Adaptive Model Selection for Regression, Classification and Density Estimation

By SATOSHI MIYATA
Genome Center, Japanese Foundation for Cancer Research
satoshi.miyata@jfcr.or.jp

1. Introduction

Suppose that $Y_1, Y_2, \ldots, Y_n$ are independently distributed according to the probability distributions $f_1, f_2, \ldots, f_n$. The distributions of Y's are estimated by various models in statistics, including the regression, the classification and the density estimation. The performance of the statistical model estimation is determined by the tuning parameters of the estimators and to capture the local property of the data structure, a fine tuning of the tuning parameters is necessary.

As the measurement of the fitness of the model, the adaptive model selection criterion (AMSC) (Shen and Ye (2002) and Shen, et al. (2004)) is adopted in this paper. The AMSC was originally introduced as the model selection criterion for the exponential family, and it will be shown that it is the best estimator of the Kullback-Leibler (KL) loss for the statistical modelings, which minimizes the L_2 distance between the KL loss for the model and its loss estimator. In this article, the underlying distribution f's of the statistical models are basically supposed to be a member of the exponential family.

The minimization problem of the above model selection criteria with respect to the tuning parameters is usually highly complex problem. In this article, we utilize the stochastic optimization procedure to minize the model selectors. This optimization procedure is a version of the Evolutionary Algorithm and its global convergence property have been proved in Miyata and Shen (2003).

2. Adaptive Model Selection Criterion for the Exponential Family

Suppose that $\{(\mathbf{Y}_i, \mathbf{x}_i)\}_{i=1}^n$ are sampled, where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ik})$ is a response and $\mathbf{x}_i$ is a p-dimesional vector of covariates, and the components of $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)'$ are independently distributed according to an exponential family:

$$p(\mathbf{y}_i|\boldsymbol{\mu}_i) = \exp\{\varphi(\boldsymbol{\mu}_i)'\mathbf{y}_i + \alpha(\boldsymbol{\mu}_i) + m(\mathbf{y}_i)\}\ ,\ \boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ik})',\ \varphi(\boldsymbol{\mu}_i) = (\varphi(\mu_{i1}), \ldots, \varphi(\mu_{ik}))'$$

with the mean $\boldsymbol{\mu}_i = E(\mathbf{Y}_i)$ and the variance $\sigma_i^2(\boldsymbol{\mu}_i) = \mathrm{var}(\mathbf{Y}_i)$. Here, φ, α, m may depend on a dispersion parameter ψ, and suppose ψ may and may not be known. The espected value of $\mathbf{Y}$, $\boldsymbol{\mu}$, is estimated by $\hat{\boldsymbol{\mu}}$.

In this section, we introduce the adaptive model selection criterion (AMSC) for the exponential family according to Shen, et al. (2004). The performance of $\hat{\boldsymbol{\mu}}_i$ is evaluated by the individual Kullback-Leibler loss of $\boldsymbol{\mu}_i$ with respect to $\hat{\boldsymbol{\mu}}_i$: $\int p(\mathbf{y}_i|\boldsymbol{\mu}_i)\log(p(\mathbf{y}_i|\boldsymbol{\mu}_i)/p(\mathbf{y}_i|\hat{\boldsymbol{\mu}}_i)) = ((\varphi(\boldsymbol{\mu}_i)'\boldsymbol{\mu}_i + \alpha(\boldsymbol{\mu}_i)) - (\varphi(\hat{\boldsymbol{\mu}}_i)'\boldsymbol{\mu}_i + \alpha(\hat{\boldsymbol{\mu}}_i)))$. Since the first term of the right side is constant, it will be omitted and the comparative Kullback-Leibler loss of $\boldsymbol{\mu}$ with respect to $\hat{\boldsymbol{\mu}}$ is defined as follows:

$$K(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = -n^{-1}\sum_{i=1}^{n}(\varphi(\hat{\boldsymbol{\mu}}_i)'\boldsymbol{\mu}_i + \alpha(\hat{\boldsymbol{\mu}}_i) + m(\mathbf{y}_i)) = n^{-1}\sum_{i=1}^{n}(\log p(\mathbf{y}_i|\hat{\boldsymbol{\mu}}_i) + \varphi(\hat{\boldsymbol{\mu}}_i)'(\mathbf{y}_i - \boldsymbol{\mu}_i))$$

