Variational Bayes and beyond: Foundations of scalable Bayesian inference

Tamara Broderick
Associate Professor
MIT

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

![Graph showing eccentricity vs mass with annotations for Grimm et al. 2018 and Gillon et al. 2017]
Bayesian inference

[Abbott et al 2016a,b] [ESO/L. Calçada/M. Kornmesser 2017] [Gillon et al 2017] [Grimm et al 2018]
Bayesian inference
Bayesian inference

[Grimm et al 2018]

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

[Woodard et al 2017]

[Stone et al 2014]
Bayesian inference
Bayesian inference


Bayesian inference


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>
<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]
Variational Bayes

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

<table>
<thead>
<tr>
<th>&quot;Arts&quot;</th>
<th>&quot;Budgets&quot;</th>
<th>&quot;Children&quot;</th>
<th>&quot;Education&quot;</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>MANAGT</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, the foundation’s president. 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

[Variational Bayes]

Variational Bayes

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

[Blei et al 2003] [Stegle et al 2010] [Gershman et al 2014] [Airoldi et al 2008] [Blei et al 2018]
Variational Bayes

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

“Arts” “Budgets” “Children” “Education”
NEW MILLION CHILDREN SCHOOL [Blei et al 2003]
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

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolita

an 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 support

of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000

donation, too.
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

\[ \theta \rightarrow \text{parameters} \]
Bayesian inference

$p(\theta)$

prior

parameters
Bayesian inference

\[ p(\theta) \]

prior

parameters

\[ \theta \]

Parameters and prior distribution in Bayesian inference.
Bayesian inference

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

likelihood prior

\[ \theta \]
Bayesian inference

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

likelihood prior
Bayesian inference

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

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

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

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
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

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

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
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?
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
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?
     - 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

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

- posterior
- likelihood
- prior
- 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
Bayesian inference

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

Posterior, likelihood, prior, 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 $q^*$

[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 $q^*$

$\Pr(\theta|y)$

$\text{NICE}$

[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 $q^*$

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

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

Instead: an optimization approach

- Approximate posterior with \( q^* \)

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

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

---

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

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

- 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

4

[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)
Why KL?

- Variational Bayes

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

- Variational Bayes

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

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

- Variational Bayes

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

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

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

- Variational Bayes

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

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

- Variational Bayes

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

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

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

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

- Variational Bayes

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

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

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

- Variational Bayes

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

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

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

- Variational Bayes

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

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

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

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

- Variational Bayes

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

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

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

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

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

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

- Variational Bayes

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

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

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

- Variational Bayes

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

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

"Evidence lower bound" (ELBO)
Why KL?

- Variational Bayes

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

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

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

“Evidence lower bound” (ELBO)
Why KL?

- Variational Bayes

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

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

- Exercise: Show \( \text{KL} \geq 0 \) \footnote{Bishop 2006, Sec 1.6.1}
Why KL?

- Variational Bayes

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

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

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

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

- \( KL \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)
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]
- \( \operatorname{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)
- \( q^* = \operatorname{argmax}_{q \in Q} \text{ELBO}(q) \)
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

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

Choose “NICE” distributions
Variational Bayes

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

Choose “NICE” distributions
Variational Bayes

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

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

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

\[ p(\theta | y) \]

\[ q^*(\theta) \]

NICE

FAR
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
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

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

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

[Bishop 2006]
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

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

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 $p(\cdot | y)$

Optimization

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

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

Optimization

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

Variational Bayes

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

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

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

Approximate Bayesian inference
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

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

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]
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?
Air pollution: Particulate matter

[Krongut 2020]
Air pollution: Particulate matter

- Sensor readings of log PM2.5 \( y = (y_1, \ldots, y_N) \)

- Model:
  \[
P(y|\theta) : \ y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)
  \]
Air pollution: Particulate matter

