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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

  where $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ is the response vector, $\boldsymbol{X}$ is the $n \times p$ predictor matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the unknown coefficient vector, and $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}_n, \sigma^2 \boldsymbol{I}_n)$ with known σ^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 β cannot be identifiable.
- Let p^* be the number of non-zero elements in β.
- 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 β;

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

- 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

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

where $\mathcal{P}_\lambda(\cdot)$ is a deterministic penalty function with a tuning parameter λ (≥ 0) controlling the degree of penalization.

PLS estimation with ℓ_0-norm penalty

- Akaike (1974) and Schwarz (1978) introduced the ℓ_0-norm penalization as follows:

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

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

- Since the ℓ_0-norm penalty directly restricts the number of non-zero coefficients, it successfully induces the sparsity for $\hat{\boldsymbol{\beta}}_{\ell_0}$.

- However, due to non-convexity and discontinuity of the ℓ_0-norm penalty, finding the minimum is challenging especially when p is large.

PLS estimation with ℓ_1-norm penalty

- As an alternative, the ℓ_1-norm penalization, called the lasso (Tibshirani, 1996), was proposed:

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

  - ▷ Since the lasso leads a continuous and convex optimization, it addresses the computational drawback of the ℓ_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,

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

  where $\lambda_j = \lambda/|\hat{\beta}_j|^\gamma$, $\gamma > 0$, and $\hat{\beta}_j$ is a $\sqrt{n}$–consistent estimator for β_j.

  - ▷ The adaptive lasso remedies the bias problem in the lasso.
  - ▷ However, in the high-dimensional setup, it is challenging to find good λ_j.

Approximation of ℓ_0-norm penalty

- Recently, Goh et al. (2017) and Goh and Dey (2018) introduced an approximation of the ℓ_0-norm penalty,

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

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

- Note that as τ goes to zero, the new penalty approaches the ℓ_0 penalty:

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

Approximation of ℓ_0-norm penalty

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

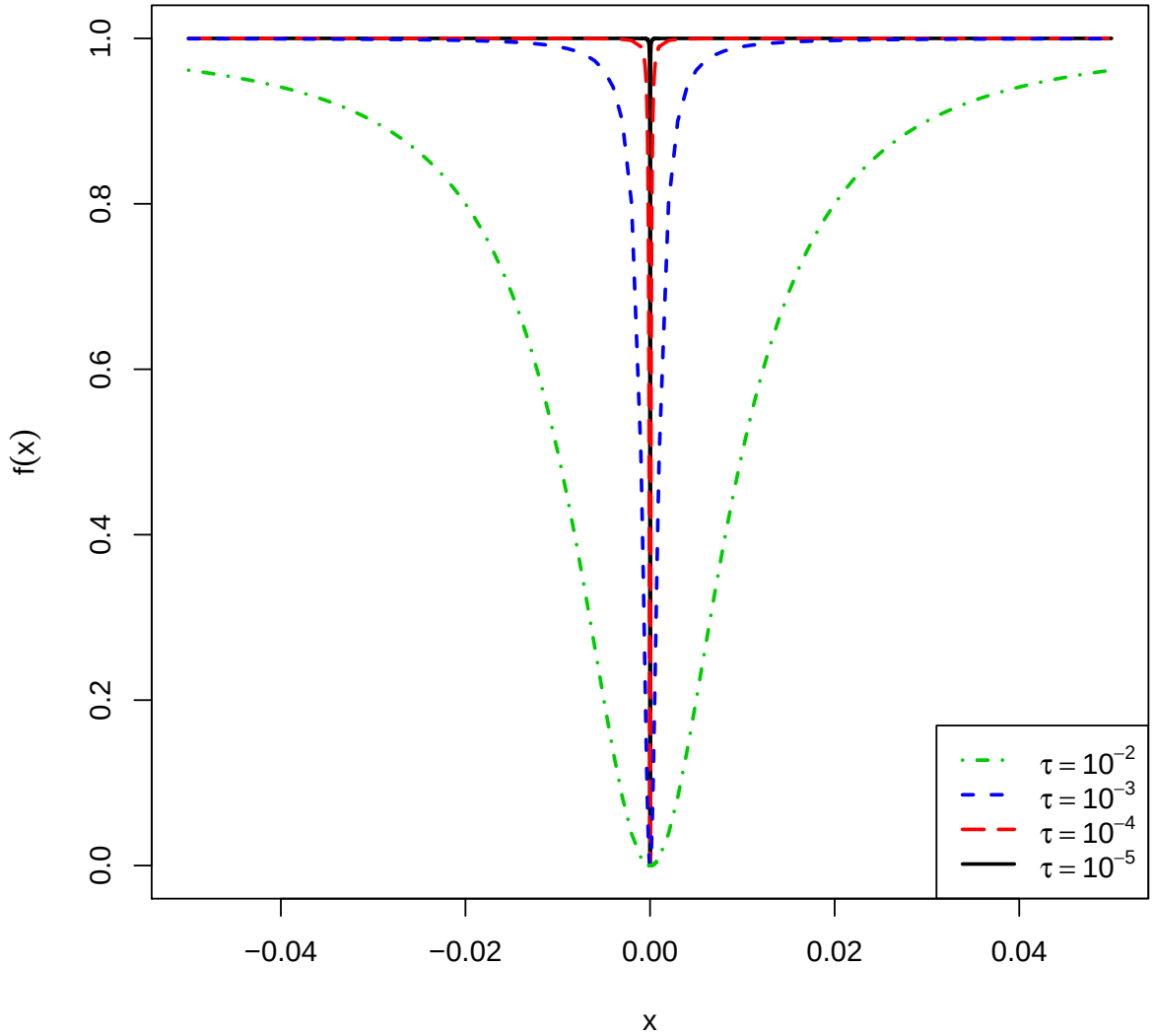

Figure: graphs of $f(x) = x^2/(\tau^2 + x^2)$ for $\tau = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$

Connection to the adaptive lasso

- 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 λ_j^* is automatically determined by β_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

$$\begin{aligned} \pi(\beta \mid \boldsymbol{y}, \lambda) &\propto f(\boldsymbol{y} \mid \beta)\pi(\beta \mid \lambda) \\ &\propto \exp\left(-\frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\beta||^2\right) \exp\left\{-\frac{1}{2}\mathcal{P}_\lambda(\beta)\right\}, \end{aligned}$$

  where
  – $f(\boldsymbol{y} \mid \beta)$: the likelihood function;
  – $\pi(\beta \mid \lambda)$: the prior for β given the hyperparameter λ.

- Then, it is easy to check that

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

Merit of Bayesian approach

- Let $\hat{\beta}_{PLS}$ be the PLS estimator, i.e.,

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

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

- Let $\hat{\beta}_{MAP}$ be the MAP estimator, i.e., $\hat{\beta}_{MAP} = \arg\max_{\beta} \pi(\beta \mid \boldsymbol{y}, \lambda)$.

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

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

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

where $L_2(\boldsymbol{x}_1, \boldsymbol{x}_2) = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2$ is the squared Euclidean loss and $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}$ 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

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

where $\mathrm{BD}_{\psi}(\cdot,\cdot)$ denotes the Bregman divergence.

Bregman divergence (Bregman, 1967)

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 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$ the Bregman divergence with respect to ψ is defined as

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

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

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

Graphical Illustration

- The Bregman divergence measures the *ordinate distance* between the value of the convex function at $\boldsymbol{x}$ and its tangent at $\boldsymbol{y}$.

