Approximating high-dimensional posteriors with nuisance parameters via integrated rotated Gaussian approximation (IRGA)

Willem van den Boom
Department of Statistics and Applied Probability
National University of Singapore

Bayesian Computation for High-Dimensional Statistical Models
Institute of Mathematical Sciences
National University of Singapore
September 10, 2018
Collaborators

David B. Dunson

Galen Reeves

Department of Statistical Science
Duke University
• IRGA with Bayesian variable selection as an example
• Approximation accuracy
• IRGA in a more general setup
IRGA with Bayesian variable selection as an example
Standard linear model

\[ y = X \beta + \epsilon \]

- \( n \times p \) matrix
- \( n \) observations
- \( p \) unknown parameters
- Gaussian errors \( \text{N}(0, \sigma^2 I) \)

The entries of \( \beta \) are conditionally independent given hyperparameters \( \theta \):

\[
p(\beta | \theta) = \prod_{j=1}^{p} p(\beta_j | \theta)
\]
Example: Bayesian Variable Selection

Entries of $\beta$ conditionally independent given hyperparameters $\theta$

$$p(\beta | \theta) = \prod_{j=1}^{p} p(\beta_j | \theta)$$

Mixed discrete-continuous distribution for marginal prior

$$p(\beta_j | \theta) = (1 - \lambda)\delta(\beta_j) + \lambda N(\beta_j | 0, \psi), \quad \theta = (\lambda, \psi)$$
Inference

**Goal:** compute posterior marginal distribution of first entry

\[
p(\beta_1 | y) = \int p(\beta | y) d\beta_2^p \quad \text{e.g. for } P(\beta_1 \neq 0 | y)
\]

where

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

- Treat \( \beta_2^p \) as nuisance parameters
- Number of possible subsets \( \{j \mid \beta_j \neq 0\} \) grows exponentially with number of variables \( p \)
- Posterior is a mixture of \( 2^p \) Gaussians
- Intractable for large \( p \). Therefore, approximations are used for scalable inference.
Approximation methods

Most popular approximation methods fit into two groups:

• **Sampling based:** MCMC, Gibbs sampling, Stochastic Search Variable Selection (George & McCulloch, 1993), or Bayesian Adaptive Sampling (Clyde et al., 2011).
  ‣ Exact solution asymptotically
  ‣ Hard to establish convergence due to exponential and discrete nature of the posterior

• **Deterministic:** Variational Bayes (Carbonetto & Stephens, 2012, Ormerod et al., 2017), Expectation propagation (Hernández-Lobato et al., 2015), or EM variable selection (Ročková & George, 2014).
  ‣ Often unclear what quality of approximation is achieved

• IRGA provides interpretable approximations to posterior marginals and can overcome some of these issues.
Overview of IRGA

• Goal: Marginal posterior, such as

\[ p(\beta_1 \mid y) = \int \pi(\beta \mid y) \, d\beta_2^p \]

• Achieved by approximately integrating out \( \beta_2^p \)
  
  ‣ Based on a data rotation
  ‣ Transparent approximation allows for theoretical analysis
Overview of IRGA

1. Rotate the data to isolate the parameter of interest $\beta_1$
   ‣ Introduce an auxiliary variable which summarizes the influence of the nuisance parameters $\beta_2^p$ on $\beta_1$

2. Use any means possible to compute/estimate the posterior mean and posterior variance of the auxiliary variable

3. Apply a Gaussian approximation to the auxiliary variable and solve the one-dimensional integration problem to obtain the posterior approximation for $\beta_1$
1: Rotate

- Apply rotation matrix $Q$ to the data which zeros out all but the first entry in the first column of the data
  - $\tilde{y} = Qy$ and $\tilde{X} = QX$

  \[
  \tilde{y} = \begin{bmatrix}
  \tilde{x}_{1,1} & \tilde{x}_{1,2} & \cdots & \tilde{x}_{1,p} \\
  0 & \tilde{x}_{2,2} & \cdots & \tilde{x}_{2,p} \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & \tilde{x}_{n,2} & \cdots & \tilde{x}_{n,p}
  \end{bmatrix} \beta + \tilde{\epsilon}
  \]

