A variational Bayes approach to debiased inference in high-dimensional linear regression

Luke Travis
Joint work with Ismaël Castillo, Alice L'Huillier and Kolyan Ray

ISNPS 28th June 2024

Motivation

Problem Setup

Problem Setup

- Consider the linear regression model

$$Y = X\beta + \varepsilon,$$

with $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ observed, $\beta \in \mathbb{R}^p$ and $\varepsilon \sim \mathcal{N}_p(0, I_n)$.

Problem Setup

- Consider the linear regression model

$$Y = X\beta + \varepsilon,$$

with $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ observed, $\beta \in \mathbb{R}^p$ and $\varepsilon \sim \mathcal{N}_p(0, I_n)$.

- We consider the **high-dimensional** case where $n < p$.

Problem Setup

- Consider the linear regression model

$$Y = X\beta + \varepsilon,$$

with $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ observed, $\beta \in \mathbb{R}^p$ and $\varepsilon \sim \mathcal{N}_p(0, I_n)$.

- We consider the **high-dimensional** case where $n < p$.

- Assume that β is s_0-**sparse** $-$ only $s_0 \ll p$ coordinates are non-zero.

Problem Setup

- Consider the linear regression model

$$Y = X\beta + \varepsilon,$$

with $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ observed, $\beta \in \mathbb{R}^p$ and $\varepsilon \sim \mathcal{N}_p(0, I_n)$.

- We consider the **high-dimensional** case where $n < p$.

- Assume that β is s_0-**sparse** $-$ only $s_0 \ll p$ coordinates are non-zero.

- **Goal:** perform **inference** on β_K, for some set $K \subset \{1, \ldots, p\}$.

  (w.l.o.g. assume we want to perform inference on $\beta_{1:k}$ for some $k \geq 1$).

Bayesian Inference

Bayesian Inference

In our context, the Bayesian approach to inference is the following.

Bayesian Inference

In our context, the Bayesian approach to inference is the following.

1. Place a prior, Π, on the parameter β.

Bayesian Inference

In our context, the Bayesian approach to inference is the following.

1. Place a prior, Π, on the parameter β.

2. Observe data $\{(X_i, Y_i)\}_{i=1}^n$ from the model with likelihood $L(X, Y \mid \beta)$.

Bayesian Inference

In our context, the Bayesian approach to inference is the following.

1. Place a prior, Π, on the parameter β.
2. Observe data $\{(X_i, Y_i)\}_{i=1}^n$ from the model with likelihood $L(X, Y \mid \beta)$.
3. Combine the likelihood and the prior to arrive at a posterior distribution

$$\Pi(\beta \in B \mid X, Y) = \frac{\int_B L(X, Y \mid \beta) d\Pi(\beta)}{\int L(X, Y \mid \beta) d\Pi(\beta)}.$$

Bayesian Inference

In our context, the Bayesian approach to inference is the following.

1. Place a prior, Π, on the parameter β.

2. Observe data $\{(X_i, Y_i)\}_{i=1}^n$ from the model with likelihood $L(X, Y \mid \beta)$.

3. Combine the likelihood and the prior to arrive at a posterior distribution

$$\Pi(\beta \in B \mid X, Y) = \frac{\int_B L(X, Y \mid \beta) d\Pi(\beta)}{\int L(X, Y \mid \beta) d\Pi(\beta)}.$$

4. Use the posterior distribution to form estimates and credible regions for the parameter of interest, e.g.

$$\hat{\beta} = E_{\Pi(\beta \mid X, Y)}(\beta)$$

$$C_{0.95} \text{ s.t. } \Pi(\beta \in C_{0.95} \mid X, Y) = 0.95$$

Model Selection Prior

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* AOS, 2015.

Model Selection Prior

- The model selection prior considered in Castillo et. al (2015) results in a posterior with many desirable properties.

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* AOS, 2015.

Model Selection Prior

- The model selection prior considered in Castillo et. al (2015) results in a posterior with many desirable properties.

- It is given as follows:

   1. The sparsity s of β is drawn according to some ν a distribution on $\{1, \ldots, p\}$. E.g. $\nu = \mathrm{Binomial}(p, q)$, for some probability of inclusion q.

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* AOS, 2015.

Model Selection Prior

- The model selection prior considered in Castillo et. al (2015) results in a posterior with many desirable properties.

- It is given as follows:

  1. The sparsity s of β is drawn according to some ν a distribution on $\{1,\ldots,p\}$. E.g. $\nu = \mathrm{Binomial}(p, q)$, for some probability of inclusion q.

  2. The active set S given $|S| = s$ of β is drawn uniformly from the $\binom{p}{s}$ subsets of $\{1,\ldots,p\}$ of size s.

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* AOS, 2015.

Model Selection Prior

- The model selection prior considered in Castillo et. al (2015) results in a posterior with many desirable properties.

- It is given as follows:

  1. The sparsity s of β is drawn according to some ν a distribution on $\{1,\ldots,p\}$. E.g. $\nu = \mathrm{Binomial}(p, q)$, for some probability of inclusion q.

  2. The active set S given $|S| = s$ of β is drawn uniformly from the $\binom{p}{s}$ subsets of $\{1,\ldots,p\}$ of size s.

  3. $\beta_i | S \overset{ind}{\sim} \begin{cases} \text{Lap}(\lambda), & i \in S, \\ \delta_0, & i \notin S\,. \end{cases}$

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* AOS, 2015.

Model Selection Prior

- The model selection prior considered in Castillo et. al (2015) results in a posterior with many desirable properties.

- It is given as follows:

  1. The sparsity s of β is drawn according to some ν a distribution on $\{1,\ldots,p\}$. E.g. $\nu = \mathrm{Binomial}(p,q)$, for some probability of inclusion q.

  2. The active set S given $|S| = s$ of β is drawn uniformly from the $\binom{p}{s}$ subsets of $\{1,\ldots,p\}$ of size s.

  3. $\beta_i | S \overset{ind}{\sim} \begin{cases} \text{Lap}(\lambda), & i \in S, \\ \delta_0, & i \notin S\,. \end{cases}$

- We will denote this prior by $\Pi(\beta) = MS_p(\nu, \lambda)(\beta)$.

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* AOS, 2015.

Model Selection Prior

Model Selection Prior

- The model selection prior is *very* well tailored to the problem at hand.

Model Selection Prior

- The model selection prior is *very* well tailored to the problem at hand.
- There exist a wealth of appealing theoretical results for the resulting posterior.

Model Selection Prior

- The model selection prior is *very* well tailored to the problem at hand.
- There exist a wealth of appealing theoretical results for the resulting posterior.
- However, the posterior is difficult to compute, requiring 2^p integrations to evaluate the denominator.

Model Selection Prior

- The model selection prior is *very* well tailored to the problem at hand.
- There exist a wealth of appealing theoretical results for the resulting posterior.
- However, the posterior is difficult to compute, requiring 2^p integrations to evaluate the denominator.
- Approximate posterior sampling (e.g. via MCMC) is non-trivial and slow.

Variational Inference

Variational Inference

- If the posterior is hard to compute/sample from, VI aims to **approximate** the posterior with a simpler class of distributions.

Variational Inference

- If the posterior is hard to compute/sample from, VI aims to **approximate** the posterior with a simpler class of distributions.

- The following **M**ean-**F**ield **(MF)** class is commonly used:

$$\mathcal{Q} = \left\{ Q_{\mu,\tau,q} = \bigotimes_{i=1}^{p} q_i \, \mathcal{N}(\mu_i, \tau_i^2) + (1 - q_i)\delta_0 : q_i \in [0,1], \mu_i \in \mathbb{R}, \tau_i \in \mathbb{R}^+ \right\},$$

a product of **spike-and-slab** distributions with 'variational parameters' $\{(\mu_i, \tau_i, q_i)\}_{i=1}^{p}$.

Variational Inference

- If the posterior is hard to compute/sample from, VI aims to **approximate** the posterior with a simpler class of distributions.