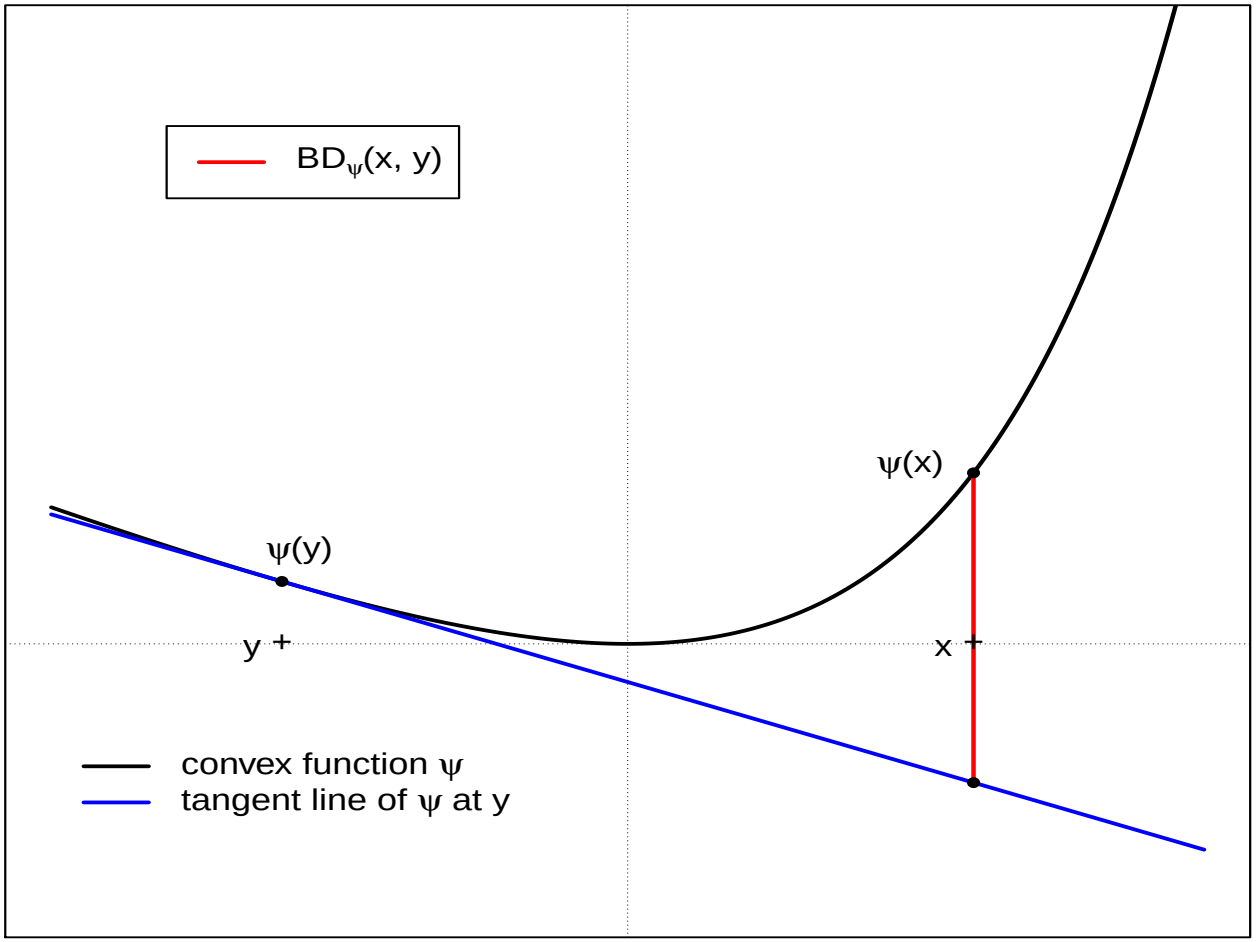

Figure: Bregman divergence with $\psi(x) = e^{cx} - cx - 1$, $c = 0.5$.

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, ψ's.

$\psi(\boldsymbol{x})$	Bregman divergence
$\|\|\boldsymbol{x}\|\|^2$	Squared error loss
$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{x}$	Mahalanobis distance
$\sum_{i=1}^{n} x_i \log x_i$	Kullback-Leibler divergence
$\sum_{i=1}^{n} -\log x_i$	Itakura-Saito distance
$\sum_{i=1}^{n} e^{cx_i}$	Weighted Linex loss

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 $\psi(\boldsymbol{y}) = \sum_{i=1}^{n} \{y_i \log y_i\}$, then our likelihood reduces to the Poisson likelihood,

$$
\begin{aligned}
f(\boldsymbol{y}|\beta) &\propto \exp\left[-\mathrm{BD}_{\psi}\left(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{X}\beta)\right)\right] \\
&\propto \exp\left[-\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{h(\boldsymbol{x}_i^{\mathrm{T}}\beta)}\right) - \left(y_i - h(\boldsymbol{x}_i^{\mathrm{T}}\beta)\right)\right\}\right] \\
&\propto \prod_{i=1}^{n}\left[e^{-h(\boldsymbol{x}_i^{\mathrm{T}}\beta)}\left\{h(\boldsymbol{x}_i^{\mathrm{T}}\beta)\right\}^{y_i}\right].
\end{aligned}
$$

Bregman divergence and natural exponential family

Table: Examples of Bregman divergence and related distributions in the natural exponential family.

