Statistical and computational trade-offs in Bayesian learning

Tamara Broderick
ITT Career Development Assistant Professor, MIT

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

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

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

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

• Bayesian inference
  • modular, complex models
  • all information about the parameter in the posterior

• Approximating the posterior can be computationally expensive

• Computational/statistical gains for trading off some posterior knowledge
  • point estimates
Statistical/computational trade-offs

• Bayesian inference
  • modular, complex models
  • all information about the parameter in the posterior

• Approximating the posterior can be computationally expensive

• Computational/statistical gains for trading off some posterior knowledge
  • point estimates
  • covariances, coherent estimates of uncertainty
Clustering
Clusters in clustering.
Clustering

“clusters”
## Clustering

<table>
<thead>
<tr>
<th></th>
<th>Arts</th>
<th>Sports</th>
<th>Economics</th>
<th>Science</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document 1</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 2</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 3</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 4</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 5</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Document 6</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Document 7</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Feature allocation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>Sports</td>
<td>Economics</td>
<td>Science</td>
<td>Technology</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“features”
Feature allocation

Many other possible latent structures in data
How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward
- Only works for K clusters
How do we learn latent structure?

K-means
How do we learn latent structure?

K-means

- Fast
How do we learn latent structure?

K-means

- Fast
- Can parallelize
How do we learn latent structure?

K-means
- Fast
- Can parallelize
- Straightforward
How do we learn latent structure?

K-means
- Fast
- Can parallelize
- Straightforward
- Only works for K clusters
How do we learn latent structure?

<table>
<thead>
<tr>
<th>K-means</th>
<th>Nonparametric Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ Fast</td>
<td></td>
</tr>
<tr>
<td>▪ Can parallelize</td>
<td></td>
</tr>
<tr>
<td>▪ Straightforward</td>
<td></td>
</tr>
<tr>
<td>▪ Only works for K clusters</td>
<td></td>
</tr>
</tbody>
</table>
How do we learn latent structure?

K-means
- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes
- Modular (general latent structure)
How do we learn latent structure?

K-means
- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes
- Modular (general latent structure)
- Flexible (K can grow as data grows)
How do we learn latent structure?

K-means
- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes
- Modular (general latent structure)
- Flexible (K can grow as data grows)
- Coherent treatment of uncertainty

But...
- E.g., Silicon Valley: can have petabytes of data
- Practitioners turn to what runs
How do we learn latent structure?

<table>
<thead>
<tr>
<th><strong>K-means</strong></th>
<th><strong>Nonparametric Bayes</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>■ Fast</td>
<td>■ Modular (general latent structure)</td>
</tr>
<tr>
<td>■ Can parallelize</td>
<td>■ Flexible (K can grow as data grows)</td>
</tr>
<tr>
<td>■ Straightforward</td>
<td>■ Coherent treatment of uncertainty</td>
</tr>
<tr>
<td>■ Only works for K clusters</td>
<td></td>
</tr>
</tbody>
</table>

But...

■ E.g., Silicon Valley: can have petabytes of data
■ Practitioners turn to what runs
MAD-Bayes Perspectives

- Bayesian nonparametrics assists the optimization-based inference community
MAD-Bayes Perspectives

- Bayesian nonparametrics assists the optimization-based inference community
  - New, modular, flexible, nonparametric objectives & regularizers
MAD-Bayes Perspectives

- Bayesian nonparametrics assists the optimization-based inference community
  - New, modular, flexible, nonparametric objectives & regularizers
  - Alternative perspective: fast initialization
MAD-Bayes Perspectives

- Bayesian nonparametrics assists the optimization-based inference community
  - New, modular, flexible, nonparametric objectives & regularizers
  - Alternative perspective: fast initialization

Inspiration
- Consider a finite Gaussian mixture model
MAD-Bayes Perspectives

- Bayesian nonparametrics assists the optimization-based inference community
  - New, modular, flexible, nonparametric objectives & regularizers
  - Alternative perspective: fast initialization

Inspiration
- Consider a finite Gaussian mixture model
- The steps of the EM algorithm limit to the steps of the K-means algorithm as the Gaussian variance is taken to 0
The MAD-Bayes idea
- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective
The MAD-Bayes idea

- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective
The MAD-Bayes idea
- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective
K-means