- The following **M**ean-**F**ield **(MF)** class is commonly used:

$$\mathcal{Q} = \left\{ Q_{\mu,\tau,q} = \bigotimes_{i=1}^{p} q_i \, \mathcal{N}(\mu_i, \tau_i^2) + (1 - q_i)\delta_0 : q_i \in [0,1], \mu_i \in \mathbb{R}, \tau_i \in \mathbb{R}^+ \right\},$$

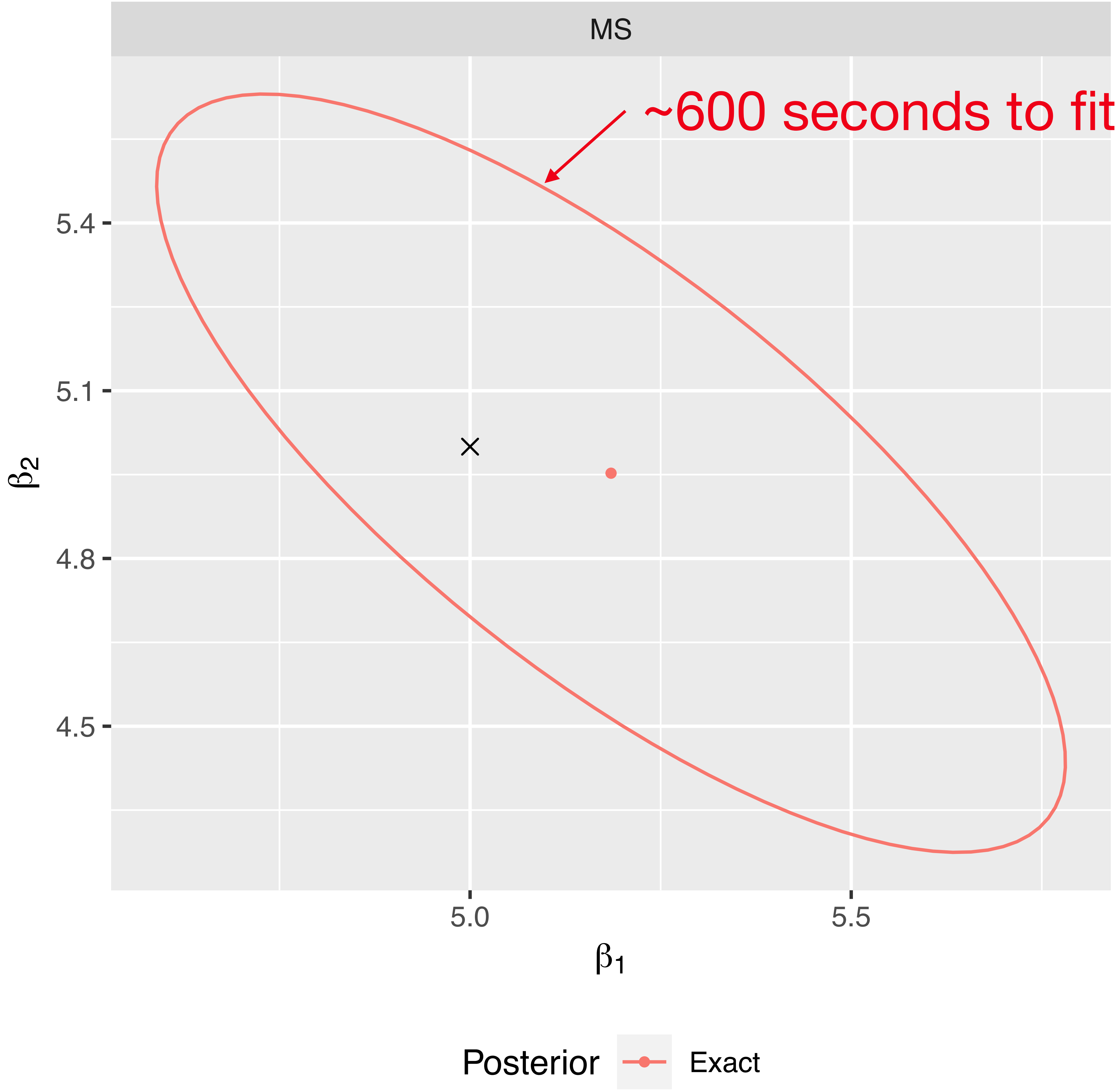
a product of **spike-and-slab** distributions with 'variational parameters' $\{(\mu_i, \tau_i, q_i)\}_{i=1}^{p}$.

- One then aims to find

$$\hat{Q} = \mathrm{argmin}_{Q \in \mathcal{Q}} \, \mathrm{KL}(Q \,|\, \Pi(\,\cdot\,|\, Y)),$$

which is commonly achieved using Coordinate Ascent Variational Inference (CAVI).

Demonstration



Left:

A region $C_{0.95}$ such that

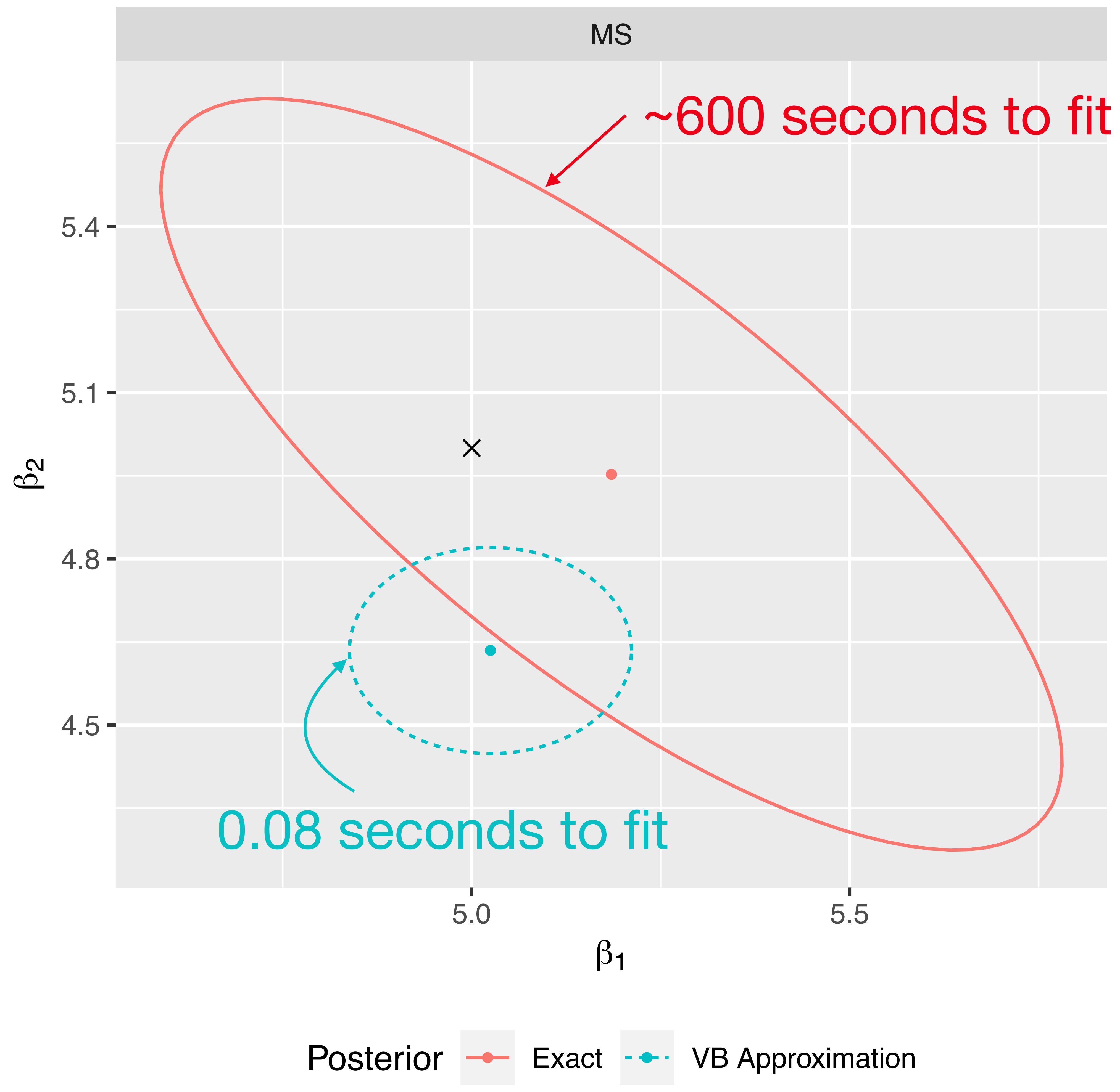
$$\Pi((\beta_1, \beta_2)^T \in C_{0.95} \,|\, X, Y) \approx 0.95.$$

Where:

- $X \in \mathbb{R}^{200 \times 400}$ has been drawn with $X_{i\cdot} \sim^{iid} \mathcal{N}_p(0, \Sigma_\rho)$,
- $\beta^0 = (\underbrace{5, \ldots, 5}_{s_0}, \underbrace{0, \ \ldots, 0}_{p-s_0})$ with sparsity $s_0 = 10$
- $Y = X\beta^0 + \varepsilon$, with $\varepsilon \sim \mathcal{N}_n(0, I_n)$.

