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

Kolyan Ray

Joint work with Ismaël Castillo, Alice L'Huillier & Luke Travis

8 November 2024
Workshop on Advances in high/infinite-dimensional inference
Verona

Imperial College London

Sparsity

- Often have many variables, but only a few are relevant, e.g. finding subsets of genes responsible for a disease.
- Can model this via sparsity.

Sparsity

- Often have many variables, but only a few are relevant, e.g. finding subsets of genes responsible for a disease.
- Can model this via sparsity.
- Consider high-dimensional linear regression

$$Y = X\theta_0 + \sigma Z, \qquad Z \sim N_n(0, I_n),$$

where $X \in \mathbb{R}^{n\times p}$, $\theta_0 \in \mathbb{R}^p$ and $\sigma > 0$.

- We assume θ_0 is s_0-sparse:

$$s_0 = \#\{i : \theta_i \neq 0\}.$$

Interested in the case $p \gg n$ and $s_0 \ll p$.

Debiased inference

$$Y = X\theta_0 + \sigma Z, \qquad\qquad Z \sim N_n(0, I_n),$$

- **Goal:** statistical inference for a single or few coordinates $\theta_{1:k} = (\theta_1, \dots, \theta_k)^T$ of $\theta = (\theta_1, \dots, \theta_p)^T \in \mathbb{R}^p$.

Debiased inference

$$Y = X\theta_0 + \sigma Z, \qquad\qquad Z \sim N_n(0, I_n),$$

- **Goal:** statistical inference for a single or few coordinates $\theta_{1:k} = (\theta_1, \ldots, \theta_k)^T$ of $\theta = (\theta_1, \ldots, \theta_p)^T \in \mathbb{R}^p$.
- The LASSO

$$\hat{\theta}^{LASSO} = \arg\min_{\theta \in \mathbb{R}^p} \|Y - X\theta\|_2^2 + \lambda\|\theta\|_1$$

  is well-known to give biased inference for θ_1.
- Reason: it shrinks all coefficients to perform regularization.

Debiased inference

$$Y = X\theta_0 + \sigma Z, \qquad\qquad Z \sim N_n(0, I_n),$$

- Can debias the LASSO:

$$\hat{\theta}^d = \hat{\theta}^{LASSO} + \frac{1}{n} M X^T (Y - X\hat{\theta}^{LASSO}).$$

- Last term is estimate of bias.

Debiased inference

$$Y = X\theta_0 + \sigma Z, \qquad Z \sim N_n(0, I_n),$$

- Can debias the LASSO:

$$\hat{\theta}^d = \hat{\theta}^{LASSO} + \frac{1}{n} M X^T (Y - X\hat{\theta}^{LASSO}).$$

- Last term is estimate of bias.
- If M is sufficiently close to precision matrix of the covariates and $s_0 \ll \sqrt{n}/(\log p)$ then

$$\hat{\theta}_1^d \approx^d N(\theta_1, \sigma^2/n)$$

e.g. Zhang & Zhang (JRSSB 2014), van de Geer et al. (AOS 2014), Javanmard & Montanari (AOS 2018).

Debiased inference

$$Y = X\theta_0 + \sigma Z, \qquad\qquad Z \sim N_n(0, I_n),$$

- Can debias the LASSO:
$$\hat{\theta}^d = \hat{\theta}^{LASSO} + \frac{1}{n} M X^T (Y - X\hat{\theta}^{LASSO}).$$
- Last term is estimate of bias.
- If M is sufficiently close to precision matrix of the covariates and $s_0 \ll \sqrt{n}/(\log p)$ then
$$\hat{\theta}^d_1 \approx^d N(\theta_1, \sigma^2/n)$$
e.g. Zhang & Zhang (JRSSB 2014), van de Geer et al. (AOS 2014), Javanmard & Montanari (AOS 2018).
- Can be used to construct confidence intervals
$$P_{\theta_0}(\theta_{0,1} \in J_1(\alpha)) \geq 1 - \alpha - o(1).$$
- Can we do this in a scalable Bayesian way?

Model selection priors

$$Y = X\theta_0 + \sigma Z, \qquad\qquad Z \sim N_n(0, I_n)$$

- For the Bayesian: natural to model sparsity via the prior Π.
- Common priors:
  - Model selection priors
  - Shrinkage priors

Model selection priors

$$Y = X\theta_0 + \sigma Z, \qquad\qquad Z \sim N_n(0, I_n)$$

- For the Bayesian: natural to model sparsity via the prior Π.
- Common priors:
    - Model selection priors
    - Shrinkage priors
- Consider the spike and slab prior:

$$\theta_i \sim^{iid} w\varphi + (1-w)\delta_0$$

for $w \in [0,1]$ and a density φ.

