Bayesian Sparse Linear Regression with Unknown Symmetric Error

Minwoo Chae

Joint work with

Lizhen Lin  
David B. Dunson

1Department of Mathematics, The University of Texas at Austin

2Department of Statistics and Data Sciences, The University of Texas at Austin

3Department of Statistical Science, Duke University

June 17, 2016

BU-KEIO 2016 Workshop
Outline

1 Introduction

2 Sparse linear model

3 Linear model with unknown error distribution

4 Asymptotic results
Outline

1. Introduction
2. Sparse linear model
3. Linear model with unknown error distribution
4. Asymptotic results
Symmetric location problem

\[ Y_i = \mu + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \eta(\cdot) \text{ (unknown)} \]

If \( \eta \) is symmetric, efficient and adaptive estimation of \( \mu \) is possible. [Beran, 1974; Stone 1975; ...]

Linear regression [Bickel, 1982]:

\[ \mu = x_i^T \theta, \quad \theta \in \mathbb{R}^p, \quad i = 1, \ldots, n. \]

For Bayesian, the semi-parametric Bernstein-von Mises (BvM) theorem holds. [Chae, Kim and Kleijn, 2016]

We study a Bayesian approach when \( p \) is large.
Bayesian paradigm

A parameter $\theta$ is generated according to a prior distribution $\Pi$.

Conditional on $\theta$, the data $X$ is generated according to a density $p_\theta$.

For given observed data $X$, statistical inferences are based on the posterior distribution:

$$d\Pi(\theta|X) \propto p_\theta(X)d\Pi(\theta).$$

Typically, the posterior distribution can be approximated via MCMC.
Bayesian asymptotics

A frequentist would like to know their performance in a frequentist viewpoint.

Assume that the data $X_1, \ldots, X_n$ is generated according to a given parameter $\theta_0$ and consider the posterior $\Pi(\theta \in \cdot | X_1, \ldots, X_n)$.

For large enough $n$, we want $\Pi(\theta \in \cdot | X_1, \ldots, X_n)$ to put most of its mass near $\theta_0$ for most $X_1, \ldots, X_n$. 
Assume that a parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is regular and $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta_0}$, where $\theta_0 \in \Theta$.

**THEOREM** (Bernstein-von Mises) [Le Cam and Yang, 1990] For any prior with positive density around $\theta_0$,

$$\left\| \Pi(\cdot|X_1, \ldots, X_n) - N(\hat{\theta}_n, I_{\theta_0}^{-1}/n) \right\|_{TV} \xrightarrow{P} 0,$$

where $\hat{\theta}_n$ is an efficient estimator for $\theta$ and $I_{\theta_0}$ is the Fisher information matrix.

The Bayesian credible interval is a standard confidence interval.
\[ \theta \sim \text{Beta}(5, 1), \quad X_1, \ldots, X_n | \theta \overset{iid}{\sim} \text{Bernoulli}(\theta), \quad \theta_0 = 1/2 \]
A frequentist would like to know their performance in a frequentist viewpoint.

Assume that the data $X_1, \ldots, X_n$ is generated according to a given parameter $\theta_0$ and consider the posterior $\Pi(\theta \in \cdot | X_1, \ldots, X_n)$.

For large enough $n$, we want $\Pi(\theta \in \cdot | X_1, \ldots, X_n)$ to put most of its mass near $\theta_0$ for most $X_1, \ldots, X_n$.

For infinite dimensional $\theta$, the choice of the prior is important.
Semi-parametric BvM (fixed $p$)

$$Y_i = x_i^T \theta + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \eta(\cdot) \text{ (unknown)}$$

Put a symmetrized Dirichlet process (DP) mixture prior on $\eta$.

**THEOREM** [Chae, Kim and Kleijn, 2016] For any prior on $\theta$, with positive density around $\theta_0$,

$$\left\| \Pi(\theta \in \cdot | X_1, \ldots, X_n) - N(\hat{\theta}_n, I_{\theta_0, \eta_0}^{-1}/n) \right\|_{TV} \xrightarrow{P} 0,$$

where $\hat{\theta}_n$ is an efficient estimator for $\theta$ and $I_{\theta_0, \eta_0}$ is the efficient information matrix.

