L^γ Penalty Models

Computation And Applications

Part IV

Wenjiang Fu

wfu@stat.tamu.edu

http://stat.tamu.edu/$\sim$wfu.

Department of Statistics, Texas A&M University

BAYESIAN APPROACH

Bayesian interpretation

Linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,
where $\boldsymbol{Y}$: $\boldsymbol{n}$–vector of responses, $\boldsymbol{X}$: regression matrix, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^T$ $\boldsymbol{p}$–vector of parameters, $\boldsymbol{\varepsilon}$: $\boldsymbol{n}$–vector of random errors with mean $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathbf{var}(\boldsymbol{\varepsilon})\boldsymbol{\sigma}_0^2$.

$\boldsymbol{L}^\gamma$ penalty has a Bayesian prior interpretation (Lectures 1–2). Assume $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma}_0^2)$. Let $\boldsymbol{\beta}_j \sim \boldsymbol{L}^\gamma$ prior, $\boldsymbol{j} = \mathbf{1}, \ldots, \boldsymbol{p}$. Study posterior $\boldsymbol{\pi}(\boldsymbol{\beta}|\boldsymbol{y})$ for given data $\boldsymbol{y}$.

However, it is difficult to compute the posterior due to lack of conjugate property in general.

Notice that two members of the $\boldsymbol{L}^\gamma$ family are special and play a major role: lasso ($\gamma = 1$) and ridge ($\gamma = 2$), which correspond to Laplacian and Gaussian priors, respectively.

Laplacian prior

We study a novel family of priors including Laplacian and Gaussian as special cases.

Why Laplacian prior?

◇ Achieve variable selection. Same idea for Lasso.

◇ Representation by simple distributions.

$$\text{Lap}(1) \overset{d}{=} \boldsymbol{N}(\mathbf{0}, \mathbf{2\Lambda}) \text{ with } \boldsymbol{\Lambda} \sim \mathbf{Exp}(\mathbf{1}) \text{ (Kotz et al. 2000).}$$

◇ Studied for variable selection with applications to microarray studies (Bae and Mallick 2004).

Extension to a Bayesian prior family

Special properties of $\mathrm{Gamma}(\lambda, k)$:

Mean $\mu = \lambda/k$ and variance $\sigma^2 = \lambda/k^2$.

How to achieve $N(0, 2C)$?

Consider $\mathrm{Gamma}(1 + Ct, 1 + t)$ with $t \geq 0$ and constant $C > 0$.

Two special cases:

◇ $t = 0$. $\beta_j \sim \mathrm{Lap}(1) \stackrel{d}{=} \mathrm{Gamma}(1, 1)$.

◇ $t \longrightarrow \infty$, $\mathrm{Gamma}(1 + Ct, 1 + t) \xrightarrow{p} C$. $\beta_j \sim N(0, 2C)$.

For $\mathrm{Gamma}(\lambda, k)$, $\mu = \lambda/k$, $\sigma^2 = \lambda/k^2$.

Bayesian Approach

Laplacian – Gaussian mixture (LGM) prior

For given $t \geq 0$ and constant $C > 0$.

Consider $\beta_j \sim N(0, 2\Lambda) \cdot \mathrm{Gamma}(1 + Ct, 1 + t)$,

the Laplacian–Gaussian mixture prior.