$\psi(z)$	$\mathrm{BD}_\psi(z_1, z_2)$	Distribution
$\frac{1}{2\sigma^2} z^2$	$\frac{1}{2\sigma^2}(z_1 - z_2)^2$	Gaussian
$z \log z$	$z_1 \log\left(\frac{z_1}{z_2}\right) - (z_1 - z_2)$	Poisson
$-\log z$	$\frac{z_1}{z_2} - \log\left(\frac{z_1}{z_2}\right) - 1$	Exponential
$z \log z + (1-z)\log(1-z)$	$z_1 \log\left(\frac{z_1}{z_2}\right) + (1-z_1)\log\left(\frac{1-z_1}{1-z_2}\right)$	Bernoulli

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 $\psi(\boldsymbol{y}) = \sum_{i=1}^{n} \int_{-\infty}^{y_i} \frac{y_i - s}{V(s)} ds$, where $V(\cdot)$ is a positive known function.
- Then, it can be shown that our likelihood reduces to the quasi-likelihood,

$$\begin{aligned} f(\boldsymbol{y}|\beta) &\propto \exp\left[-\mathrm{BD}_{\psi}\left(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{X}\beta)\right)\right] \\ &\propto \exp\left\{\sum_{i=1}^{n} \int_{-\infty}^{h(\boldsymbol{x}_i^{\mathrm{T}}\beta)} \frac{y_i - s}{V(s)} ds\right\}. \end{aligned}$$

GD prior

- Now, we define our new prior, called GD prior, by

$$\pi_{\text{GD}}(\boldsymbol{\beta}, \mathbf{d}) \quad \propto \quad \pi_{\text{G}}(\boldsymbol{\beta}|\mathbf{d})\pi_{\text{D}}(\mathbf{d}),$$

such that

$$\pi_{\text{G}}(\boldsymbol{\beta}|\mathbf{d}) \quad \propto \quad \prod_{j=1}^{p} \left\{ d_j^{1/2} \exp\left(-\frac{d_j}{2}\beta_j^2\right)\right\} \quad \text{(Gaussian)},$$

$$\pi_{\text{D}}(\mathbf{d}) \quad \propto \quad \prod_{j=1}^{p} \left\{ d_j^{\lambda-1/2} \exp\left(-\frac{\tau^2}{2} d_j\right)\right\} \quad \text{(Diffused-gamma)},$$

where $\tau(>0)$ is determined to be sufficiently small.

Connection to ℓ_0-norm penalization

- Recall that a valid approximation of the ℓ_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

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

  where $\hat{\mathbf{d}} = \arg\max_{\mathbf{d}} \{\max_{\boldsymbol{\beta}} f(\boldsymbol{y}|\boldsymbol{\beta}) \pi_{\mathrm{GD}}(\boldsymbol{\beta}, \mathbf{d})\}$.

- For $\tau \approx 0$, our MAP estimator of $\boldsymbol{\beta}$ approximates the penalized Bregman divergence estimator with the ℓ_0-norm penalty.

Maximum A Posteriori (MAP) estimation

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

$$(\hat{\beta}, \hat{\mathbf{d}}) = \arg\max_{\beta, \mathbf{d}} \{f(\boldsymbol{y}|\beta)\pi_{\mathrm{G}}(\beta|\mathbf{d})\pi_{\mathrm{D}}(\mathbf{d})\}\,.$$

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

$$\begin{aligned} \hat{\mathbf{d}} &\leftarrow \arg\max_{\mathbf{d}} \left\{\pi_{\mathrm{G}}(\hat{\beta}|\mathbf{d})\pi_{\mathrm{D}}(\mathbf{d})\right\}; \\ \hat{\beta} &\leftarrow \arg\max_{\beta} \left\{f(\boldsymbol{y}|\beta)\pi_{\mathrm{G}}(\beta|\hat{\mathbf{d}})\right\}; \end{aligned}$$

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 $j = 1, 2, \ldots, p$;

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

until convergence.

