Penalized least squares with non quadratic penalties

Seminar LEAR - Grenoble

A. Antoniadis
LJK-Université Joseph Fourier

Montbonnot, 28 Mai 2007

Summary

- Model formulation and basic notation
- Penalties
- Shrinkage estimation
- A closer look at Lasso, Bridge and SCAD estimators
- Some computational issues
- Asymptotics

Least squares

Consider the standard linear regression model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i \\ &= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \quad (i = 1, \ldots, n). \end{aligned}$$

Assume that the predictors are centered, so we can estimate β_0 by $\bar{Y}$ and focus on estimation of remaining parameters $\boldsymbol{\beta}$.

These parameters can be estimated by least squares (LS) or possibly some other more robust method.

Minimize $\|\mathbf{Y} - X\boldsymbol{\beta}\|^2$. The solution is known to be

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}.$$

- a possibility of collinearity in the design; this leads to increased variability in estimation.
- large number of predictors (relative to number of observations); this increases the possibility of overfitting.

A shrinkage approach will often result in estimates of the regression coefficients that, while biased, are lower in mean squared error and are more close to the true parameters.

How?

A good approach to shrinkage is penalized least squares estimation. The use of a criterion function with penalty has a long history which goes back to Whittaker (1929) and Tikhonov (1963).

A general form of penalized least squares is

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{p}\rho_\lambda(|\beta_j|)$$

From the least squares loss a so-called 'penalty' is added, that discourages regression coefficients to become large.

Penalty functions

Several penalty functions have been used in the literature.

- The L_2 penalty $\rho_\lambda(\beta) = \lambda|\beta|^2$ yields a ridge type regression
- The L_1 penalty $\rho_\lambda(\beta) = \lambda|\beta|$ results in LASSO (first proposed by Donoho and Johnstone (1994) in the wavelet setting and extended by Tibshirani (1996) for general least squares settings).
- More generally, the L_q $(0 \leq q \leq 1)$ leads to bridge regression (see Frank and Friedman (1993), Ruppert and Carroll (1997), Fu (1998), Knight and Fu (2000), Yu and Ruppert (2001)).

Conditions on ρ

Usually, the penalty function ρ is chosen to be symmetric and increasing on $[0, +\infty)$. Furthermore, ρ can be convex or non-convex, smooth or non-smooth.

A good penalty function should result in a estimator with the following three properties (Antoniadis & Fan, 2001):

- Unbiasedness : The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid excessive estimation bias
- Sparsity : Estimating a small coefficient as zero, to reduce model complexity
- Continuity : The resulting estimator is continuous in the data to avoid instability in model prediction

Generalities

Convex penalties (e.g. quadratic penalties)

- make trade-offs between bias and variance
- can create unnecessary biases when the true parameters are large
- parsimonious models cannot be produced

Nonconcave penalities

- select variables and estimate coefficients of variables simultaneously
- e.g. hard thresholding penalty (HARD, Antoniadis 1997)

$$\rho_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$$

Discussion

In the orthogonal design case, and for penalties that are symmetric and increasing on $[0, +\infty)$, differentiable everywhere *except perhaps at* $\beta = 0$ some necessary conditions for unbiasedness, sparsity and stability have been derived by Nikolova (2000) and Antoniadis and Fan (2001).

- unbiasedness $\leftrightarrow \dot{\rho}(|\beta|) = 0$ for large $|\beta|$
- sparsity $\leftrightarrow |\beta| + \lambda\dot{\rho}(|\beta|) \geq 0$
- stability $\leftrightarrow \text{argmin}\{|\beta| + \lambda\dot{\rho}(|\beta|)\} = 0$

From the above, a penalty satisfying the conditions on sparsity and stability must be non-smooth at 0.

Why?