Model selection priors

$$Y = X\theta_0 + \sigma Z, \qquad\qquad Z \sim N_n(0, I_n)$$

- For the Bayesian: natural to model sparsity via the prior Π.
- Common priors:
  - Model selection priors
  - Shrinkage priors
- Consider the spike and slab prior:

$$\theta_i \sim^{iid} w\varphi + (1-w)\delta_0$$

  for $w \in [0,1]$ and a density φ.
- We take Laplace slab $\varphi(x) = \frac{\lambda}{2}e^{-\lambda|x|}$ and hyperprior

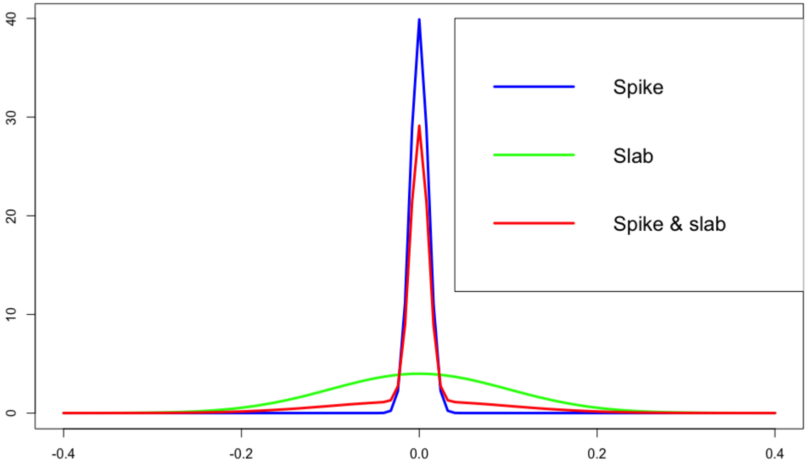
$$w \sim \mathrm{Beta}(a_0, b_0)$$

  to adapt to the unknown sparsity level s_0.

Model selection priors

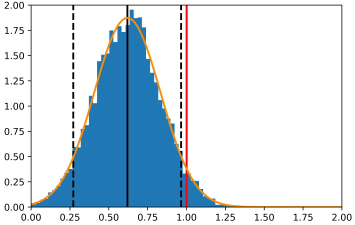
Spike and slab: e.g.

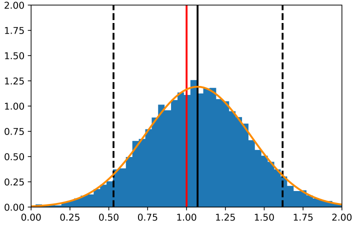
$$\theta_i \sim^{iid} 0.3 \times N(0, \sigma^2_{large}) + 0.7 \times N(0, \sigma^2_{small}).$$



- **Recall:** goal is to estimate single coordinate θ_1.

- **Recall:** goal is to estimate single coordinate θ_1.





- In high-dimensions the marginal posterior can pick up regularization bias $\implies$ bad UQ.
- Similar issues occur with "plug-in" estimators, e.g. double debiased machine learning methods (Chernozhukov et al.).

- A strong form of limit distribution is the (parametric) Bernstein-von Mises theorem. If $Y_1, \ldots, Y_n \sim^{iid} P_{\theta_0}$, then

$$\theta | Y_1, \ldots, Y_n \approx^d N\left(\hat{\theta}_n, \frac{1}{n} I_{\theta_0}^{-1}\right)$$

as $n \to \infty$, with $\hat{\theta}_n$ an efficient estimator and I_{θ_0} is the Fisher information.

- A strong form of limit distribution is the (parametric) Bernstein-von Mises theorem. If $Y_1, \ldots, Y_n \sim^{iid} P_{\theta_0}$, then

$$\theta | Y_1, \ldots, Y_n \approx^d N\left(\hat{\theta}_n, \frac{1}{n} I_{\theta_0}^{-1}\right)$$

as $n \to \infty$, with $\hat{\theta}_n$ an efficient estimator and I_{θ_0} is the Fisher information.

- Says posterior is asymptotically optimal from a frequentist perspective.
- Bayesian credible sets are frequentist confidence sets of optimal size.

- A strong form of limit distribution is the (parametric) Bernstein-von Mises theorem. If $Y_1, \dots, Y_n \sim^{iid} P_{\theta_0}$, then

$$\theta | Y_1, \dots, Y_n \approx^d N\left(\hat{\theta}_n, \frac{1}{n} I_{\theta_0}^{-1}\right)$$

as $n \to \infty$, with $\hat{\theta}_n$ an efficient estimator and I_{θ_0} is the Fisher information.

- Says posterior is asymptotically optimal from a frequentist perspective.
- Bayesian credible sets are frequentist confidence sets of optimal size.

- For sparse priors, can get posterior normality for entire θ under strong signal-to-noise conditions (Castillo et al. AOS 2015), e.g. strong model selection.

- Let $\gamma_i = X_1^T X_i / n$ be the (rescaled) correlation.
- Let θ have a model selection (e.g. spike and slab) prior with θ_1 only slab.

Theorem (Castillo, van der Pas, Ray, van der Vaart & Vuursteen (in preparation))

- Let $\gamma_i = X_1^T X_i / n$ be the (rescaled) correlation.
- Let θ have a model selection (e.g. spike and slab) prior with θ_1 only slab.

