Bayesian modeling of sparse high-dimensional data using divergence measures

Dipak K. Dey

Department of Statistics
University of Connecticut, Storrs, USA

Joint work with Gyuhyeong Goh, Kansas State University.
High-dimensional problems

Consider

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

where \( y = (y_1, \ldots, y_n)^T \) is the response vector, \( X \) is the \( n \times p \) predictor matrix, \( \beta = (\beta_1, \ldots, \beta_p)^T \) is the unknown coefficient vector, and \( \epsilon \sim N(0_n, \sigma^2 I_n) \) with known \( \sigma^2 \).

In many practical situations, we encounter that the number of coefficients, \( p \), is very large, i.e., \( p \approx n, p > n, \) or \( p \gg n \).

For instance, the genomics study investigates numerous genes that are possibly related to a certain phenotype.
Sparsity in high-dimensional problems

- In high-dimensional regression models, the sparsity assumption for the coefficient vector is necessary, otherwise $\beta$ cannot be identifiable.

- Let $p^*$ be the number of non-zero elements in $\beta$.

- The sparsity assumption implies that $p^* \ll n$.

- In fact, this assumption is practical.

  ▶ A genome-wide association study looks at millions of single nucleotide polymorphisms (SNPs) to identify a few relevant genes to a certain phenotype.
Ordinary least squares (OLS) estimation

- In the classical regression analysis, ordinary least squares (OLS) is the most popular method to estimate $\beta$;

\[ \hat{\beta}_{\text{OLS}} = \arg \min_{\beta} \left[ \| y - X\beta \|^{2} \right]. \]

- In sparse high-dimensional problems, however, OLS estimator behaves poorly;
  - Extremely large variance when $p$ is large;
  - No unique solution in the case of $p > n$ or in presence of multicollinearity.
Penalized Least Squares (PLS) estimation

- In sparse high-dimensional estimation, penalized least squares (PLS) has played a key role.

- PLS estimator is defined as

$$\hat{\beta}_{PLS} = \arg \min_{\beta} \left[ \| y - X\beta \|^2 + P_{\lambda}(\beta) \right],$$

where $P_{\lambda}(\cdot)$ is a deterministic penalty function with a tuning parameter $\lambda \geq 0$ controlling the degree of penalization.
PLS estimation with $\ell_0$-norm penalty

- Akaike (1974) and Schwarz (1978) introduced the $\ell_0$-norm penalization as follows:

$$\hat{\beta}_{\ell_0} = \arg \min_{\beta} \left[ \| y - X\beta \|^2 + \lambda \sum_{j=1}^{p} \mathbb{I}\{\beta_j \neq 0\} \right],$$

where $\mathbb{I}\{\cdot\}$ denotes an indicator function.

- Since the $\ell_0$-norm penalty directly restricts the number of non-zero coefficients, it successfully induces the sparsity for $\hat{\beta}_{\ell_0}$.

- However, due to non-convexity and discontinuity of the $\ell_0$-norm penalty, finding the minimum is challenging especially when $p$ is large.
As an alternative, the $\ell_1$-norm penalization, called the lasso (Tibshirani, 1996), was proposed:

$$\hat{\beta}_{\text{Lasso}} = \arg \min_{\beta} \left[ \| y - X\beta \|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right].$$

Since the lasso leads a continuous and convex optimization, it addresses the computational drawback of the $\ell_0$-norm.

However, the lasso often leads to undesirable bias in the resulting estimators, because it imposes the same degree of penalization for both zero and non-zero coefficients (Zou, 2006).
Adaptive Lasso

- Zou (2006) introduced the adaptive lasso,

\[
\hat{\beta}_{\text{Alasso}} = \arg \min_{\beta} \left[ \| y - X\beta \|^2 + \sum_{j=1}^{p} \lambda_j |\beta_j| \right],
\]

where \( \lambda_j = \lambda / |\hat{\beta}_j|^\gamma \), \( \gamma > 0 \), and \( \hat{\beta}_j \) is a \( \sqrt{n} \)-consistent estimator for \( \beta_j \).

- The adaptive lasso remedies the bias problem in the lasso.