Roughly the penalized estimator minimizes

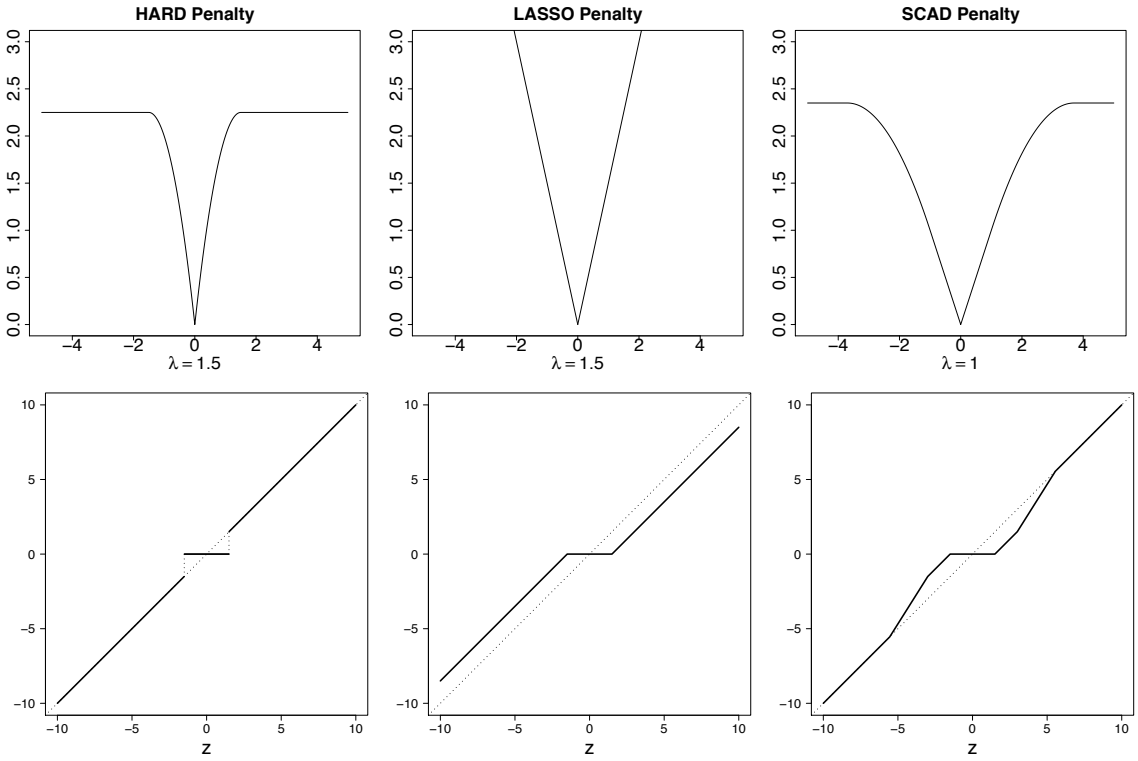
$$(z-\theta)^2/2 + \lambda\rho(|\theta|)$$

By the assumptions, the solution $\hat{\theta}$ is an antisymmetric function of z and can be located either at $\theta = 0$ or at a zero $\theta = \tau$ of the derivative of the criterion, i.e.

$$z = \tau + \lambda\dot{\rho}(\tau)$$

Because τ and $\dot{\rho}(\tau)$ have the same sign, we have $|\tau| \leq |z|$ (shrinkage). Moreover $\tau = z + \lambda\dot{\rho}(|\beta|) + o(\dot{\rho}(|\beta|))$ as $|\beta| \to \infty$.

SCAD



Shrinkage (related approaches)

Many penalization methods developed recently achieve shrinkage and variable selection.

- Nonnegative garrote (Breiman, 1995), which minimizes $\sum(y_i - \beta_0 - \sum_j c_j \beta_j x_{ij})^2$ under the constraint $\sum c_j \leq s$. The solution may be written as $\hat{\beta} = C\hat{\beta}_{ols}$ where $C \geq 0$ and diagonal and $\text{Trace}(C) \leq s$. Making s small will cause some coefficients to be exactly zero. However the solution depends on both the sign and the magnitude of the OLS coefficients.
- Elastic net (Zou & Hastie, 2005), where the penalty is a convex combination of the lasso and ridge penalty.
- Relaxed Lasso (Meinshausen, 2005).

