Variational Bayes and beyond: Foundations of scalable Bayesian inference

Tamara Broderick
Associate Professor, Electrical Engineering & Computer Science
MIT

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

[Grimm et al 2018]  

[ESO/ L. Calçada M. Kornmesser 2017] [Abbott et al 2016a,b]

[Woodard et al 2017]
Bayesian inference
Bayesian inference

[Gillon et al 2017]

[ESO/ L. Calçada M. Kornmesser 2017] [Abbott et al 2016a,b]

[Woodard et al 2017]

[amcharts.com 2016][Meager 2018a,b]

[Chati, Balakrishnan [Julian Hertzog 2016] 2017]
Bayesian inference
Goals: good point estimates, uncertainty estimates
Bayesian inference

- Goals: good point estimates, uncertainty estimates
- More: interpretable, modular, expert info
Bayesian inference

- Goals: good point estimates, uncertainty estimates
- More: interpretable, modular, expert info
Bayesian inference

- Goals: good point estimates, uncertainty estimates
- More: interpretable, modular, expert info
- Challenge: speed (compute, user), reliable inference
Bayesian inference

- Goals: good point estimates, uncertainty estimates
- More: interpretable, modular, expert info
- Challenge: speed (compute, user), reliable inference
- Uncertainty doesn’t have to disappear in large data sets
Variational Bayes
Variational Bayes

- Modern problems: often large data, large dimensions
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast
Variational Bayes

• Modern problems: often large data, large dimensions
• Variational Bayes can be very fast

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

[Blei et al 2003]

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.
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

![Table Example](image)

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. Every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

[Blei et al 2003]

[Airoldi et al 2008]
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

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

[Blei et al 2003]

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.
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

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

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,” said Randolph A. Hearst, who established the foundation in 1944. “The grants 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 receive $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated, has made its usual annual $100,000 donation, too.

[Blei et al 2003]

[Blei et al 2018]

[Airoldi et al 2008]

[Gershman et al 2014]

[Xing et al 2004]

[Xing 2003]

[Stegle et al 2010]
Roadmap

- Bayes & Approximate Bayes review
Roadmap

• Bayes & Approximate Bayes review
• What is:
  • Variational Bayes (VB)
  • Mean-field variational Bayes (MFVB)
Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB?
Roadmap

• Bayes & Approximate Bayes review
• What is:
  • Variational Bayes (VB)
  • Mean-field variational Bayes (MFVB)
• Why use VB?
• When can we trust VB?
Roadmap

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

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

\[ p(\theta) \]

prior

parameters
Bayesian inference

$p(\theta)$

prior

parameters
Bayesian inference

\[ p(y_1:N | \theta) p(\theta) \]

likelihood prior

parameters

\[ \theta \]
Bayesian inference

\[ p(y_{1:N} | \theta)p(\theta) \]

likelihood prior

\[ \theta \]

data parameters
Bayesian inference

\[ p(\theta | y_{1:N}) \propto p(y_{1:N} | \theta) p(\theta) \]

posterior, likelihood, prior
Bayesian inference

\[ p(\theta|y_{1:N}) \propto p(y_{1:N}|\theta)p(\theta) \]

posterior \quad likelihood \quad prior

Bayes Theorem
Bayesian inference

\[ p(\theta|y_{1:N}) \propto p(y_{1:N}|\theta)p(\theta) \]

posterior  likelihood  prior

Bayes Theorem
Bayesian inference

\[ p(\theta | y_{1:N}) \propto p(y_{1:N} | \theta)p(\theta) \]

posterior \quad likelihood \quad prior

1. Build a model: choose prior & choose likelihood
Bayesian inference

\[ p(\theta | y_{1:N}) \propto p(y_{1:N} | \theta) p(\theta) \]

data \hspace{1cm} \text{parameters}

posterior \hspace{1cm} \text{likelihood} \hspace{1cm} \text{prior}

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
Bayesian inference

\[ p(\theta | y_{1:N}) \propto p(y_{1:N} | \theta) p(\theta) \]

posterior likelihood prior

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
Bayesian inference

$$p(\theta|y_{1:N}) \propto p(y_{1:N}|\theta)p(\theta)$$

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   • Why are steps 2 and 3 hard?
Bayesian inference

Bayes Theorem

$$p(\theta|y_{1:N}) \propto \theta \ p(y_{1:N} | \theta)p(\theta)$$

posterior  likelihood  prior

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   • Why are steps 2 and 3 hard?
     • Typically no closed form
Bayesian inference

\[ p(\theta|y_{1:N}) \propto p(y_{1:N} | \theta)p(\theta) \]

posterior \quad likelihood \quad prior

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   - Why are steps 2 and 3 hard?
     - Typically no closed form, high-dimensional integration
Bayesian inference

\[ p(\theta|y_{1:N}) = \frac{p(y_{1:N}|\theta)p(\theta)}{p(y_{1:N})} \]

posterior likelihood prior

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   - Why are steps 2 and 3 hard?
     - Typically no closed form, high-dimensional integration
Bayesian inference

$$p(\theta|y_{1:N}) = \frac{p(y_{1:N} | \theta)p(\theta)}{p(y_{1:N})}$$

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   - Why are steps 2 and 3 hard?
     - Typically no closed form, high-dimensional integration
Bayesian inference

\[ p(\theta | y_{1:N}) = \frac{p(y_{1:N} | \theta)p(\theta)}{\int p(y_{1:N}, \theta)d\theta} \]

posterior \quad likelihood \quad prior \quad evidence

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   - Why are steps 2 and 3 hard?
     - Typically no closed form, high-dimensional integration
Approximate Bayesian Inference
Approximate Bayesian Inference