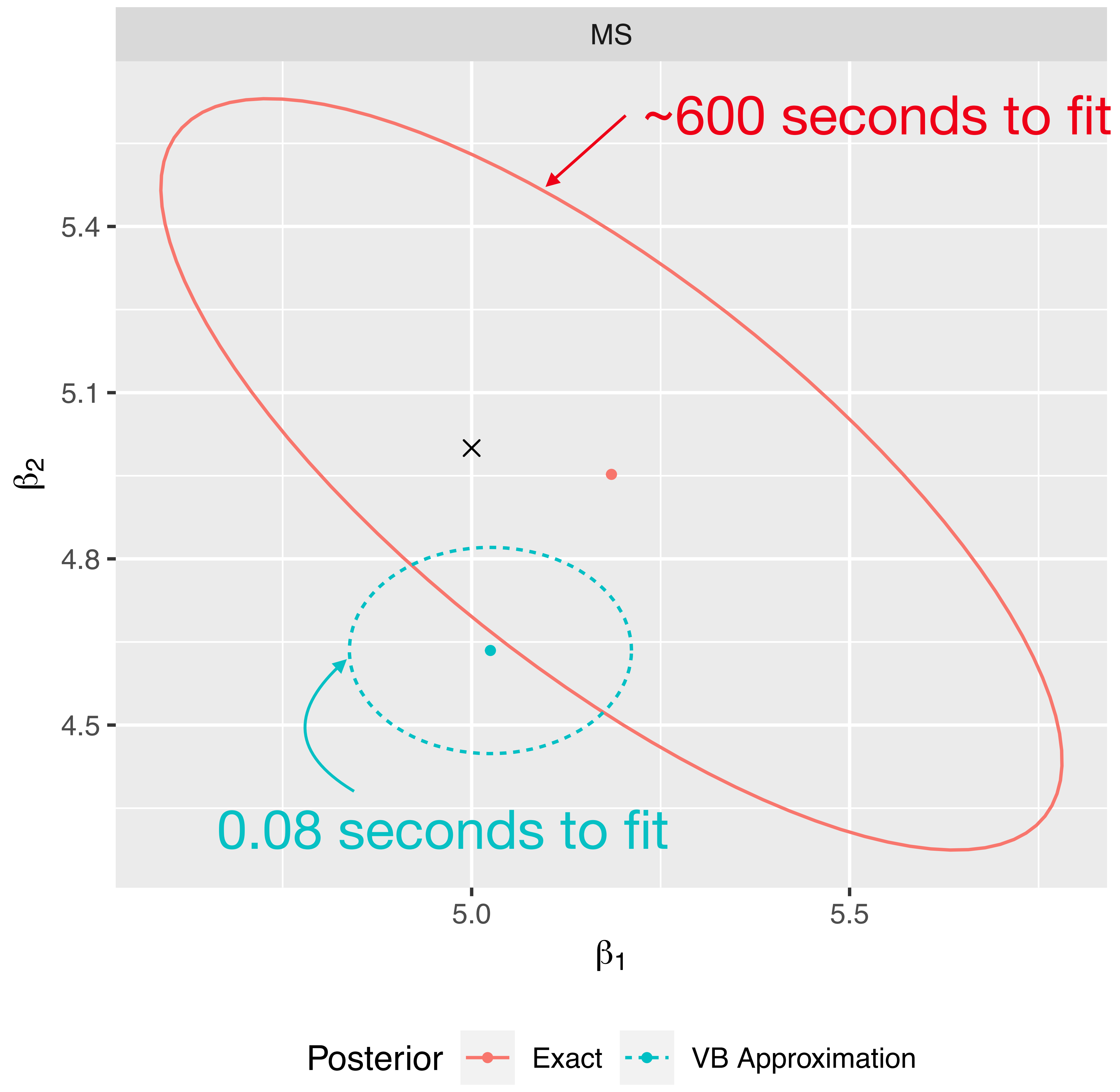
Obtained via a Gibbs' sampler implemented in C++.

Demonstration



Ray, K., Szabó, B.. *Variational Bayes for High-Dimensional Linear Regression with Sparse Priors.* JASA, 2020.

Demonstration

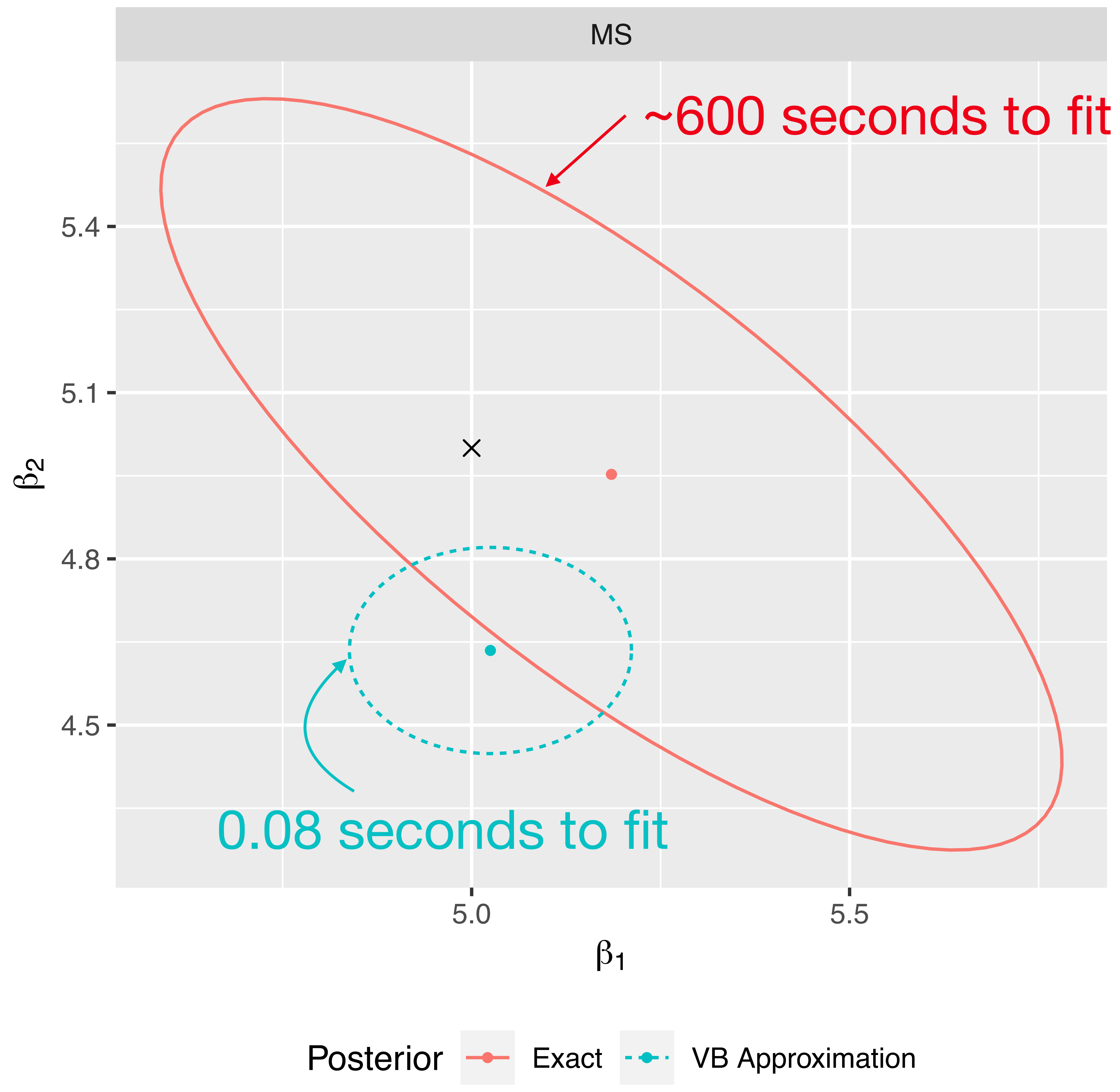


A MF VB approximation to the posterior (Ray and Szabó, 2020):

- Is **much faster**.
- Is **good** at **point estimation**.