Smoothly Clipped Absolute Deviation

To overcome LASSO's limitations Fan (1997) proposed the SCAD penalty function defined by

$$\dot{\rho}_\lambda(|\theta|) = \lambda \left\{ I(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| > \lambda) \right\}, \quad a > 2$$

with thresholding rule

$$\hat{\theta}(z) = \begin{cases} \operatorname{sgn}(z)(|z| - \lambda)_+, & |z| \leq 2\lambda \\ \{(a-1)z - \operatorname{sgn}(z)a\lambda\}/(a-2), & 2\lambda < |z| \leq a\lambda \\ z, & |z| > a\lambda. \end{cases}$$

It satisfies all three requirements (unbiasedness, sparsity and continuity).

SCAD

The SCAD penalty corresponds to a quadratic spline

$$\rho_\lambda(|\theta|) = \begin{cases} \lambda|\theta|, & |\theta| \leq \lambda \\ -\frac{(|\theta|^2 - 2a\lambda|\theta| + \lambda^2)}{2(a-1)}, & \lambda < |\theta| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2}, & |\theta| > a\lambda. \end{cases}$$

- Computation of the SCAD estimates can be done via Newton-Raphson.
- The SCAD function has a similar form as the L_1-penalty for small coefficients, but for larger coefficients, SCAD applies a constant penalty in contrast to the LASSO penalty which increases linearly with the coefficient.

Nonparametric regression

Regularization/shrinkage estimation is also common in nonparametric regression; for example, assume the model

$$Y_i = g(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where g is assumed to be smooth.

Assume that g can be approximated by a linear combination of basis functions (e.g. B-splines, wavelets, ...):

$$g(x) \simeq \beta_0 + \sum_{k=1}^{p} \beta_k \phi_k(x)$$

To avoid overfitting, one then adds a penalty term to the LS criterion

$$\sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_k \beta_k \phi_k(x_i) \right)^2 + \rho_\lambda(\boldsymbol{\beta}).$$

LASSO and BRIDGE

For some $\lambda > 0$ and $\gamma > 0$, $\hat{\boldsymbol{\beta}}$ minimizes

$$\sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma$$

I will concentrate on $0 < \gamma \leq 1$:

- $\gamma = 1$ (LASSO)
- $\gamma \downarrow 0$ (Model selection methods, e.g. AIC, BIC).

The objective function is non-convex for $\gamma < 1$, and if λ is sufficiently large exact zero estimates will result.

Computational issues

Problem: How to minimize

$$\sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma$$

efficiently?

$\gamma = 1$. Several algorithms are available:

- quadratic programming algorithms (Tibshirani, 1996).
- primal-dual algorithm (Osborne, Presnell & Turlach, 1998)

$0 < \gamma < 1$. The problem seems to become much more difficult because

- objective function is not differentiable everywhere.
- multiple local minima can exist (because of nonconvexity).

The one variable problem

But... the one variable problem is feasible to solve: Define

$$h(x) = x^2 - 2bx + \lambda|x|^\gamma.$$

Then $\operatorname{argmin}(h) \in [0, b]$. Moreover $\operatorname{argmin}(h) = 0$ iff $\lambda \geq \lambda_{crit}(\gamma, b)$. Othewise, $\hat{x} = \operatorname{argmin}(h)$ satisfies

$$\dot{h}(\hat{x}) = 2\hat{x} - 2b + \lambda\gamma\frac{|\hat{x}|^\gamma}{\hat{x}} = 0$$

which can be solved by Newton-Raphson or fixed-point iteration.

Example

Consider the function

$$h(x) = x^2 - 2bx + |x|^{1/2}.$$

Then $\operatorname{argmin}(h) = 0$ if $b < (27/32)^{1/3} = 0.9449408$.