K-means clustering problem
K-means clustering problem

minimize \( \text{(sum of square distances from data points to cluster centers)} \)
K-means clustering problem

minimize

(sum of square distances from data points to cluster centers)
K-means clustering problem

minimize \( \sum_{n=1}^{N} ||x_n - center_n||^2 \)
K-means clustering problem

minimize \( \sum_{n=1}^{N} \| x_n - \text{center}_n \|^2 \)

K-means
K-means clustering problem

minimize \[ \sum_{n=1}^{N} \| x_n - center_n \|^2 \]
K-means

K-means objective

minimize

\[
\sum_{n=1}^{N} \| x_n - \text{center}_n \|^2
\]
Lloyd’s algorithm

Iterate until no changes:
1. For $n = 1, \ldots, N$
   - Assign point $n$ to a cluster
2. Update cluster means
Lloyd’s algorithm

Iterate until no changes:
1. For $n = 1, \ldots, N$
   - Assign point $n$ to a cluster
2. Update cluster means
Lloyd’s algorithm

Iterate until no changes:
1. For \( n = 1, \ldots, N \)
   - Assign point \( n \) to a cluster
2. Update cluster means
Lloyd’s algorithm

Iterate until no changes:
1. For $n = 1, \ldots, N$
   - Assign point $n$ to a cluster
2. Update cluster means
Lloyd’s algorithm

Iterate until no changes:
1. For $n = 1, ..., N$
   - Assign point $n$ to a cluster
2. Update cluster means
Lloyd’s algorithm

Iterate until no changes:
1. For $n = 1, \ldots, N$
   - Assign point $n$ to a cluster
2. Update cluster means
Lloyd’s algorithm

Iterate until no changes:
1. For $n = 1, ..., N$
   - Assign point $n$ to a cluster
2. Update cluster means
Lloyd’s algorithm

Iterate until no changes:
1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means
Lloyd’s algorithm

Iterate until no changes:
1. For $n = 1, \ldots, N$
   - Assign point $n$ to a cluster
2. Update cluster means
The MAD-Bayes idea
- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective
MAD-Bayes

The MAD-Bayes idea
- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective
Bayesian model
Bayesian model
Bayesian model
Bayesian model
Bayesian model

Nonparametric

- number of parameters can grow with the number of data points
The MAD-Bayes idea

- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective
The MAD-Bayes idea

- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective
MAD-Bayes
MAD-Bayes

- Maximum a Posteriori (MAP) is an optimization problem

\[ \arg \max_{\text{parameters}} \mathbb{P}(\text{parameters}|\text{data}) \]
MAD-Bayes

- **Maximum a Posteriori (MAP)** is an optimization problem

\[ \text{argmax}_{\text{parameters}} P(\text{parameters} | \text{data}) \]

- We take a limit of the objective (posterior) and get one like K-means
MAD-Bayes

- Maximum a Posteriori (MAP) is an optimization problem

\[
\arg\max_{\text{parameters}} \mathbb{P}(\text{parameters}|\text{data})
\]

- We take a limit of the objective (posterior) and get one like K-means
  - “Small-variance asymptotics”
MAD-Bayes

Bayesian posterior | K-means-like objectives
MAD-Bayes

Bayesian posterior  \hspace{1cm}  K-means-like objectives

\hspace{1cm}

Mixture of K Gaussians  \hspace{1cm}  K-means
MAD-Bayes

Bayesian posterior  K-means-like objectives

Mixture of K Gaussians  K-means

Dirichlet process mixture  Unbounded number of clusters
MAD-Bayes

**Bayesian posterior**

- Mixture of K Gaussians
- Dirichlet process mixture
- Hierarchical Dirichlet process

**K-means-like objectives**

- K-means
- Unbounded number of clusters
- Multiple data sets share cluster centers
MAD-Bayes

Bayesian posterior  \[ \rightarrow \]  K-means-like objectives

- Mixture of K Gaussians  \[ \rightarrow \]  K-means
- Dirichlet process mixture  \[ \rightarrow \]  Unbounded number of clusters
- Hierarchical Dirichlet process  \[ \rightarrow \]  Multiple data sets share cluster centers