- However, in the high-dimensional setup, it is challenging to find good \( \lambda_j \).
Approximation of $\ell_0$-norm penalty

• Recently, Goh et al. (2017) and Goh and Dey (2018) introduced an approximation of the $\ell_0$-norm penalty,

$$\tilde{\ell}_{0,\tau}(\beta) = \lambda \sum_{j=1}^{p} \frac{\beta_j^2}{\tau^2 + \beta_j^2},$$

where $\tau$ is a deterministic constant (e.g., $\tau = 10^{-5}$).

• Note that as $\tau$ goes to zero, the new penalty approaches the $\ell_0$ penalty:

$$\lim_{\tau \to 0} \ell_{0,\tau}(\beta) = \lambda \sum_{j=1}^{p} \mathbb{I}\{\beta_j \neq 0\}.$$
Approximation of $\ell_0$-norm penalty

The figure illustrates that as $\tau \to 0$, $\frac{x^2}{\tau^2 + x^2} \to \mathbb{I}\{x \neq 0\}$.

**Figure**: graphs of $f(x) = \frac{x^2}{(\tau^2 + x^2)}$ for $\tau = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$
The new penalty can be viewed as the adaptive lasso in the sense that

$$\lambda \sum_{j=1}^{p} \frac{\beta_j^2}{\tau^2 + \beta_j^2} = \sum_{j=1}^{p} \frac{\lambda|\beta_j|}{\tau^2 + \beta_j^2} |\beta_j| = \sum_{j=1}^{p} \lambda_j^*|\beta_j|,$$

where $\lambda_j^* = \frac{\lambda|\beta_j|}{\tau^2 + \beta_j^2}$, but $\lambda_j^*$ is automatically determined by $\beta_j$. 
Bayesian perspective

- The PLS estimator can be viewed as the maximum a posteriori (MAP) estimator or posterior mode (Tibshirani, 1996; Park and Casella, 2008; Kyung et al., 2010).

- Define

\[
\pi(\beta \mid y, \lambda) \propto f(y \mid \beta)\pi(\beta \mid \lambda)
\]

\[
\propto \exp \left( -\frac{1}{2} \|y - X\beta\|^2 \right) \exp \left\{ -\frac{1}{2} \mathcal{P}_\lambda(\beta) \right\},
\]

where

- \( f(y \mid \beta) \): the likelihood function;
- \( \pi(\beta \mid \lambda) \): the prior for \( \beta \) given the hyperparameter \( \lambda \).

- Then, it is easy to check that

\[
\arg \max_{\beta} \pi(\beta \mid y, \lambda) = \arg \min_{\beta} \left[ \|y - X\beta\|^2 + \mathcal{P}_\lambda(\beta) \right].
\]
Merit of Bayesian approach

Let \( \hat{\beta}_{PLS} \) be the PLS estimator, i.e.,

\[
\hat{\beta}_{PLS} = \arg \min_{\beta} \left[ \| y - X\beta \|^2 + P_\lambda(\beta) \right].
\]

In general, it is hard to estimate \( \text{var}(\hat{\beta}_{PLS}) \), unless the sample size \( n \) is very large.

Let \( \hat{\beta}_{MAP} \) be the MAP estimator, i.e.,

\[
\hat{\beta}_{MAP} = \arg \max_{\beta} \pi(\beta \mid y, \lambda).
\]

Under a Bayesian framework, the uncertainty associated with \( \hat{\beta}_{MAP} \) can be easily quantified by the posterior distribution \( \pi(\beta \mid y, \lambda) \).
Duality property

- The relationship between the loss function and the likelihood, called *duality property*, was originally discussed by Bernardo and Smith (1994).

- The duality property states that the negative log-likelihood function can be viewed as a loss function.

- For example, the standard normal likelihood can be viewed as

\[
f(y|\beta) \propto \exp \left[ -\frac{1}{2} L_2(y, h(X\beta)) \right],
\]

where \( L_2(x_1, x_2) = \|x_1 - x_2\|^2 \) is the squared Euclidean loss and \( h(x) = x \) is the identity link function.
Development of likelihood function via duality property

- Now, we consider the case in which the data-generating distribution satisfies the duality property.

- To provide a general framework, we suppose that the negative log-likelihood function belongs to a general class of divergence measures, called *Bregman divergence*.