• Gold standard: Markov Chain Monte Carlo (MCMC) [Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
- Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

[23x697]Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$
Approximate Bayesian Inference

• Gold standard: Markov Chain Monte Carlo (MCMC)
  • Eventually accurate but can be slow

Instead: an optimization approach

• Approximate posterior with $q^*$

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

References:
[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach
- Approximate posterior with $q^*$

[Anonymous, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with \( q^* \)

\[
p(\theta|y) \rightarrow q^*(\theta)
\]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
- Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
- Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$
Approximate Bayesian Inference

• Gold standard: Markov Chain Monte Carlo (MCMC)
  • Eventually accurate but can be slow

Instead: an optimization approach

• Approximate posterior with $q^*$

$$q^* = \arg\min_{q \in \mathcal{Q}} f(q(\cdot), p(\cdot|y))$$

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$
  \[
  q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))
  \]

- Variational Bayes (VB): $f$ is Kullback-Leibler divergence
  \[
  KL(q(\cdot)||p(\cdot|y))
  \]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
- Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$

- Variational Bayes (VB): $f$ is Kullback-Leibler divergence

$$KL(q(\cdot)\|p(\cdot|y))$$
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
- Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with \( q^* \)
  \[
  q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))
  \]

- Variational Bayes (VB): \( f \) is Kullback-Leibler divergence
  \[
  KL(q(\cdot)||p(\cdot|y))
  \]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot | y))$$

- Variational Bayes (VB): $f$ is Kullback-Leibler divergence

$$KL(q(\cdot) \| p(\cdot | y))$$
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

\[
q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))
\]

- Variational Bayes (VB): $f$ is Kullback-Leibler divergence

\[
KL(q(\cdot)||p(\cdot|y))
\]

- VB practical success

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$
  
  $$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$

- Variational Bayes (VB): $f$ is Kullback-Leibler divergence

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

- VB practical success: point estimates and prediction

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$
  $$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$

- Variational Bayes (VB): $f$ is Kullback-Leibler divergence
  $$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success: point estimates and prediction, fast

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with \( q^* \)
  \[
  q^* = \arg\min_{q \in \mathcal{Q}} f(q(\cdot), p(\cdot|y))
  \]

- Variational Bayes (VB): \( f \) is Kullback-Leibler divergence
  \[
  KL(q(\cdot)||p(\cdot|y))
  \]

- VB practical success: point estimates and prediction, fast, streaming, distributed (3.6M Wikipedia, 350K Nature)

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

[Bardenet, Doucet, Holmes 2017]
Why KL?

- Variational Bayes

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

- Variational Bayes

\[ q^* = \arg \min_{q \in Q} \text{KL} (q(\cdot) \| p(\cdot|y)) \]

\[
\text{KL} (q(\cdot) \| p(\cdot|y)) := \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta
\]
Why KL?

- Variational Bayes

\[ q^* = \arg\min_{q \in Q} \text{KL} (q(\cdot) || p(\cdot | y)) \]

\[
\text{KL} (q(\cdot) || p(\cdot | y)) \\
:= \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta \\
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta
\]
Why KL?

- Variational Bayes

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

\[
\text{KL} (q(\cdot) \| p(\cdot \mid y)) \\
= \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid y)} d\theta \\
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta
\]
Why KL?

- Variational Bayes

\[ q^* = \arg \min_{q \in Q} KL(q(\cdot) || p(\cdot | y)) \]

\[
KL(q(\cdot) || p(\cdot | y)) = \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta = \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta
\]
Why KL?

- Variational Bayes

\[
q^* = \text{argmin}_{q \in Q} \text{KL} \left( q(\cdot) \parallel p(\cdot|y) \right)
\]

\[
\text{KL} \left( q(\cdot) \parallel p(\cdot|y) \right) := \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta
\]

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta
\]
Why KL?

- Variational Bayes

\[ q^* = \arg\min_{q \in Q} \text{KL} (q(\cdot) \| p(\cdot | y)) \]

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[ \log p(y) + \log \frac{q(\theta)}{p(\theta, y)} \right] d\theta
\]
Why KL?

• Variational Bayes

\[ q^* = \arg \min_{q \in Q} KL (q(\cdot) \| p(\cdot \mid y)) \]

\[
KL (q(\cdot) \| p(\cdot \mid y)) = \int q(\theta) \log \frac{q(\theta) p(y)}{p(\theta \mid y)} d\theta = \log p(y) + \int q(\theta) \log \frac{q(\theta)}{p(\theta, y)} d\theta
\]
Why KL?