Return $\hat{\beta}$.

Prior specification

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

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

$$m(\boldsymbol{y}|\lambda) = \int f(\boldsymbol{y}|\beta)\pi_G(\beta|\mathbf{d}_\lambda)d\beta,$$

  where $\mathbf{d}_\lambda$ denotes the MAP of $\mathbf{d}$ given λ.

- The optimal value of λ can be defined as

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

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):

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

- 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 $\mu_i = h_1(\boldsymbol{x}_i^{\mathrm{T}}\beta)$, where $h_1(x) = x$, $\beta = (\text{rep}(2,5), \text{rep}(0,10), \text{rep}(-2,5), \text{rep}(0, p-20))^{\mathrm{T}}$, and $\boldsymbol{x}_i \overset{iid}{\sim} \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\Sigma_{ij})_{p\times p}$ and $\Sigma_{ij} = \rho^{|i-j|}$.

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

M3. Generate $y_i \overset{iid}{\sim} \textit{Poisson}(\mu_i)$ with $\mu_i = h_3(\boldsymbol{x}_i^{\mathrm{T}}\beta)$, where $\beta = (\text{rep}(2,3), \text{rep}(0,10), \text{rep}(-2,3), \text{rep}(0, p-16))^{\mathrm{T}}$, $h_3(x) = \exp(x)$, $\boldsymbol{x}_i = \boldsymbol{\Phi}(\mathbf{z}_i) - 0.5\mathbf{1}_p$, $\boldsymbol{\Phi}(\mathbf{z}_i) = (\Phi(z_{i1}), \ldots, \Phi(z_{ip}))^{\mathrm{T}}$, $\Phi(\cdot)$ is the CDF of standard normal distribution, and $\mathbf{z}_i \overset{iid}{\sim} \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\Sigma_{ij})_{p\times p}$ and $\Sigma_{ij} = \rho^{|i-j|}$.

Simulation studies

Bregman divergence specification

- To specify the Bregman divergence likelihood function, we define

$$\begin{aligned}
\textbf{M1}: \ \psi(\boldsymbol{x}) &= \sum_{i=1}^{n} \left\{ \frac{x_i^2}{2} \right\}, \\
\textbf{M2}: \ \psi(\boldsymbol{x}) &= \sum_{i=1}^{n} \{x_i \log x_i + (1 - x_i) \log(x_i - 1)\}, \\
\textbf{M3}: \ \psi(\boldsymbol{x}) &= \sum_{i=1}^{n} \{x_i \log x_i\},
\end{aligned}$$

Simulation studies

Result

Table: Simulation results for continuous data (**M1**).

Case	(n, p, ρ)	Method	MSE_{est}	MSE_{pred}	FPR%	FNR%
M1	(100,100,0.5)	Oracle	0.0018	0.1028	0.0000	0.0000
		GD	0.0019	0.1070	0.0000	0.0000
		a-LASSO	0.0133	1.6622	4.6000	0.0000
		E-net	0.0120	1.6084	5.6667	0.0000
		Lasso	0.0138	1.7399	4.2333	0.0000
		MCP	0.0040	0.2136	0.5889	0.0000
		SCAD	0.0046	0.2393	0.6333	0.0000
	(100,500,0.5)	Oracle	0.0004	0.0995	0.0000	0.0000
		GD	0.0004	0.1052	0.0000	0.0000
		a-LASSO	0.0025	1.5074	1.6612	0.0000
		E-net	0.0024	1.5681	2.1796	0.0000
		Lasso	0.0025	1.5404	1.7204	0.0000
		MCP	0.0009	0.2254	0.3714	0.0000
		SCAD	0.0010	0.2516	0.4122	0.0000

Oracle: MLE under the true model.

Simulation studies

Result

Table: Simulation results for binary data (**M2**).