...
MAD-Bayes

Bayesian posterior  K-means-like objectives

Mixture of K Gaussians  K-means

Dirichlet process mixture  Unbounded number of clusters

Hierarchical Dirichlet process  Multiple data sets share cluster centers

Beta process  Features
Features

<table>
<thead>
<tr>
<th>Point 1</th>
<th>Point 2</th>
<th>Point 3</th>
<th>Point 4</th>
<th>Point 5</th>
<th>Point 6</th>
<th>Point 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature 1</td>
<td>Feature 2</td>
<td>Feature 3</td>
<td>Feature 4</td>
<td>Feature 5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Z
### Features

<table>
<thead>
<tr>
<th>Point</th>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Feature 3</th>
<th>Feature 4</th>
<th>Feature 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Z**

**A**
### Features

<table>
<thead>
<tr>
<th>Z</th>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Feature 3</th>
<th>Feature 4</th>
<th>Feature 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point 1</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
</tr>
<tr>
<td>Point 2</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
</tr>
<tr>
<td>Point 3</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
</tr>
<tr>
<td>Point 4</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
</tr>
<tr>
<td>Point 5</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
</tr>
<tr>
<td>Point 6</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
</tr>
<tr>
<td>Point 7</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
<td>🔴</td>
</tr>
</tbody>
</table>

A
MAD-Bayes

Bayesian posterior

\[ P(Z, A|X) \]

\[ \propto \frac{1}{(2\pi \sigma^2)^{ND/2}} \exp \left\{ -\frac{1}{2\sigma^2} \text{tr}((X - ZA)'(X - ZA)) \right\} \]

\[ \gamma^{K^+} \exp \left\{ -\sum_{n=1}^{N} \frac{\gamma}{n} \right\} \frac{K^+}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!} \]

\[ \cdot \frac{1}{(2\pi \rho^2)^{K+D/2}} \exp \left\{ -\frac{1}{2\rho^2} \text{tr}(A'A) \right\} . \]
MAD-Bayes

Bayesian posterior

\[ \mathbb{P}(Z, A|X) \]

\[ \propto \frac{1}{(2\pi \sigma^2)^{ND/2}} \exp \left\{ -\frac{1}{2\sigma^2} \text{tr}((X - ZA)'(X - ZA)) \right\} \]

\[ \gamma^{K^+} \exp \left\{ -\sum_{n=1}^{N} \frac{\gamma}{n} \right\} \frac{1}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!} \]

\[ \cdot \frac{1}{(2\pi \rho^2)^{K^+D/2}} \exp \left\{ -\frac{1}{2\rho^2} \text{tr}(A'A) \right\} . \]
MAD-Bayes

Bayesian posterior

\[ P(Z, A | X) \]

\[ \propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp \left\{ -\frac{1}{2\sigma^2} \text{tr}((X - ZA)'(X - ZA)) \right\} \]

\[ \gamma^{K^+} \exp \left\{ -\sum_{n=1}^{N} \frac{\gamma}{n} \right\} \prod_{h=1}^{H} \tilde{K}_h! \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!} \]

\[ \cdot \frac{1}{(2\pi\rho^2)^{K^+D/2}} \exp \left\{ -\frac{1}{2\rho^2} \text{tr}(A'A) \right\} . \]
MAD-Bayes

Bayesian posterior

\[ P(Z, A | X) \]

\[
\propto \frac{1}{(2\pi \sigma^2)^{ND/2}} \exp \left\{ -\frac{1}{2\sigma^2} \text{tr}((X - ZA)'(X - ZA)) \right\} \\
\cdot \gamma^{K^+} \exp \left\{ -\sum_{n=1}^{N} \frac{\gamma}{n} \right\} \prod_{h=1}^{H} \tilde{K}_h! \\
\cdot \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)! (N - S_{N,k})!}{N!} \\
\cdot \frac{1}{(2\pi \rho^2)^{K^+D/2}} \exp \left\{ -\frac{1}{2\rho^2} \text{tr}(A'A) \right\} .
\]
MAD-Bayes

Bayesian posterior

\[ P(Z, A | X) \]

\[
\propto \frac{1}{(2\pi \sigma^2)^{ND/2}} \exp \left\{ -\frac{1}{2\sigma^2} \text{tr}((X - ZA)'(X - ZA)) \right\} \\
\gamma^{K^+} \exp \left\{ - \sum_{n=1}^{N} \frac{\gamma}{n} \right\} \prod_{h=1}^{H} \tilde{K}_h! \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!} \\
\cdot \frac{1}{(2\pi \rho^2)^{K^+D/2}} \exp \left\{ -\frac{1}{2\rho^2} \text{tr}(A'A) \right\} .
\]
MAD-Bayes