- Variational Bayes

\[ q^* = \arg\min_{q \in Q} KL (q(\cdot) || p(\cdot | y)) \]

\[
KL (q(\cdot) || p(\cdot | y)) \\
:= \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta \\
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) + \int q(\theta) \log \frac{q(\theta)}{p(\theta, y)} d\theta
\]
Why KL?

- Variational Bayes

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

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\]
Why KL?

- Variational Bayes
  
  \[ q^* = \text{argmin}_{q \in Q} \text{KL} \left( q(\cdot) \parallel p(\cdot | y) \right) \]

  \[
  \text{KL} \left( q(\cdot) \parallel p(\cdot | y) \right) := \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta \\
  = \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
  \]
Why KL?

- Variational Bayes

\[ q^* = \text{argmin}_{q \in Q} KL (q(\cdot) \mid \mid p(\cdot \mid y)) \]

\[
KL \left( q(\cdot) \mid \mid p(\cdot \mid y) \right) := \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid y)} d\theta
\]

\[ = \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta \]
Why KL?

- Variational Bayes

\[ q^* = \arg\min_{q \in Q} KL (q(\cdot) \parallel p(\cdot | y)) \]

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\]
Why KL?

- Variational Bayes

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

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\]

"Evidence lower bound" (ELBO)
Why KL?

- Variational Bayes

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

\[
\text{KL} (q(\cdot) \| p(\cdot | y)) \]

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta
= \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\]

“Evidence lower bound” (ELBO)
Why KL?

- Variational Bayes
  
  \[ q^* = \arg\min_{q \in Q} \text{KL} (q(\cdot) \| p(\cdot|y)) \]

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

  \[
  = \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
  \]

- Exercise: Show \( \text{KL} \geq 0 \) [Bishop 2006, Sec 1.6.1]

  “Evidence lower bound” (ELBO)
Why KL?

- Variational Bayes

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

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\]

- Exercise: Show \( \text{KL} \geq 0 \) [Bishop 2006, Sec 1.6.1]

- \( \text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)
Why KL?

- Variational Bayes
  
  \[ q^* = \text{argmin}_{q \in Q} \text{KL} (q(\cdot) || p(\cdot|y)) \]

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

  \[ = \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta \]

- Exercise: Show \( \text{KL} \geq 0 \) [Bishop 2006, Sec 1.6.1]
- \( \text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)
- \( q^* = \text{argmax}_{q \in Q} \text{ELBO}(q) \)
Why KL?

- Variational Bayes
  \[ q^* = \text{argmin}_{q \in Q} \text{KL}(q(\cdot)||p(\cdot|y)) \]

\[
\text{KL}(q(\cdot)||p(\cdot|y)) \\
:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta \\
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\]

- Exercise: Show \( \text{KL} \geq 0 \) [Bishop 2006, Sec 1.6.1]
- \( \text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)
- \( q^* = \text{argmax}_{q \in Q} \text{ELBO}(q) \)
- Why KL (in this direction)?
Variational Bayes

\[ q^* = \arg\min_{q \in Q} \text{KL} (q(\cdot) || p(\cdot | y)) \]
Variational Bayes

\[ q^* = \arg\min_{q \in Q} \text{KL} (q(\cdot) \| p(\cdot | y)) \]

Choose “NICE” distributions
Variational Bayes

\[ q^* = \arg\min_{q \in \mathcal{Q}} \text{KL} (q(\cdot) \| p(\cdot|y)) \]

Choose “NICE” distributions
Variational Bayes

\[ q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|y)) \]

Choose “NICE” distributions
Variational Bayes

\[ q^* = \arg\min_{q \in Q} KL (q(\cdot) || p(\cdot|y)) \]

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

\[ Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\} \]
Variational Bayes

\[ q^* = \arg \min_{q \in Q} \text{KL} \left( q(\cdot) \| p(\cdot | y) \right) \]

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

\[ Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\} \]

- Often also exponential family
Variational Bayes

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

Choose “NICE” distributions
- Mean-field variational Bayes (MFVB)

\[ Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\} \]

- Often also exponential family
- Not a modeling assumption
Variational Bayes

\[ q^* = \arg\min_{q \in Q} KL(q(\cdot) || p(\cdot | y)) \]

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

\[ Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\} \]

- Often also exponential family
- Not a modeling assumption

[Bishop 2006]
Variational Bayes

\[ q^* = \arg\min_{q \in Q} KL (q(\cdot) || p(\cdot | y)) \]

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

\[ Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\} \]

- Often also exponential family
- *Not* a modeling assumption

Now we have an optimization problem; how to solve it?
Variational Bayes

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

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

- Often also exponential family
- Not a modeling assumption

Now we have an optimization problem; how to solve it?