If $b > (27/32)^{1/3} = 0.9449408$ then $\hat{x} = \operatorname{argmin}(h)$ satisfies

$$2\hat{x} - 2b + \frac{|\hat{x}|^{1/2}}{2\hat{x}} = 0$$

which can be solved via

$$\hat{x}^{(0)} = b$$

$$\hat{x}^{(k)} = b - \frac{|\hat{x}^{(k-1)}|^{1/2}}{4\hat{x}^{(k-1)}}, \quad k = 1, 2, \ldots.$$

Backfitting

Recall we want to minimize $g(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma$
A possible solution is to minimize g iteratively on variable at a time (backfitting). Assume for simplicity that $\bar{Y} = 0$ (or replace Y_i by $Y_i - \bar{Y}$ below).

(0) Initialize: Centre and scale covariates to have mean 0 and variance 1.
Using standardized covariates, define initial $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{ols}$; set $k \leftarrow 1$.

(1) Define:

$$g_k(\boldsymbol{\beta}_k) = \sum_{i=1}^{n} \left(Y_i - \sum_{j \neq k} \hat{\beta}_j x_{ji} - \beta_k x_{ki} \right)^2 + \lambda |\beta_k|^\gamma$$

and set $\hat{\beta}_k = \text{argmin}(g_k)$.

(2) If $k < p$, set $k \leftarrow k + 1$; else set $k \leftarrow 1$.

(3) Repeat (1), (2) until convergence occurs.

Remarks

- This algorithm works very well if the design is not too collinear. Otherwise, it can get stuck in local minima - estimates get send to 0 too quickly and then can't get out.
- Non-convergence can be resolved by either trying multiple starting points or by introducing relaxation factors to send estimates to 0 more slowly.

Asymptotics (fixed p)

Consider asymptotic distributions of estimators $\hat{\boldsymbol{\beta}}_n$ minimizing

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + n\sum_{j=1}^{p} p_\lambda(|\beta_j|).$$

Several basic reasons to consider asymptotics of estimators:

- gives some insight to the properties of the estimators;
- provides a basis for inference;
- suggests approaches to choosing λ.

In order to get non-trivial results, we need to assume that $\lambda \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$.

Design conditions

Assume that

$$C_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \to C$$

with C non-singular.

Moreover, if $\boldsymbol{\beta}_0$ denotes the true value of the parameter, let $\lambda_n \to 0$ as $n \to \infty$ and let

$$a_n = \lambda_n \max\{|\dot{\psi}(|\beta_{0j}|)|; \beta_{0j} \neq 0\} \quad \text{and} \quad b_n = \lambda_n \max\{|\ddot{\psi}(|\beta_{0j}|)|; \beta_{0j} \neq 0\}$$

Then if $b_n \to 0$, then there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} + a_n).$$

It is clear that by choosing λ_n approprietely, there exists a root-n consistent estimator.

Proof

Let $\alpha_n = n^{-1/2} + a_n$. The result will follow if for any $\epsilon > 0$, there exists a large enough constant C_ϵ such that

$$\mathbb{P}\{\inf_{\|\mathbf{u}\|=C_\epsilon} Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) > Q(\boldsymbol{\beta}_0)\} \geq 1 - \epsilon.$$

Let

$$W_n(\mathbf{u}) := Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - Q(\boldsymbol{\beta}_0).$$

Recall that $\psi_\lambda(0) = 0$.

Proof (next)

A Taylor's expansion of ψ gives :

$$\begin{aligned} W_n(\mathbf{u}) \quad &\geq \quad \frac{1}{2} n\alpha_n^2 \mathbf{u}^T C\mathbf{u} - \alpha_n \mathbf{u}^T X^T(\mathbf{Y} - X\boldsymbol{\beta}_0) \\ &\quad + n\lambda_n \sum_{j=1}^{s} \{\psi_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - \psi_{\lambda_n}(|\beta_{j0}|)\}, \end{aligned}$$

where s denotes the number of non-zero components of $\boldsymbol{\beta}_0$.