- That is, the likelihood is expressed as

  \[ f(y|\beta) \propto \exp \left[ -\text{BD}_\psi (y, h(X\beta)) \right], \]

  where \( \text{BD}_\psi (\cdot, \cdot) \) denotes the Bregman divergence.
**Definition**

Let $\psi : \Omega \to \mathbb{R}$ be a strictly convex function on a convex set $\Omega \subseteq \mathbb{R}^m$, assumed to be nonempty and differentiable. Then for $x, y \in \mathbb{R}^m$ the Bregman divergence with respect to $\psi$ is defined as

$$\text{BD}_\psi (x, y) = \psi(x) - \psi(y) - (x - y)^T \nabla \psi(y),$$

where $\nabla \psi$ represents the gradient vector of $\psi$.

- The Bregman divergence can be interpreted as the difference between the value of the convex function at $x$ and its first order Taylor’s expansion at $y$. 

---

Dipak K. Dey (Department of Statistics, University of Connecticut, Storrs, USA) Joint work with Gyuhyeong Goh, Kansas State University.

Bayesian modeling of sparse high-dimensional data using divergence measures
Graphical Illustration

- The Bregman divergence measures the *ordinate distance* between the value of the convex function at \( x \) and its tangent at \( y \).

\[
\text{Bregman divergence with } \psi(x) = e^{cx} - cx - 1, \ c = 0.5.
\]

**Figure:** Bregman divergence with \( \psi(x) = e^{cx} - cx - 1, \ c = 0.5 \).
Examples of Bregman divergences

- The Bregman divergence includes a large class of well-known loss functions.

**Table: Examples of the Bregman divergence generated by some convex functions, $\psi$’s.**

<table>
<thead>
<tr>
<th>$\psi(x)$</th>
<th>Bregman divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$|x|^2$</td>
<td>Squared error loss</td>
</tr>
<tr>
<td>$x^T W x$</td>
<td>Mahalanobis distance</td>
</tr>
<tr>
<td>$\sum_{i=1}^{n} x_i \log x_i$</td>
<td>Kullback-Leibler divergence</td>
</tr>
<tr>
<td>$\sum_{i=1}^{n} - \log x_i$</td>
<td>Itakura-Saito distance</td>
</tr>
<tr>
<td>$\sum_{i=1}^{n} e^{cx_i}$</td>
<td>Weighted Linex loss</td>
</tr>
</tbody>
</table>
Likelihood with Bregman divergence

- One may wonder, “what is the corresponding distribution family to Bregman divergence?”
- Banerjee et al. (2005) showed that any member of the natural exponential family corresponds to a unique and distinct member of Bregman divergence.
- This implies that the developed class of likelihood functions by Bregman divergence contains the natural exponential family as a subset.
- For example, if we define $\psi(y) = \sum_{i=1}^{n} \{y_i \log y_i\}$, then our likelihood reduces to the Poisson likelihood,

$$f(y|\beta) \propto \exp \left[ -\text{BD}_\psi (y, h(X\beta)) \right]$$

$$\propto \exp \left[ -\sum_{i=1}^{n} \left\{ y_i \log \left( \frac{y_i}{h(x_i^T \beta)} \right) - (y_i - h(x_i^T \beta)) \right\} \right]$$

$$\propto \prod_{i=1}^{n} \left[ e^{-h(x_i^T \beta)} \{h(x_i^T \beta)\}^{y_i} \right].$$
Table: Examples of Bregman divergence and related distributions in the natural exponential family.

<table>
<thead>
<tr>
<th>$\psi(z)$</th>
<th>$BD_\psi(z_1, z_2)$</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{2\sigma^2} z^2$</td>
<td>$\frac{1}{2\sigma^2} (z_1 - z_2)^2$</td>
<td>Gaussian</td>
</tr>
<tr>
<td>$z \log z$</td>
<td>$z_1 \log \left( \frac{z_1}{z_2} \right) - (z_1 - z_2)$</td>
<td>Poisson</td>
</tr>
<tr>
<td>$- \log z$</td>
<td>$\frac{z_1}{z_2} - \log \left( \frac{z_1}{z_2} \right) - 1$</td>
<td>Exponential</td>
</tr>
<tr>
<td>$z \log z + (1 - z) \log(1 - z)$</td>
<td>$z_1 \log \left( \frac{z_1}{z_2} \right) + (1 - z_1) \log \left( \frac{1-z_1}{1-z_2} \right)$</td>
<td>Bernoulli</td>
</tr>
</tbody>
</table>
Pseudo-likelihood with Bregman divergence

- In fact, our Bregman divergence approach encompasses a wide range of likelihood functions.