Theorem (Castillo, van der Pas, Ray, van der Vaart & Vuursteen (in preparation))

Let $\theta_0 \in \mathbb{R}^p$ be s_0-sparse. Assume that

- $\max_{2 \leq i \leq p} |\gamma_i| \leq c\sqrt{(\log p)/n}$ *(not too much correlation).*
- *X satisfies a compatibility condition.*
- $s_0 = o(\sqrt{n}/\log p)$.

- Let $\gamma_i = X_1^T X_i / n$ be the (rescaled) correlation.
- Let θ have a model selection (e.g. spike and slab) prior with θ_1 only slab.

Theorem (Castillo, van der Pas, Ray, van der Vaart & Vuursteen (in preparation))

Let $\theta_0 \in \mathbb{R}^p$ be s_0-sparse. Assume that

- $\max_{2 \le i \le p} |\gamma_i| \le c\sqrt{(\log p)/n}$ *(not too much correlation).*
- *X satisfies a compatibility condition.*
- $s_0 = o(\sqrt{n}/\log p)$.

Then the posterior distribution for θ_1 satisfies

$$\mathcal{L}(\sqrt{n}(\theta_1 - \hat{\theta}_1)|Y) \to^{P_{\theta_0}} N(0,1)$$

as $n \to \infty$, where $\hat{\theta}_1$ is an efficient estimator for θ_1.

- The condition
$$\max_{2\leq i\leq p}\left|\frac{X_1^T X_i}{n}\right| \leq c\sqrt{\frac{\log p}{n}}$$
ensures there is not too much correlation between X_1 and X_i.
- Model selection priors satisfy a semiparametric BvM for θ_1 under significantly weaker conditions.

- The condition

$$\max_{2 \le i \le p} \left| \frac{X_1^T X_i}{n} \right| \le c \sqrt{\frac{\log p}{n}}$$

ensures there is not too much correlation between X_1 and X_i.
- Model selection priors satisfy a semiparametric BvM for θ_1 under significantly weaker conditions.
- Set

$$\mathcal{S}_0 := \left\{ i > 1 : |\theta_{0i}| \gtrsim \frac{s_0 \sqrt{\log p}}{\sqrt{n}} \right\}$$

to be the large coordinates (easy to detect).
- We allow large correlation between θ_1 and $\mathcal{S}_0$, and need

$$\max_{j \in \mathcal{S}_0^c} \left| \frac{(X_1 - X_{\mathcal{S}_0} \hat{\beta}_{\mathcal{S}_0})^T X_j}{n} \right| \le c \sqrt{\frac{\log p}{n}}$$

where $\hat{\beta}_{\mathcal{S}_0}$ is the least square estimator for $X_1 = X_{\mathcal{S}_0} \beta_{\mathcal{S}_0} + \epsilon$.

- The condition

$$\max_{2\leq i\leq p}\left|\frac{X_1^T X_i}{n}\right| \leq c\sqrt{\frac{\log p}{n}}$$

ensures there is not too much correlation between X_1 and X_i.

- Model selection priors satisfy a semiparametric BvM for θ_1 under significantly weaker conditions.
- Set

$$\mathcal{S}_0 := \left\{ i > 1 : |\theta_{0i}| \gtrsim \frac{s_0\sqrt{\log p}}{\sqrt{n}} \right\}$$

to be the large coordinates (easy to detect).

- We allow large correlation between θ_1 and $\mathcal{S}_0$, and need

$$\max_{j\in\mathcal{S}_0^c}\left|\frac{(X_1 - X_{\mathcal{S}_0}\hat{\beta}_{\mathcal{S}_0})^T X_j}{n}\right| \leq c\sqrt{\frac{\log p}{n}}$$

where $\hat{\beta}_{\mathcal{S}_0}$ is the least square estimator for $X_1 = X_{\mathcal{S}_0}\beta_{\mathcal{S}_0} + \epsilon$.

- The condition

$$\max_{2\le i\le p}\left|\frac{X_1^T X_i}{n}\right| \le c\sqrt{\frac{\log p}{n}}$$

ensures there is not too much correlation between X_1 and X_i.
- Model selection priors satisfy a semiparametric BvM for θ_1 under significantly weaker conditions.
- Set

$$\mathcal{S}_0 := \left\{ i > 1 : |\theta_{0i}| \gtrsim \frac{s_0\sqrt{\log p}}{\sqrt{n}} \right\}$$

to be the large coordinates (easy to detect).
- We allow large correlation between θ_1 and $\mathcal{S}_0$, and need

$$\max_{j\in\mathcal{S}_0^c}\left|\frac{(X_1 - X_{\mathcal{S}_0}\hat{\beta}_{\mathcal{S}_0})^T X_j}{n}\right| \le c\sqrt{\frac{\log p}{n}}$$

where $\hat{\beta}_{\mathcal{S}_0}$ is the least square estimator for $X_1 = X_{\mathcal{S}_0}\beta_{\mathcal{S}_0} + \epsilon$.
- Suggests true Bayesian methods may already be quite good at debiasing.