Case	(n, p, ρ)	Method	MSE_{est}	MSE_{pred}	FPR%	FNR%
M2	(100,100,0.5)	Oracle	0.0149	0.0039	0.0000	0.0000
		GD	0.0238	0.0110	0.6495	0.0000
		a-LASSO	0.0542	0.0301	1.7732	0.0000
		E-net	0.0765	0.0545	1.8351	2.0000
		Lasso	0.0542	0.0301	1.7732	0.0000
		MCP	0.0447	0.0235	2.1237	0.0000
		SCAD	0.0480	0.0256	2.2165	0.0000
	(100,500,0.5)	Oracle	0.0003	0.0005	0.0000	0.0000
		GD	0.0017	0.0026	0.0342	0.0000
		a-LASSO	0.0019	0.0062	0.0201	0.3333
		E-net	0.0025	0.0107	0.0241	0.6667
		Lasso	0.0019	0.0062	0.0201	0.3333
		MCP	0.0017	0.0051	0.0262	0.3333
		SCAD	0.0018	0.0058	0.0302	0.3333

Simulation studies

Result

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

Case	(n, p, ρ)	Method	MSE_{est}	MSE_{pred}	FPR%	FNR%
M3	(100,100,0.5)	Oracle	0.0040	1.3862	0.0000	0.0000
		GD	0.0049	1.5019	0.0638	0.0000
		a-LASSO	0.0103	4.1607	3.2128	0.0000
		E-net	0.0126	5.1199	8.6809	0.0000
		Lasso	0.0103	4.1607	3.2128	0.0000
		MCP	0.0077	5.4540	0.9255	0.5000
		SCAD	0.0084	5.4122	1.0213	0.8333
	(100,500,0.5)	Oracle	0.0008	1.3339	0.0000	0.0000
		GD	0.0011	1.5134	0.0263	0.0000
		a-LASSO	0.0031	6.1365	1.0891	0.0000
		E-net	0.0054	6.6374	3.5789	0.0000
		Lasso	0.0031	6.1365	1.0891	0.0000
		MCP	0.0086	8.7068	0.6943	6.6667
		SCAD	0.0065	7.4857	0.7126	5.0000

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 $\boldsymbol{x}_i$ 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 =Probability$(y_i = 1)$.

Set-up

- The link function h is defined as

$$p_i = h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) = \left[1 + \exp\left(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right]^{-1},$$

  i.e., logit link.

- To specify the convex function, we define

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

  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.

Analysis result

- The GD method provided the largest AUC using the smallest number of genes.

Table: Average of AUC and Number of selected predictors

Method	AUC	Number of predictors
GD	0.9853	1.36
a-LASSO	0.9800	12.68
E-net	0.9812	28.57
Lasso	0.9826	12.86
MCP	0.9787	8.70
SCAD	0.9790	11.35

Concluding remarks

- From a Bayesian perspective, we have developed a new approach to sparse high-dimensional problems using Bregman divergence and a valid ℓ_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

Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005), "Clustering with Bregman Divergences," *Journal of Machine Learning Research*, 6, 1705–1749.

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, London: Wiley.

Bregman, L. M. (1967), "The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming," *USSR Computational Mathematics and Mathematical Physics*, 7, 200–217.

Fan, J., and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society, Series B*, 70, 849–911.

Fan, J., and Song, R. (2010), "Sure independence screening in generalized linear models with NP–dimensionality," *The Annals of Statistics*, 38, 3217–3841.

Goh, G., and Dey, D. K. (2018), "Bayesian MAP estimation using Gaussian and diffused-gamma prior," *The Canadian Journal of Statistics*, Accepted.

Goh, G., Dey, D. K., and Chen, K. (2017), "Bayesian sparse reduced rank multivariate regression," *Journal of Multivariate Analysis*, 157, 14 – 28.

Kulis, B., Sustik, M. A., and Dhillon, I. S. (2009), "Low–Rank Kernel Learning with Bregman Matrix Divergences," *Journal of Machine Learning Research*, 10, 341–376.

Kyung, M., Gilly, J., Ghosh, M., and Casella, G. (2010), "Penalized Regression, Standard Errors, and Bayesian Lassos," *Bayesian Analysis*, 5, 369–412.

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Wedderburn, R. W. M. (1974), "Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method," *Biometrika*, 61, 439–447.

Zhang, C., Jiang, Y., and Shang, Z. (2009), "New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation," *Canadian Journal of Statistics*, 37, 119–139.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Q/A

Thank you!