Ray, K., Szabó, B.. *Variational Bayes for High-Dimensional Linear Regression with Sparse Priors.* JASA, 2020.

Demonstration



A MF VB approximation to the posterior (Ray and Szabó, 2020):

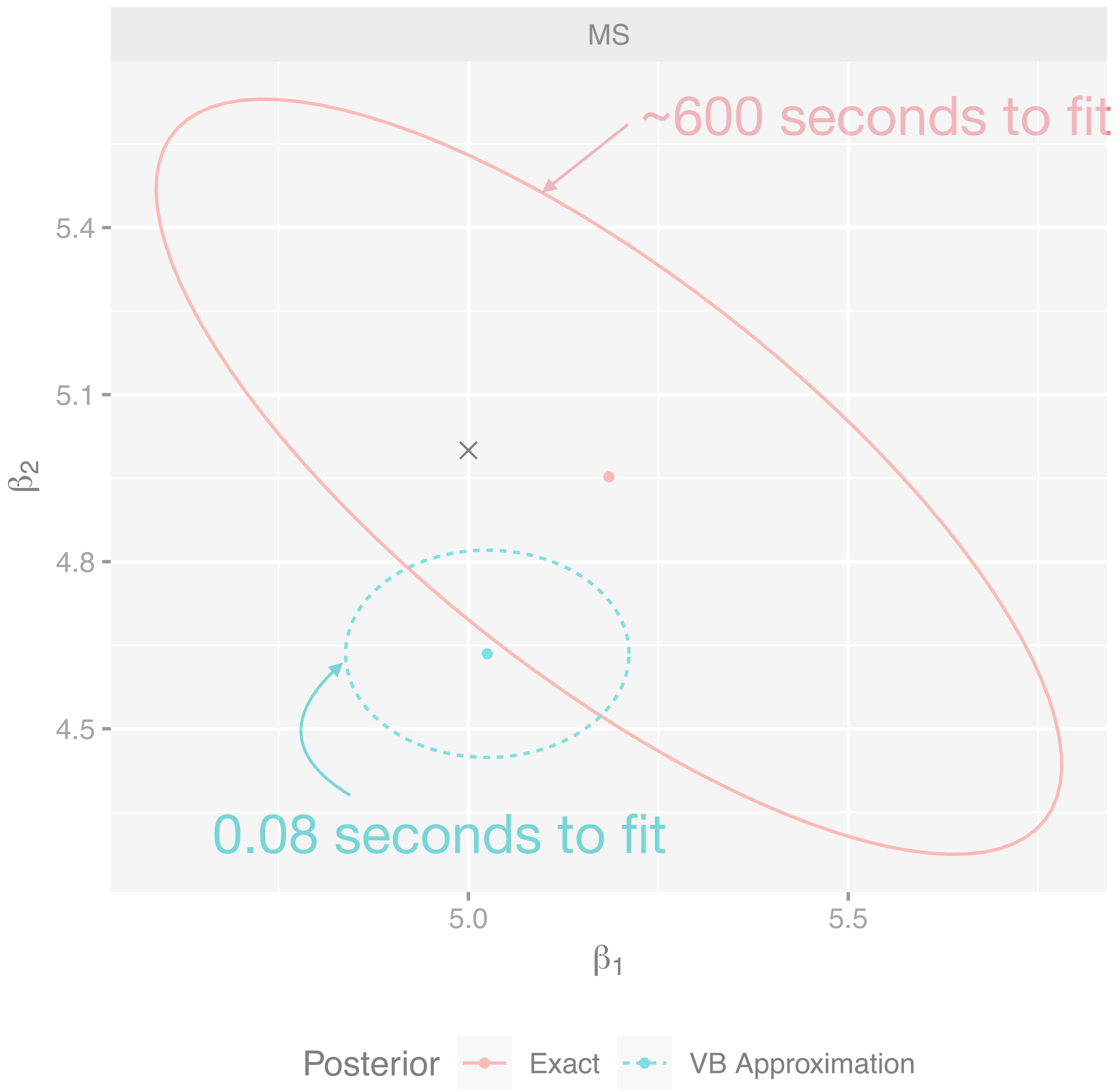
- Is **much faster**.
- Is **good** at **point estimation**.

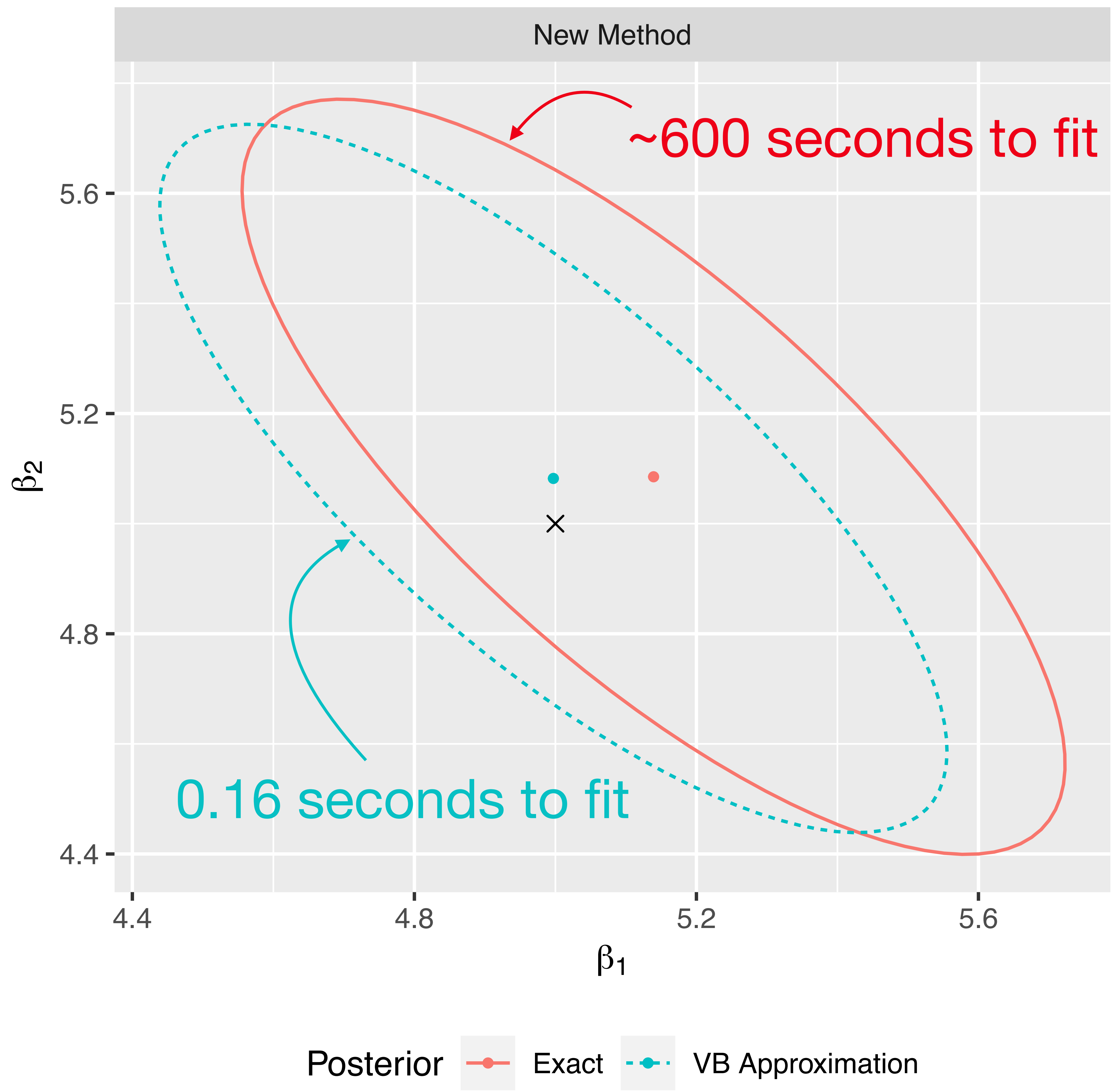
But:

- **Underestimates** the **posterior variance**.
- Has the **wrong covariance structure**.

