LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

Brian Trippe, Jonathan Huggins, Raj Agrawal, and Tamara Broderick
LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

Brian Trippe, Jonathan Huggins, Raj Agrawal, and Tamara Broderick

Genomic Study (motivating example)
- **Goal:** Understand relationship between genomic variation & disease outcome
- **N**=20,000 samples — **D**=500,000 SNPs

https://www.ebi.ac.uk/training/
LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations
Brian Trippe, Jonathan Huggins, Raj Agrawal, and Tamara Broderick

Genomic Study (motivating example)
- **Goal:** Understand relationship between genomic variation & disease outcome
- N=20,000 samples — D=500,000 SNPs

Generalized Linear Models (GLMs)
- Interpretability
- E.g. Logistic/Poisson/Negative Binomial Regression

https://www.ebi.ac.uk/training/
Generalized Linear Models (GLMs)
- Interpretability
- E.g. Logistic/Poisson/Negative Binomial Regression

Bayesian Modeling & Inference
- Coherent uncertainty quantification

Genomic Study (motivating example)
- Goal: Understand relationship between genomic variation & disease outcome
- N=20,000 samples — D=500,000 SNPs

LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations
Brian Trippe, Jonathan Huggins, Raj Agrawal, and Tamara Broderick

https://www.ebi.ac.uk/training/
LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

Brian Trippe, Jonathan Huggins, Raj Agrawal, and Tamara Broderick

Genomic Study (motivating example)
- **Goal:** Understand relationship between genomic variation & disease outcome
- N=20,000 samples — D=500,000 SNPs

Generalized Linear Models (GLMs)
- Interpretability
- E.g. Logistic/Poisson/Negative Binomial Regression

Bayesian Modeling & Inference
- Coherent uncertainty quantification

Problem: Super-linear scaling with D

https://www.ebi.ac.uk/training/
LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

Brian Trippe, Jonathan Huggins, Raj Agrawal, and Tamara Broderick

Genomic Study (motivating example)
- **Goal:** Understand relationship between genomic variation & disease outcome
- N=20,000 samples — D=500,000 SNPs

Generalized Linear Models (GLMs)
- Interpretability
- E.g. Logistic/Poisson/Negative Binomial Regression

Bayesian Modeling & Inference
- Coherent uncertainty quantification

Problem: Super-linear scaling with D

We present **LR-GLM**, a method with linear scaling in D and theoretical guarantees on quality
How does it work?
How does it work?

Cartoon Example
- Logistic Regression with two correlated features
How does it work?

**Cartoon Example**
- Logistic Regression with two correlated features

![Diagram showing feature effect sizes](image)
**How does it work?**

**Cartoon Example**
- Logistic Regression with two correlated features

**Uncertainty in Effect Sizes**

- **Exact Posterior**
- **True Effect Sizes**
How does it work?

Cartoon Example
- Logistic Regression with two correlated features

Uncertainty in Effect Sizes

Lots of Information
Little Information

Exact Posterior
True Effect Sizes
How does it work?

**Cartoon Example**
- Logistic Regression with two correlated features

**Uncertainty in Effect Sizes**

**The LR-GLM Approximation**
We ignore the least informative directions

\[
p(y_i | x_i^T \beta) \approx p(y_i | x_i UU^T \beta)
\]
How does it work?

**Cartoon Example**
- Logistic Regression with two correlated features

---

**Uncertainty in Effect Sizes**

- Lots of Information
- Little Information

---

**The LR-GLM Approximation**

We ignore the least informative directions

\[
p(y_i | x_i^T \beta) \approx p(y_i | x_i UU^T \beta)
\]
How does it work?

**Cartoon Example**
- Logistic Regression with two correlated features

**Uncertainty in Effect Sizes**

**The LR-GLM Approximation**
We ignore the least informative directions

\[
p(y_i | x_i^T \beta) \approx p(y_i | x_i UU^T \beta)
\]

**Approximation Quality**
- Exact when data are low rank
How does it work?

**Cartoon Example**
- Logistic Regression with two correlated features

**Uncertainty in Effect Sizes**

- **Lots of Information**
- **Little Information**

**The LR-GLM Approximation**
We ignore the least informative directions

\[ p(y_i|x_i^T \beta) \approx p(y_i|x_iUU^T \beta) \]

**Approximation Quality**
- **Exact when data are low rank**
- We prove: Approximation is close when the data are approximately low rank
Does it Work?
Does it Work?

Evaluate by comparing exact means and uncertainties (*slow*) against our approximation (*fast*)
Does it Work?

Evaluate by comparing exact means and uncertainties (*slow*) against our approximation (*fast*)

**Post. Mean Estimation**

**Post. Uncertainty Estimation**

- No Error
- LR-Laplace (Our Method)
Does it Work?

Evaluate by comparing exact means and uncertainties \textit{(slow)} against our approximation \textit{(fast)}

\begin{itemize}
  \item Rank of approximation defines a computational-statistical trade-o\textsuperscript{f}
  \item The approximation is conservative (overestimates uncertainty)
  \item For high-dimensional, correlated data, LR-GLM closely approximates the exact posterior up to 5X faster!
\end{itemize}

We rigorously show…
- Rank of approximation defines a computational-statistical trade-off
Does it Work?

Evaluate by comparing exact means and uncertainties (slow) against our approximation (fast)

We rigorously show…
- Rank of approximation defines a computational-statistical trade-off
Evaluate by comparing exact means and uncertainties \textit{(slow)} against our approximation \textit{(fast)}

We rigorously show...
- Rank of approximation defines a computational-statistical trade-off
- The approximation is conservative (overestimates uncertainty)
We rigorously show...
- Rank of approximation defines a computational-statistical trade-off
- The approximation is conservative (overestimates uncertainty)
- For high-dimensional, correlated data, **LR-GLM closely approximates the exact posterior** up to 5X faster!
We rigorously show…
- Rank of approximation defines a computational-statistical trade-off
- The approximation is conservative (overestimates uncertainty)
- For high-dimensional, correlated data, **LR-GLM closely approximates the exact posterior up to 5X faster!**