- One option: Coordinate descent in $q_1, \ldots, q_J$
Approximate Bayesian inference
Approximate Bayesian inference

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

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

Optimization

$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot \mid y))$
Approximate Bayesian inference

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

Optimization

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot | y))$$

Variational Bayes

$$q^* = \arg\min_{q \in Q} KL(q(\cdot) || p(\cdot | y))$$
Approximate Bayesian inference

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

Optimization

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Variational Bayes

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)\|p(\cdot|y))$$

Mean-field variational Bayes

$$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)\|p(\cdot|y))$$
Approximate Bayesian inference

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

Optimization

$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot | y))$

Variational Bayes

$q^* = \arg\min_{q \in Q} KL(q(\cdot) || p(\cdot | y))$

Mean-field variational Bayes

$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot) || p(\cdot | y))$
Approximate Bayesian inference

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

Optimization
$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot | y))$$

Variational Bayes
$$q^* = \arg\min_{q \in Q} KL(q(\cdot) || p(\cdot | y))$$

Mean-field variational Bayes
$$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot) || p(\cdot | y))$$

- Coordinate descent
Approximate Bayesian inference

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

Optimization

$\begin{align*}
q^* &= \arg\min_{q \in Q} f(q(\cdot), p(\cdot | y)) \\
\end{align*}$

Variational Bayes

$\begin{align*}
q^* &= \arg\min_{q \in Q} KL(q(\cdot) || p(\cdot | y)) \\
\end{align*}$

Mean-field variational Bayes

$\begin{align*}
q^* &= \arg\min_{q \in Q_{MFVB}} KL(q(\cdot) || p(\cdot | y)) \\
\end{align*}$

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

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

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

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

Optimization

\[
q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))
\]

Variational Bayes

\[
q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|y))
\]

Mean-field variational Bayes

\[
q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))
\]

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

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

Optimization

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot | y))$$

Variational Bayes

$$q^* = \arg\min_{q \in Q} KL(q(\cdot) || p(\cdot | y))$$

Mean-field variational Bayes

$$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot) || p(\cdot | y))$$

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

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

Optimization

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Variational Bayes

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

Mean-field variational Bayes

$$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))$$

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

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

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

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

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

- Model:
  \[
p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N
  \]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and variance
- Model:
  \[ p(y|\theta) : y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N \]
Midge wing length

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

\[
p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N
\]
\[
p(\theta) : \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)
\]
\[
\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0\sigma^2)
\]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and variance
- Model (conjugate prior):
  \[
  \begin{align*}
  p(y|\theta) & : \ y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N \\
  p(\theta) & : \ (\sigma^2)^{-1} \sim \Gamma(a_0, b_0) \\
  \mu|\sigma^2 & \sim N(\mu_0, \lambda_0\sigma^2)
  \end{align*}
  \]

[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
- Model (conjugate prior): [Exercise: find the posterior] 
  \[
  \begin{align*}
  p(y|\theta) : & \\ y_n & \sim \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N \\
  p(\theta) : & \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0) \\
  \mu|\sigma^2 & \sim \mathcal{N}(\mu_0, \lambda_0\sigma^2)
  \end{align*}
  \]
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 (conjugate prior): [Exercise: find the posterior]
  \[
  p(y|\theta) : \quad y_n \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)
  \]
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): [Exercise: find the posterior]

\[
p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N
\]
\[
p(\theta) : \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)
\]
\[
\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)
\]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision
- Model (conjugate prior): [Exercise: find the posterior]
  \[ p(y|\theta) : \quad y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N \]
  \[ p(\theta) : \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0) \]
  \[ \mu | \sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2) \]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision
- Model (conjugate prior): [Exercise: find the posterior]
  \[
  p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
  \]
  \[
  p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
  \]
  \[
  \mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})
  \]

[Exercise: find the posterior]

[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
- Model (conjugate prior): [Exercise: find the posterior] \( \theta = (\mu, \tau) \)
  \[
p(y|\theta) : \quad y_n \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})
  
\]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision
- Model (conjugate prior): [Exercise: find the posterior] \( \theta = (\mu, \tau) \)
  
  \[
  p(y|\theta) : \quad y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
  \]
  
  \[
  p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
  \]
  
  \[
  \mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})
  \]

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

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision
- Model (conjugate prior): \( \theta = (\mu, \tau) \)
  \[
p(y|\theta) : \quad y_n \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})
  
\]
- Exercise: check \( p(\mu, \tau|y) \neq f_1(\mu, y)f_2(\tau, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \tau) = q^*_\mu(\mu)q^*_\tau(\tau) = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))
  \]
Midge wing length

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

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

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

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

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

\[ q^*(\mu, \tau) = q^*_\mu(\mu) q^*_\tau(\tau) = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y)) \]

- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]
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): [Exercise: find the posterior]
  \[
p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
  \]
  \[
p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
  \]
  \[
  \mu | \tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})
  \]
- Exercise: check \( p(\mu, \tau|y) \neq f_1(\mu, y)f_2(\tau, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \tau) = q^*_\mu(\mu)q^*_\tau(\tau) = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))
  \]
- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]
  \[
  q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \quad q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)
  \]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision
- Model (conjugate prior): [Exercise: find the posterior] \( \theta = (\mu, \tau) \)
  \[
p(y|\theta) : \quad y_n \sim \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
  \]
  \[
p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
  \mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})
  \]
- Exercise: check \( p(\mu, \tau|y) \neq f_1(\mu, y)f_2(\tau, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \tau) = q^*_\mu(\mu)q^*_\tau(\tau) = \arg\min_{q\in Q_{\text{MFVB}}} KL(q(\cdot)\|p(\cdot|y))
  \]
- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]
  \[
  q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \quad q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)
  \]
  “variational parameters”