- The condition
$$\max_{2\le i\le p}\left|\frac{X_1^T X_i}{n}\right| \le c\sqrt{\frac{\log p}{n}}$$
ensures there is not too much correlation between X_1 and X_i.
- Model selection priors satisfy a semiparametric BvM for θ_1 under significantly weaker conditions.
- Set
$$\mathcal{S}_0 := \left\{ i > 1 : |\theta_{0i}| \gtrsim \frac{s_0\sqrt{\log p}}{\sqrt{n}} \right\}$$
to be the large coordinates (easy to detect).
- We allow large correlation between θ_1 and $\mathcal{S}_0$, and need
$$\max_{j\in\mathcal{S}_0^c}\left|\frac{(X_1 - X_{\mathcal{S}_0}\hat\beta_{\mathcal{S}_0})^T X_j}{n}\right| \le c\sqrt{\frac{\log p}{n}}$$
where $\hat\beta_{\mathcal{S}_0}$ is the least square estimator for $X_1 = X_{\mathcal{S}_0}\beta_{\mathcal{S}_0} + \epsilon$.
- Suggests true Bayesian methods may already be quite good at debiasing.
- **Problem**: posterior is expensive to compute.

Computation

- Using Bayes formula:

$$\Pi(B|Y) = \frac{\int_B e^{-\frac{1}{2}\|Y-X\theta\|_2^2} d\Pi(\theta)}{\int_{\mathbb{R}^p} e^{-\frac{1}{2}\|Y-X\theta\|_2^2} d\Pi(\theta)}.$$

Computation

- Using Bayes formula:

$$\Pi(B|Y) = \frac{\int_B e^{-\frac{1}{2}\|Y-X\theta\|_2^2} d\Pi(\theta)}{\int_{\mathbb{R}^p} e^{-\frac{1}{2}\|Y-X\theta\|_2^2} d\Pi(\theta)}.$$

- **Problem:** full posterior is expensive to compute since model space has size $O(2^p)$.
- Standard MCMC methods are slow for p large (1000's).
- Discrete structure & high-dimensional multi-modal posterior $\implies$ difficult mixing.

Computation

- Using Bayes formula:

$$\Pi(B|Y) = \frac{\int_B e^{-\frac{1}{2}\|Y-X\theta\|_2^2} d\Pi(\theta)}{\int_{\mathbb{R}^p} e^{-\frac{1}{2}\|Y-X\theta\|_2^2} d\Pi(\theta)}.$$

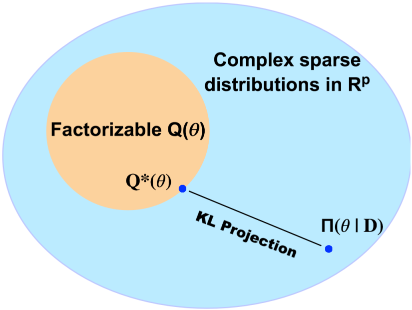
- **Problem:** full posterior is expensive to compute since model space has size $O(2^p)$.
- Standard MCMC methods are slow for p large (1000's).
- Discrete structure & high-dimensional multi-modal posterior $\Longrightarrow$ difficult mixing.

- **Alternative**: in variational Bayes (VB), propose a family of tractable distributions $\mathcal{Q}$ for θ.
- Solve the following optimization problem:

$$Q^* = \arg\min_{Q\in\mathcal{Q}} \mathrm{KL}(Q||\Pi(\cdot|Y)), \qquad \mathrm{KL}(q||p) = \int q \log\frac{q}{p}.$$

e.g. using gradient descent, coordinate descent.

Variational Bayes



- Tradeoff: simple vs complex class $\iff$ speed vs accuracy.
- Typically much faster than standard MCMC methods.

Variational Bayes

- Common choice is mean-field (factorizable) distributions:

$$Q(\theta) = Q_1(\theta_1) \otimes \cdots \otimes Q_p(\theta_p)$$

- Underestimates posterior variance/uncertainty:

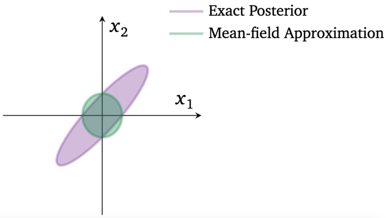


Figure: Figure 1 from Blei et al. (JASA 2017).

- Common choice is mean-field (factorizable) distributions:

$$Q(\theta) = Q_1(\theta_1) \otimes \cdots \otimes Q_p(\theta_p)$$

- Underestimates posterior variance/uncertainty:

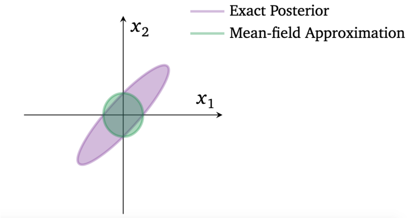


Figure: Figure 1 from Blei et al. (JASA 2017).

- **Cause:** correlation in posterior.
- **One solution:** use approximation that is 'mean-field' in a transformed space that decorrelates parameter of interest θ_1.

Variational Bayes for sparsity

- Want a family that preserves properties of spike and slab.