What if $p$ is large?
Outline

1. Introduction

2. Sparse linear model

3. Linear model with unknown error distribution

4. Asymptotic results
Sparse linear model

Consider the linear regression model

\[ Y_i = x_i^T \theta + \epsilon_i, \quad i = 1, \ldots, n \]

where \( \theta = (\theta_1, \ldots, \theta_p)^T \) and possibly \( p \gg n \).

Simply, \( Y = X\theta + \epsilon \).

A sparse model assumes that most of \( \theta_i \)'s are (nearly) zero.

We apply full Bayesian procedures, and express the sparsity in priors.
A prior $\Pi_{\Theta}$ for $\theta \in \mathbb{R}^p$ can be constructed as follows:

1. **(Dimension)** Choose $s$ from prior $\pi_p$ on $\{0, 1, \ldots, p\}$.
2. **(Model)** Choose $S \subset \{0, 1, \ldots, p\}$ of size $|S| = s$ at random.
3. **(Nonzero coeff.)** Choose $\theta_S = (\theta_i)_{i \in S}$ from density $g_S$ on $\mathbb{R}^{|S|}$ and set $\theta_{Sc} = 0$.

Formally, 

$$(S, \theta) \mapsto \pi_p(s) \frac{1}{\binom{p}{s}} g_S(\theta_S) \delta_0(\theta_{Sc}).$$

Prior $\pi_p$ on the dimension controls the level of sparsity.
Sparse prior: example

Spike and slab [Ishwaran and Rao 2005; and many authors]

\[ s \sim \text{Binomial}(p, r) \]

for some \( r \in (0, 1) \), similarly,

\[ \theta_i \sim (1 - r)\delta_0 + rG, \quad \forall i \leq p \]

for some continuous distribution \( G \).

Good asymptotic properties if \( r \sim \text{Beta}(1, p^u) \) for some \( u > 1 \) and tail of \( G \) is as thick as Laplace. [Castillo and van der Vaart, 2015]
Sparse prior: example

**Complexity prior** [Castillo and van der Vaart, 2012]

\[ \pi_p(s) \propto c^{-s} p^{-as}, \quad s = 0, 1, \ldots, p \]

for some constants \( a, c > 0 \).

Roughly,

\[ \pi_p(s) \propto \binom{p}{s}^{-1}, \quad \text{for } s \ll p. \]
Continuous shrinkage priors that peaks near zero.

Typically, scale mixtures of normals: for $i = 1, \ldots, p,$

$$\theta_i | \tau^2, \lambda_i^2 \sim N(0, \tau^2 \lambda_i^2), \quad \lambda_i^2 \sim \pi_\lambda(\lambda_i^2), \quad \tau^2 \sim \pi_\tau(\tau^2).$$

1. Bayesian Lasso [Park and Casella, 2008]
2. Horseshoe [Carvalho, Polson and Scott, 2010]
3. Normal-gamma [Griffin and Brown, 2010]
5. Dirichlet-Laplace [Bhattacharya et al., 2016]
6. ...
Outline

1. Introduction
2. Sparse linear model
3. Linear model with unknown error distribution
4. Asymptotic results
Gaussian model

\[ Y_i = x_i^T \theta + \epsilon_i, \quad i = 1, \ldots, n. \]

Assume that \( \epsilon_i \stackrel{i.i.d.}{\sim} \eta \) for some density \( \eta \in \mathcal{H} \).

Usually it is assumed that \( \eta(y) = \phi_\sigma(y) \) because of

1. computational simplicity, and
2. good theoretical properties.

Some properties (e.g. consistency and rate) tend to be robust to misspecification.
Key problems

\[ Y_i = x_i^T \theta + \epsilon_i, \quad i = 1, \ldots, n. \]

Assume that \( \epsilon_i \)'s are not really normally distributed.