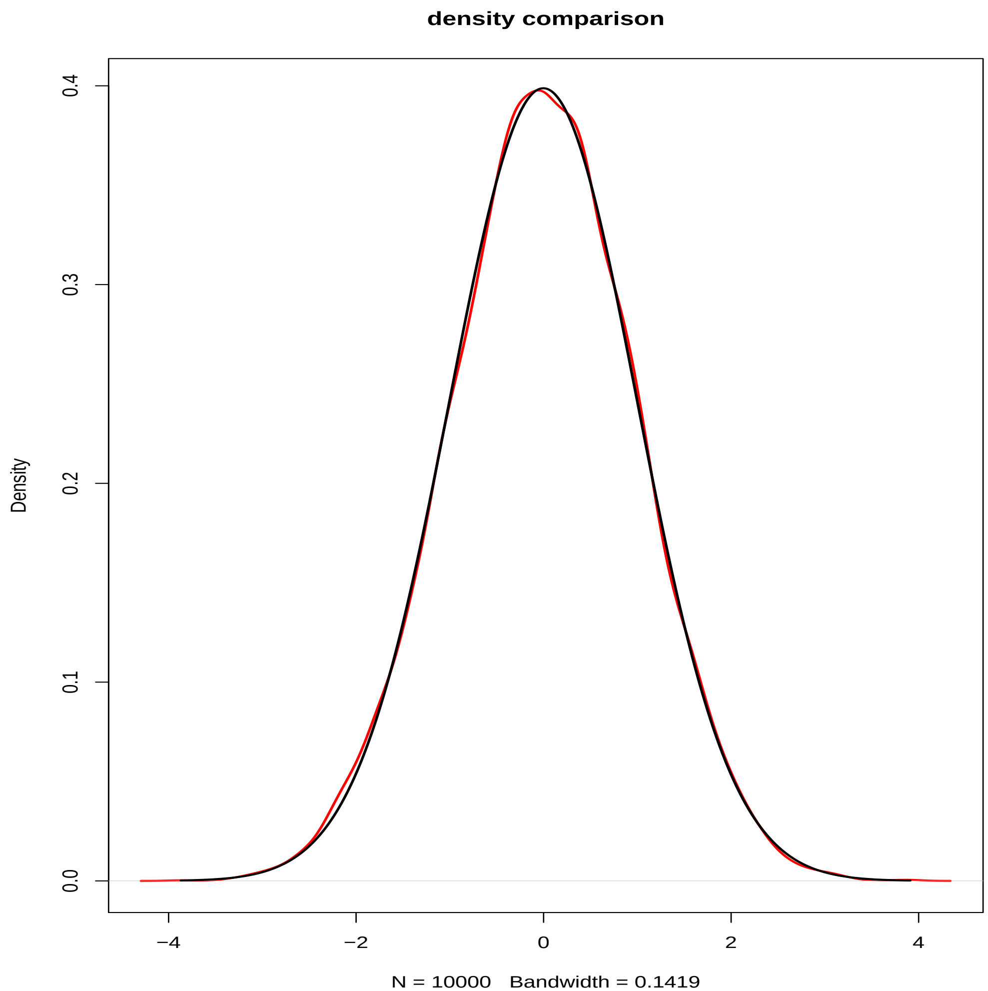
◇ LGM is a natural extension of Laplacian to a family including Gaussian.

◇ Posterior computation will take the advantage of $\beta_j \sim N(0, 2\Lambda)$.

◇ L^γ is another one, but the posterior is difficult to handle. If use $\beta_j \sim \exp(-\lambda|\beta|^\gamma)$, it will involve stable distribution, complicated and difficult.

Laplacian – Gaussian mixture (LGM) prior

LGM behaves like a Gaussian for $t \geq 300$.



Black curve: $N(0, 1)$; Red curve: LGM $t = 300$

Advantages of LGM prior

- **Prior on hyperparameter $t \geq 0$: $\pi(t)$.**

◇ $\pi(t)$: point mass at $t = 0$, Laplacian prior.

◇ $\pi(t)$: point mass at $t = t_0$ large $(t_0 \geq 300)$, Gaussian prior.

◇ $\pi(t)$: point mass at $t = 0$ and $t = t_0 > 0$ large.
Combines Lasso and ridge, Elastic Net (Zou and Hastie 2004).

◇ $\pi(t)$: continuous $t \geq 0$, Bayesian model averaging.

Posterior with LGM prior

Denote $\Lambda^{-1} = \mathrm{diag}(\Lambda_1^{-1}, \ldots, \Lambda_p^{-1})$, the inverse of the diagonal matrix of elements $(\Lambda_1, \ldots, \Lambda_p)$.

$$\begin{aligned} \pi(\beta, \Lambda | y, t) &\propto \exp[-\frac{1}{2\sigma_0^2}(y - X\beta)^T (y - X\beta)] \\ &\times \exp(-\frac{1}{4}\beta^T \Lambda^{-1} \beta) \frac{(1+t)^{p(1+ct)}}{[\Gamma(1+ct)]^p} \\ &\times \Lambda_1^t \ldots \Lambda_p^t \exp[-(1+t)(\Lambda_1 + \ldots + \Lambda_p)]. \end{aligned}$$

$$\pi(\beta | y, t) = \int_\Lambda \pi(\beta, \Lambda | y, t)\, d\Lambda$$

Bayesian variable selection (SSVS)

SSVS (Stochastic search variable selection) (George and McCulloch 1993)

For given γ_j Bernoulli (0 or 1) – index for variable selection.
$\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2)$, $\quad c_j > 0$ large,
$P(\gamma_j = 1) = 1 - P(\gamma_j = 0) = p_j$.