- Sensor readings of log PM2.5 \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: PM2.5 mean and variance
- Model:
  \[
p(y|\theta) : y_n \sim iid N(\mu, \sigma^2)
\]
Air pollution: Particulate matter

- Sensor readings of log PM2.5  \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: PM2.5 mean and variance
- Model:
  
  \[
  p(y|\theta) \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \\
  p(\theta) \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0) \\
  \mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0\sigma^2)
  \]

\( \theta = (\mu, \sigma^2) \)
Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and variance
- Model (conjugate prior):
  
  $p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2),$
  
  $p(\theta) : \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$
  
  $\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0\sigma^2)$

$\theta = (\mu, \sigma^2)$

[Krongut 2020]

[MacKay 2003; Bishop 2006]
Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and variance
- Model (conjugate prior): [Exercise: find the posterior]  
  $p(y|\theta): y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$,  
  $p(\theta): (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$  
  $\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0\sigma^2)$

[MacKay 2003; Bishop 2006]
Air pollution: Particulate matter

- Sensor readings of log PM2.5 \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: PM2.5 mean and variance \( \theta = (\mu, \sigma^2) \)
- Model (conjugate prior): [Exercise: find the posterior]
  \[
  p(y|\theta) : \quad y_n \overset{iid}{\sim} N(\mu, \sigma^2),
  \]
  \[
  p(\theta) : \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)
  \]
  \[
  \mu|\sigma^2 \sim N(\mu_0, \lambda_0\sigma^2)
  \]

[MacKay 2003; Bishop 2006]
Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]
  \[ p(y|\theta) : y_n \sim \mathcal{N}(\mu, \sigma^2), \]
  \[ p(\theta) : (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0) \]
  \[ \mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0\sigma^2) \]

[MacKay 2003; Bishop 2006] [Krongut 2020]
Air pollution: Particulate matter

- Sensor readings of log PM2.5 \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: PM2.5 mean and precision
- Model (conjugate prior): [Exercise: find the posterior] \( \theta = (\mu, \tau) \)

\[
p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2),
\]

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

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

[MacKay 2003; Bishop 2006]
Air pollution: Particulate matter

- Sensor readings of log PM2.5 \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: PM2.5 mean and precision
- Model (conjugate prior): [Exercise: find the posterior]

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

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

[Krongut 2020] [MacKay 2003; Bishop 2006]
Air pollution: Particulate matter

- Sensor readings of log PM2.5 \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: PM2.5 mean and precision
- Model (conjugate prior): [Exercise: find the posterior]
  \[
p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \\
p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0) \\
\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})
\]

\[\theta = (\mu, \tau)\] [MacKay 2003; Bishop 2006] [Krongut 2020]
Air pollution: Particulate matter

- Sensor readings of log PM2.5 \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: PM2.5 mean and precision
- Model (conjugate prior): [Exercise: find the posterior] \( \theta = (\mu, \tau) \)
  \[
p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}),
  \]
  \[
p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
  \]
  \[
  \mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})
  \]
- Exercise: check
  \[
p(\mu, \tau|y) \neq f_1(\mu, y)f_2(\tau, y)
  \]

\[\text{[MacKay 2003; Bishop 2006]}\]
Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision
- Model (conjugate prior): $p(y|\theta)$: $y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1})$,
  $p(\theta)$: $\tau \sim \text{Gamma}(a_0, b_0)$
  $\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})$
- Exercise: check $p(\mu, \tau|y) \neq f_1(\mu, y)f_2(\tau, y)$
- MFVB approximation:
  $q^*(\mu, \tau) = q^*_\mu(\mu)q^*_\tau(\tau) = \arg\min_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$
Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision
- Model (conjugate prior): [Exercise: find the posterior]
  $$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}),$$
  $$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$
  $$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})$$
- 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]
Air pollution: Particulate matter

