Variational Bayes and beyond: Bayesian inference for big data

Tamara Broderick
ITT Career Development
Assistant Professor, MIT

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

![Diagram](image)

[Gillon et al 2017]

[Grimm et al 2018]
Bayesian inference

[Abbott et al 2016a,b]

[ESO/
L. Calçada/
M. Kornmesser
2017]

[Gillon et al 2017]

[Grimm et al 2018]
Bayesian inference

[Image 1x334 to 1014x597]

[Abbott et al 2016a,b]
[ESO/ L. Calçada/ M. Kornmesser 2017]  [Gillon et al 2017]
[Stone et al 2014]
Bayesian inference

[Woodard et al 2017]

[Gillon et al 2017]

[Grimm et al 2018]

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

[Stone et al 2014]
Bayesian inference
Bayesian inference
Bayesian inference
Bayesian inference

- Analysis goals: Point estimates, coherent uncertainties
- Interpretable, complex, modular; expert information
Bayesian inference

- Analysis goals: Point estimates, coherent uncertainties
- Interpretable, complex, modular; expert information
Bayesian inference

- Analysis goals: Point estimates, coherent uncertainties
- Interpretable, complex, modular; expert information

- Challenge: fast (compute, user), reliable inference
Bayesian inference

- Analysis goals: Point estimates, coherent uncertainties
- Interpretable, complex, modular; expert information

- Challenge: fast (compute, user), reliable inference
- Uncertainty doesn’t have to disappear in large data sets

[Image 1: Quote from Gillon et al. 2017]
[Image 2: Quote from Grimm et al. 2018]
[Image 3: Quote from Stone et al. 2014]
[Image 4: Quote from Woodard et al. 2017]
[Image 5: Quote from Chati, Balakrishnan 2017]
[Image 6: Quote from Baltic Salmon Fund 2014]
[Image 7: Quote from mc-stan.org 2016]
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

[Blei et al 2003]
Variational Bayes

• Modern problems: often large data, large dimensions

• Variational Bayes can be very fast

[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.

[Airoldi et al 2008]
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.

[Airoldi et al 2008] [Gershman et al 2014] [Blei et al 2018]
Variational Bayes

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

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

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

[Blei et al 2003]

[Stegle et al 2010]

[Xing et al 2004]

[Xing 2003]

[2]

[Airoldi et al 2008]

[Gershman et al 2014]

[Blei et al 2018]
Roadmap

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

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

$p(\theta)$

prior

parameters
Bayesian inference

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

likelihood prior

parameters
Bayesian inference

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

likelihood prior

data parameters
Bayesian inference

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

posterior likelihood prior

data parameters

\[ \theta \]

\[ y_{1:N} \]
Bayesian inference

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

posterior likelihood prior

Bayes Theorem

Bayesian inference

data

parameters
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 likelihood prior

1. Build a model: choose prior & choose likelihood

Bayes Theorem
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
Bayesian inference

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

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

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

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?
Bayesian inference

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

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
Bayesian inference

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

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, 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

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

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

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 $q^*$

[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 $q^*$

[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 $q^*$

[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 $q^*$

[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 $q^*$

[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 $q^*$

[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 $q^*$

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

[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 $q^*$

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

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

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

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

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

- Variational Bayes

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

• Variational Bayes

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

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

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

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

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

- Variational Bayes

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

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

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

- Variational Bayes

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

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

- Variational Bayes

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

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

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

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

"Evidence lower bound" (ELBO)
Why KL?

- Variational Bayes

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

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

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

"Evidence lower bound" (ELBO)
Why KL?

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

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

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

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

“Evidence lower bound” (ELBO)
Why KL?

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

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

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

- **Exercise:** Show \( \operatorname{KL} \geq 0 \) [Bishop 2006, Sec 1.6.1]
- \( \operatorname{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)
Why KL?

- Variational Bayes

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

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

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

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

- Variational Bayes

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

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

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

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

-  \( q^* = \operatorname{argmax}_{q \in Q} \text{ELBO}(q) \)

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

Variational Bayes
Variational Bayes

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

Choose “NICE” distributions
Variational Bayes

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

Choose “NICE” distributions
Variational Bayes

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

Choose “NICE” distributions
Variational Bayes

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

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

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

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

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

- Often also exponential family
Variational Bayes

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

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

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

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?

[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?