- For instance, Zhang et al. (2009) verified that the quasi-likelihood function (Wedderburn, 1974) belongs to the class of Bregman divergence.

- Let \( \psi(y) = \sum_{i=1}^{n} \int_{-\infty}^{y_i} \frac{y_i - s}{V(s)} ds \), where \( V(\cdot) \) is a positive known function.

- Then, it can be shown that our likelihood reduces to the quasi-likelihood,

\[
\begin{align*}
  f(y|\beta) &\propto \exp \left[ -BD_\psi (y, h(X\beta)) \right] \\
  &\propto \exp \left\{ \sum_{i=1}^{n} \int_{-\infty}^{h(x_i^T\beta)} \frac{y_i - s}{V(s)} ds \right\} .
\end{align*}
\]
Now, we define our new prior, called GD prior, by

\[ \pi_{GD}(\beta, d) \propto \pi_G(\beta|d)\pi_D(d), \]

such that

\[ \pi_G(\beta|d) \propto \prod_{j=1}^{p} \left\{ d_j^{1/2} \exp \left( -\frac{d_j}{2} \beta_j^2 \right) \right\} \] (Gaussian),

\[ \pi_D(d) \propto \prod_{j=1}^{p} \left\{ d_j^{\lambda-1/2} \exp \left( -\frac{\tau^2}{2} d_j \right) \right\} \] (Diffused-gamma),

where \( \tau(>0) \) is determined to be sufficiently small.
Recall that a valid approximation of the $\ell_0$-norm penalty was defined as

$$\tilde{\ell}_{0,\tau}(\beta) = \lambda \sum_{j=1}^{p} \frac{\beta_j^2}{\tau^2 + \beta_j^2}.$$ 

It can be shown that

$$\arg\max_{\beta} \left\{ f(y|\beta)\pi_{GD}(\beta, \hat{d}) \right\} = \arg\min_{\beta} \left[ BD_\psi \{ y, h(X\beta) \} + \tilde{\ell}_{0,\tau}(\beta) \right],$$

where $\hat{d} = \arg\max_d \left\{ \max_{\beta} f(y|\beta)\pi_{GD}(\beta, d) \right\}.$

For $\tau \approx 0$, our MAP estimator of $\beta$ approximates the penalized Bregman divergence estimator with the $\ell_0$-norm penalty.
Maximum A Posteriori (MAP) estimation

The MAP estimator, say $\hat{\beta}$, can be obtained by

$$(\hat{\beta}, \hat{d}) = \arg \max_{\beta, d} \{ f(y|\beta)\pi_G(\beta|d)\pi_D(d) \}.$$ 

Using the Iterated Conditional Modes (ICM) algorithm, $\hat{\beta}$ can be obtained by iteratively updating the current $\hat{\beta}$ as follows:

$$\hat{d} \leftarrow \arg \max_d \left\{ \pi_G(\hat{\beta}|d)\pi_D(d) \right\};$$

$$\hat{\beta} \leftarrow \arg \max_\beta \left\{ f(y|\beta)\pi_G(\beta|\hat{d}) \right\};$$

until convergence.
ICM algorithm

- Our component-wise updating ICM algorithm can be summarized as follows:

  **Set** an initial value $\hat{\beta} = \beta^{(0)}$.

  **Update** $\hat{\beta}$ as follows: for $j = 1, 2, \ldots, p$;

  \[
  \hat{d}_j \leftarrow \frac{2\lambda}{\tau_0^2 + (\hat{\beta}_j)^2}; \\
  \tilde{\beta}_j \leftarrow \arg \min_{\hat{\beta}_j} \left[ \text{BD}_{\psi} \{ y, \mathbf{h}(X\hat{\beta}) \} + \frac{\hat{d}_j}{2} \hat{\beta}_j^2 \right]; \\
  \xi_j \leftarrow \frac{2}{\sqrt{\frac{2\lambda}{\tau_0^2 + (\tilde{\beta}_j)^2}}}; \\
  \hat{\beta}_j \leftarrow \tilde{\beta}_j 1\{|\tilde{\beta}_j| > \xi_j\};
  \]

  until convergence.

  **Return** $\hat{\beta}$. 
Prior specification

- In our Bayesian approach, the determination of the hyperparameter $\lambda$ is important because it controls the degree of the sparsity of our MAP estimator.