Priors: $\beta|\gamma \sim N_p(0, D_\gamma R D_\gamma)$,
Variance component: $\quad \sigma^2|\gamma \sim \mathrm{IG}(\nu_\gamma/2, \nu_\gamma\lambda_\gamma/2)$,

$$\gamma \sim f(\gamma) = \Pi\, p_j^{\gamma_j}(1-p_j)^{(1-\gamma_j)}.$$

Computational Methods for SSVS

Gibbs sampling for best subset:

$$\boldsymbol{\beta}^0, \boldsymbol{\sigma}^0, \gamma^0, \boldsymbol{\beta}^1, \boldsymbol{\sigma}^1, \gamma^1, \ldots, \boldsymbol{\beta}^m, \boldsymbol{\sigma}^m, \gamma^m, \ldots,$$

Variable selection by determining posterior distribution of γ.

Computationally intensive !

Metropolis – Hastings search.

Brown, Vannucci and Fearn (1998, 2002),
Brown, Fearn and Vannucci (1999),
Vannucci, Brown and Fearn (2001),
Lee, Sha, Dougherty, Vannucci and Mallick (2003).

Representation of Laplace Distribution

Theorem 7 (Kotz 2000)
A standard classical Laplace r.v. X has the representation $X \stackrel{d}{=} \sqrt{2W}Z$, where the r.v.s W and Z have the standard exponential and normal distributions, respectively.

The moment generating function of exponential W is $M_w(t) = (1-t)^{-1}, t < 1$. Characteristic function of normal Z is $\exp(-t^2/2)$. The characteristic function of $\sqrt{2W}Z$

$$
\begin{aligned}
\mathrm{E}[\exp(it\sqrt{2W}Z)] &= \mathrm{E}\{\mathrm{E}[\exp(it\sqrt{2W}Z)|W]\} \\
&= \mathrm{E}[\phi_z(t\sqrt{2W})] = \mathrm{E}[\exp(-t^2W)] \\
&= M_W(-t^2) = (1+t^2)^{-1},
\end{aligned}
$$

where $\phi_z(t) = \exp(-t^2/2)$ is the characteristic function of standard normal.

Representation of Laplace Distribution

The density of $X = \sqrt{2W}Z$ is given by

$$\int_0^\infty \frac{1}{2\sqrt{\pi w}} \exp[-\frac{1}{2}(\frac{x^2}{2w} + 2w)]dw.$$

Consider transformation $Y_1 = W, Y_2 = \sqrt{2W}Z$. Calculate joint density and the marginal of Y_2 by integrating out Y_1.

Fused Lasso

Fused Lasso (Tibshirani, Saunders, Rosset and Zhu 2005) Consider linear model

$$\boldsymbol{y_i = \sum X_{ij}\beta_j + \varepsilon_i},$$

where $\boldsymbol{x_j = (x_{1j}, \ldots, x_{nj})}$ are standardized and ordered variables (Protein mass spectroscopy data: time of flight with mass/charge m/z).

Idea Penalize both the parameters $|\boldsymbol{\beta_j}|$ and their differences $|\boldsymbol{\beta_j - \beta_{j-1}}|$.

Min RSS subject to $\sum |\boldsymbol{\beta_j}| \le \boldsymbol{s_1}$ and $\sum |\boldsymbol{\beta_j - \beta_{j-1}}| \le \boldsymbol{s_2}$.

Goal Achieve sparsity and smoothness.

Two special cases:

1). Lasso: $\boldsymbol{s_2}$ is large; 2). Fusion: $\boldsymbol{s_1}$ is large (Land and Friedman 1996).

FUSED LASSO

Performance using prostate cancer protein mass spectroscopy data with random split of training 216 + test 108 samples (total 157 healthy + 167 cancer) (Tibshirani 2005)

Model	test err	df	sites	s_1	s_2
Lasso	6/108	116	116	144	262
Fusion	19/108	168	171	175	200
Fused Lasso	6/108	122	344	184	222

Fused Lasso

Comparison with leukemia classification using microarrays (training: 27+11; test: 34) (Tibshirani 2005)

Method	s_1	s_2	10-FdCV	Test err	genes
Golub (50 genes)			3/38	4/34	50
Lasso 37 df	0.65	1.32	1/38	1/34	37
Fused Lasso 38 df	1.08	0.71	1/38	2/34	135
Fused Lasso 20 df	1.35	1.01	1/38	4/34	737

Elastic Net Model

Elastic Net (ENet) (Zou and Hastie 2004)

$$L(\beta, \lambda_1, \lambda_2) = (y - X\beta)^T(y - X\beta) + \lambda_1|\beta|_1 + \lambda_2|\beta|_2^2,$$

where $|\beta|_1 = \sum|\beta_j|$, $|\beta|_2^2 = \sum \beta_j^2$.