- One option: Coordinate descent in \( q_1, \ldots, q_J \)

[Bishop 2006]
Approximate Bayesian inference
Approximate Bayesian inference

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

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

Optimization

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

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

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

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]
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]
Roadmap

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

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

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

[CSIRO 2004]
Midge wing length

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

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

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

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

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

[CSIRO 2004]
Midge wing length

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

[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:
  
  \[
  p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
  \]
  
  \[
  p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
  \]
  
  \[
  \mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})
  \]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision
- Model:
  \[
  \begin{align*}
  p(y|\theta) : & \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N \\
  p(\theta) : & \quad \tau \sim \text{Gamma}(a_0, b_0) \\
  & \quad \mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})
  \end{align*}
  \]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision \( \theta = (\mu, \tau) \)
- Model:
  \[
p(y|\theta) : \quad y_n \sim iid \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
\]
  \[
p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
\]
  \[
\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})
\]
- 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 \( \theta = (\mu, \tau) \)
- Model:
  \[
p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
  
p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
  
  \mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})
\]
- Exercise: check \( p(\mu, \tau|y) \neq f_1(\mu, y)f_2(\tau, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \tau) = q^*_\mu(\mu)q^*_\tau(\tau) = \arg\min_{q\in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))
  \]

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

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

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision \( \theta = (\mu, \tau) \)
- Model:
  \[
p(y|\theta) : \quad y_n \sim \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
  \]
  \[
p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
  \]
  \[
  \mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})
  \]
- 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]
  \[
  q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \quad q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)
  \]
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model:
  
  $p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N$

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

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

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

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

- 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

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision \( \theta = (\mu, \tau) \)
- Model:
  \[
  p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N \\
  p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0) \\
  \mu | \tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})
  \]
- 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]
  \[
  q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \\
  q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)
  \]
- Iterate: \( (\mu_N, \rho_N) = f(a_N, b_N) \)
  \[
  (a_N, b_N) = g(\mu_N, \rho_N)
  \]

[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] \( \theta = (\mu, \tau) \)

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

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

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

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

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

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

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

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

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

"variational parameters"

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

\[ \tau \]

exact posterior

\[ \mu \]

[Bishop 2006]
Midge wing length approximation

\[ \tau \]

exact posterior

\[ \mu \]
Midge wing length approximation

exact posterior

[Reference: Bishop 2006]
Midge wing length approximation

exact posterior
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 ~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)
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:

$$y_{kn}$$
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:

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

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

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

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

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

profit

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:

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

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:

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

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:

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

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:

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

- Priors and hyperpriors:
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site (~900 to ~17K)
- Profit of $n$th business at $k$th site:
  \[
  y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)
  \]
  \[
  \mu_k, \tau_k \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, C\right)
  \]
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:
  \[
  \begin{pmatrix}
  \mu_k \\
  \tau_k
  \end{pmatrix}
  \overset{iid}{\sim} \mathcal{N}
  \left(
  \begin{pmatrix}
  \mu \\
  \tau
  \end{pmatrix},
  C
  \right)
  \\
  \sigma_k^{-2} \overset{iid}{\sim} \Gamma(a, b)
  \]
  1 if microcredit

profit

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

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

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

- Priors and hyperpriors:

\[
\begin{align*}
\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} & \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, C\right) \\
\begin{pmatrix} \mu \\ \tau \end{pmatrix} & \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu_0 \\ \tau_0 \end{pmatrix}, \Lambda^{-1}\right) \\
\sigma_k^{-2} & \overset{\text{iid}}{\sim} \Gamma(a, b) \\
C & \sim \text{Sep&LKJ}(\eta, c, d)
\end{align*}
\]
Microcredit

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

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

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

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

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

Criteo Online Ads Experiment

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

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

Criteo Online Ads Experiment

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

[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

Global parameters (-τ)

- MAP: $12 \text{ s}$
- MFVB: $57 \text{ s}$
Criteo Online Ads Experiment

- **MAP:** 12 s
- **MFVB:** 57 s
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 MFVB?
• When can we trust MFVB?
• Where do we go from here?
Roadmap

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

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

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

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

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

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

• 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?