Bayesian posterior

\[
\mathbb{P}(Z, A | X) \\
\propto \frac{1}{(2\pi \sigma^2)^{ND/2}} \exp \left\{ -\frac{1}{2\sigma^2} \text{tr}((X - ZA)'(X - ZA)) \right\} \\
\gamma^{K^+} \exp \left\{ -\sum_{n=1}^{N} \frac{\gamma}{n} \right\} \prod_{h=1}^{H} \tilde{K}_h! \prod_{k=1}^{K^+} (S_{N,k} - 1)! (N - S_{N,k})! \frac{N!}{N!} \\
\cdot \frac{1}{(2\pi \rho^2)^{K^+D/2}} \exp \left\{ -\frac{1}{2\rho^2} \text{tr}(A'A) \right\}.
\]
MAD-Bayes

Bayesian posterior

\[
P(Z, A | X) \propto \frac{1}{(2\pi \sigma^2)^{ND/2}} \exp \left\{ -\frac{1}{2\sigma^2} \text{tr}((X - ZA)'(X - ZA)) \right\} \cdot \frac{\gamma^{K^+}}{\prod_{h=1}^{H} \tilde{K}_h!} \cdot \frac{1}{(2\pi \rho^2)^{K^+D/2}} \exp \left\{ -\frac{1}{2\rho^2} \text{tr}(A'A) \right\}.
\]
MAD-Bayes

BP-means objective

$$\arg\min_{K^+, Z, A} \text{tr}[(X - ZA)'(X - ZA)] + K^+ \lambda^2.$$
MAD-Bayes

BP-means objective

$$\arg \min_{K^+, Z, A} \text{tr}[ (X - ZA)'(X - ZA) ] + K^+ \lambda^2.$$
MAD-Bayes

BP-means objective

$$\arg\min_{K^+,Z,A} \text{tr}[(X - ZA)'(X - ZA)] + K^+ \lambda^2.$$
MAD-Bayes

BP-means objective

$$\operatorname{argmin}_{K^+, Z, A} \operatorname{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$ 

BP-means algorithm

Iterate until no changes:
1. For n = 1, ..., N
   - Assign point n to features
   - Create a new feature if it lowers the objective
2. Update feature means
   $$A \leftarrow (Z'Z)^{-1}Z'X$$
MAD-Bayes

BP-means objective

$$\arg\min_{K+, Z, A} \text{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$ 

BP-means algorithm

Iterate until no changes:

1. For \( n = 1, \ldots, N \)
   - Assign point \( n \) to features
   - Create a new feature if it lowers the objective
2. Update feature means
   $$A \leftarrow (Z'Z)^{-1}Z'X$$
MAD-Bayes

BP-means objective

$$\arg\min_{K+,Z,A} \text{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$ 

BP-means algorithm

Iterate until no changes:
1. For n = 1, ..., N
   - Assign point n to features
   - Create a new feature if it lowers the objective
2. Update feature means $$A \leftarrow (Z'Z)^{-1}Z'X$$
**MAD-Bayes**

**BP-means objective**

$$\arg\min_{K+, Z, A} \text{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$  

**BP-means algorithm**

Iterate until no changes:

1. For \(n = 1, \ldots, N\)
   - Assign point \(n\) to features
   - Create a new feature if it lowers the objective

2. Update feature means  
   \[ A \leftarrow (Z'Z)^{-1}Z'X \]
MAD-Bayes
Griffiths & Ghahramani (2006) computer vision problem “tabletop data”
MAD-Bayes

Griffiths & Ghahramani (2006) computer vision problem “tabletop data”
MAD-Bayes

BP-means features: table and four objects
MAD-Bayes

BP-means features: table and four objects
MAD-Bayes

Griffiths & Ghahramani (2006) computer vision problem “tabletop data”