- To select the optimal $\lambda$, we utilize the marginal likelihood (or equivalently the Bayes factor) as follows:

$$m(y|\lambda) = \int f(y|\beta)\pi_G(\beta|d_\lambda) d\beta,$$

where $d_\lambda$ denotes the MAP of $d$ given $\lambda$.

- The optimal value of $\lambda$ can be defined as

$$\hat{\lambda} = \arg \max_\lambda m(y|\lambda).$$
Simulation studies

- We assessed the performance of GD method using a Monte Carlo simulation study.

- For the purpose of comparison, we also considered widely-used PL methods, Elastic-net, LASSO, adaptive LASSO, SCAD, and MCP.

- We measured the estimation accuracy using the following two types of mean squared error (MSE):

\[
\text{MSE}_{\text{est}} = \frac{1}{p} \| \hat{\beta} - \beta \|^2; \quad \text{MSE}_{\text{pred}} = \frac{1}{n} \| X \hat{\beta} - X \beta \|^2.
\]

- To assess the variable selection performance, we calculated False Positive Rate (FPR) and False Negative Rate (FNR) as follows:

\[
\text{FPR\%} = 100 \times \frac{\text{FP}}{\text{TN} + \text{FP}}; \quad \text{FNR\%} = 100 \times \frac{\text{FN}}{\text{TP} + \text{FN}},
\]

where TP, FP, TN and FN denote the numbers of true non-zeros, false non-zeros, true zeros and false zeros, respectively.
Simulation studies

Set-up

- We generate 1000 data sets from each of the following three cases: for
  $i = 1, \ldots, n$,

**M1.** Generate $y_i \overset{iid}{\sim} N(\mu_i, 1)$ with $\mu_i = h_1(x_i^T \beta)$, where $h_1(x) = x$, $\beta = (\text{rep}(2, 5), \text{rep}(0, 10), \text{rep}(-2, 5), \text{rep}(0, p - 20))^T$, and $x_i \overset{iid}{\sim} N_p(\mathbf{0}, \Sigma)$ with $\Sigma = (\Sigma_{ij})_{p \times p}$ and $\Sigma_{ij} = \rho|i-j|$.

**M2.** Generate $y_i \overset{iid}{\sim} \text{Bernoulli}(p_i)$ with $p_i = h_2(x_i^T \beta)$, where $h_2(x) = \frac{1}{1 + \exp(-x)}$, $\beta = (\text{rep}(2, 2), \text{rep}(0, 10), \text{rep}(-2, 1), \text{rep}(0, p - 13))^T$, and $x_i \overset{iid}{\sim} N_p(\mathbf{0}, \Sigma)$ with $\Sigma = (\Sigma_{ij})_{p \times p}$ and $\Sigma_{ij} = \rho|i-j|$.

**M3.** Generate $y_i \overset{iid}{\sim} \text{Poisson}(\mu_i)$ with $\mu_i = h_3(x_i^T \beta)$, where $\beta = (\text{rep}(2, 3), \text{rep}(0, 10), \text{rep}(-2, 3), \text{rep}(0, p - 16))^T$, $h_3(x) = \exp(x)$, $x_i = \Phi(z_i) - 0.51p$, $\Phi(z_i) = (\Phi(z_{i1}), \ldots, \Phi(z_{ip}))^T$, $\Phi(\cdot)$ is the CDF of standard normal distribution, and $z_i \overset{iid}{\sim} N_p(\mathbf{0}, \Sigma)$ with $\Sigma = (\Sigma_{ij})_{p \times p}$ and $\Sigma_{ij} = \rho|i-j|$.
Simulation studies

Bregman divergence specification

To specify the Bregman divergence likelihood function, we define

\[
\text{M1} : \psi(x) = \sum_{i=1}^{n} \left\{ \frac{x_i^2}{2} \right\}, \\
\text{M2} : \psi(x) = \sum_{i=1}^{n} \left\{ x_i \log x_i + (1 - x_i) \log(x_i - 1) \right\}, \\
\text{M3} : \psi(x) = \sum_{i=1}^{n} \left\{ x_i \log x_i \right\},
\]
**Table:** Simulation results for continuous data (**M1**).