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

approximation

exact posterior

[ Bishop 2006 ]
Midge wing length

approximation

exact posterior

\[ \tau \]

\[ \mu \]
Midge wing length approximation

Exact posterior
Midge wing length

approximation

exact posterior

\[ \tau \]

\[ \mu \]

[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 \( k \)th site (\(~900\) to \(~17\text{K}\))
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th 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 \( k \)th site (~900 to ~17K)
- Profit of \( n \)th business at \( k \)th site:
Microcredit Experiment

• Simplified from Meager (2018a)
• $K = 7$ 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:

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

- Simplified from Meager (2018a)
- $K = 7$ 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:

  \[ y_{kn} \sim \mathcal{N}(\mu_k + T_{kn} \times k, 2) \]
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ 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:

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

- Simplified from Meager (2018a)
- \( K = 7 \) 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 (2018a)
- \( K = 7 \) 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:

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

- Simplified from Meager (2018a)
- $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site (~900 to ~17K)
- Profit of $n$th business at $k$th site:

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

1 if microcredit
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ 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:

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

- Simplified from Meager (2018a)
- $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site (~900 to ~17K)
- Profit of $n$th business at $k$th site:

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

• Simplified from Meager (2018a)
• \( K = 7 \) 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:

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

• Priors and hyperpriors:

1 if microcredit

profit

\( y_{kn} \)
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ 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:

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

- Priors and hyperpriors:

$$
\begin{pmatrix}
\mu_k \\
\tau_k
\end{pmatrix}
\overset{\text{iid}}{\sim} \mathcal{N}
\begin{pmatrix}
\mu \\
\tau
\end{pmatrix}, C
$$
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ 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:

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

- 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 $k$th site (~900 to ~17K)
- Profit of $n$th business at $k$th site:

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

- Priors and hyperpriors:

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

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

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

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

MFVB: Do we need to check the output?
Microcredit

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

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

[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
Microcredit

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

Criteo Online Ads Experiment

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

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

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

Criteo Online Ads Experiment

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

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

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

Criteo Online Ads Experiment

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

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

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

Criteo Online Ads Experiment

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

[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2018]
Criteo Online Ads Experiment
Criteo Online Ads Experiment

- MAP: 12 s
Criteo Online Ads Experiment

- MAP: **12 s**
Criteo Online Ads Experiment

- MAP: 12 s
- MFVB: 57 s
Criteo Online Ads Experiment

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

[Giordano, Broderick, Jordan 2018]
Criteo Online Ads Experiment

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

[Giordano, Broderick, Jordan 2018]
Roadmap

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

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

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

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

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

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

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

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

- Conjugate linear regression

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

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

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

- Conjugate linear regression
- Bayesian central limit theorem

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

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

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

- Conjugate linear regression
- Bayesian central limit theorem
  
  [Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]
What about uncertainty?

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

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

- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]
What about uncertainty?

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

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

- Conjugate linear regression
- Bayesian central limit theorem
  
  [Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

- Underestimates variance (sometimes severely)
What about uncertainty?

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

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

- Conjugate linear regression
- Bayesian central limit theorem
  
  [Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

- Underestimates variance (sometimes severely)
What about uncertainty?

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

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

- Conjugate linear regression
- Bayesian central limit theorem

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

- Underestimates variance (sometimes severely)
- No covariance estimates
What about uncertainty?

• Microcredit
What about uncertainty?

- Microcredit
What about uncertainty?

- Microcredit effect
- \( \tau \) mean: 3.08 USD PPP
What about uncertainty?

- Microcredit effect
- $\tau$ mean: 3.08 USD PPP
- $\tau$ std dev: 1.83 USD PPP

[Giordano, Broderick, Meager, Huggins, Jordan 2016]
What about uncertainty?

- Microcredit effect
- \( \tau \) mean: 3.08 USD PPP
- \( \tau \) std dev: 1.83 USD PPP
- Mean is 1.68 std dev from 0

[Giordano, Broderick, Meager, Huggins, Jordan 2016]
What about uncertainty?

- Microcredit effect
- \( \tau \) mean: 3.08 USD PPP
- \( \tau \) std dev: 1.83 USD PPP
- Mean is 1.68 std dev from 0

- Criteo online ads experiment

[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2018]
What about means?

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

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

![Scatter plot of Means vs. MCMC](Fosdick 2013, Ch 4, Fig 4.3)
What about means?