Key problems caused from model misspecification:

1. **Efficiency** Asymptotic variance of \( \sqrt{n}(\hat{\theta}_i - \theta_i) \) can be large.
2. **Uncertainty quantification** Credible sets do not give valid confidence. [Kleijn and van der Vaart, 2012]
3. **Selection** Misspecification might result in serious overfitting. [Grünwald and Ommen, 2014]
Key problems: example

[Grünwald and Ommen, 2014]

\[ Y_i = \theta_{\text{int}} + \theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_p x_i^p + \epsilon_i, \quad \theta_0 = 0 \in \mathbb{R}^{p+1} \]
Key problems

\[ Y_i = x_i^T \theta + \epsilon_i, \quad i = 1, \ldots, n. \]

Assume that \( \epsilon_i \)'s are not really normally distributed.

Key problems caused from model misspecification:

1. **(Efficiency)** Asymptotic variance of \( \sqrt{n}(\hat{\theta}_i - \theta_i) \) can be large.
2. **(Uncertainty quantification)** Credible sets do not give valid confidence. [Kleijn and van der Vaart, 2012]
3. **(Selection)** Misspecification might result in serious overfitting. [Grünwald and Ommen, 2014]

Good remedy: semi-parametric modelling.
Frequentist’s method for fixed $p$

$$Y_i = x_i^T \theta + \epsilon_i, \quad \epsilon_i \sim \eta.$$  

There is an efficient estimator for $\theta$. [Bickel, 1982]

One way to get an efficient estimator is:

1. Find an initial $n^{-1/2}$-consistent estimator $\tilde{\theta}_n$.
2. Estimate the score function with perturbed sample

   $$\tilde{\epsilon}_i = Y_i - \tilde{\theta}_n^T X_i.$$

3. Solve the score equation using one step Newton-Raphson iteration.

Does it work if $p \gg n$?
Bayesian method for fixed $p$

$$Y_i = x_i^T \theta + \epsilon_i, \quad \epsilon_i \sim \eta.$$ 

Put a symmetrized DP mixture prior $\Pi_{H}$ on $\eta$:

$$\eta(y) = \int \phi_\sigma(y - z) d\bar{F}(z, \sigma), \quad F \sim \text{DP}(\alpha),$$

and

$$d\bar{F}(z, \sigma) = \frac{dF(z, \sigma) + dF(-z, \sigma)}{2}.$$ 

Then, the BvM theorem holds. [Chae, Kim and Kleijn, 2016]

Inference: Gibbs sampler algorithm
Bayesian inference

\[ Y_i = x_i^T \theta + \epsilon_i \quad \Leftrightarrow \quad Y_i = x_i^T \theta + z_i + \sigma_i \tilde{\epsilon}_i \]

\[ \epsilon_i \sim \eta \quad (z_i, \sigma_i) \sim F, \quad \tilde{\epsilon}_i \sim N(0, 1) \]

Inference can be done through Gibbs sampler algorithm:

1. For given \((z_i, \sigma_i)_{i \leq n}\), \(\theta\) can be sampled as in the Gaussian model.
2. For given \(\theta\), \((z_i, \sigma_i)_{i \leq n}\) can be sampled as in the DPM model.

Additional computational burden by semi-parametric modelling depends only on \(n\). \(\Rightarrow\) Feasible when \(p \gg n\)!
Outline

1. Introduction
2. Sparse linear model
3. Linear model with unknown error distribution
4. Asymptotic results
Goal: frequentist properties \((p \gg n)\)

Assume fixed design \(X\), and response vector \(Y\) is really generated from a given \((\theta_0, \eta_0)\), possibly \(p \gg n\).