  $N(0, \sigma^2 I)$

- Only the first observation depends on $\beta_1$:
  \[
  \tilde{y}_1 = \tilde{x}_{1,1} \beta_1 + \sum_{j=2}^{p} \tilde{x}_{1,j} \beta_j + \tilde{\epsilon}_1
  \]

  auxiliary variable captures influence of nuisance parameters
1: Rotate

\[ p \text{ unknown parameters} \quad n \text{ observations} \quad (\text{the data}) \]
1: Rotate
1: Rotate

\[ y_1 \]

\[ \beta_1 \]

\[ y_2 \]
\[ \beta_2 \]

\[ y_3 \]
\[ \beta_3 \]

\[ \ldots \]
\[ \beta_4 \]

\[ y_n \]
\[ \beta_p \]
1: Rotate
1: Rotate

\[ \zeta = \sum_{j=2}^{p} \tilde{x}_{1,j} \beta_j \]

auxiliary variable encapsulates influence of nuisance parameters
2: Estimate / compute

• Compute the posterior mean and variance of the auxiliary variable:
  \[ E[\zeta \mid \tilde{y}_2^n] \quad \text{Cov}[\zeta \mid \tilde{y}_2^n] \]

  ▶ Vector Approximate Message Passing (VAMP, Rangan et al., 2017)
  ▶ LASSO
  ▶ Variational Bayes
  ▶ [your favorite method]

• The quantities are independent of the target parameter \( \beta_1 \)
3: Approximate

- Apply Gaussian approximation to auxiliary variable to compute posterior approximation

\[ p(\beta_1 \mid y) \propto \int p(\tilde{y}_1 \mid \zeta, \beta_1) p(\beta_1) p(\zeta \mid \tilde{y}_2^p) d\zeta \]

Gaussian by assumption on noise

replace with Gaussian using mean and variance from previous step

- Approximation \( \hat{p}(\zeta \mid \tilde{y}_2^p) \) can be accurate even if the prior and posterior are highly non-Gaussian:
  - Most low-dimensional projections of high-dimensional distributions are close to Gaussian (projection pursuit, Diaconis & Freedman, 1984).
  - Additionally, CLT or Bernstein-von Mises: \( \zeta = \sum_{j=2}^{p} \tilde{x}_{1,j} \beta_j \)
Overview of IRGA

1. Rotate the data to isolate the parameter of interest $\beta_1$
   - Introduce an auxiliary variable $\zeta$ which summarizes the influence of the nuisance parameters $\beta_2^p$ on $\beta_1$

2. Use any means possible to compute/estimate the posterior mean and posterior variance of $p(\zeta \mid \tilde{y}_2^p)$

3. Apply a Gaussian approximation to $p(\zeta \mid \tilde{y}_2^p)$ and solve the one-dimensional integration problem to obtain the approximation for $p(\beta_1 \mid y)$
Approximation accuracy
Gaussian approximation accuracy

Q-Q plots of $p(\zeta \mid \tilde{y}^n_2)$ from simulated data with $X$ equal to 3,571 highly correlated SNPs from Friedman et al. (2010).
Quadratic Wasserstein distance

\[ W_2 \left\{ p_1(x_1), p_2(x_2) \right\} = \inf E \left( \|x_1 - x_2\|^2 \right)^{\frac{1}{2}} \]

where the infimum is over all joint distributions on \((x_1, x_2)\) such that \(x_1 \sim p_1(\cdot)\) and \(x_2 \sim p_2(\cdot)\).
Theorem
If \( p(\beta_2^p \mid \tilde{y}_2^p) \) is nearly isotropic and its distance from its mean concentrates, then the Wasserstein distance

\[
W_2 \left\{ p(\zeta \mid \tilde{y}_2^n), \hat{p}(\zeta \mid \tilde{y}_2^n) \right\}
\]

is small for most vectors \( (\tilde{x}_{1,j})_{j=2}^p \).