- Sensor readings of log PM2.5 \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: PM2.5 mean and precision
- Model (conjugate prior): [Exercise: find the posterior] \( \theta = (\mu, \tau) \)
  \[
p(y|\theta) : \quad y_n \sim iid \mathcal{N}(\mu, \tau^{-1}),
  
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] \([\text{Bishop 2006, Sec 10.1.3}]\)
  \[
  q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1})
  
  q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)
  
  [\text{MacKay 2003; Bishop 2006}]\]
Air pollution: Particulate matter

- Sensor readings of log PM2.5 $y = (y_1, \ldots, y_N)$
- Parameters of interest: PM2.5 mean and precision
- Model (conjugate prior): \[ p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \]
  \[ p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0) \]
  \[ \mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1}) \]
- 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]
  \[ q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \quad \text{“variational parameters”} \]
  \[ q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N) \]

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

[Krongut 2020] [MacKay 2003; Bishop 2006]
Air pollution: Particulate matter

approximation

exact posterior
Air pollution: Particulate matter

approximation

exact posterior

μ

τ

[Bishop 2006]
Air pollution: Particulate matter

approximation

exact posterior
Air pollution: Particulate matter

approximation

exact posterior
Microcredit Experiment
Microcredit Experiment

• Simplified from Meager (2019)
Microcredit Experiment

- Simplified from Meager (2019)
- \( K = 7 \) microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
Microcredit Experiment

- Simplified from Meager (2019)
- $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site (~900 to ~17K)
Microcredit Experiment

• Simplified from Meager (2019)
• $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
• $N_k$ businesses in $k$th site (~900 to ~17K)
Microcredit Experiment

- Simplified from Meager (2019)
- \( K = 7 \) 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 (2019)
• $K = 7$ 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 (2019)
- $K = 7$ 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 (2019)
- $K = 7$ 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:
  \[
  y_{kn} \overset{\text{indep}}{\sim} \mathcal{N}(\mu_k, \sigma^2)
  \]
Microcredit Experiment

• Simplified from Meager (2019)
• $K = 7$ 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:

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

- Simplified from Meager (2019)
- $K = 7$ 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:

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn} \tau_k, 1)$$

1 if microcredit

profit
Microcredit Experiment

- Simplified from Meager (2019)
- $K = 7$ 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:

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

- Simplified from Meager (2019)
- $K = 7$ 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:

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

- Simplified from Meager (2019)
- \( K = 7 \) 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:
  \[
  y_{kn} \sim \text{iid} \mathcal{N}(\mu_k + T_{kn} \tau_k, \sigma_k^2)
  \]
  1 if microcredit
Microcredit Experiment

- Simplified from Meager (2019)
- \( K = 7 \) 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:
Microcredit Experiment

• Simplified from Meager (2019)
• $K = 7$ 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:

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

1 if microcredit

• Priors and hyperpriors:

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

- Simplified from Meager (2019)
- $K = 7$ 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:
  \[ y_{kn} \sim \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2) \]

- Priors and hyperpriors:
  \[
  \begin{pmatrix}
  \mu_k \\
  \tau_k
  \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}
  \mu \\
  \tau
  \end{pmatrix}, C\right)
  \]
  \[
  \sigma_k^{-2} \sim \Gamma(a, b)
  \]
Microcredit Experiment

- Simplified from Meager (2019)
- \( K = 7 \) 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:

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

- Priors and hyperpriors:

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

\[
\sigma_k^{-2} \sim \Gamma(a, b) \quad C \sim \text{Sep&LKJ}(\eta, c, d)
\]
MFVB: Do we need to check the output?
Microcredit

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

Means

Parameter
- \mu
- \mu_k
- \tau
- \tau_k
- \log(\sigma^2)

MFVB

MCMC (ground truth)
Microcredit

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

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

Global parameters (-τ)

- MAP: 12 s
- MFVB: 57 s

Global parameter τ

Local parameters

Global parameters (all)

Local parameters
Criteo Online Ads Experiment

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

- Topic discovery

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

[Blei et al 2003]
Why use MFVB?