Ray, K., Szabó, B.. *Variational Bayes for High-Dimensional Linear Regression with Sparse Priors.* JASA, 2020.

Our work


MS
~600 seconds to fit
0.08 seconds to fit
5.4
5.1
4.8
4.5
5.0
5.5
β_2
β_1
Posterior
Exact
VB Approximation


New Method
~600 seconds to fit
0.16 seconds to fit
5.6
5.2
4.8
4.4
4.4
4.8
5.2
5.6
β_2
β_1
Posterior
Exact
VB Approximation

Methodology

Decomposition of the likelihood

Yang, D.. *Posterior asymptotic normality for an individual coordinate in high-dimensional linear regression.* EJS, 2019.

Decomposition of the likelihood

- Suppose for the moment we are interested in the case $k = 1$, inference on the first coordinate of β.

Yang, D.. *Posterior asymptotic normality for an individual coordinate in high-dimensional linear regression.* EJS, 2019.

Decomposition of the likelihood

- Suppose for the moment we are interested in the case $k = 1$, inference on the first coordinate of β.
- Write $H = X_1 X_1^T / \|X_1\|_2^2$, the projection matrix onto $\mathrm{span}(X_1)$, and $\gamma_i := X_1^T X_i / \|X_1\|_2^2$ a rescaled correlation between the i^{th} and 1^{st} columns of X.

Yang, D.. *Posterior asymptotic normality for an individual coordinate in high-dimensional linear regression.* EJS, 2019.

Decomposition of the likelihood

- Suppose for the moment we are interested in the case $k = 1$, inference on the first coordinate of β.
- Write $H = X_1 X_1^T / \|X_1\|_2^2$, the projection matrix onto span(X_1), and $\gamma_i := X_1^T X_i / \|X_1\|_2^2$ a rescaled correlation between the i^{th} and 1^{st} columns of X.
- We make use of the following **decomposition of the likelihood** presented in Yang (2019).

Yang, D.. *Posterior asymptotic normality for an individual coordinate in high-dimensional linear regression.* EJS, 2019.

Decomposition of the likelihood

- Suppose for the moment we are interested in the case $k = 1$, inference on the first coordinate of β.
- Write $H = X_1 X_1^T / \|X_1\|_2^2$, the projection matrix onto span(X_1), and $\gamma_i := X_1^T X_i / \|X_1\|_2^2$ a rescaled correlation between the i^{th} and 1^{st} columns of X.
- We make use of the following **decomposition of the likelihood** presented in Yang (2019).

$$L(X, Y \,|\, \beta) \propto \exp\left\{-\frac{1}{2}\|Y - X\beta\|_2^2\right\}$$

$$\propto \underbrace{\exp\left\{-\frac{1}{2}\|HY - X_1\beta_1^*\|_2^2\right\}}_{L(X,Y|\beta_1^*)} \underbrace{\exp\left\{-\frac{1}{2}\|(I-H)Y - (I-H)X_{-1}\beta_{-1}\|_2^2\right\}}_{L(X,Y|\beta_{-1})}$$

where $\beta_1^* = \beta_1 + \sum_{i=2}^{p} \gamma_i \beta_i$.

Yang, D.. *Posterior asymptotic normality for an individual coordinate in high-dimensional linear regression.* EJS, 2019.

Definition of the prior

Definition of the prior

- On β_1^* we place a prior defined by a function g:

$$d\Pi(\beta_1^*) = g(\beta_1^*)d\beta_1^*,$$

where we may not require g to be a density (an improper prior).

Definition of the prior

- On β_1^* we place a prior defined by a function g:

$$d\Pi(\beta_1^*) = g(\beta_1^*)d\beta_1^*,$$

where we may not require g to be a density (an improper prior).

- On β_{-1} we will place a $(p-1)-$dimensional model selection prior:

$$\Pi(\beta_{-1}) = MS_{p-1}(\nu, \lambda)(\beta_{-1}).$$

Specifically, we use the spike-and-slab interpretation where $\nu \sim \mathrm{Binomial}(p-1, \theta)$ for some probability of inclusion θ.

Definition of the prior

- On β_1^* we place a prior defined by a function g:

$$d\Pi(\beta_1^*) = g(\beta_1^*)d\beta_1^*,$$

  where we may not require g to be a density (an improper prior).

- On β_{-1} we will place a $(p-1)-$dimensional model selection prior:

$$\Pi(\beta_{-1}) = MS_{p-1}(\nu, \lambda)(\beta_{-1}).$$

  Specifically, we use the spike-and-slab interpretation where $\nu \sim \mathrm{Binomial}(p-1, \theta)$ for some probability of inclusion θ.

- Since the **priors are independent**, and the likelihood decomposes as a **product of likelihoods** on β_1^* and β_{-1}, the **posterior distributions on each will also be independent.**

Posterior

Posterior

- The posterior distribution on β_1^* and β_{-1} is then given by

$$d\pi(\beta_{-1} \,|\, Y) \propto e^{-\frac{1}{2}\|\check{Y}-\check{W}\beta_{-1}\|_2^2} dMS_{p-1}(\nu, \lambda),$$

$$d\pi(\beta_1^* \,|\, Y) \propto e^{-\frac{1}{2}\|X_1\|_2^2\left(\beta_1^* - \frac{X_1^T Y}{\|X_1\|_2^2}\right)^2} g(\beta_1^*) d\beta_1^*,$$

independently.

Posterior

Transformed linear model.

- The posterior distribution on β_1^* and β_{-1} is then given by

$$d\pi(\beta_{-1}\,|\,Y) \propto \boxed{e^{-\frac{1}{2}\|\check{Y}-\check{W}\beta_{-1}\|_2^2}} dMS_{p-1}(\nu,\lambda),$$

$$d\pi(\beta_1^*\,|\,Y) \propto e^{-\frac{1}{2}\|X_1\|_2^2\left(\beta_1^* - \frac{X_1^TY}{\|X_1\|_2^2}\right)^2} g(\beta_1^*)d\beta_1^*,$$

independently.

Posterior

Transformed linear model.

- The posterior distribution on β_1^* and β_{-1} is then given by

$$d\pi(\beta_{-1}\,|\,Y) \propto \boxed{e^{-\frac{1}{2}\|\check{Y}-\check{W}\beta_{-1}\|_2^2}}\, dMS_{p-1}(\nu,\lambda),$$

1D Gaussian likelihood

$$d\pi(\beta_1^*\,|\,Y) \propto \boxed{e^{-\frac{1}{2}\|X_1\|_2^2\left(\beta_1^* - \frac{X_1^TY}{\|X_1\|_2^2}\right)^2}}\, g(\beta_1^*)d\beta_1^*,$$

independently.

Posterior

Transformed linear model.

- The posterior distribution on β_1^* and β_{-1} is then given by

$$d\pi(\beta_{-1} \mid Y) \propto \boxed{e^{-\frac{1}{2}\|\check{Y}-\check{W}\beta_{-1}\|_2^2}} dMS_{p-1}(\nu, \lambda),$$

1D Gaussian likelihood

$$d\pi(\beta_1^* \mid Y) \propto \boxed{e^{-\frac{1}{2}\|X_1\|_2^2\left(\beta_1^* - \frac{X_1^T Y}{\|X_1\|_2^2}\right)^2}} g(\beta_1^*) d\beta_1^*,$$

independently.

- Together, these induce a posterior distribution on β_1 which can be sampled from in the following way:

  1. Sample $\beta_{-1} \sim \Pi(\beta_{-1} \mid Y)$.
  2. Sample $\beta_1^* \sim \Pi(\beta_1^* \mid Y)$.
  3. Compute $\beta_1 = \beta_1^* - \sum_{i=2}^{p} \gamma_i \beta_i$

Sampling from $\Pi(\beta \mid Y)$

Sampling from $\Pi(\beta \mid Y)$

- $\Pi(\beta_1^* \mid Y)$ is 1-dimensional and **easy to sample from**.

Sampling from $\Pi(\beta \,|\, Y)$

- $\Pi(\beta_1^* \,|\, Y)$ is 1-dimensional and **easy to sample from**.
    - ⋆ We typically use choices of g for which the posterior is a Gaussian (e.g. if g a Gaussian density or $g \equiv 1$), or a mixture of truncated Gaussians (e.g. if g is a Laplace density).

