \left[ \prod_{k=1}^{K} q(\mu_k)q(\Lambda_k)q(\pi_k) \right] \prod_{n=1}^{N} q(z_n)
  \]

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, \( \text{R \ bayesm} \) package (function \( \text{rnmixGibbs} \); at least 500 effective samples)
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 \( \pi, \mu, \Lambda \)

• MFVB assumption:
  \[ \prod_{k=1}^{K} q(\mu_k)q(\Lambda_k)q(\pi_k) \prod_{n=1}^{N} q(z_n) \]

• 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, \texttt{R bayesm} package (function \texttt{rnmixGibbs}; at least 500 effective samples)

• MNIST digits: 12,665 0s and 1s; PCA for 25 dimensions
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 \( \pi, \mu, \Lambda \)

- MFVB assumption:
  \[ \left[ \prod_{k=1}^{K} q(\mu_k)q(\Lambda_k)q(\pi_k) \right] \prod_{n=1}^{N} q(z_n) \]

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R \texttt{bayesm} package (function \texttt{rnmixGibbs}; at least 500 effective samples)

- MNIST digits: 12,665 0s and 1s; PCA for 25 dimensions
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 \( \pi, \mu, \Lambda \)

- MFVB assumption:
  \[ \left[ \prod_{k=1}^{K} q(\mu_k) q(\Lambda_k) q(\pi_k) \right] \prod_{n=1}^{N} q(z_n) \]

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, \( R \) \texttt{bayesm} package (function \texttt{rnmixGibbs}; at least 500 effective samples)

- MNIST digits: 12,665 0s and 1s; PCA for 25 dimensions
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 \( \pi, \mu, \Lambda \)

- MFVB assumption:
  \[ \prod_{k=1}^{K} q(\mu_k)q(\Lambda_k)q(\pi_k) \prod_{n=1}^{N} q(z_n) \]

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R \texttt{bayesm} package (function \texttt{rnmixGibbs}; at least 500 effective samples)

- MNIST digits: 12,665 0s and 1s; PCA for 25 dimensions
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 \( \pi, \mu, \Lambda \)

- MFVB assumption:
  \[ \left[ \prod_{k=1}^{K} q(\mu_k)q(\Lambda_k)q(\pi_k) \right] \prod_{n=1}^{N} q(z_n) \]

- 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, R \texttt{bayesm} package (function \texttt{rnmixGibbs}; at least 500 effective samples)

- MNIST digits: 12,665 0s and 1s; PCA for 25 dimensions
1. Derive *Linear Response Variational Bayes* (LRVB) variance/covariance correction

2. Accuracy experiments

3. Scalability experiments
Experiments
Experiments

• Scaling: Gaussian mixture model ($K$ components, $P$ dimensions, $N$ data points)
Experiments

• Scaling: Gaussian mixture model ($K$ components, $P$ dimensions, $N$ data points)
• The number of parameters in $\mu, \pi, \Lambda$ grows as $O(KP^2)$
Experiments

• Scaling: Gaussian mixture model ($K$ components, $P$ dimensions, $N$ data points)
• The number of parameters in $\mu, \pi, \Lambda$ grows as $O(KP^2)$
• The number of parameters in $z$ grows as $O(KN)$
Experiments

- Scaling: Gaussian mixture model ($K$ components, $P$ dimensions, $N$ data points)
- The number of parameters in $\mu, \pi, \Lambda$ grows as $O(KP^2)$
- The number of parameters in $z$ grows as $O(KN)$
- Worst case scaling: $O(K^3), O(P^6), O(N)$
Experiments

• Scaling: Gaussian mixture model ($K$ components, $P$ dimensions, $N$ data points)
• The number of parameters in $\mu, \pi, \Lambda$ grows as $O(KP^2)$
• The number of parameters in $z$ grows as $O(KN)$
• Worst case scaling: $O(K^3), O(P^6), O(N)$
Experiments

• Scaling: Gaussian mixture model ($K$ components, $P$ dimensions, $N$ data points)
• The number of parameters in $\mu, \pi, \Lambda$ grows as $O(KP^2)$
• The number of parameters in $z$ grows as $O(KN)$
• Worst case scaling: $O(K^3), O(P^6), O(N)$

Experiments
Experiments

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

LRVB, Gibbs
Conclusions, etc

• LRVB covariance correction: in many cases, accurate covariance estimates for VB

• Next steps:
  • Scaling in parameter cardinality
  • Mean correction
  • Bayesian nonparametrics
  • MFVB $q$ not in exponential family

• Targeting other posterior statistics besides point estimates and covariance
Conclusions, etc

• LRVB covariance correction: in many cases, accurate covariance estimates for VB
Conclusions, etc

- LRVB covariance correction: in many cases, accurate covariance estimates for VB

- Next steps:
Conclusions, etc

• LRVB covariance correction: in many cases, accurate covariance estimates for VB

• Next steps:
  • Scaling in parameter cardinality
Conclusions, etc

• LRVB covariance correction: in many cases, accurate covariance estimates for VB

• Next steps:
  • Scaling in parameter cardinality
  • Mean correction
Conclusions, etc

• LRVB covariance correction: in many cases, accurate covariance estimates for VB

• Next steps:
  • Scaling in parameter cardinality
  • Mean correction
  • Bayesian nonparametrics
Conclusions, etc

• LRVB covariance correction: in many cases, accurate covariance estimates for VB

• Next steps:
  • Scaling in parameter cardinality
  • Mean correction
  • Bayesian nonparametrics
  • MFVB $q$ not in exponential family
Conclusions, etc

• LRVB covariance correction: in many cases, accurate covariance estimates for VB

• Next steps:
  • Scaling in parameter cardinality
  • Mean correction
  • Bayesian nonparametrics
  • MFVB $q$ not in exponential family
  • Targeting other posterior statistics besides point estimates and covariance
References

R Bardenet, A Doucet, and C Holmes. On Markov chain Monte Carlo methods for tall data. arXiv:1505.02827 (v1), 2015.

CM Bishop. *Pattern Recognition and Machine Learning*. Springer, 2010.

T Broderick, B Kulis, and MI Jordan. MAD-Bayes: MAP-based asymptotic derivations from Bayes. In *ICML*, 2013.

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

D Dunson. Robust and scalable approach to Bayesian inference. Talk at *ISBA* 2014.

R Giordano, T Broderick, and MI Jordan. Linear response methods for accurate covariance estimates from mean field variational Bayes. Submitted, http://arxiv.org/abs/1506.04088, 2015.

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

RE Turner and M Sahani. Two problems with variational expectation maximisation for time-series models. In D Barber, AT Cemgil, and S Chiappa, editors, *Bayesian Time Series Models*, 2011.

B Wang and M Titterington. Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *AISTATS*, 2004.