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

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

- Underestimates variance (sometimes severely)
- Conjugate linear regression
- Bayesian central limit theorem

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

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

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

- Underestimates variance (sometimes severely)
- Conjugate linear regression
- Bayesian central limit theorem

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

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

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

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

- Underestimates variance (sometimes severely)
- No covariance estimates
- Conjugate linear regression
- Bayesian central limit theorem

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

[Image of graphs showing means comparison between MFVB and MCMC]

[Fosdick 2013, Ch 4, Fig 4.3]
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 $r_n$
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: \( y_n | \beta, z, \sigma^2 \overset{\text{indep}}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2) \)
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:
  \[ y_n | \beta, z, \sigma^2 \sim \mathcal{N}(\beta^T x_n + z_k(n) r_n, \sigma^2) \]
  \[ z_k | \rho^2 \sim \mathcal{N}(0, \rho^2) \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2}) \]
  \[ \beta \sim \mathcal{N}(0, \Sigma) \quad (\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2}) \]
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: $y_n | \beta, z, \sigma^2 \sim \text{N}(\beta^T x_n + z_k(n) r_n, \sigma^2)$
  
  $z_k | \rho^2 \sim \text{N}(0, \rho^2)$
  
  $(\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$
  
  $\beta \sim \text{N}(0, \Sigma)$
  
  $(\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$

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

[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: $y_n | \beta, z, \sigma^2 \overset{\text{indep}}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$
  
  $z_k | \rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2)$
  
  $(\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$
  
  $\beta \sim \mathcal{N}(0, \Sigma)$
  
  $(\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$

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

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

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

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

Can have small KL and arbitrarily bad variance estimate

**Proposition (HKCB).** For any $t > 1$, there exist zero-mean, unimodal distributions $q^*$ and $p$ such that

$$KL(q^* || p) < 0.802 \quad \text{but also} \quad \sigma_p^2 \geq t \sigma_{q^*}^2$$
Is it just MFVB?

Can have small KL and arbitrarily bad variance estimate

**Proposition (HKCB).** For any $t > 1$, there exist zero-mean, unimodal distributions $q^*$ and $p$ such that

$$KL(q^* || p) < 0.802 \quad \text{but also} \quad \sigma^2_p \geq t\sigma^2_{q^*}$$

Can have small KL and坏mean estimate

**Proposition (HKCB).** Take any $\delta > 0$. There exist Gaussian distributions $q^*$ and $p$ such that

$$KL(q^* || p) = \delta \quad \text{but also} \quad (\mu_p - \mu_{q^*})^2 = \sigma^2_{q^*} \{\exp(2\delta) - 1\}$$
Is it just MFVB?

Can have small KL and arbitrarily bad variance estimate

**Proposition (HKCB).** For any $t > 1$, there exist zero-mean, unimodal distributions $q^*$ and $p$ such that

$$KL(q^*||p) < 0.802$$

but also

$$\sigma_p^2 \geq t\sigma_{q^*}^2$$

Can have small KL and bad mean estimate

**Proposition (HKCB).** Take any $\delta > 0$. There exist Gaussian distributions $q^*$ and $p$ such that

$$KL(q^*||p) = \delta$$

but also

$$(\mu_p - \mu_{q^*})^2 = \sigma_{q^*}^2 \{\exp(2\delta) - 1\}$$
Roadmap

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

- Reliable diagnostics
What can we do?

- Reliable diagnostics
- KL vs ELBO
What can we do?

- Reliable diagnostics
- KL vs ELBO
What can we do?

- Reliable diagnostics
- KL vs ELBO
What can we do?

• Reliable diagnostics

• KL vs 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
What can we do?

- Reliable diagnostics
- KL vs 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

Optimize: closest nice distr.

Variational Bayes

Mean-field variational Bayes

Approximate posterior
What can we do?

- Reliable diagnostics
- KL vs 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

- Richer “nice” set; alternative divergences
What can we do?

- Reliable diagnostics

- KL vs 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

- Richer “nice” set; alternative divergences
  
  [Turner, Sahani 2011]
What can we do?

- Reliable diagnostics
- KL vs ELBO

[Turner, Sahani 2011]

“Yes, but did it work? Evaluating variational inference” ICML 2018

- Richer “nice” set; alternative divergences

[Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]

Approximate posterior
Optimize: closest nice distr.

Variational Bayes
Mean-field variational Bayes
What can we do?

- Reliable diagnostics
- KL vs 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

- Richer “nice” set; alternative divergences
  - [Turner, Sahani 2011]
  - [Huggins, Kasprzak, Campbell, Broderick, 2018]

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

- Reliable diagnostics
- KL vs ELBO
  - Richer “nice” set; alternative divergences
  - Corrections
  - Theoretical guarantees on finite-data quality [seminar]

“Yes, but did it work? Evaluating variational inference” ICML 2018

Approximate posterior
Optimize: closest nice distr.

Variational Bayes
Mean-field variational Bayes

[Turner, Sahani 2011]
[Huggins, Kasprzak, Campbell, Broderick, 2018]
[Giordano, Broderick, Jordan 2018]
What to read next

Textbooks and Reviews


More Experiments


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)