Sampling from $\Pi(\beta \mid Y)$

- $\Pi(\beta_1^* \mid Y)$ is 1-dimensional and **easy to sample from**.

  - ⋆ We typically use choices of g for which the posterior is a Gaussian (e.g. if g a Gaussian density or $g \equiv 1$), or a mixture of truncated Gaussians (e.g. if g is a Laplace density).

  - ⋆ Even if the posterior is not analytically available — it is reasonably quick to perform approximate sampling in 1D.

Sampling from $\Pi(\beta \mid Y)$

- $\Pi(\beta_1^* \mid Y)$ is 1-dimensional and **easy to sample from**.
  - ⋆ We typically use choices of g for which the posterior is a Gaussian (e.g. if g a Gaussian density or $g \equiv 1$), or a mixture of truncated Gaussians (e.g. if g is a Laplace density).
  - ⋆ Even if the posterior is not analytically available — it is reasonably quick to perform approximate sampling in 1D.
- $\Pi(\beta_{-1} \mid Y)$ can be recognised as the posterior distribution using the model selection prior of Castillo et. al (2015), but with the transformed data $\check{Y}$, $\check{W}$ instead of Y, X.

Sampling from $\Pi(\beta \mid Y)$

- $\Pi(\beta_1^* \mid Y)$ is 1-dimensional and **easy to sample from**.
  - ⋆ We typically use choices of g for which the posterior is a Gaussian (e.g. if g a Gaussian density or $g \equiv 1$), or a mixture of truncated Gaussians (e.g. if g is a Laplace density).
  - ⋆ Even if the posterior is not analytically available — it is reasonably quick to perform approximate sampling in 1D.
- $\Pi(\beta_{-1} \mid Y)$ can be recognised as the posterior distribution using the model selection prior of Castillo et. al (2015), but with the transformed data $\check{Y}, \check{W}$ instead of Y, X.
  - ✤ We are presented with the same problem as before — **this posterior is difficult to sample from**.

Sampling from $\Pi(\beta \mid Y)$

- $\Pi(\beta_1^* \mid Y)$ is 1-dimensional and **easy to sample from**.
  - ⋆ We typically use choices of g for which the posterior is a Gaussian (e.g. if g a Gaussian density or $g \equiv 1$), or a mixture of truncated Gaussians (e.g. if g is a Laplace density).
  - ⋆ Even if the posterior is not analytically available — it is reasonably quick to perform approximate sampling in 1D.
- $\Pi(\beta_{-1} \mid Y)$ can be recognised as the posterior distribution using the model selection prior of Castillo et. al (2015), but with the transformed data $\check{Y}$, $\check{W}$ instead of Y, X.
  - ✤ We are presented with the same problem as before — **this posterior is difficult to sample from**.
  - ✤ For this reason, we opt to **approximate this marginal posterior** by a mean-field variational class.

Approximating $\Pi(\beta_{-1} \mid Y)$

Clara, G., Szabó, B., Ray, K.. *sparsevb: Spike-and-Slab Variational Bayes for Linear and Logistic Regression.* CRAN, 2021.

Approximating $\Pi(\beta_{-1} \mid Y)$

- We approximate $\Pi(\beta_{-1} \mid Y)$ with the mean-field variational class

$$\mathcal{Q}_{-1} = \left\{ Q_{\mu,\tau,q} = \bigotimes_{i=2}^{p} q_i \, \mathcal{N}(\mu_i, \tau_i^2) + (1 - q_i)\delta_0 : q_i \in [0,1], \mu_i \in \mathbb{R}, \tau_i \in \mathbb{R}^+ \right\}.$$

Clara, G., Szabó, B., Ray, K.. *sparsevb: Spike-and-Slab Variational Bayes for Linear and Logistic Regression.* CRAN, 2021.

Approximating $\Pi(\beta_{-1} \,|\, Y)$

- We approximate $\Pi(\beta_{-1} \,|\, Y)$ with the mean-field variational class

$$\mathcal{Q}_{-1} = \left\{ Q_{\mu,\tau,q} = \bigotimes_{i=2}^{p} q_i \, \mathcal{N}(\mu_i, \tau_i^2) + (1 - q_i)\delta_0 : q_i \in [0,1], \mu_i \in \mathbb{R}, \tau_i \in \mathbb{R}^+ \right\}.$$

- Taking

$$\hat{Q}_{-1} = \mathrm{argmin}_{Q_{-1} \in \mathcal{Q}_{-1}} \mathrm{KL}(Q_{-1} \,|\, \Pi(\beta_{-1} \,|\, Y)).$$

Clara, G., Szabó, B., Ray, K.. *sparsevb: Spike-and-Slab Variational Bayes for Linear and Logistic Regression.* CRAN, 2021.

Approximating $\Pi(\beta_{-1} \mid Y)$

- We approximate $\Pi(\beta_{-1} \mid Y)$ with the mean-field variational class

$$\mathcal{Q}_{-1} = \left\{ Q_{\mu,\tau,q} = \bigotimes_{i=2}^{p} q_i \, \mathcal{N}(\mu_i, \tau_i^2) + (1 - q_i)\delta_0 : q_i \in [0,1], \mu_i \in \mathbb{R}, \tau_i \in \mathbb{R}^+ \right\}.$$

- Taking

$$\hat{Q}_{-1} = \operatorname{argmin}_{Q_{-1} \in \mathcal{Q}_{-1}} \mathrm{KL}(Q_{-1} \mid \Pi(\beta_{-1} \mid Y)).$$

- Performing the **optimisation** via Coordinate Ascent Variational Inference (CAVI).

- This approximation has been studied already in Ray and Szabó (2020), and implemented in the `sparsevb` package (Clara, Szabo and Ray, 2021).

Clara, G., Szabó, B., Ray, K.. *sparsevb: Spike-and-Slab Variational Bayes for Linear and Logistic Regression.* CRAN, 2021.

Our Variational Posterior

Our Variational Posterior

- The marginal posterior $\Pi(\beta_1^* \mid Y)$, coupled with the variational posterior $\hat{Q}_{-1}(\beta_{-1})$ allows us to define a **variational posterior distribution** on β_1:

Our Variational Posterior

- The marginal posterior $\Pi(\beta_1^* \mid Y)$, coupled with the variational posterior $\hat{Q}_{-1}(\beta_{-1})$ allows us to define a **variational posterior distribution** on β_1:

$$\beta_1^* \sim \Pi(\beta_1^* \mid Y), \qquad \beta_{-1} \sim \hat{Q}_{-1}(\beta_{-1}), \text{ independently,}$$

$$\beta_1 = \beta_1^* - \sum_{i=2}^{p} \gamma_i \beta_i.$$

Our Variational Posterior

- The marginal posterior $\Pi(\beta_1^* \,|\, Y)$, coupled with the variational posterior $\hat{Q}_{-1}(\beta_{-1})$ allows us to define a **variational posterior distribution** on β_1:

$$\beta_1^* \sim \Pi(\beta_1^* \,|\, Y), \qquad \beta_{-1} \sim \hat{Q}_{-1}(\beta_{-1}), \text{ independently,}$$

$$\beta_1 = \beta_1^* - \sum_{i=2}^{p} \gamma_i \beta_i.$$

- From which it is clear how to draw samples of β_1.

Our Variational Posterior

- The marginal posterior $\Pi(\beta_1^* \mid Y)$, coupled with the variational posterior $\hat{Q}_{-1}(\beta_{-1})$ allows us to define a **variational posterior distribution** on β_1:

$$\beta_1^* \sim \Pi(\beta_1^* \mid Y), \qquad \beta_{-1} \sim \hat{Q}_{-1}(\beta_{-1}), \text{ independently,}$$

$$\beta_1 = \beta_1^* - \sum_{i=2}^{p} \gamma_i \beta_i.$$

- From which it is clear how to draw samples of β_1.

- We refer to inferences drawn using this posterior as <u>**I-SVB**</u>, **G-SVB** and **L-SVB**, where the first letter gives the distribution of g (Improper, Gaussian or Laplace).

Extension to higher dimensions

Extension to higher dimensions

- Our variational posterior on β_1 extends naturally to $\beta_{1:k}$, a $k-$dimensional sub parameter.

Extension to higher dimensions

- Our variational posterior on β_1 extends naturally to $\beta_{1:k}$, a $k-$dimensional sub parameter.