We want (marginal) posterior \(\Pi(\theta \in \cdot | Y)\):

1. **(Recovery)** to put most of its mass around \(\theta_0\)
2. **(Uncertainty quantification)** to express remaining uncertainty
3. **(Selection)** to find the true nonzero set \(S_0\) of \(\theta_0\)
4. **(Adaptation)** to adapt unknown sparsity level and error density with high \(P_{\theta_0, \eta_0}\)-probability.
Prior for $\theta$

The probability $\pi_p(s)$ decrease exponentially:
[Castillo and van der Vaart, 2012; 2015]

(i) for some constants $A_1, A_2, A_3, A_4 > 0$,

$$A_1 p^{-A_3} \pi_p(s - 1) \leq \pi_p(s) \leq A_2 p^{-A_4} \pi_p(s - 1), \quad s = 1, \ldots, p$$

Tails of nonzero coeff. are as thick as Laplace distribution:
[Castillo and van der Vaart, 2012; van der Pas et al., 2016]

(ii) $g_S(\theta) = \otimes_{i \in S} g(\theta_i)$, $g(\theta_i) \propto e^{\lambda|\theta_i|}$ and $\lambda$ satisfies

$$\frac{\sqrt{n}}{p} \leq \lambda \leq \sqrt{n \log p}.$$
Prior for $\eta$

Put a symmetrized DP mixture prior $\Pi_{\mathcal{H}}$ on $\eta$ [Chae, Kim and Kleijn, 2016]:

$$
\eta(y) = \int \phi_{\sigma}(y - z)d\overline{F}(z, \sigma), \quad F \sim \text{DP}(\alpha),
$$

and

$$
d\overline{F}(z, \sigma) = \frac{dF(z, \sigma) + dF(-z, \sigma)}{2}.
$$

Assume that $\text{supp}(\alpha) \subset [-M, M] \times [\sigma_1, \sigma_2]$ for some positive constants $M$ and $\sigma_1 < \sigma_2$. 
Design matrix

Assume uniformly bounded covariates: $|x_{ij}| \lesssim 1$.

Define uniform compatibility numbers

$$
\phi^2(s) = \inf \left\{ \frac{s_\theta \|X\theta\|_2^2}{n\|\theta\|_1^2} : 0 < s_\theta \leq s \right\}
$$

and restricted eigenvalues

$$
\psi^2(s) = \inf \left\{ \frac{\|X\theta\|_2^2}{n\|\theta\|_2^2} : 0 < s_\theta \leq s \right\}.
$$

$\phi(Ks_0) \gtrsim 1$ ($\psi(Ks_0) \gtrsim 1$, resp.) for some const. $K > 1$ is sufficient for the recovery of $\theta$ in $\ell_1$- ($\ell_2$-, resp.) norm.
By C-S inequality, $\phi(s) \geq \psi(s)$.

$\psi(s) \gtrsim 1$ in many examples:

1. Typically, $\psi(s) \geq \text{const.} - s \max_{i \neq j} \text{corr}(x_i, x_j)$. [Lounici, 2008]
2. If $x_{ij}$’s are i.i.d. random variables, then $\psi(s) \gtrsim 1$ with high probability for $s \lesssim \sqrt{n/\log p}$. [Cai and Jiang, 2011]
3. If $p = n$ and $\text{corr}(x_i, x_j) = \rho^{|i-j|}$ for some $\rho \in (0, 1)$, then $\psi(p) \gtrsim 1$. [Zhao and Yu, 2006]

There are some examples such that $\phi(s) \gtrsim 1$ but not for $\psi(s)$. [van de Geer and Bühlmann, 2009]
Asymptotic: dimension

**THEOREM** [Chae, Lin and Dunson, 2016] If $\lambda \|\theta_0\|_1 \lesssim s_0 \log p$ and $s_0 \log p \ll n$, then

$$E \Pi (s_\theta > Ks_0 \mid Y) \to 0$$

for some constant $K > 1$.