- Model for relational data with covariates
- When 1000+ nodes, MCMC > 1 day

[Fosdick 2013, Ch 4]
Posterior means: revisited

- Want to predict college GPA $y_n$
Posterior means: revisited

• Want to predict college GPA $y_n$
• Collect: standardized test scores (e.g., SAT, ACT) $x_n$
Posterior means: revisited

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

[Giordano, Broderick, Jordan 2015]
Posterior means: revisited

- Want to predict college GPA $y_n$
- Collect: standardized test scores (e.g., SAT, ACT) $x_n$
- Collect: regional test scores $r_n$
- Model: $y_n | \beta, z, \sigma^2 \sim N(\beta^T x_n + z_k(n)r_n, \sigma^2)$
Posterior means: revisited

- Want to predict college GPA \( y_n \)
- Collect: standardized test scores (e.g., SAT, ACT) \( x_n \)
- Collect: regional test scores \( r_n \)
- Model:
  \[
  y_n | \beta, z, \sigma^2 \sim \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2) \\
  z_k | \rho^2 \sim \mathcal{N}(0, \rho^2) \\
  \sigma^2 \sim \Gamma(a_{\sigma^2}, b_{\sigma^2}) \\
  \rho^2 \sim \Gamma(a_{\rho^2}, b_{\rho^2})
  \]

[Giordano, Broderick, Jordan 2015]
Posterior means: revisited

- Want to predict college GPA $y_n$
- Collect: standardized test scores (e.g., SAT, ACT) $x_n$
- Collect: regional test scores $r_n$
- Model: $y_n | \beta, z, \sigma^2 \overset{\text{indep}}{\sim} \mathcal{N}(\beta^T x_n + z_k(n) r_n, \sigma^2)$
  
  $z_k | \rho^2 \overset{\text{iid}}{\sim} \mathcal{N}(0, \rho^2)$

  $(\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$

  $(\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$

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

  ![Graph showing the relationship between MFVB mean and MCMC mean](image)
Posterior means: revisited

- Want to predict college GPA $y_n$
- Collect: standardized test scores (e.g., SAT, ACT) $x_n$
- Collect: regional test scores $r_n$
- Model:
  \[
  y_n | \beta, z, \sigma^2 \overset{\text{indep}}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)
  \]
  \[
  z_k | \rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2)
  \]
  \[
  \rho^2 \overset{iid}{\sim} \text{Gamma}(a_{\rho^2}, b_{\rho^2})
  \]
  \[
  (\sigma^2)^{-1} \overset{\text{Gamma}}{\sim} \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})
  \]

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

- Want to predict college GPA $y_n$
- Collect: standardized test scores (e.g., SAT, ACT) $x_n$
- Collect: regional test scores $r_n$
- Model: $y_n | \beta, z, \sigma^2 \sim \mathcal{N}(\beta^T x_n + z_k(n) r_n, \sigma^2)$
  $z_k | \rho^2 \sim \mathcal{N}(0, \rho^2)$
  $(\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$
  $\beta \sim \mathcal{N}(0, \Sigma)$
  $(\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$

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

• Want to predict college GPA \( y_n \)
• Collect: standardized test scores (e.g., SAT, ACT) \( x_n \)
• Collect: regional test scores \( r_n \)
• Model:
  \[
  \begin{align*}
  y_n | \beta, z, \sigma^2 & \sim \mathcal{N}(\beta^T x_n + z_k(n)r_n, \sigma^2) \\
  z_k | \rho^2 & \sim \mathcal{N}(0, \rho^2) \\
  \beta & \sim \mathcal{N}(0, \Sigma) \\
  (\sigma^2)^{-1} & \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2}) \\
  (\rho^2)^{-1} & \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})
  \end{align*}
  \]

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

[Giordano, Broderick, Jordan 2015]
Approximate Bayesian inference

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

Optimization

\[ q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot | y)) \]

Variational Bayes

\[ q^* = \arg\min_{q \in Q} KL(q(\cdot) || p(\cdot | y)) \]

Mean-field variational Bayes

\[ q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot) || p(\cdot | y)) \]

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

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

Optimization

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot | y))$$

Variational Bayes

$$q^* = \arg\min_{q \in Q} KL(q(\cdot) || p(\cdot | y))$$

Mean-field variational Bayes

$$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot) || p(\cdot | y))$$

Algorithm

How deep is the issue?
Optimization
\[ q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y)) \]

Variational Bayes
\[ q^* = \arg\min_{q \in Q} KL(q(\cdot) || p(\cdot|y)) \]

Mean-field variational Bayes
\[ q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot) || p(\cdot|y)) \]

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

How deep is the issue?
Approximate Bayesian inference

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

Optimization

$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$

Variational Bayes

$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|y))$

Mean-field variational Bayes

$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))$

Algorithm

Implementation

How deep is the issue?
Approximate Bayesian inference

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

Optimization

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Variational Bayes

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

Mean-field variational Bayes

$$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))$$

How deep is the issue?

Algorithm

Implementation

Gaussian example was exact

Implementation
Approximate Bayesian inference

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

Optimization

$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$

Variational Bayes

$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|y))$

Mean-field variational Bayes

$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))$

Algorithm

Implementation

Gaussian example was exact

How deep is the issue?
Approximate Bayesian inference

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

Optimization

$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot | y))$

Variational Bayes

$q^* = \arg\min_{q \in Q} KL(q(\cdot) || p(\cdot | y))$

Mean-field variational Bayes

$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot) || p(\cdot | y))$

Algorithm

Implementation

How deep is the issue?