<table>
<thead>
<tr>
<th>Case</th>
<th>Method</th>
<th>$\text{MSE}_{\text{est}}$</th>
<th>$\text{MSE}_{\text{pred}}$</th>
<th>FPR%</th>
<th>FNR%</th>
</tr>
</thead>
<tbody>
<tr>
<td>(100,100,0.5)</td>
<td>Oracle</td>
<td>0.0018</td>
<td>0.1028</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>GD</td>
<td>0.0019</td>
<td>0.1070</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>a-LASSO</td>
<td>0.0133</td>
<td>1.6622</td>
<td>4.6000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>E-net</td>
<td>0.0120</td>
<td>1.6084</td>
<td>5.6667</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>0.0138</td>
<td>1.7399</td>
<td>4.2333</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>MCP</td>
<td>0.0040</td>
<td>0.2136</td>
<td>0.5889</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>SCAD</td>
<td>0.0046</td>
<td>0.2393</td>
<td>0.6333</td>
<td>0.0000</td>
</tr>
<tr>
<td>(100,500,0.5)</td>
<td>Oracle</td>
<td>0.0004</td>
<td>0.0995</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>GD</td>
<td>0.0004</td>
<td>0.1052</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>a-LASSO</td>
<td>0.0025</td>
<td>1.5074</td>
<td>1.6612</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>E-net</td>
<td>0.0024</td>
<td>1.5681</td>
<td>2.1796</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>0.0025</td>
<td>1.5404</td>
<td>1.7204</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>MCP</td>
<td>0.0009</td>
<td>0.2254</td>
<td>0.3714</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>SCAD</td>
<td>0.0010</td>
<td>0.2516</td>
<td>0.4122</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Oracle: MLE under the true model.
Simulation studies

Result

<table>
<thead>
<tr>
<th>Case</th>
<th>Method</th>
<th>$\text{MSE}_{\text{est}}$</th>
<th>$\text{MSE}_{\text{pred}}$</th>
<th>FPR%</th>
<th>FNR%</th>
</tr>
</thead>
<tbody>
<tr>
<td>(100,100,0.5)</td>
<td>Oracle</td>
<td>0.0149</td>
<td>0.0039</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>GD</td>
<td>0.0238</td>
<td>0.0110</td>
<td>0.6495</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>a-LASSO</td>
<td>0.0542</td>
<td>0.0301</td>
<td>1.7732</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>E-net</td>
<td>0.0765</td>
<td>0.0545</td>
<td>1.8351</td>
<td>2.0000</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>0.0542</td>
<td>0.0301</td>
<td>1.7732</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>MCP</td>
<td>0.0447</td>
<td>0.0235</td>
<td>2.1237</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>SCAD</td>
<td>0.0480</td>
<td>0.0256</td>
<td>2.2165</td>
<td>0.0000</td>
</tr>
<tr>
<td>(100,500,0.5)</td>
<td>Oracle</td>
<td>0.0003</td>
<td>0.0005</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>GD</td>
<td>0.0017</td>
<td>0.0026</td>
<td>0.0342</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>a-LASSO</td>
<td>0.0019</td>
<td>0.0062</td>
<td>0.0201</td>
<td>0.3333</td>
</tr>
<tr>
<td></td>
<td>E-net</td>
<td>0.0025</td>
<td>0.0107</td>
<td>0.0241</td>
<td>0.6667</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>0.0019</td>
<td>0.0062</td>
<td>0.0201</td>
<td>0.3333</td>
</tr>
<tr>
<td></td>
<td>MCP</td>
<td>0.0017</td>
<td>0.0051</td>
<td>0.0262</td>
<td>0.3333</td>
</tr>
<tr>
<td></td>
<td>SCAD</td>
<td>0.0018</td>
<td>0.0058</td>
<td>0.0302</td>
<td>0.3333</td>
</tr>
</tbody>
</table>
## Simulation studies