**Isotropy** is when independent draws from \( p(\beta_2^p \mid \tilde{y}_2^p) \) are not too correlated.
Kullback-Leibler divergence

\[ D \{ p_1(x_1) \parallel p_2(x_2) \} = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} \, dx \]
Theorem

\[ E \left[ D \left\{ p(\beta_1 \mid y) \parallel \hat{p}(\beta_1 \mid y) \right\} \mid \tilde{y}_2^n \right] \leq \frac{1}{2\sigma^2} W_2^2 \left\{ p(\zeta \mid \tilde{y}_2^n), \hat{p}(\zeta \mid \tilde{y}_2^n) \right\} \]

where \( y \mid \beta \sim \mathcal{N} \left( X\beta, \sigma^2 I_n \right) \) with \( \beta \sim p(\beta) \)
Theorem

\[
E \left[ D \left\{ p(\beta_1 \mid y) \parallel \hat{p}(\beta_1 \mid y) \right\} \mid \tilde{y}_2^n \right] \leq \frac{1}{2\sigma^2} W_2^2 \left\{ p(\zeta \mid \tilde{y}_2^n), \hat{p}(\zeta \mid \tilde{y}_2^n) \right\}
\]

where \( y \mid \beta \sim \mathcal{N}(X\beta, \sigma^2 I_n) \) with \( \beta \sim p(\beta) \)
Application

Difference from Gibbs estimate

- Integrated rotated Gaussian approximation
- Expectation propagation (Hernández-Lobato et al., 2015)
- Variational Bayes (Carbonetto & Stephens, 2012)

Diabetes data (Efron et al., 2004)
IRGA in a more general setup
Standard linear model

\[ y = X \beta + \epsilon \]

- \( n \times p \) matrix
- \( p \) unknown parameters
Linear model with nuisance parameter

\[ y = X \beta + F(Z) + \epsilon \]

- \( n \times p \) matrix
- \( n \times q \) matrix
- \( p < n \) unknown parameters
- Nuisance parameter \( F : \mathbb{R}^{n \times q} \rightarrow \mathbb{R}^n \)
Overview of IRGA

- Apply rotation matrix \( Q = (R, S) \) where \( R (n \times p) \) is an orthonormal basis for the column space of \( X \) and \( S (n \times (n-p)) \) for the orthogonal complement.
  Then, the likelihood for \( Q^T y \) factorizes as

\[
R^T y \mid \beta, F \sim \mathcal{N} \left\{ R^T X \beta + R^T F(Z), \sigma^2 I_p \right\} \\
S^T y \mid \beta, F \sim \mathcal{N} \left\{ S^T F(Z), \sigma^2 I_{n-p} \right\}
\]

- Only the first distribution involves \( \beta \)
- Auxiliary variable \( \zeta \) captures influence of \( F \) on \( \beta \)

- Approximate \( p(\zeta \mid S^T y) \) by a Gaussian

- Compute \( \hat{p}(\beta \mid y) \propto \int p(R^T y \mid \beta, \zeta) p(\beta) \hat{p}(\zeta \mid S^T y) d\zeta \)
Theorem

IRGA approximation

\[ E \left[ D \{ p(\beta \mid y) \parallel \hat{p}(\beta \mid y) \} \mid S^T y \right] \leq \frac{1}{2\sigma^2} W^2 \{ p(\zeta \mid S^T y), \hat{p}(\zeta \mid S^T y) \} \]

Gaussian approximation

where \( y \mid \beta, F \sim \mathcal{N} \{ X\beta + F(Z), \sigma^2 I_n \} \) with \( (\beta, F) \sim p(\beta)p(F) \)
Advantages of IRGA

• Approximation employed is transparent and allows for analysis yielding theoretical guarantees

• Can leverage your favorite method (e.g. lasso, approximate message passing, etc.) to produce accurate approximations of marginal posteriors


References