By the law of large numbers we have that

$$X^T(\mathbf{Y} - X\boldsymbol{\beta}_0) = O_P(\sqrt{n}).$$

Proof (next)

The first term on the right hand side of the above equality is of the order $O_P(n^{1/2}\alpha_n)$ and the second term of the order $O_P(n\alpha_n^2)$. By choosing a sufficiently large C_ϵ the first term dominates the second one, uniformly in $\mathbf{u}$ such that $\|\mathbf{u}\| = C_\epsilon$.

Now the third term is bounded above by

$$\sqrt{s}n\|\mathbf{u}\|a_n\alpha_n + n\alpha_n^2 b_n\|\mathbf{u}\|^2,$$

which is also dominated by the first term of order $O_P(n^{1/2}\alpha_n)$.

By choosing therefore a large enough C_ϵ the result follows •

Oracle Property

Assume that the true vector of coefficients $\boldsymbol{\beta}_0$ is sparse. Without loss of generality write $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ with $\boldsymbol{\beta}_2 = 0$.

Assume that $\sqrt{n}\lambda_n \to +\infty$, then again there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that

$$\hat{\boldsymbol{\beta}}_2 = 0$$

and

$$\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\| = O_P(n^{-1/2} + a_n).$$

Moreover the estimator is asymptotically normal.

Asymptotics when $p \to \infty$

Allowing the dimension to grow as the sample size increases allows a better control of the approximation bias.

(a) $\liminf_{\beta \to 0^+} \dot{\psi}(\beta) > 0$

(b) $a_n = O(n^{-1/2})$

(c) $a_n = o\left((np_n)^{-1/2}\right)$

(d) $b_n = \max_{1 \le j \le p_n} \{|\ddot{\psi}(|\beta_j|)|; \beta_j \neq 0\} \to 0$

(e) $b_n = o_P(p_n^{-1/2})$

(f) There exist C and D such that when x_1 and $x_2 > C\lambda_n$,

$$\lambda_n |\ddot{\psi}(x_1) - \ddot{\psi}(x_2)| \le D|x_1 - x_2|.$$

Under such conditions all results extend to the case with $p_n \to \infty$.

Choosing the hyperparameters

Goal : Choose λ (eventually a also)

- SCAD penalty : $\boldsymbol{\theta} = (\lambda, a)$
- LASSO penalty : $\theta = \lambda$

Five Fold Cross-Validation : Minimize with respect to $\boldsymbol{\theta}$

$$CV(\boldsymbol{\theta}) = \sum_{\nu=1}^{5} \sum_{(\mathbf{x}_k, y_k) \in T^\nu} \{y_k - \mathbf{x}_k^T \hat{\boldsymbol{\beta}}^{(\nu)}(\boldsymbol{\theta})\}^2$$

Generalized Cross-Validation : Minimize with respect to $\boldsymbol{\theta}$

$$GCV(\boldsymbol{\theta}) = \frac{\|\mathbf{Y} - X\boldsymbol{\beta}(\boldsymbol{\theta})\|^2}{n(1 - e(\boldsymbol{\theta})/n)^2}$$

where $e(\boldsymbol{\theta}) = \text{Trace}(X(X^T X + nV(\boldsymbol{\beta}(\boldsymbol{\theta}))^{-1} X^T)$

Penalized Model-Based Clustering

Variable selection in clustering analysis, especially for "high dimension, low sample size" data, is both challenging and important.