- Topic discovery
- Latent Dirichlet allocation (LDA)

<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 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.
Why use MFVB?

• Topic discovery

• Latent Dirichlet allocation (LDA): 31,000+ citations

<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></td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td></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></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 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.
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?

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

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

[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

[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.]

[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.]

- Underestimates variance (sometimes severely)

[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.]

- 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]

[Diagram: Scatter plot showing the relationship between MFVB and MCMC with points scattered along the diagonal line.]

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

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

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

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

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

![Diagram showing mean comparison](image-url)
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:

\[
y_n | \beta, z, \sigma^2 \; \text{iid} \sim \mathcal{N}(\beta^T x_n + z_k(n)r_n, \sigma^2)
\]

\[
z_k | \rho^2 \; \text{iid} \sim \mathcal{N}(0, \rho^2)
\]

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

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

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

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

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

Optimization

$\minimize_{q \in Q} f(q(\cdot), p(\cdot|y))$

Variational Bayes

$\minimize_{q \in Q} KL(q(\cdot)||p(\cdot|y))$

Mean-field variational Bayes

$\minimize_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))$

How deep is the issue?

Algorithm

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

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

How deep is the issue?

Algorithm

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

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

Variational Bayes

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

Mean-field variational Bayes

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

Algorithm

Implementation

Gaussian example was exact

How deep is the issue?
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

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

How deep is the issue?

Algorithm

Implementation

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

[NICE′][Baqué et al 2017; Huggins, Karsprzak, Campbell, Broderick 2019]

\[ p(\theta|y) \quad q^*(\theta) \]

NICE

NICE′
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
  [Baqué et al 2017; Huggins et al 2020]
Is it just MFVB?

• Some KL values seen in practice:
  ~1 to ~70, 0.5 to 3
  [Baqué et al 2017; Huggins et al 2020]
• Take any constant $c$
Is it just MFVB?

• Some KL values seen in practice: ~1 to ~70, 0.5 to 3
  [Baqué et al 2017; Huggins et al 2020]

• 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 2020]
- 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 2020]

• 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 2020]
- 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
- 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 2020]
- 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 2020]
- 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
  
  - 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 $\sigma_p^2$

[Huggins, Karsprzak, Campbell, Broderick 2020]
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

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

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

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

[Baqué et al 2017; Huggins et al 2020]
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 $\sigma_p^2$

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

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

[Baqué et al 2017; Huggins et al 2020]

[Huggins, Karsprzak, Campbell, Broderick 2020]
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 $m_p$

$q$: 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 2020]
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 $m_q$

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

$\sigma_p^2$: 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 2020]
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

$$\sigma^2_p \geq c\sigma^2_q$$

$p$: Student's t, variance $\sigma^2_q$. $q$: Gaussian, variance $\sigma^2_q$.

$p$: Weibull, mean $m_p$

$q$: Weibull, mean $m_q$

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

$$\left(m_p - m_q\right)^2 \geq c\sigma^2_p$$

[Huggins, Karsprzak, Campbell, Broderick 2020]
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

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

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

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

[Baqué et al 2017; Huggins et al 2020]
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
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
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?

- “Linear response” (LRVB) corrections fix the variance
  [Giordano, Broderick, Jordan 2015, 2018]
What can we do?

• “Linear response” (LRVB) corrections fix the variance [Giordano, Broderick, Jordan 2015, 2018]

• “Hilbert coresets” allow theoretical guarantees on finite-data quality [Huggins, Campbell, Broderick 2016; Huggins, Campbell, Kasprzak, Broderick, 2018; Campbell, Broderick 2018, 2019]
What can we do?

- “Linear response” (LRVB) corrections fix the variance [Giordano, Broderick, Jordan 2015, 2018]
- “Hilbert coresets” allow 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?