Variational Bayes for sparsity

- Want a family that preserves properties of spike and slab.
- Pick mean-field (factorizable) variational family $\mathcal{Q}$:

$$\theta_i \sim^{ind} \gamma_i N(\mu_i, \sigma_i^2) + (1 - \gamma_i)\delta_0,$$

  $\mu_i \in \mathbb{R}$, $\sigma_i^2 > 0$ and $\gamma_i \in [0, 1]$.
- Reduces posterior model size from $O(2^p)$ to $O(p)$.
- Mimic prior not posterior - breaks dependencies.

Variational Bayes for sparsity

- Want a family that preserves properties of spike and slab.
- Pick mean-field (factorizable) variational family $\mathcal{Q}$:

$$\theta_i \sim^{ind} \gamma_i N(\mu_i, \sigma_i^2) + (1 - \gamma_i)\delta_0,$$

  $\mu_i \in \mathbb{R}$, $\sigma_i^2 > 0$ and $\gamma_i \in [0, 1]$.
- Reduces posterior model size from $O(2^p)$ to $O(p)$.
- Mimic prior not posterior - breaks dependencies.
- Can be computed numerically using coordinate descent (non-convex optimization problem).

- Consider Gaussian design matrices with correlation ρ: $X_i \sim N_p(0, \Sigma)$ with

$$\Sigma_{jk} = \begin{cases} 1 & \text{if } j = k \\ \rho & \text{if } j \neq k. \end{cases}$$

ρ	MF VB			Proposed VB		
	Coverage	Length	Error	Coverage	Length	Error
0.00	0.92	0.40	0.01	0.96	0.40	0.01
0.25	0.92	0.39	0.02	0.94	0.45	0.01
0.50	0.80	0.39	0.02	0.97	0.59	0.01
0.90	0.05	0.39	0.73	0.97	1.43	0.01

- MF VB gets both bias and variance wrong.
- MF VB ignores correlation in credible interval lengths.

- Let $H = X_1 X_1^T / n$ denote the projection matrix onto $\text{span}(X_1)$ and $\gamma_i = X_1^T X_i / n$ the covariate correlation.

- Let $H = X_1 X_1^T/n$ denote the projection matrix onto $\text{span}(X_1)$ and $\gamma_i = X_1^T X_i/n$ the covariate correlation.
- **Intuition:** likelihood factorizes ($\approx$ 'independence'):

$$
\begin{aligned}
e^{\ell_n(Y)} &\propto e^{-\frac{1}{2}\|Y - X\theta\|_2^2} \\
&\propto e^{-\frac{1}{2}\|HY - X_1\theta_1^*\|_2^2} e^{-\frac{1}{2}\|(I-H)Y - (I-H)X_{-1}\theta_{-1}\|_2^2},
\end{aligned}
$$

where

$$
\theta_1^* = \theta_1 + \sum_{i \geq 2} \gamma_i \theta_i, \qquad \theta_{-1} = (\theta_2, \dots, \theta_p)^T.
$$

- Let $H = X_1 X_1^T/n$ denote the projection matrix onto span(X_1) and $\gamma_i = X_1^T X_i/n$ the covariate correlation.
- **Intuition:** likelihood factorizes ($\approx$ 'independence'):

$$
\begin{aligned}
e^{\ell_n(Y)} &\propto e^{-\frac{1}{2}\|Y - X\theta\|_2^2} \\
&\propto e^{-\frac{1}{2}\|HY - X_1\theta_1^*\|_2^2} e^{-\frac{1}{2}\|(I-H)Y - (I-H)X_{-1}\theta_{-1}\|_2^2},
\end{aligned}
$$

where

$$\theta_1^* = \theta_1 + \sum_{i\geq 2} \gamma_i \theta_i, \qquad \theta_{-1} = (\theta_2, \ldots, \theta_p)^T.$$

- $(\theta_1^*, \theta_{-1})$ less correlated compared to (θ_1, θ_{-1}) under the posterior.
- **Idea:** use a mean-field approximation for $(\theta_1^*, \theta_2, \ldots, \theta_p)$ **not** $(\theta_1, \theta_2, \ldots, \theta_p)$.

- To speed up computation, we use the prior introduced by Yang (EJS 2019) for this problem:

$$\theta_1^* \sim g, \qquad \qquad \theta_{-1} \sim \text{model selection prior}$$

  independent, where g is a slab distribution.
- True posterior still computationally expensive due to θ_{-1} part.