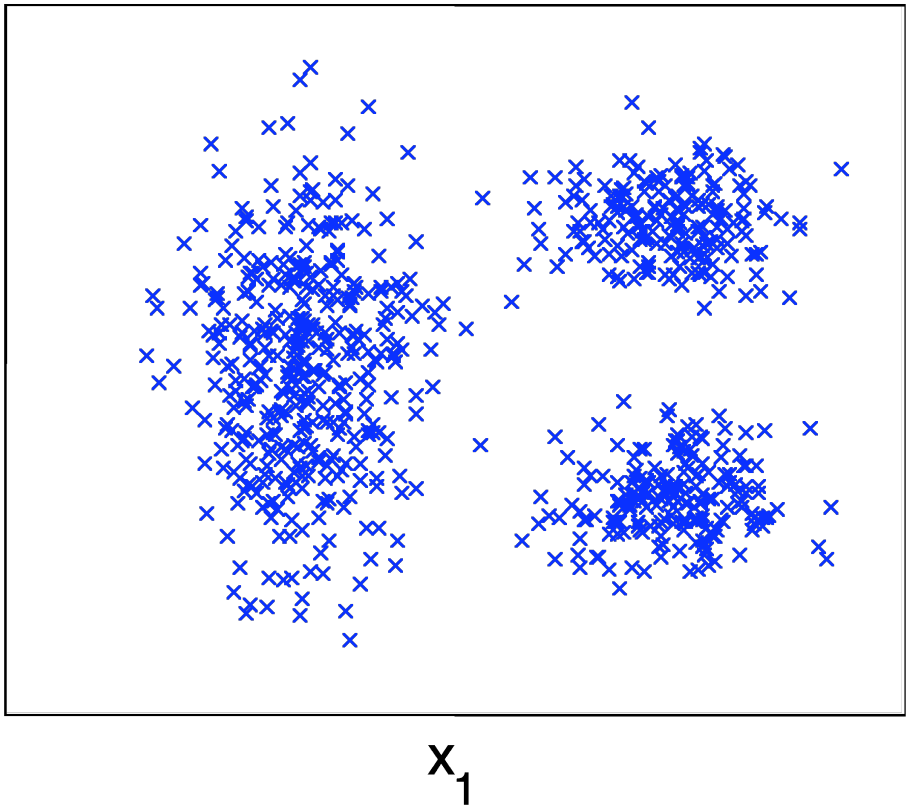
- Clustering applications with large number of features: Text categorization, genomic microarray analysis
- Noisy features can lead to misleading clusters
- There is no clear-cut criterion function for feature selection in unsupervised learning

Setup

Specifically, given n p-dimensional observations $\mathbf{x}_j = (x_{j1}, \ldots, x_{jp})^T$ for $j = 1, \ldots, n$ we aim to group the data into a few, say K , clusters such that the observations in the same cluster are more similar to each other than those from different clusters.

In this context, some of the attributes x_{jk}'s of $\mathbf{x}_j$ may not be relevant: use of such attributes only introduces noise, and may impede uncovering the clustering structure of interest. In addition, removing non-informative attributes may largely enhance interpretability.

Example



Optimal feature subset is inter-related with the number of clusters. The optimal feature subset is $\{x_1, x_2\}$, $\{x_2\}$, $\{x_1\}$ if we assume there are 3, 2 and 1 cluster(s), respectively.

Model-based clustering

Model-based clustering (McLachlan and Peel, 2002; Fraley and Raftery, 2002) assumes that data come from a finite mixture model with each component corresponding to a cluster.

Each observation $\mathbf{x}$ is drawn from a finite mixture distribution

$$f(x, \Theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}, \boldsymbol{\theta}_k),$$

with the mixing proportion π_k, component specific distribution f_k and its parameters $\boldsymbol{\theta}_k$.

Denote by $\Theta = \{(\pi_k, \boldsymbol{\theta}_k), k = 1, \ldots, K\}$ all unknown parameters, with restriction that $0 \leq \pi_k \leq 1$ and $\sum \pi_k = 1$.

Each component of the mixture distribution corresponds to a cluster. The number of clusters, K, has to be determined in practice. In the sequel, we focus on a mixture of Gaussians for clustering.

The mixture density

We assume that each observation $\mathbf{x}_j$, $j = 1, \ldots, n$, is drawn from a finite Gaussian mixture

$$f(\mathbf{x}_j) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_j; \boldsymbol{\mu}_k, \Sigma_k),$$

where $\boldsymbol{\mu}_k$ is the mean vector of the Gaussian distribution characterizing the kth cluster and Σ_k is the corresponding covariance matrix.

We will assume that features are conditionally independent given the component label, i.e. that each Σ_k is a diagonal matrix, and that the Σ_k's are the same across different clusters. A theoretical justification of such an assumption can be found in Bickel and Levina (2004).

Advantages of this approach is that there is no need to specify the number of components, K.