Gaussian example was exact
Is it just MFVB?
Is it just MFVB?
Is it just MFVB?
Is it just MFVB?

$p(\theta|y)$

$q^*(\theta)$

NICE
Is it just MFVB?

• Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates
Is it just MFVB?

- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates.
Is it just MFVB?

- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates
Is it just MFVB?

- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates

- Exercise: Show, with a simple example, that a smaller KL does not imply better mean and variance estimates

[Baqué et al 2017; Huggins, Karsprzak, Campbell, Broderick 2019]
Is it just MFVB?

- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates

- Exercise: Show, with a simple example, that a smaller KL does not imply better mean and variance estimates

- But how much worse can the estimates be? And could it have just been the implementation?

[Baqué et al 2017; Huggins, Karsprzak, Campbell, Broderick 2019]
Is it just MFVB?
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3

[Huggins, Karsprzak, Campbell, Broderick 2019]

[Baqué et al 2017; Huggins et al 2019]
Is it just MFVB?

• Some KL values seen in practice: ~1 to ~70, 0.5 to 3
  [Baqué et al 2017; Huggins et al 2019]
• Take any constant $c$
Is it just MFVB?

• Some KL values seen in practice:
  ~1 to ~70, 0.5 to 3
  \[\text{[Baqué et al 2017; Huggins et al 2019]}\]
• Take any constant \(c\)

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

\[\sigma_p^2 \geq c\sigma_q^2\]
Is it just MFVB?

- Some KL values seen in practice: \(\sim 1\) to \(\sim 70\), \(0.5\) to \(3\) [Baqué et al 2017; Huggins et al 2019]
- Take any constant \(c\)

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

\[
\sigma_p^2 \geq c\sigma_q^2
\]
Is it just MFVB?

- Some KL values seen in practice: \(\sim 1\) to \(\sim 70\), 0.5 to 3
  \[\text{[Baqué et al 2017; Huggins et al 2019]}\]
- Take any constant \(c\)

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

\[
\sigma_p^2 \geq c \sigma_q^2
\]

\(\rho: \text{Student's t. variance } \sigma_p^2\)

[Huggins, Karsprzak, Campbell, Broderick 2019]
Is it just MFVB?

- Some KL values seen in practice: 
  ~1 to ~70, 0.5 to 3
- Take any constant $c$

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate 

$$\sigma_p^2 \geq c\sigma_q^2$$

$q$: Gaussian, variance $\sigma_q^2$

$p$: Student's t, variance $\sigma_p^2$
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
- Take any constant $c$

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

$$\sigma_p^2 \geq c\sigma_q^2$$

$q$: Gaussian, variance $\sigma_q^2$

$p$: Student's t, variance $\sigma_p^2$
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
- Take any constant $c$

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

$$\sigma_p^2 \geq c \sigma_q^2$$
Is it just MFVB?

- Some KL values seen in practice: \( \sim 1 \) to \( \sim 70 \), 0.5 to 3
- Take any constant \( c \)

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

\[
\sigma_p^2 \geq c\sigma_q^2
\]
Is it just MFVB?

• Some KL values seen in practice: ~1 to ~70, 0.5 to 3
  [Baqué et al 2017; Huggins et al 2019]
• Take any constant $c$

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

$$\sigma_p^2 \geq c\sigma_q^2$$
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
  [Baqué et al 2017; Huggins et al 2019]
- Take any constant $c$

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

$\sigma_p^2 \geq c\sigma_q^2$

$q$: Gaussian, variance $\sigma_q^2$

$p$: Student's t, variance $\sigma_p^2$

**Proposition.** Can have small KL (<0.9) and arbitrarily bad mean estimate

$(m_p - m_q)^2 \geq c\sigma_p^2$

[Huggins, Karsprzak, Campbell, Broderick 2019]
Is it just MFVB?

- Some KL values seen in practice: \( \sim 1 \) to \( \sim 70 \), 0.5 to 3
- Take any constant \( c \)

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

\[
\sigma_p^2 \geq c \sigma_q^2
\]

[Baqué et al 2017; Huggins et al 2019]

\[
p: \text{Student's t, variance } \sigma_p^2
\]

\[
q: \text{Gaussian, variance } \sigma_q^2
\]

**Proposition.** Can have small KL (<0.9) and arbitrarily bad mean estimate

\[
(m_p - m_q)^2 \geq c \sigma_p^2
\]

[Huggins, Karsprzak, Campbell, Broderick 2019]
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
- Take any constant $c$

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

$$
\sigma_p^2 \geq c\sigma_q^2
$$

$q$: Gaussian, variance $\sigma_q^2$

$p$: Student's t, variance $\sigma_p^2$

$p$: Weibull, mean $m_p$

**Proposition.** Can have small KL (<0.9) and arbitrarily bad mean estimate

$$
(m_p - m_q)^2 \geq c\sigma_p^2
$$

[Huggins, Karsprzak, Campbell, Broderick 2019]
Is it just MFVB?