### Result

**Table:** Simulation results for count data (M3).

<table>
<thead>
<tr>
<th>Case</th>
<th>(n, p, ρ)</th>
<th>Method</th>
<th>$\text{MSE}_{\text{est}}$</th>
<th>$\text{MSE}_{\text{pred}}$</th>
<th>FPR%</th>
<th>FNR%</th>
</tr>
</thead>
<tbody>
<tr>
<td>M3</td>
<td>(100,100,0.5)</td>
<td>Oracle</td>
<td>0.0040</td>
<td>1.3862</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GD</td>
<td>0.0049</td>
<td>1.5019</td>
<td>0.0638</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a-LASSO</td>
<td>0.0103</td>
<td>4.1607</td>
<td>3.2128</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E-net</td>
<td>0.0126</td>
<td>5.1199</td>
<td>8.6809</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lasso</td>
<td>0.0103</td>
<td>4.1607</td>
<td>3.2128</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MCP</td>
<td>0.0077</td>
<td>5.4540</td>
<td>0.9255</td>
<td>0.5000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SCAD</td>
<td>0.0084</td>
<td>5.4122</td>
<td>1.0213</td>
<td>0.8333</td>
</tr>
<tr>
<td>(100,500,0.5)</td>
<td>Oracle</td>
<td>0.0008</td>
<td>1.3339</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GD</td>
<td>0.0011</td>
<td>1.5134</td>
<td>0.0263</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a-LASSO</td>
<td>0.0031</td>
<td>6.1365</td>
<td>1.0891</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E-net</td>
<td>0.0054</td>
<td>6.6374</td>
<td>3.5789</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lasso</td>
<td>0.0031</td>
<td>6.1365</td>
<td>1.0891</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MCP</td>
<td>0.0086</td>
<td>8.7068</td>
<td>0.6943</td>
<td>6.6667</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SCAD</td>
<td>0.0065</td>
<td>7.4857</td>
<td>0.7126</td>
<td>5.0000</td>
</tr>
</tbody>
</table>
Summary of simulation result

- The result clearly shows that our GD method always performs better than all the PL methods.

- Furthermore, our GD method is comparable to the Oracle method (MLE under the true model).
Real data analysis: predictive binary classification

- In practice, especially in genetics study, a researcher conducts a pre-screening procedure such as Sure Independence Screening (SIS) (Fan and Lv, 2008; Fan and Song, 2010) to reduce the ultra-high dimensionality \( n \ll p \) prior to the estimation.

- We studied collaborative performance of our proposed method with SIS for classification problem using *Leukemia data* (Fan and Lv, 2008).
Data description: Leukemia data

- Leukemia data are available in R package SIS.
- This data set consists of 72 samples with 7,129 genes.
- For the $i^{th}$ observation, the response variable $y_i$ is a binary outcome, indicating the types of acute leukemia (Acute Lymphoblastic Leukemia = 0 and Acute Myeloid Leukemia = 1).
- The predictor vector $x_i$ gives the expression levels of 7,129 genes.
- Define the probability of being Acute Myeloid Leukemia (AML) for the $i^{th}$ sample as $p_i = \text{Probability}(y_i = 1)$.
Set-up

- The link function $h$ is defined as

$$p_i = h(x_i^T \beta) = \frac{1 + \exp(-x_i^T \beta)}{1},$$

i.e., logit link.

- To specify the convex function, we define

$$\psi(x) = \sum_{i=1}^{n} \{x_i \log x_i + (1 - x_i) \log(x_i - 1)\},$$

which induces the Bernoulli likelihood.
Analysis procedure

- We randomly split the data into training set (of size 38) and test set (of size 34).

- First, we conduct a pre-screening procedure (SIS) to reduce the ultra-high-dimensionality on the training set.

- Using SIS, we select the top $152 (= 4n)$ genes, then analyze the reduced training data set using the GD method and the PL methods used in the simulation study.

- Using the test set, we compute Area Under Curve (AUC) for each method.

- We repeat the above procedure 100 times.
The GD method provided the largest AUC using the smallest number of genes.

Table: Average of AUC and Number of selected predictors

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>Number of predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>GD</td>
<td>0.9853</td>
<td>1.36</td>
</tr>
<tr>
<td>a-LASSO</td>
<td>0.9800</td>
<td>12.68</td>
</tr>
<tr>
<td>E-net</td>
<td>0.9812</td>
<td>28.57</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.9826</td>
<td>12.86</td>
</tr>
<tr>
<td>MCP</td>
<td>0.9787</td>
<td>8.70</td>
</tr>
<tr>
<td>SCAD</td>
<td>0.9790</td>
<td>11.35</td>
</tr>
</tbody>
</table>
Concluding remarks

- From a Bayesian perspective, we have developed a new approach to sparse high-dimensional problems using Bregman divergence and a valid $\ell_0$-norm approximation.

- One advantage of our divergence-based approach is that many extensions can be easily developed by replacing a new divergence measure in the likelihood function.

- For example, using Bregman matrix divergence (Kulis et al., 2009), our model can be adapted to multivariate regression models.
REFERENCES


Thank you!