Bayesian posterior
Gibbs sampler
BP-means algorithm

8.5 * 10^3 sec  0.36 sec

Still faster by order of magnitude if restart 1000 times
Face data

Pre-aligned faces

Samples
Face data

Pre-aligned faces

Samples

3 features (BP-means)
Face data

Pre-aligned faces

Samples

3 features
(BP-means)
Face data

Pre-aligned faces

Samples

3 features (BP-means)
Face data

Pre-aligned faces

Samples

3 features
(BP-means)
**Face data**

Pre-aligned faces

Samples

3 features (BP-means)
Face data

Pre-aligned faces

Samples

4 clusters (K-means, K=4)
Face data

Pre-aligned faces

Samples

4 clusters (K-means, K=4)
Face data

Pre-aligned faces

Samples

4 clusters
(K-means, K=4)
MAD-Bayes

Parallelism and optimistic concurrency control

<table>
<thead>
<tr>
<th></th>
<th>DP-means alg.</th>
<th>BP-means alg.</th>
</tr>
</thead>
<tbody>
<tr>
<td># data points</td>
<td>134M</td>
<td>8M</td>
</tr>
<tr>
<td>time per iteration</td>
<td>5.5 min</td>
<td>4.3 min</td>
</tr>
</tbody>
</table>
MAD-Bayes

Bayesian posterior  \quad  K\text{-}means\text{-}like\ objectives

Mixture of K Gaussians  \quad  K\text{-}means

Dirichlet process mixture  \quad  Unbounded number of clusters

Hierarchical Dirichlet process  \quad  Multiple data sets share cluster centers

Beta process  \quad  Features
MAD-Bayes conclusions
MAD-Bayes conclusions

- We provide new optimization objectives and regularizers
MAD-Bayes conclusions

- We provide new optimization objectives and regularizers
  - In fact, general means of obtaining more
MAD-Bayes conclusions

- We provide new optimization objectives and regularizers
  - In fact, general means of obtaining more
  - Straightforward, fast algorithms
What about uncertainty?

• Variational Bayes (VB)
  • Approximation for posterior
  • Minimize Kullback-Liebler (KL) divergence:

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

• VB practical success
  • Point estimates and prediction
  • Fast
What about uncertainty?

- Variational Bayes (VB)
What about uncertainty?

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

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

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

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

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

\[
KL(q\|p(\cdot|x))
\]
What about uncertainty?

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

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

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

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

- VB practical success
  - point estimates and prediction
  - fast

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

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

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

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

- Variational Bayes
What about uncertainty?

- Variational Bayes

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

- Variational Bayes

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

- Variational Bayes

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

- Mean-field variational Bayes (MFVB)

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

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

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

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

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

- Underestimates variance (sometimes severely)
What about uncertainty?

- Variational Bayes
  \[ KL(q||p(\cdot|x)) = \int_{\theta} q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta \]
- Mean-field variational Bayes (MFVB)
  \[ q(\theta) = \prod_{j=1}^{J} q(\theta_j) \]
- Underestimates variance (sometimes severely)
- No covariance estimates
What about uncertainty?

- Variational Bayes
  \[ KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta \]
- Mean-field variational Bayes (MFVB)
  \[ q(\theta) = \prod_{j=1}^{J} q(\theta_j) \]
- Underestimates variance (sometimes severely)
- No covariance estimates

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

- Variational Bayes

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

- Mean-field variational Bayes (MFVB)

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

- Underestimates variance (sometimes severely)

- No covariance estimates

[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011]
[Dunson 2014; Bardenet, Doucet, Holmes 2015]
1. Derive *Linear Response Variational Bayes* (LRVB) variance/covariance correction

2. Accuracy experiments

3. Scalability
Linear response
Linear response

- Cumulant-generating function
Linear response

• Cumulant-generating function

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

- Cumulant-generating function

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

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

- Cumulant-generating function

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

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

- True posterior covariance

[Image of a contour plot labeled \( p(\theta | x) \)]
Linear response

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

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

[Bishop 2006]
Linear response

- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

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

• Cumulant-generating function

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

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

• True posterior covariance vs MFVB covariance

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

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

• Cumulant-generating function

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

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

• True posterior covariance vs MFVB covariance

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

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

• “Linear response”
Linear response

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

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

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

• Cumulant-generating function

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

• True posterior covariance vs MFVB covariance

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

• “Linear response”

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

- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

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

- “Linear response”

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

- Cumulant-generating function

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

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

- True posterior covariance vs MFVB covariance

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

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

- “Linear response”