- To speed up computation, we use the prior introduced by Yang (EJS 2019) for this problem:

$$\theta_1^* \sim g, \qquad \theta_{-1} \sim \text{model selection prior}$$

  independent, where g is a slab distribution.
- True posterior still computationally expensive due to θ_{-1} part.
- We use a mean-field approximation for $(\theta_1^*, \theta_{-1})$, not $(\theta_1, \ldots, \theta_p)$:

$$\theta_i \sim^{ind} \gamma_i N(\mu_i, \sigma_i^2) + (1-\gamma_i)\delta_0, \qquad 2 \leq i \leq p$$
$$\theta_1^* \sim^{ind} q$$

- By posterior factorization, the KL minimizing q is simply the true posterior for θ_1^*.

$$e^{\ell_n(Y)} \propto e^{-\frac{1}{2}\|HY - X_1\theta_1^*\|_2^2} e^{-\frac{1}{2}\|(I-H)Y-(I-H)X_{-1}\theta_{-1}\|_2^2},$$

Independent priors on $(\theta_1^*, \theta_{-1}) \implies$ independent posteriors on $(\theta_1^*, \theta_{-1})$.

$$e^{\ell_n(Y)} \propto e^{-\frac{1}{2}\|HY - X_1\theta_1^*\|_2^2} e^{-\frac{1}{2}\|(I-H)Y - (I-H)X_{-1}\theta_{-1}\|_2^2},$$

Independent priors on $(\theta_1^*, \theta_{-1}) \implies$ independent posteriors on $(\theta_1^*, \theta_{-1})$.

1. Compute the true 1d posterior for θ_1^* based on likelihood

$$HY|\theta_1^* \sim N_n(X_1\theta_1^*, I_n).$$

2. Compute the MF VB approximation for θ_{-1} based on likelihood

$$(I-H)Y|\theta_{-1} \sim N_n((I-H)X_{-1}\theta_{-1}, I_n).$$

$$e^{\ell_n(Y)} \propto e^{-\frac{1}{2}\|HY - X_1\theta_1^*\|_2^2} e^{-\frac{1}{2}\|(I-H)Y-(I-H)X_{-1}\theta_{-1}\|_2^2},$$

Independent priors on $(\theta_1^*, \theta_{-1}) \implies$ independent posteriors on $(\theta_1^*, \theta_{-1})$.

1. Compute the true 1d posterior for θ_1^* based on likelihood

$$HY|\theta_1^* \sim N_n(X_1\theta_1^*, I_n).$$

2. Compute the MF VB approximation for θ_{-1} based on likelihood

$$(I-H)Y|\theta_{-1} \sim N_n((I-H)X_{-1}\theta_{-1}, I_n).$$

3. Sample $(\theta_1^*, \theta_{-1})$ independently and compute VB draw

$$\theta_1 = \theta_1^* - \sum_{i \geq 2} \gamma_i \theta_i.$$

Allows to plug-in standard computational tools, e.g. conjugacy, MCMC (Step 1), coordinate descent (Step 2).

- Preferable to use heavier tailed slabs for θ_1^*.
- Compare with debiased LASSO methods of Zhang & Zhang (2014) and Javanmard & Montanari (2014).
- Consider again Gaussian design matrices with correlation ρ.

	$(n, p, \rho) = (100, 1000, 0.5)$				$(200, 800, 0.9)$			
Method	Cov.	Len.	MAE	Time	Cov.	Len.	MAE	Time
I-SVB	0.94	2.24	0.44	0.39	1.00	1.87	0.18	0.71
MF	0.71	1.32	0.52	0.32	0.01	0.28	3.63	1.06
ZZ	0.84	2.82	0.65	0.40	0.94	1.06	0.22	0.63
JM	0.84	3.01	0.93	1.48	0.26	1.44	1.69	9.93

- Generally performs at least as well as frequentist methods.

Conditions on design matrix

$$Y = X\theta_0 + \sigma Z, \qquad\qquad X \in \mathbb{R}^{n\times p}.$$

- If $p > n$, θ_0 is not generally identifiable.
- e.g. if $X\theta_1 = X\theta_2$, how can the likelihood tell θ_1 and θ_2 apart?
- If θ_0 sparse, then 'local invertibility' of $X^T X$ is enough.

Conditions on design matrix

$$Y = X\theta_0 + \sigma Z, \qquad X \in \mathbb{R}^{n\times p}.$$

- If $p > n$, θ_0 is not generally identifiable.
- e.g. if $X\theta_1 = X\theta_2$, how can the likelihood tell θ_1 and θ_2 apart?
- If θ_0 sparse, then 'local invertibility' of $X^T X$ is enough.

Assumption (smallest scaled sparse singular value)

Assume there exists $\phi(s) > 0$ such that for all s-sparse vectors:

$$\|X\theta\|_2 \geq \phi(s)\|X\|\|\theta\|_2,$$

where $\|X\| = \max_{1\leq j\leq p}\|X_{\cdot j}\|$ is the maximal Euclidean column norm. $\phi(s)$ is called the smallest scaled singular value of dimension s.

- For s-sparse vectors:

$$\|X(\theta_1 - \theta_2)\|_2 \geq \phi(s)\|X\|\|\theta_1 - \theta_2\|_2.$$

- e.g. orthogonal matrices, i.i.d. random matrices.

Let $\gamma_i = X_1^T X_i/n$ be the (rescaled) correlation.

Theorem (Castillo, L'Huillier, Ray, Travis)

Let $\theta_0 \in \mathbb{R}^p$ be s_0-sparse. Assume that

- $$\max_{2 \le i \le p} |\gamma_i| s_0 \sqrt{\log p} \to 0$$

  (enough sparsity and not too much correlation).
- *X satisfies a compatibility condition.*

Theoretical guarantees

Let $\gamma_i = X_1^T X_i / n$ be the (rescaled) correlation.