- Letting $\Sigma_k = (X_k^T X_k)^{-1}$, we now place a $k-$dimensional prior distribution on

$$\beta^*_{1:k} = \beta_{1:k} + \Sigma_k X_{1:k}^T X_{-k} \beta_{-k},$$

and an independent model selection prior on β_{-k}.

Extension to higher dimensions

- Our variational posterior on β_1 extends naturally to $\beta_{1:k}$, a $k-$dimensional sub parameter.

- Letting $\Sigma_k = (X_k^T X_k)^{-1}$, we now place a $k-$dimensional prior distribution on

$$\beta^*_{1:k} = \beta_{1:k} + \Sigma_k X^T_{1:k} X_{-k} \beta_{-k},$$

  and an independent model selection prior on β_{-k}.

- We once again use the actual posterior distribution on $\beta^*_{1:k}$ and approximate the posterior distribution of β_{-k} with the MF variational class.

Empirical Results

Empirical Performance (1D)

Javanmard, A., Montanari, A.. *Confidence Intervals and Hypothesis Testing for High-Dimensional Regression.* JMLR, 2014.

Zhang, C., Zhang, S.. *Confidence Intervals for Low-Dimensional Parameters in High-Dimensional Linear Models.* JRSSB, 2013.

Empirical Performance (1D)

- For our method here we use the improper prior $g \equiv 1$, (I-SVB), our supplement contains a comparison of the other methods.

Javanmard, A., Montanari, A.. *Confidence Intervals and Hypothesis Testing for High-Dimensional Regression.* JMLR, 2014.

Zhang, C., Zhang, S.. *Confidence Intervals for Low-Dimensional Parameters in High-Dimensional Linear Models.* JRSSB, 2013.

Empirical Performance (1D)

- For our method here we use the improper prior $g \equiv 1$, (I-SVB), our supplement contains a comparison of the other methods.
- We compare to:
  - Mean-field VI with the model selection prior (MF),
  - Javanmard and Montanari (2014) (JM),
  - Zhang and Zhang (2013) (ZZ),
  - The 'oracle' — the least squares estimate if one knows the true support S_0.

Javanmard, A., Montanari, A.. *Confidence Intervals and Hypothesis Testing for High-Dimensional Regression.* JMLR, 2014.

Zhang, C., Zhang, S.. *Confidence Intervals for Low-Dimensional Parameters in High-Dimensional Linear Models.* JRSSB, 2013.

Empirical Performance (1D)

- E.g. $(n, p, s_0) = (400, 1500, 32)$, $\mathrm{Corr}(X_i, X_j) = \rho$ for all $i \neq j$.

ρ	Method	Cov.	MAE	Length	Time
0	I-SVB	0.926	**0.045 ± 0.034**	**0.203 ± 0.009**	1.299 ± 0.238
	MF	0.818	0.059 ± 0.052	0.196 ± 0.007	0.688 ± 0.137
	ZZ	0.822	0.059 ± 0.045	0.201 ± 0.012	1.396 ± 0.259
	JM	0.724	0.145 ± 0.105	0.369 ± 0.015	17.77 ± 1.712
	Oracle	0.946	0.035 ± 0.029	0.206 ± 0.003	0.002 ± 0.001

Empirical Performance (1D)

- E.g. $(n, p, s_0) = (400, 1500, 32)$, $\mathrm{Corr}(X_i, X_j) = \rho$ for all $i \neq j$.

ρ	Method	Cov.	MAE	Length	Time
0	I-SVB	0.926	**0.045 ± 0.034**	**0.203 ± 0.009**	1.299 ± 0.238
	MF	0.818	0.059 ± 0.052	0.196 ± 0.007	0.688 ± 0.137
	ZZ	0.822	0.059 ± 0.045	0.201 ± 0.012	1.396 ± 0.259
	JM	0.724	0.145 ± 0.105	0.369 ± 0.015	17.77 ± 1.712
	Oracle	0.946	0.035 ± 0.029	0.206 ± 0.003	0.002 ± 0.001
0.5	I-SVB	0.992	**0.051 ± 0.038**	**0.331 ± 0.024**	1.384 ± 0.238
	MF	0.752	0.072 ± 0.064	0.196 ± 0.007	0.913 ± 0.159
	ZZ	0.846	0.069 ± 0.051	0.254 ± 0.017	1.396 ± 0.232
	JM	0.636	0.289 ± 0.217	0.655 ± 0.13	33.673 ± 5.227
	Oracle	0.934	0.031 ± 0.019	0.236 ± 0.011	0.001 ± 0.000

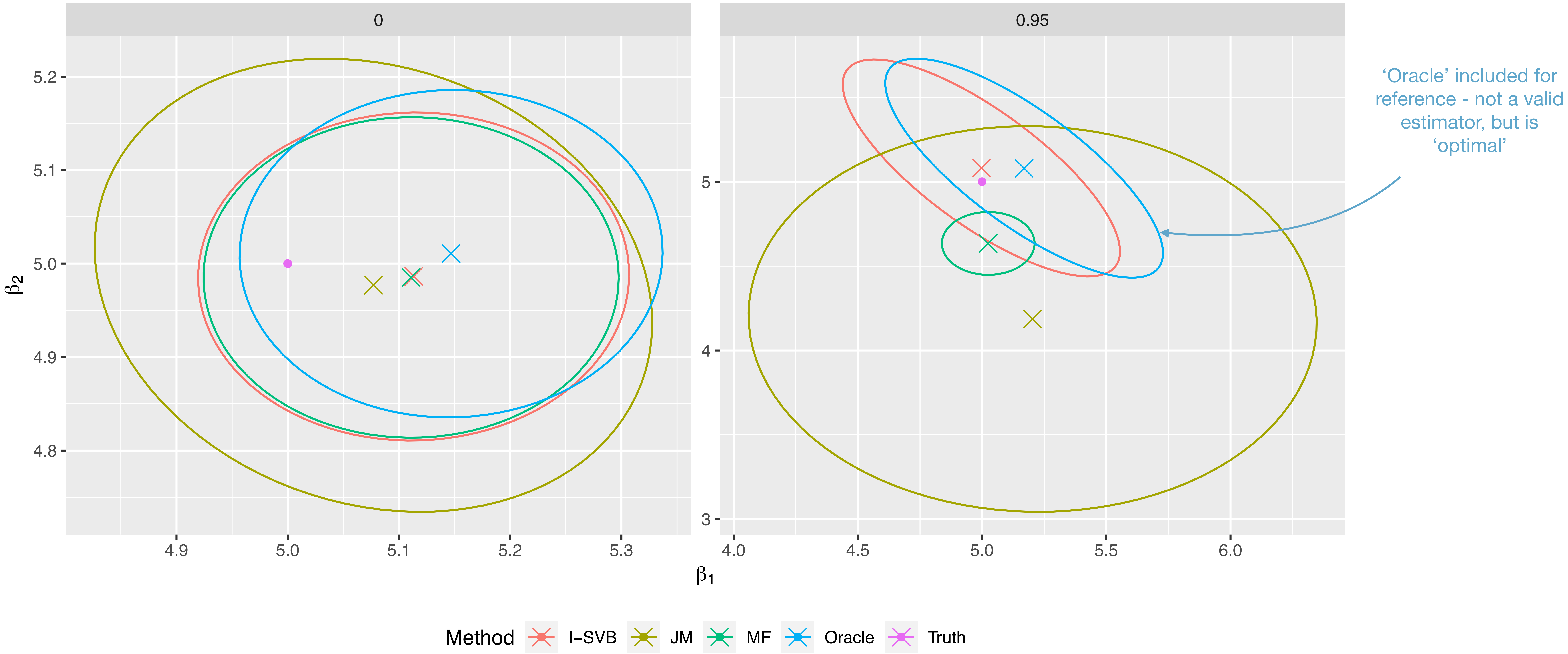
Empirical Performance (1D)

- E.g. $(n, p, s_0) = (400, 1500, 32)$, $\mathrm{Corr}(X_i, X_j) = \rho$ for all $i \neq j$.