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

- Cumulant-generating function

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

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

- True posterior covariance vs MFVB covariance

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

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

- “Linear response”

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

[Bishop 2006]
Linear response

- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

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

- "Linear response"

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

- The LRVB approximation

[Bishop 2006]
Linear response

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

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

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

- The LRVB approximation
  \[ \Sigma = \left. \frac{d}{dt^T} \left[ \frac{d}{dt} C_p(\cdot|x)(t) \right] \right|_{t=0} \]
Linear response

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

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

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

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

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

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

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

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

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

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

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

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

- Cumulant-generating function

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

- True posterior covariance vs MFVB covariance

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

- "Linear response"

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

- The LRVB approximation

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

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

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

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

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

[Bishop 2006]
• LRVB covariance estimate

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

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

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

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

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

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

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

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

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

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

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

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

• KL optimization: fixed point equation in the mean params

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

- LRVB covariance estimate $\hat{\Sigma} := \frac{d}{dt^T} m^*_t \bigg|_{t=0}$
- Suppose $q_t$ exponential family with mean parametrization $m_t$
- KL optimization: fixed point equation in the mean params

$$m^*_t = \frac{\partial}{\partial m_t} KL_t \bigg|_{m_t=m^*_t} + m^*_t$$
Getting rid of $t$

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

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

- KL optimization: fixed point equation in the mean params

\[
m^*_t = \left. \frac{\partial}{\partial m_t} KL_t \right|_{m_t = m^*_t} + m^*_t
\]

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

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

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

- KL optimization: fixed point equation in the mean params

\[
m_t^* = \frac{\partial}{\partial m_t} KL \bigg|_{m_t=m_t^*} + m_t^*
\]

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

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

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

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

- KL optimization: fixed point equation in the mean params

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

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

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

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

- LRVB covariance estimate $\hat{\Sigma} := \frac{d}{dt^T} m_t^* \bigg|_{t=0}$
- Suppose $q_t$ exponential family with mean parametrization $m_t$
- KL optimization: fixed point equation in the mean params

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

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

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

\[ \hat{\Sigma} = (V^{-1} - H)^{-1} \quad \text{for} \quad H := \frac{\partial^2 L}{\partial m \partial m^T} \bigg|_{m=m^*} \]
Getting rid of $t$

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

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

- KL optimization: fixed point equation in the mean params

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

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

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

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

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

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

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

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

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

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

\[
\hat{\Sigma} = (V^{-1} - H)^{-1} = (I - VH)^{-1}V \quad \text{for} \quad H := \frac{\partial^2 L}{\partial m \partial m^T} \bigg|_{m=m^*}
\]
LRVB estimator

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

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

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

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

- LRVB covariance estimate

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

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

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

• LRVB covariance estimate

\[ \hat{\Sigma} := \left. \frac{d}{dt} \left( \mu_t \right) \right|_{t=0} \]

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

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

- LRVB covariance estimate

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

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

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

- LRVB covariance estimate

\[ \hat{\Sigma} := \left. \frac{d}{dt} \mathbb{E}_{q_t^*} \theta \right|_{t=0} \]

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

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

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

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

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

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

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

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

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

- Symmetric and positive definite at local min of KL

- The LRVB assumption: \[ \mathbb{E}_{p_t} \theta \approx \mathbb{E}_{q^*_t} \theta \]

[Bishop 2006]
LRVB estimator

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

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

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

- Symmetric and positive definite at local min of KL

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

[Bishop 2006]
LRVB estimator

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

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

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

• Symmetric and positive definite at local min of KL

• The LRVB assumption: $E_{p_t} \theta \approx E_{q_t^*} \theta$

[Bishop 2006]
LRVB estimator

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

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

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

- Symmetric and positive definite at local min of KL

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

- LRVB estimate is exact when VB gives exact mean (e.g. multivariate normal)
1. Derive *Linear Response Variational Bayes* (LRVB) variance/covariance correction