Theorem (Castillo, L'Huillier, Ray, Travis)

Let $\theta_0 \in \mathbb{R}^p$ be s_0-sparse. Assume that

- $$\max_{2 \le i \le p} |\gamma_i| s_0 \sqrt{\log p} \to 0$$

  (enough sparsity and not too much correlation).
- *X satisfies a compatibility condition.*

Then under the VB method,

$$\mathcal{L}(\sqrt{n}(\theta_1 - \hat{\theta}_1)) \to^{P_{\theta_0}} N(0,1)$$

as $n \to \infty$, where $\hat{\theta}_1$ is an efficient estimator for θ_1.

- Conditions broadly similar to Yang (EJS 2019).
- We need additional conditions for lighter tailed distributions g.

Idea of the proof:

1. Likelihood factorizes in $(\theta_1^*, \theta_{-1})$, so independent priors give independent posteriors.
2. Use parametric Bernstein-von Mises techniques to get asymptotic normality of $\sqrt{n}(\theta_1^* - \hat{\theta}_1^*)$ under the variational distribution for θ_1^*.
3. Relate

$$(\theta_1 - \hat{\theta}_1)|Y = (\theta_1^* - \hat{\theta}_1^*)|Y + (\theta_1 - \theta_1^*)|Y - (\hat{\theta}_1 - \hat{\theta}_1^*)$$

Difference roughly reduces to controlling $\|\theta_{-1} - \theta_{0,-1}\|_1$ to prevent bias. $\implies$ use contraction rates for sparse VB method (Ray & Szabó JASA 2022).

- Extends straightforwardly to the multidimensional case where we are interested in inference on $\theta_{1:k} = (\theta_1, \ldots, \theta_k)^T$.

- Extends straightforwardly to the multidimensional case where we are interested in inference on $\theta_{1:k} = (\theta_1, \ldots, \theta_k)^T$.
- Set

$$\theta^*_{1:k} = \theta_{1:k} + A(X_1, \ldots, X_k)\theta_{-k} \qquad \left(= \theta_1 + \sum_{i \geq 2} \gamma_i \theta_i \right).$$

- Prior:

$$\theta^*_{1:k} \sim g, \qquad \theta_{-k} \sim \text{model selection prior}$$

- Extends straightforwardly to the multidimensional case where we are interested in inference on $\theta_{1:k} = (\theta_1, \dots, \theta_k)^T$.
- Set

$$\theta^*_{1:k} = \theta_{1:k} + A(X_1, \dots, X_k)\theta_{-k} \qquad \left(= \theta_1 + \sum_{i \geq 2} \gamma_i \theta_i \right).$$

- Prior:

$$\theta^*_{1:k} \sim g, \qquad \theta_{-k} \sim \text{model selection prior}$$

- Variational approximation:

$$\theta^*_{1:k} \sim \text{posterior} \qquad \theta_{-k} \sim \text{mean field spike and slab.}$$

Theorem (Castillo, L'Huillier, Ray, Travis)

Under the analogous k-dimensional conditions to before, the VB method satisfies

$$\theta_{1:k} \approx^d N_k\left(\hat{\theta}_{1:k}, (X_{1:k}^T X_{1:k})^{-1}\right)$$

as $n \to \infty$, where $\hat{\theta}_{1:k}$ is an efficient estimator for $\theta_{1:k}$.

Theorem (Castillo, L'Huillier, Ray, Travis)

Under the analogous k-dimensional conditions to before, the VB method satisfies

$$\theta_{1:k} \approx^d N_k \left(\hat{\theta}_{1:k}, (X_{1:k}^T X_{1:k})^{-1} \right)$$

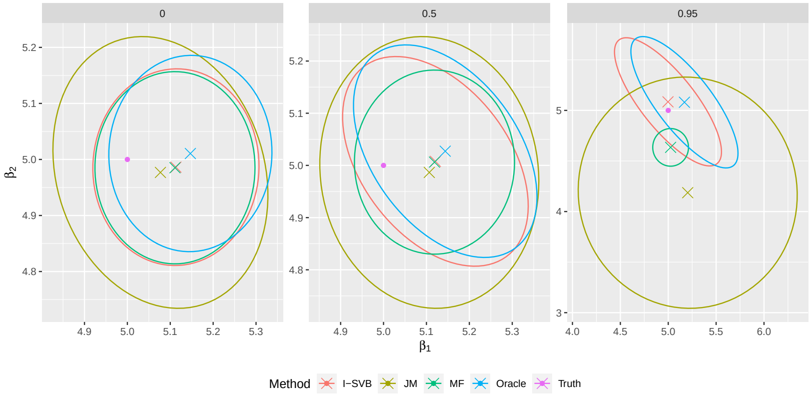
as $n \to \infty$, where $\hat{\theta}_{1:k}$ is an efficient estimator for $\theta_{1:k}$.