Small value of $\lambda$ is preferred for large $\|\theta_0\|_1$. 
Asymptotic: consistency

\[ d_n^2((\theta, \eta), (\theta_0, \eta_0)) = \frac{1}{n} \sum_{i=1}^{n} d_H^2(p_{\theta,i}, p_{\theta_0,i}). \]

Mean Hellinger distance \( d_n \) allows to construct certain exponentially consistent tests for independent observations. [Birgé, 1983; Ghosal and van der Vaart 2007]

**THEOREM** [Chae, Lin and Dunson, 2016] If, furthermore, 
\( \phi(Ks_0) \gtrsim p^{-1} \), then

\[ \mathbb{E} \Pi \left( d_n((\theta, \eta), (\theta_0, \eta_0)) \gtrsim \sqrt{\frac{s_0 \log p}{n}} \mid Y \right) \to 0. \]
THEOREM [Chae, Lin and Dunson, 2016] Under the previous conditions,

$$\mathbb{E} \Pi \left( d_H(\eta, \eta_0) \gtrsim \sqrt{\frac{s_0 \log p}{n}} \mid Y \right) \rightarrow 0.$$ 

If, furthermore, $s_0^2 \log p / \phi^2(Ks_0) \ll n$, then

$$\mathbb{E} \Pi \left( \|\theta - \theta_0\|_1 \gtrsim \frac{s_0}{\phi(Ks_0)} \sqrt{\frac{\log p}{n}} \mid Y \right) \rightarrow 0$$

$$\mathbb{E} \Pi \left( \|\theta - \theta_0\|_2 \gtrsim \frac{1}{\psi(Ks_0)} \sqrt{\frac{s_0 \log p}{n}} \mid Y \right) \rightarrow 0$$

$$\mathbb{E} \Pi \left( \|X(\theta - \theta_0)\|_2 \gtrsim \sqrt{s_0 \log p} \mid Y \right) \rightarrow 0.$$
Asymptotic: LAN

\[ r_n(\theta, \eta) = L_n(\theta, \eta) - L_n(\theta_0, \eta) \]

\[- \left\{ \sqrt{n}(\theta - \theta_0)^T \mathbb{G}_n \ell_{\theta_0, \eta_0} - \frac{n}{2} (\theta - \theta_0)^T V_{n, \eta_0} (\theta - \theta_0) \right\} \]

**THEOREM** [Chae, Lin and Dunson, 2016] If \( s_0 \log p \ll n^{1/6} \), then

\[ \sup_{\theta \in \Theta_n} \sup_{\eta \in \mathcal{H}_n} |r_n(\theta, \eta)| = o_P(1), \]

where \( \Pi(\Theta_n \times \mathcal{H}_n | Y) \to 1 \) in probability.
Asymptotic: BvM theorem

Let $\mathcal{N}_{n,S}$ be the $|S|$-dimensional normal dist’n to which an efficient estimator $\sqrt{n}(\hat{\theta}_S - \theta_0^S)$ converges in dist’n.

**THEOREM** [Chae, Lin and Dunson, 2016] If, furthermore, $\lambda s_0 \sqrt{\log p} \ll \sqrt{n}$ and $\psi(Ks_0) \gtrsim 1$, then

$$\sup_{S \in S_n} \sup_B \left| \Pi(\sqrt{n}(\theta_S - \theta_{0,S}) \in B | Y, S_\theta = S) - \mathcal{N}_{n,S}(B) \right| = o_P(1),$$

where $\Pi(S_\theta \in S_n | Y) \to 1$ in probability.

Posterior dist’n of nonzero coeff. is asymptotically a mixture of normal dist’n.
Asymptotic: selection

**THEOREM** [Chae, Lin and Dunson, 2016] Under the previous conditions,

\[ \Pi(S_\theta \supseteq S_0 | Y) \to 0 \]

in probability.

The true non-zero coeff. can be selected if every non-zero coeff. is not very small (beta-min condition).
Discussion

- Condition $s_0 \log p \ll n^{1/6}$ is required due to semi-parametric bias.
- If $\eta$ is known (may not be a Gaussian) and $p = s_0$, the condition may be reduced to $s_0 \ll n^{1/3}$, and this cannot be improved. [Panov and Spokoiny, 2015]
- In some parametric models, $s_0 \ll n^{1/6}$ is required for BvM theorem. [Ghosal, 2000]
- Results can be extended to more general prior, i.e., $M, \sigma_1 \to \infty$ and $\sigma_1 \to 0$, but sub-Gaussian tail of $\ell_{\eta_0}$ is (maybe) essential in selection. [Kim and Jeon, 2016]
Selected references