- “Linear response” (LRVB) corrections fix the variance [Giordano, Broderick, Jordan 2015, 2018]
- “Hilbert coresets” allow theoretical guarantees on finite-data quality [Huggins, Campbell, Broderick 2016; Huggins, Campbell, Kasprzak, Broderick, 2018; Campbell, Broderick 2018, 2019]
- Reliable diagnostics

Optimize: closest nice distr.

Variational Bayes

Mean-field variational Bayes

Approximate posterior
What can we do?

• “Linear response” (LRVB) corrections fix the variance [Giordano, Broderick, Jordan 2015, 2018]
• “Hilbert coresets” allow 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?

- “Linear response” (LRVB) corrections fix the variance [Giordano, Broderick, Jordan 2015, 2018]

- “Hilbert coresets” allow theoretical guarantees on finite-data quality [Huggins, Campbell, Broderick 2016; Huggins, Campbell, Kasprzak, Broderick, 2018; Campbell, Broderick 2018, 2019]

- Reliable diagnostics
  - cf. KL

Approximate posterior
Optimize: closest nice distr.

Variational Bayes
Mean-field variational Bayes

KL
iteration

0
What can we do?

• “Linear response” (LRVB) corrections fix the variance
  [Giordano, Broderick, Jordan 2015, 2018]

• “Hilbert coresets” allow 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

0

iteration
What can we do?

• “Linear response” (LRVB) corrections fix the variance [Giordano, Broderick, Jordan 2015, 2018]

• “Hilbert coresets” allow 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
0 18,067
iteration iteration
What can we do?

• “Linear response” (LRVB) corrections fix the variance
  [Giordano, Broderick, Jordan 2015, 2018]

• “Hilbert coresets” allow 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
What can we do?

- “Linear response” (LRVB) corrections fix the variance [Giordano, Broderick, Jordan 2015, 2018]
- “Hilbert coresets” allow theoretical guarantees on finite-data quality [Huggins, Campbell, Broderick 2016; Huggins, Campbell, Kasprzak, Broderick, 2018; Campbell, Broderick 2018, 2019]
- Reliable diagnostics
  - “Yes, but did it work? Evaluating variational inference” ICML 2018
- Diagnostics & workflow with theoretical guarantees
  - “Validated Variational Inference via Practical Posterior Error Bounds” [Huggins, Kasprzak, Campbell, Broderick, 2020]
“Core” of the data set
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011]
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011]
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

- Theoretical guarantees on quality
- How to develop **coresets for Bayes**?

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs

---

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]
“Core” of the data set

• Observe: redundancies can exist even if data isn’t “tall”
• Coresets: pre-process data to get a smaller, weighted data set

• Theoretical guarantees on quality
• How to develop coresets for diverse tasks/geometries?
• Previous heuristics: data squashing, big data GPs
• Compare to subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]
"Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]
Uniform subsampling
Uniform subsampling
Uniform subsampling
Uniform subsampling
Uniform subsampling

Benign

Malicious
Uniform subsampling

- Benign
- Malicious

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Benign
- Malicious

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Benign
- Malicious

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
- Noisy estimates
Uniform subsampling

- Might miss important data
- Noisy estimates

$M = 10$

[Campbell, Broderick 2018, 2019]
Uniform subsampling

- Might miss important data
- Noisy estimates

$M = 10$

$M = 100$

$M = 1000$

[Campbell, Broderick 2018, 2019]
Data summarization alternatives

Uniform subsampling

$M = 10$

$M = 100$

$M = 1000$

[Campbell, Broderick 2018, 2019]
Data summarization alternatives

Uniform subsampling

Importance sampling

\[ M = 10 \quad M = 100 \quad M = 1000 \]

[Campbell, Broderick 2018, 2019]
Data summarization alternatives

Uniform subsampling

Importance sampling

Bayesian/Hilbert coresets

\[ M = 10 \quad M = 100 \quad M = 1000 \]

[Campbell, Broderick 2018, 2019]
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

- 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)