ρ	Method	Cov.	MAE	Length	Time
0	I-SVB	0.926	**0.045 ± 0.034**	**0.203 ± 0.009**	1.299 ± 0.238
	MF	0.818	0.059 ± 0.052	0.196 ± 0.007	0.688 ± 0.137
	ZZ	0.822	0.059 ± 0.045	0.201 ± 0.012	1.396 ± 0.259
	JM	0.724	0.145 ± 0.105	0.369 ± 0.015	17.77 ± 1.712
	Oracle	0.946	0.035 ± 0.029	0.206 ± 0.003	0.002 ± 0.001
0.5	I-SVB	0.992	**0.051 ± 0.038**	**0.331 ± 0.024**	1.384 ± 0.238
	MF	0.752	0.072 ± 0.064	0.196 ± 0.007	0.913 ± 0.159
	ZZ	0.846	0.069 ± 0.051	0.254 ± 0.017	1.396 ± 0.232
	JM	0.636	0.289 ± 0.217	0.655 ± 0.13	33.673 ± 5.227
	Oracle	0.934	0.031 ± 0.019	0.236 ± 0.011	0.001 ± 0.000

Empirical Visualisation ($k = 2$)

Realisations of 2-D credible regions. In each plot, $n = 200$, $p = 400$, $s_0 = 10$, $X_{i\cdot} \sim \mathcal{N}(0, \Sigma_\rho^{AR})$, with ρ given in the title of each facet. The 'truth' $\beta^0_{1:2}$ is $(5,5)$ and non-zero elements of the nuisance parameter are given by $5 \approx \log n$.



Empirical Performance ($k > 1$)

- Same as before, but with $k-$dimensional credible regions.

(k, ρ)	Method	Cov.	L_2-error	Rel. Volume	Time
$(2, 0)$	I-SVB	0.960	$\mathbf{0.091 \pm 0.048}$	1.007 ± 0.081	0.301 ± 0.107
	MF	0.946	$\mathbf{0.091 \pm 0.048}$	$\mathbf{0.951 \pm 0.071}$	0.162 ± 0.080
	JM	0.952	0.112 ± 0.065	1.843 ± 0.314	1.012 ± 0.163
	Oracle	0.948	0.096 ± 0.065	1.000 ± 0.073	-
$(2, 0.5)$	I-SVB	0.966	0.127 ± 0.065	$\mathbf{1.517 \pm 0.118}$	0.206 ± 0.040
	MF	0.790	$\mathbf{0.125 \pm 0.064}$	0.532 ± 0.043	0.309 ± 0.071
	JM	0.742	0.342 ± 0.527	2.975 ± 0.402	6.526 ± 0.902
	Oracle	0.950	0.129 ± 0.068	1.000 ± 0.069	-
$(6, 0.5)$	I-SVB	0.966	$\mathbf{0.163 \pm 0.052}$	$\mathbf{1.325 \pm 0.177}$	1.591 ± 0.545
	MF	0.616	0.246 ± 0.859	0.157 ± 0.016	7.984 ± 2.167
	JM	0.144	0.474 ± 0.551	7.304 ± 1.964	32.330 ± 2.805
	Oracle	0.936	0.166 ± 0.054	1.000 ± 0.088	-

Theoretical Results

Semiparametric Bernstein-von Mises

Semiparametric Bernstein-von Mises

For $\hat{\beta}_1$ a sequence satisfying $\hat{\beta}_1 = \beta_1^0 + \frac{1}{n} X_1^T \varepsilon + o_{P_0}\left(\frac{1}{\sqrt{n}}\right)$.

With $\mathscr{L}_{\hat{Q}}(\sqrt{n}(\beta_1 - \hat{\beta}_1))$ the **marginal variational posterior distribution** of $\sqrt{n}(\beta_1 - \hat{\beta}_1)$,

Semiparametric Bernstein-von Mises

For $\hat{\beta}_1$ a sequence satisfying $\hat{\beta}_1 = \beta_1^0 + \frac{1}{n} X_1^T \varepsilon + o_{P_0}\left(\frac{1}{\sqrt{n}}\right)$.

With $\mathscr{L}_{\hat{Q}}(\sqrt{n}(\beta_1 - \hat{\beta}_1))$ the **marginal variational posterior distribution** of $\sqrt{n}(\beta_1 - \hat{\beta}_1)$,

under **standard assumptions on λ,** and assuming that the design matrix X **is 'compatible'**, if

Semiparametric Bernstein-von Mises

For $\hat{\beta}_1$ a sequence satisfying $\hat{\beta}_1 = \beta_1^0 + \frac{1}{n} X_1^T \varepsilon + o_{P_0}\left(\frac{1}{\sqrt{n}}\right)$.

With $\mathscr{L}_{\hat{Q}}(\sqrt{n}(\beta_1 - \hat{\beta}_1))$ the **marginal variational posterior distribution** of $\sqrt{n}(\beta_1 - \hat{\beta}_1)$,

under **standard assumptions on λ,** and assuming that the design matrix X **is 'compatible'**, if

$$\frac{\|X_1\|_2 \max_{i=2,\ldots,p} |\gamma_i|}{\max_{i=2,\ldots,p} \|(I-H)X_i\|_2} s_0 \sqrt{\log p} \to 0,$$

then

$$d_{BL}\left(\mathscr{L}_{\hat{Q}}(\sqrt{n}(\beta_1 - \hat{\beta}_1)), \mathscr{N}(0,1)\right) \xrightarrow{P_0} 0.$$

Semiparametric Bernstein-von Mises

Semiparametric Bernstein-von Mises

For example, if $X_{ij} \sim \mathcal{N}(0,1)$ i.i.d., then the condition

$$s_0 \log p / \sqrt{n} \to 0,$$

is sufficient for the previous assumption

$$\frac{\|X_1\|_2 \max_{i=2,\ldots,p} |\gamma_i|}{\max_{i=2,\ldots,p} \|(I-H)X_i\|_2} s_0 \sqrt{\log p} \to 0,$$

to hold.

Semiparametric Bernstein-von Mises

For example, if $X_{ij} \sim \mathcal{N}(0,1)$ i.i.d., then the condition

$$s_0 \log p/\sqrt{n} \to 0,$$

is sufficient for the previous assumption

$$\frac{\|X_1\|_2 \max_{i=2,\ldots,p} |\gamma_i|}{\max_{i=2,\ldots,p} \|(I-H)X_i\|_2} s_0 \sqrt{\log p} \to 0,$$

to hold.

Have a similar result in k dimensions, where we get under similar assumptions

$$d_{BL}\left(\mathscr{L}_{\hat{Q}}(L_k^{-1}(\beta_{1:k} - \hat{\beta}_{1:k})), \mathcal{N}_k(0,I_k)\right) \xrightarrow{P_0} 0,$$

for $L_k = \Sigma_k^{1/2} = (X_{1:k}^T X_{1:k})^{-1/2}$.

Summary

Summary

We introduce a:

- **scalable** (similar fit time order to fast frequentist methods)**,**
- **theoretically justified** (via a Semiparametric BvM),
- **accurate** (in practice),
- **simple** (easy to implement with existing packages),

Bayesian method for inference on a low-dimensional sub-parameter in a high-dimensional linear regression model.

Summary

We introduce a:

- **scalable** (similar fit time order to fast frequentist methods)**,**
- **theoretically justified** (via a Semiparametric BvM),
- **accurate** (in practice),
- **simple** (easy to implement with existing packages),

Bayesian method for inference on a low-dimensional sub-parameter in a high-dimensional linear regression model.

Future work:

- Extending to models other than the linear model presented here.

References

- Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* In: AOS, 2015.
- Ray, K., Szabó, B.. *Variational Bayes for High-Dimensional Linear Regression with Sparse Priors.* In: JASA, 2020.
- Clara, G., Szabó, B., Ray, K.. *sparsevb: Spike-and-Slab Variational Bayes for Linear and Logistic Regression.* In: CRAN, 2021.
- Javanmard, A., Montanari, A.. *Confidence Intervals and Hypothesis Testing for High-Dimensional Regression.* In: JMLR, 2014.
- Zhang, C., Zhang, S.. *Confidence Intervals for Low-Dimensional Parameters in High-Dimensional Linear Models.* In: JRSSB, 2013.

Our paper:

- Castillo, I., L'Huillier, A., Ray, K., Travis, L.. *A variational Bayes approach to debiased inference for low-dimensional parameters in high-dimensional linear regression.* 2024, *arXiv preprint: https://arxiv.org/abs/2406.12659*

Code: https://github.com/lukemmtravis/Debiased-SVB/

Empirical Visualisation ($k = 2$)

The covariance structure of each component of $\beta_{1:2}$.