Naive elastic net estimator is defined as

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}}\, L(\beta, \lambda_1, \lambda_2).$$

It's equivalent to

$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}}(y - X\beta)^T(y - X\beta)$ subject to

$(1-\alpha)|\beta|_1 + \alpha|\beta|_2^2 \leq t$ for some t.

ENet combines Lasso penalty and ridge penalty.

Algorithm for Naive ENet

Given data $(\boldsymbol{y}, \boldsymbol{X})$ and fixed $(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2})$. Define artificial data set $(\boldsymbol{y}^*, \boldsymbol{X}^*)$ by

$$\boldsymbol{X}^*_{(n+p)\times p} = (1+\boldsymbol{\lambda_2})^{-1/2}\begin{pmatrix} \boldsymbol{X} \\ \sqrt{\boldsymbol{\lambda_2}}\boldsymbol{I} \end{pmatrix}, \quad \boldsymbol{y}^*_{(n+p)} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix}.$$

Let $\gamma = \boldsymbol{\lambda_1}/\sqrt{1+\boldsymbol{\lambda_2}}$ and $\boldsymbol{\beta}^* = \sqrt{1+\boldsymbol{\lambda_2}}\boldsymbol{\beta}$.

$\boldsymbol{L}^*(\boldsymbol{\beta}^*, \gamma) = (\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)^T(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*) + \gamma|\boldsymbol{\beta}^*|_1.$

Let $\widehat{\boldsymbol{\beta}}^* = \operatorname*{argmin}_{\boldsymbol{\beta}^*} \boldsymbol{L}(\boldsymbol{\beta}^*, \gamma)$, then the ENat estimator

$$\widehat{\boldsymbol{\beta}} = \frac{1}{\sqrt{1+\boldsymbol{\lambda_2}}}\widehat{\boldsymbol{\beta}}^*.$$

Relationship with Lasso estimator

For orthonormal design matrix $\boldsymbol{X}$,

$$\widehat{\boldsymbol{\beta}}_j(\mathbf{NENet}) = \frac{(|\widehat{\boldsymbol{\beta}}_j(\mathbf{ols})| - \boldsymbol{\lambda}_1/2)_+}{1+\boldsymbol{\lambda}_2}\mathbf{sign}(\widehat{\boldsymbol{\beta}}_j(\mathbf{ols})).$$

Two special cases:

1) $\boldsymbol{\lambda}_1 = 0$, ridge estimator $\widehat{\boldsymbol{\beta}}(\mathbf{ridge}) = 1/(1+\boldsymbol{\lambda}_2)\widehat{\boldsymbol{\beta}}(\mathbf{ols})$;
2) $\boldsymbol{\lambda}_2 = 0$, lasso estimator

$$\widehat{\boldsymbol{\beta}}_j(\mathbf{lasso}) = (|\widehat{\boldsymbol{\beta}}_j(\mathbf{ols})| - \boldsymbol{\lambda}_1/2)_+\mathbf{sign}(\widehat{\boldsymbol{\beta}}_j(\mathbf{ols})).$$

Grouping effect of ENet

Given data $(\boldsymbol{y}, \boldsymbol{X})$ and $(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2})$ with centered $\boldsymbol{y}$ and standardized $\boldsymbol{X}$. Let $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2})$ be the NENet estimator. Suppose $\widehat{\boldsymbol{\beta}}_i(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2})\widehat{\boldsymbol{\beta}}_j(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}) > \mathbf{0}$. Define

$$\boldsymbol{D}_{\boldsymbol{\lambda_1},\boldsymbol{\lambda_2}}(\boldsymbol{i},\boldsymbol{j}) = \frac{\mathbf{1}}{|\boldsymbol{y}|_\mathbf{1}}|\widehat{\boldsymbol{\beta}}_i(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}) - \widehat{\boldsymbol{\beta}}_j(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2})|,$$

then $\boldsymbol{D}_{\boldsymbol{\lambda_1},\boldsymbol{\lambda_2}}(\boldsymbol{i},\boldsymbol{j}) \le [\sqrt{\mathbf{2}(\mathbf{1}-\boldsymbol{\rho})}]/\boldsymbol{\lambda_2}$,
where $\boldsymbol{\rho} = \boldsymbol{x}_i^T \boldsymbol{x}_j$, the sample correlation.
Highly correlated covariates tend to be selected together.

