

Bayesian approaches for variable selection and shrinkage: An investigation of milk composition traits using infrared spectroscopy analysis

Yogasudha Veturi

&

Alessandro Ferragina

High dimensional data

- $p \gg n$; *curse of dimensionality!*
- Microarray data, deep sequencing, biomedical imaging, high-frequency data in finance
- OLS performs poorly in both prediction and interpretation
- Variable selection and/or shrinkage estimation procedure needed

Variable selection and shrinkage procedures

- Penalized
 - LASSO, ridge regression, elastic net
- Bayesian
 - BayesA, BayesB, BayesC, Bayesian LASSO
- Semi-/nonparametric
 - Neural Networks, random forests, support vector regression, reproducing kernel Hilbert spaces

Variable selection and shrinkage procedures

- Penalized
 - LASSO, ridge regression, elastic net
- Bayesian
 - BayesA, BayesB, BayesC, Bayesian LASSO
- Semi-parametric approaches
 - Neural Networks, random forests, support vector regression, reproducing kernel Hilbert spaces

Bias-variance trade-off

- $MSE(\hat{\theta}) = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$
- Using OLS, for fixed n , as p increases, variance of estimates increases \rightarrow high MSE
- If estimates are shrunk towards zero, variance of estimator is reduced, although bias may increase
- E.g. Consider $\tilde{\theta} = \alpha(\hat{\theta}) + (1 - \alpha)0$ $\alpha \in [0,1]$
- $E(\tilde{\theta}) = \alpha\theta$; $Var(\tilde{\theta}) = \alpha^2 Var(\hat{\theta})$

Penalty/prior

- Different penalties / priors -> different solutions
- Choice of penalty/prior is based on:
 - Model comparison
 - model interpretation
 - model parsimony
 - Accuracy of predictions on future data
 - Parameter estimation
- Bayesian setting is more useful for parameter estimation

Penalized methods

$$(\hat{\mu}, \hat{\beta})_{\text{argmin}} \left\{ \sum_i (y_i - \mu - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda J(\beta) \right\}$$

- $J(\beta)$ = penalty function
- λ = regularization parameter (controls trade-offs between lack of fit and model complexity)
- Choice of penalty determines extent of shrinkage and/or variable selection

Penalized Estimators

$$J(\beta)$$

Bridge Regression

$$\sum_{j=1}^p |\beta_j|^\gamma$$

LASSO

$$(\gamma = 1) \sum_{j=1}^p |\beta_j|$$

Ridge Regression

$$(\gamma = 2) \sum_{j=1}^p \beta_j^2$$

Subset selection

$$(\gamma \rightarrow 0) \sum_{j=1}^p I(\beta_j \neq 0)$$

Elastic Net

$$\alpha \sum_{j=1}^p |\beta_j| + (1 - \alpha) \sum_{j=1}^p \beta_j^2$$

LASSO

- Does both shrinkage and variable selection simultaneously

...however...

- Only selects n variables before it saturates
- Constraints on bound of the L1-norm
- Does not work well with correlated predictors - does not reveal grouping information
- Might result in low prediction power when $p \gg n$

Ridge regression

- Better bias-variance trade-off than LASSO

...however...

- Keeps all variables in the model (no variable selection); model parsimony is not achieved

Subset selection

- Produces a sparse model
- Penalizes non-zero effects regardless of magnitude

...however...

- Only selects n variables before it saturates (like the LASSO)