2. Accuracy experiments

3. Scalability experiments
Experiments
Experiments

• Non-conjugate normal-Poisson generalized linear mixed model

\[
\begin{align*}
& z_n | \tau, x_n \sim \mathcal{N}(z_n | 0, \tau) \\
& y_n | z_n, \tau \sim \text{Poisson}(\text{exp}(z_n)) \\
& \tau \sim \text{Gamma}(\tau | \alpha, \beta) \\
& q(z_n | \tau, \tau) = q(z_n | \tau) \\
& q(z_n | \tau, \tau) = \mathcal{N}(z_n | 0, \tau) \\
& q(z_n) = \mathcal{N}(z_n | 0, \tau) \\
& q(\tau) = \text{Gamma}(\tau | \alpha, \beta)
\end{align*}
\]
Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

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

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

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

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

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

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

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

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

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

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

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

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

- Non-conjugate normal-Poisson generalized linear mixed model

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

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

- MFVB assumption:

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

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

LRVB, MFVB
Experiments

- Non-conjugate normal-Poisson generalized linear mixed model

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

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

- MFVB assumption:

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

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

LRVB, MFVB
Experiments

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

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

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

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

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

- 100 simulated data sets, 500 data points each, R \texttt{MCMCg1mm} package (20,000 samples)
Experiments
Experiments
• Linear model with random effects
Experiments

• Linear model with random effects

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

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

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

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

• Linear model with random effects

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

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

• MFVB assumption:

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

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

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

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

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

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

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

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

LRVB, MFVB
Experiments
Experiments

- Gaussian mixture model
Experiments

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

- MFVB assumption:
  - 68 simulated data sets (2 components, 2 dimensions), 10,000 data points each, \texttt{rnmixGibbs} package (function \texttt{rnmixGibbs}; at least 500 effective samples)
  - MNIST digits: 12,665 0s and 1s; PCA for 25 dimensions

- Experiments "LRVB, MFVB with conjugate priors on \( \pi, \mu, \Lambda \)"
Experiments

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

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

• Gaussian mixture model

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

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

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

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

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

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

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

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

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

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

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

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

LRVB, MFVB
Experiments

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

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

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

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

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

LRVB, MFVB
Experiments

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

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

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

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

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

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

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

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

2. Accuracy experiments

3. Scalability experiments
Scaling the matrix inverse
Scaling the matrix inverse

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

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

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

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

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

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

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

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

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

- Schur complement

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

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

- Decomposition of parameter vector

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

- Schur complement

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

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

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

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

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

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

• Decomposition of parameter vector

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

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

• Schur complement

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

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

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

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

- Sparsity patterns
Scaling the matrix inverse

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

- Decomposition of parameter vector
  \[
  \theta = (\alpha^T, z^T)^T \quad H = \begin{pmatrix}
  H_\alpha & H_{\alpha z} \\
  H_{z \alpha} & H_z
  \end{pmatrix}
  \]

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

- Sparsity patterns
Scaling the matrix inverse

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

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

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

• Sparsity patterns
Scaling the matrix inverse

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

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

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

- Sparsity patterns

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

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

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

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

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

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

LRVB, Gibbs
Experiments

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

• The number of parameters in $\mu, \pi, \Lambda$ grows as $O(KP^2)$

• The number of parameters in $z$ grows as $O(KN)$

• Worst case scaling: $O(K^3), O(P^6), O(N)$
Experiments

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

- MAD-Bayes: fast point estimates
- LRVB covariance correction: in many cases, accurate covariance estimates for VB
- Open questions:
  - Mean correction
  - Global parameter scaling
  - Targeting other posterior statistics besides point estimates and covariance
- More LRVB: Bayesian robustness (work in progress)
Conclusions, etc

• MAD-Bayes: fast point estimates
Conclusions, etc

• MAD-Bayes: fast point estimates

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

• MAD-Bayes: fast point estimates

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

• Open questions:
Conclusions, etc

- MAD-Bayes: fast point estimates
- LRVB covariance correction: in many cases, accurate covariance estimates for VB
- Open questions:
  - Mean correction
Conclusions, etc

• MAD-Bayes: fast point estimates

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

• Open questions:
  • Mean correction
  • Global parameter scaling
Conclusions, etc

• MAD-Bayes: fast point estimates

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

• Open questions:
  • Mean correction
  • Global parameter scaling
  • Targeting other posterior statistics besides point estimates and covariance
Conclusions, etc

• MAD-Bayes: fast point estimates

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

• Open questions:
  • Mean correction
  • Global parameter scaling
  • Targeting other posterior statistics besides point estimates and covariance

• More LRVB: Bayesian robustness (work in progress)
References


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