ENet estimator

Given data $(\boldsymbol{y}, \boldsymbol{X})$ and $(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$, and augmented data $(\boldsymbol{y}^*, \boldsymbol{X}^*)$.
Naive ENet estimator

$$\widehat{\boldsymbol{\beta}}^* = \mathop{\mathbf{argmin}}_{\boldsymbol{\beta}^*} (\boldsymbol{y}^* - \boldsymbol{X}^* \boldsymbol{\beta}^*)^T (\boldsymbol{y}^* - \boldsymbol{X}^* \boldsymbol{\beta}^*) + \frac{\boldsymbol{\lambda}_1}{\sqrt{\mathbf{1} + \boldsymbol{\lambda}_2}} |\boldsymbol{\beta}^*|_1.$$

$$\widehat{\boldsymbol{\beta}}(\mathbf{ENet}) = \sqrt{\mathbf{1} + \boldsymbol{\lambda}_2}\, \widehat{\boldsymbol{\beta}}^*.$$

So

$$\widehat{\boldsymbol{\beta}}(\mathbf{ENet}) = (\mathbf{1} + \boldsymbol{\lambda}_2)\, \widehat{\boldsymbol{\beta}}(\mathbf{NENet})$$

possesses all properties of the Lasso.

ENet estimator

$$\widehat{\boldsymbol{\beta}}(\mathbf{ENet}) = \underset{\boldsymbol{\beta}}{\mathbf{argmin}}\, \boldsymbol{\beta}^{\boldsymbol{T}} \left(\frac{\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \boldsymbol{\lambda}_{\mathbf{2}}\boldsymbol{I}}{\mathbf{1} + \boldsymbol{\lambda}_{\mathbf{2}}} \right) \boldsymbol{\beta} - \mathbf{2}\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\lambda}_{\mathbf{1}}|\boldsymbol{\beta}|_{\mathbf{1}}.$$

$$\widehat{\boldsymbol{\beta}}(\mathbf{lasso}) = \underset{\boldsymbol{\beta}}{\mathbf{argmin}}\, \boldsymbol{\beta}^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})\boldsymbol{\beta} - \mathbf{2}\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\lambda}_{\mathbf{1}}|\boldsymbol{\beta}|_{\mathbf{1}}.$$

ELASTIC NET MODEL

Comparison with leukemia classification using microarrays

(training: 27+11; test: 34) (Zou and Hastie 2004)

Method	10-FdCV	Test err	genes
Golub	3/38	4/34	50
SVM	1/38	1/34	31
PenLogitReg	1/38	2/34	26
NSC(PAM)	2/38	2/34	21
ENet	3/38	0/34	45

References

Bae, and Mallick, B. (2004) Gene selection using a two-level hierarchical Bayesian model. *Bioinfo.* **20**, 18: 3423-3430.

Brown, P. Vannucci, M. and Fearn (1998) Multivariate Bayesian variable selection and prediction, *JRSSB*, **60**, 627-641.

Brown, P. Vannucci, M. and Fearn (2002) Bayesian model averaging with selection of regressors, *JRSSB*, **64**, 519-536.

Fu, WJ. (1998) Penalized regressions: the Bridge versus the Lasso. *J. Comp. Grap. Statist.* **7**, 3: 397-416.

George, EI. and McCulloch, R.E. (1993) Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, **88**, 881-889.

Golub,T. et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531-536.

Kotz, S. Kozubowski, TJ. and Podgorski, K. (2000) *The Laplace Distribution And Generalizations*, Birkhauser, Boston.

REFERENCES

Land, S. and Friedman, J. (1996) Varible Fusion: a new method of adaptive signal regression, Technical report, Department of Statistics, Stanford University.

Lee, KE. Sha, N. Dougherty, E. Vannucci, M. and Mallick, B. (2003) Gene selection: a Bayesian variable selection approach. *Bioinfo.*, **19**, 90-97.

Tibshirani R, Saunders M, Rosset S, Zhu, J. (2005) Sparsity and smoothness via the fused lasso, *J. Roy. Statist. Soc. B.* 67: 91-108.

Vannucci, M. Brown, P. Fearn, T. and (2001) Predictor selection for model averaging. In *Bayesian Methods with Applications to Science, Policy and Official Statistics*, (eds E.I. George and P. Nanopoulos), pp.553-562.

Zou, H. and Hastie, T. (2004) Regularization and variable selection via the elastic net, *Technical report, Stanford University*.