Elastic Net

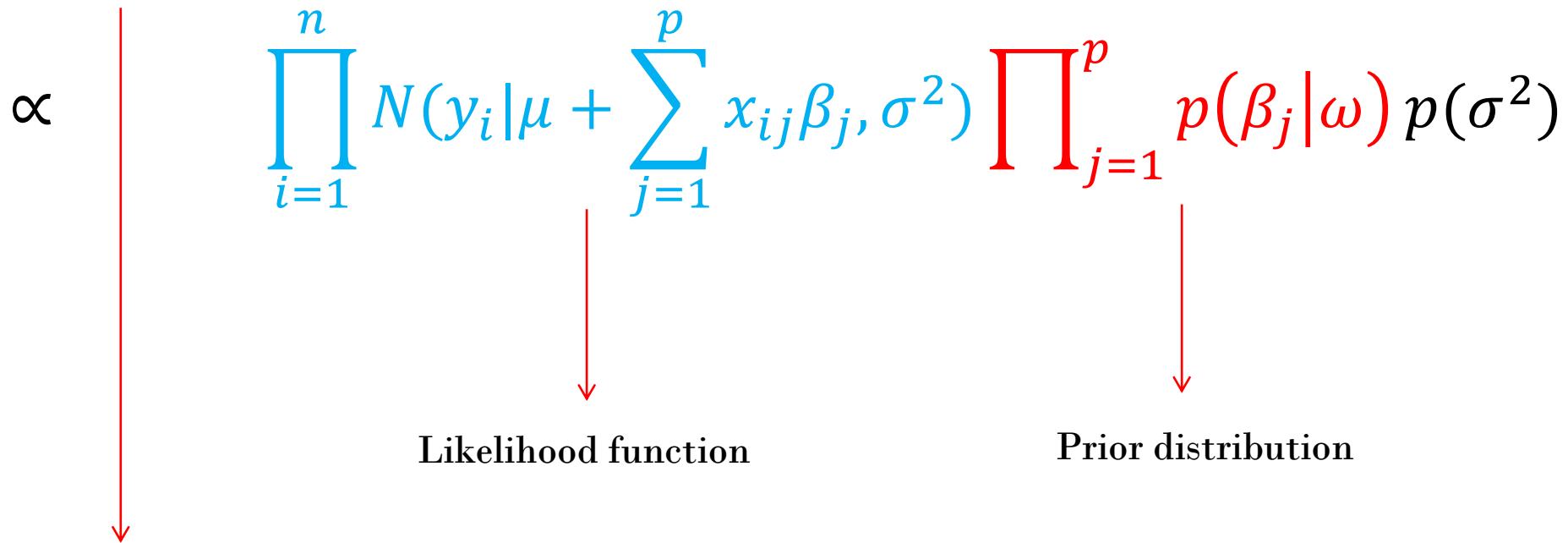
- Simultaneously does variable selection and shrinkage
- Can select groups of correlated variables; retains the “big fish”
- Algorithm “LARS-EN” can create the entire elastic net path with the computational efforts of a single OLS fit
- Also a good classifier; e.g. with microarray data, can do automatic gene selection (unlike other popular classifiers like LASSO, SVM, penalized logistic regression, nearest shrunken centroid)

...however...

Doesn't select related variables when the within-group correlations are non-extreme ($\rho \approx 0.85$)

Bayesian methods

$$p(\mu, \beta, \sigma^2 | y, \omega) \propto p(y | \mu, \beta, \sigma^2) p(\mu, \beta, \sigma^2 | \omega)$$

$$\propto \prod_{i=1}^n N(y_i | \mu + \sum_{j=1}^p x_{ij} \beta_j, \sigma^2) \prod_{j=1}^p p(\beta_j | \omega) p(\sigma^2)$$


The diagram illustrates the decomposition of the posterior distribution into its likelihood and prior components. Red arrows point from the terms in the equation to their respective labels: the likelihood function and the prior distribution.

Likelihood function

Prior distribution

Distribution of the unknowns given the data and hyper-parameters

Bayesian Estimators	$p(\beta_j \omega)$
Spike Slab models	$\pi N(\beta_j 0, \sigma_{\beta_1}^2) + (1 - \pi)N(\beta_j 0, \sigma_{\beta_2}^2)$
Bayes B	$\pi t_{df,S} + (1 - \pi)I(\beta_j = 0)$
Bayes A	$(\pi = 1) t_{df,S}$
Bayes C	$(df \rightarrow \infty) \pi N(\beta_j 0, \sigma_{\beta}^2) + (1 - \pi)I(\beta_j = 0)$
Bayesian Ridge Regression	$(\pi = 1, df \rightarrow \infty) N(\beta_j 0, \sigma_{\beta}^2)$
Bayesian LASSO	Double-exponential

Gaussian prior (Bayesian Ridge Regression)

- Multivariate normal with posterior mean **same as the RR** with $\lambda = \frac{\sigma^2}{\sigma^2_{\beta}}$.
- Homogeneous shrinkage across markers
- May not be useful when correlation patterns vary across the dataset

Thick tailed priors (Bayes A and Bayesian LASSO)

- Higher mass at zero and thicker tails
 - Bayesian LASSO has posterior mean **same as LASSO**
- Induces less shrinkage of large effect estimates than BRR
- Commonly represented as infinite mixtures of scaled normal densities
- Scaled t (2 parameters) has more flexibility than DE to control the thickness of the tails

Spike slab priors

- Mixture of two densities; one with small variance (spike) and the other with large variance (slab)
- Combines variable selection and shrinkage
- Can mix Gaussian or non-Gaussian (e.g. scaled- t and double exponential) components

Point of mass and slab priors (Bayes B and Bayes C)

- Obtained from spike-slab models when
$$(\sigma_{\beta_1}^2 \rightarrow 0)$$
- Again, induce a combination of variable selection and shrinkage

Comparison of priors