- Some KL values seen in practice: \(~1\) to \(~70\), \(0.5\) to \(3\)
- Take any constant \(c\)

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

\[
\sigma_p^2 \geq c\sigma_q^2
\]

- \(p\): Weibull, mean \(m_p\)
- \(q\): Weibull, mean \(m_q\)

**Proposition.** Can have small KL (<0.9) and arbitrarily bad mean estimate

\[
(m_p - m_q)^2 \geq c\sigma_p^2
\]
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
- Take any constant \( c \)

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate
\[
\sigma_p^2 \geq c\sigma_q^2
\]

\( p \): Student's t, variance \( \sigma_p^2 \)
\( q \): Gaussian, variance \( \sigma_q^2 \)

\( p \): Weibull, mean \( m_p \)
\( q \): Weibull, mean \( m_q \)

**Proposition.** Can have small KL (<0.9) and arbitrarily bad mean estimate
\[
(m_p - m_q)^2 \geq c\sigma_p^2
\]

[Huggins, Karsprzak, Campbell, Broderick 2019]
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
- Take any constant $c$

**Proposition.** Can have small KL ($<0.23$) & arbitrarily bad variance estimate

\[
\sigma_p^2 \geq c\sigma_q^2
\]

$q$: Gaussian, variance $\sigma_q^2$

$p$: Student's t, variance $\sigma_p^2$

$p$: Weibull, mean $m_p$

$q$: Weibull, mean $m_q$

**Proposition.** Can have small KL ($<0.9$) and arbitrarily bad mean estimate

\[
(m_p - m_q)^2 \geq c\sigma_p^2
\]

[Baque et al 2017; Huggins et al 2019]

[Huggins, Karsprzak, Campbell, Broderick 2019]
Approximate Bayesian inference

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

Optimization

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Variational Bayes

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

Mean-field variational Bayes

$$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))$$

Algorithm

Implementation

Gaussian example was exact

How deep is the issue?
Approximate Bayesian inference

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

Optimization

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Variational Bayes

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

Mean-field variational Bayes

$$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))$$

How deep is the issue?

Algorithm

Implementation

Gaussian example was exact
Roadmap

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

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

Approximate posterior

Optimize: closest nice distr.

Variational Bayes

Mean-field variational Bayes
What can we do?

- Corrections [Giordano, Broderick, Jordan 2015, 2018]
What can we do?

- Corrections [Giordano, Broderick, Jordan 2015, 2018]
- Theoretical guarantees on finite-data quality [Huggins, Campbell, Broderick 2016; Huggins, Campbell, Kasprzak, Broderick, 2018; Campbell, Broderick 2018, 2019]
What can we do?

- Corrections [Giordano, Broderick, Jordan 2015, 2018]
- Theoretical guarantees on finite-data quality [Huggins, Campbell, Broderick 2016; Huggins, Campbell, Kasprzak, Broderick, 2018; Campbell, Broderick 2018, 2019]
- Reliable diagnostics
What can we do?

• Corrections [Giordano, Broderick, Jordan 2015, 2018]

• Theoretical guarantees on finite-data quality
  [Huggins, Campbell, Broderick 2016; Huggins, Campbell, Kasprzak, Broderick, 2018; Campbell, Broderick 2018, 2019]

• Reliable diagnostics
  • cf. KL
What can we do?

- **Corrections**
  - [Giordano, Broderick, Jordan 2015, 2018]

- **Theoretical guarantees on finite-data quality**
  - [Huggins, Campbell, Broderick 2016; Huggins, Campbell, Kasprzak, Broderick, 2018; Campbell, Broderick 2018, 2019]

- **Reliable diagnostics**
  - cf. KL
What can we do?

- Corrections [Giordano, Broderick, Jordan 2015, 2018]

- Theoretical guarantees on finite-data quality [Huggins, Campbell, Broderick 2016; Huggins, Campbell, Kasprzak, Broderick, 2018; Campbell, Broderick 2018, 2019]

- Reliable diagnostics
  - cf. KL, ELBO
What can we do?

- Corrections
  [Giordano, Broderick, Jordan 2015, 2018]

- Theoretical guarantees on finite-data quality
  [Huggins, Campbell, Broderick 2016; Huggins, Campbell, Kasprzak, Broderick, 2018; Campbell, Broderick 2018, 2019]

- Reliable diagnostics
  - cf. KL, ELBO

approximate posterior
Optimize: closest nice distr.
Variational Bayes
Mean-field variational Bayes

KL
ELBO

iteration
iteration

18,067
What can we do?

• Corrections
  [Giordano, Broderick, Jordan 2015, 2018]

• Theoretical guarantees on finite-data quality
  [Huggins, Campbell, Broderick 2016; Huggins, Campbell, Kasprzak, Broderick, 2018; Campbell, Broderick 2018, 2019]

• Reliable diagnostics
  • cf. KL, ELBO
  [Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]

"Yes, but did it work? Evaluating variational inference" ICML 2018
[Huggins, Kasprzak, Campbell, Broderick, 2019]

"Practical posterior error bounds from variational objectives"
Bayesian inference

- Goals: good point estimates, uncertainty estimates
- Challenge: speed (compute, user), reliable inference
What to read next

Textbooks and Reviews


Our Experiments

References (1/6)


References (2/6)


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)