Assignment

Given an observation $\mathbf{x}^*$ one computes the probability that $\mathbf{x}^*$ is from the kth cluster

$$p_k = \frac{\pi_k}{\prod_{j=1}^{p}(2\pi\sigma_j^2)^{1/2}} \exp\left(-\sum_{j=1}^{p}\frac{(x_j^* - \mu_{kj})^2}{2\sigma_j^2}\right), \quad k = 1, \ldots, K,$$

and $\mathbf{x}^*$ will be assigned to the cluster with the largets p_k.

Given the data $\mathbf{x}_j$, $j = 1, \ldots, n$, the log-likelihood function is

$$\ell_0(\Theta) = \sum_{j=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_j; \boldsymbol{\mu}_k, \Sigma)\right).$$

Maximization of the above log-likelihood with respect to Θ is difficult, and it is common to use the EM algorithm (Dempster et al., 1977) by casting the problem in the framework of missing data.

Penalized EM

Define z_{kj} as the indicator of whether $\mathbf{x}_j$ is from component k; If the missing data z_{kj}'s could be observed, then the log-likelihood for the complete data is:

$$\ell(\Theta) = \sum_{j=1}^{n} \sum_{k=1}^{K} z_{jk}(\log \pi_k + \log f_k(\mathbf{x}_j; \boldsymbol{\mu}_k, \Sigma))$$

With the same motivation as in penalized regression, we propose a penalized model-based clustering approach resulting in automatic variable selection.

Specifically, we regularize $\ell(\Theta)$ to yield a penalized log-likelihood:

$$\ell_\psi(\Theta) = \ell(\Theta) + \sum_{k=1}^{K} \sum_{i=1}^{p} \psi_\lambda(|\mu_{ik}|)$$

where ψ_λ is a penalty function with penalization parameter λ.

The indicator variables z_{ik} are not observed and an EM algorithm for the penalized model-based clustering can be derived closely following from that for standard model-based clustering (McLachlan and Peel, 2002) and the general methodology for penalized likelihood (Green, 1990).

The only difference exists in estimating the means μ_{jk}'s in the M-step.

In practice, we need to determine the number of components, K. This is realized by first fitting a series of models with various numbers of components, and then using a model selection criterion to choose the best one. For standard model-based clustering, it is common to use Bayesian information criterion (BIC) (Schwarz, 1978).

Choosing K and λ

For penalized model-based clustering, in addition to K , we also have to choose an appropriate value of penalization parameter λ;.

One difficulty in using the BIC criterion is that it is not always clear what is the dimension of the parameter space in a penalized model.

Following a conjecture of Efron et al. (2004) and a result of Zou et al. (2004) for L_1-penalized regression, we treat this dimension as the number of non-zero parameter estimates, modifying BIC for penalized model-based clustering.

References

Antoniadis, A. (1997) Wavelets in Statistics: A Review (with discussion), *Journal of the Italian Statistical Association*, 6, 97-144.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with dis- cussion). *Journal of American Statistical Association*, 96, 939967.

P.J. Bickel, and E. Levina. (2004) Some theory for Fishers linear discriminant function, naive Bayes, and some alternatives when there are many more variables than observations. *Bernoulli*, 10:989-1010.

Efron B, Hastie T, Johnstone I, Tibshirani R. (2004). Least angle regression. Annals of Statistics 32, 407-499.

Fan, J. (1997) Comment on Wavelets in Statistics: A Review by A. Antoniadis. Journal of the Italian Statistical Association, 6, 131-138.

Fan, J., and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *JASA*, 96. 1348-1360.

McLachlan, G.J. and Basford, K.E. (1988). Mixture Models: Inference and Applications to Clustering. New York: Marcel Dekker.

Tibshirani (1996). Regression shrinkage and selection via the Lasso. JRSS-B.

Zou H, Hastie T, Tibshirani R. (2004). On the Degrees of Freedom of the Lasso. Technical report, Statistics dept, Stanford University. http://stat.stanford.edu/ hastie/pub.htm.