- Motivates using approximate k-dimensional VB credible set for $\theta_{1:k}$:

$$C_\alpha = \left\{ v \in \mathbb{R}^k : (v - \hat{\theta}_{1:k})^T \hat{\Sigma}_{1:k}^{-1} (v - \hat{\theta}_{1:k}) \leq \chi_k^2(\alpha) \right\}$$

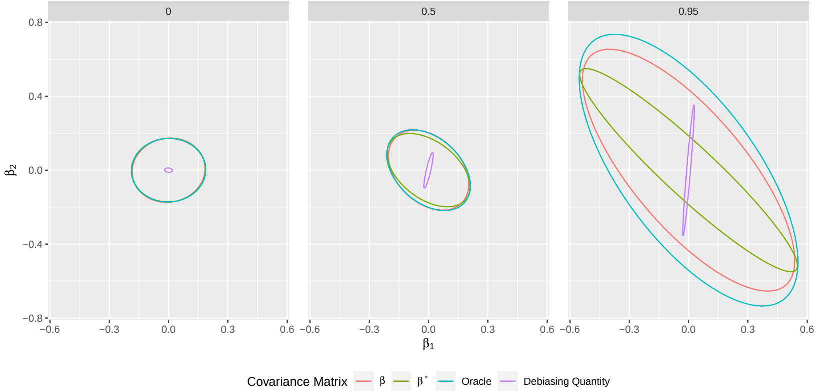
with $\chi_k^2(\alpha)$ the α-quantile of the χ_k^2 distribution, $\hat{\theta}_{1:k}$ the posterior mean and $\hat{\Sigma}_{1:k}$ the posterior covariance.

Consider estimating $\theta_{1:2} = (\theta_1, \theta_2)$ $(k = 2)$ for increasing covariate correlation ρ.

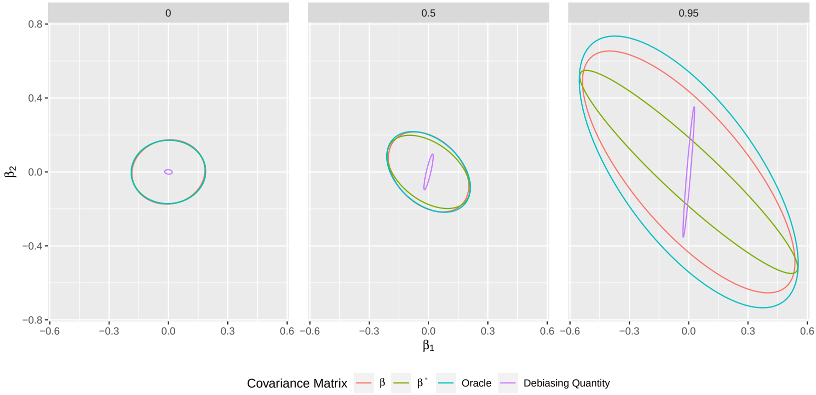


Our method seems to be close to the 'oracle' OLS based on regressing $Y = X_{S_0}\theta_{S_0} + Z$ if you *knew* the true low-dimensional support of θ_0 (note: not a valid method!).

- One can think of the difference $\beta^*_{1:2} - \beta_{1:2}$ as the 'debiasing' quantity.
- Plot its covariance contours under the VB posterior:



- One can think of the difference $\beta^*_{1:2} - \beta_{1:2}$ as the 'debiasing' quantity.
- Plot its covariance contours under the VB posterior:



- Does more than just counteract variance underestimate of MF VB: does a covariance correction.
- Seems to outperform frequentist methods in practice.

- Compare with debiased LASSO method of Javanmard & Montanari (2014) for estimating $\theta_{1:2} = (\theta_1, \theta_2)^T$.
- Consider again Gaussian design matrices with correlation ρ, $n = 200$, $p = 400$, $s_0 = 10$.

	$\rho = 0$			$\rho = 0.5$		
	Cov.	Rel. Vol.	L^2-error	Cov.	Rel. Vol.	L^2-error
I-SVB	0.96	1.01	0.09	0.97	1.51	0.13
MF	0.95	0.95	0.09	0.79	0.53	0.13
JM	0.95	1.84	0.11	0.74	2.98	0.34
Oracle	0.95	1.00	0.10	0.95	1.00	0.13

- Competitive in terms of computational time.
- Can be a bit conservative in highly correlated settings.

Summary

- Proposed a variational Bayes approach to estimating one (or several) coordinates in high-dimensional linear regression.
- **Idea:** use a factorization that decorrelates the functional of interest from high-dimensional nuisance parameter.
- Can be thought of as choosing a variational family tailored for the specific functional θ_1.

Summary

- Proposed a variational Bayes approach to estimating one (or several) coordinates in high-dimensional linear regression.
- **Idea:** use a factorization that decorrelates the functional of interest from high-dimensional nuisance parameter.
- Can be thought of as choosing a variational family tailored for the specific functional θ_1.

- Gives accurate and fast performance, which is competitive with the debiased LASSO in practice.
- Heavier tailed slabs perform best, e.g. improper priors.
- Semiparametric Bernstein-von Mises theorem justifies this procedure from a frequentist perspective.