If $\boldsymbol{\mu}$ is known, the goodness of fit of various models can be compared by $K(\cdot,\cdot)$. Since $\boldsymbol{\mu}$ is unknown, $K(\cdot,\cdot)$ must be estimated by the given data. To estimate $K(\cdot,\cdot)$, we consider a class of loss estimators of the form $-\sum_{i=1}^n \log f(\mathbf{y}_i|\hat{\boldsymbol{\eta}}_i) + \kappa$, which is characterized by the penalty parameter κ. For the estimation of $K(\cdot,\cdot)$, the optimal $\kappa = D(\mathcal{M})$ is selected so that the L_2 distance between $K(\cdot,\cdot)$ and the model selector is minimized.

$$D(\mathcal{M}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\mathrm{Cov}\left(\varphi(\hat{\mu}_{ij}), y_{ij}\right) = \sum_{i=1}^{n}\sum_{j=1}^{k}\sigma_{ij}^2(\mu_{ij})\frac{\partial}{\partial\mu_{ij}}E\left(\varphi(\hat{\mu}_{ij})\right)$$

The right hand side of (**??**) is derived based on Shen et al. (2004). Then the optimal estimator of $K(\cdot,\cdot)$ is obtained as $-\log f(\mathbf{Y}|\hat{\boldsymbol{\eta}}) + D(\mathcal{M})$. The quantity $D(\mathcal{M})$ coincides with the one called the GDF (Generalized Degree of Freedom) in Shen, et al. (2004). $D(\mathcal{M})$ still contains unknown $\boldsymbol{\mu}$ and is estimated by $\hat{D}(\mathcal{M})$ and approximated by Monte Carlo simulation as in Shen, et al. (2004).

The minimization problems in the Algorithm 1 are complex optimization problems. In this article, the stochastic optimization procedure which is a version of the one used for knot selection of regression spline in Miyata and Shen (2003, 2005) will be adopted for the optimization.

3. AMSC for Nonparametric Regression

Consider the regression problem $Y_i = f(\mathbf{x}_i) + \epsilon_i,\ \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),\ i = 1, \ldots, n$. The unknown regression function f would be estimated by various nonparametric regression models $\hat{f}$, including the regression spline, the kernel smoothers, and so on. The regression models are controled by its tuning parameters, including the knot sequence of the regression spline and the bandwidth of the kernel smoothers. For the normal distribution, the AMSC for the nonparametric regression is obtained as follows:

$$AMSC = n^{-1}\left(-\sum_{i=1}^{n} \log p(y_i|\hat{\mu}_i) + \sum_{i=1}^{n} \partial E(\hat{\mu}_i)/\partial \mu_i\right).$$

In this article, we adopt the adaptive free-knot spline to estimate f. The knot sequence of the spline is optimized so that (**??**) is minimized.

4. AMSC for Classification

Now we consider the application of the classification problem. Let $G \in \Gamma$ be the label of the class and suppse that the log odds of the posterior probabilities of the K classes are models by the covariates. K-class $g_j, j = 1, \ldots, K$ are coded via the multinomial random variable $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i(K-1)})$ where

$$P(G = k|X = \mathbf{x}) = \frac{\exp\{\boldsymbol{\beta}_k'\mathbf{x}\}}{1 + \exp\{\sum_{l=1}^{K-1} \boldsymbol{\beta}_l'\mathbf{x}\}},\ \mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i(K-1)}), Y_{ik} = \begin{cases} 1 : G = k \\ 0 : G \neq k \end{cases}$$

Suppose that the conditional expectations $E(Y_{ij}|\mathbf{x})$ is related to the predictor $\mathbf{x} = (x_1, \ldots, x_p)$ by the logit link, $\eta_{ij} = \log(\pi_{ij}/\pi_{i0}) = \sum_{r=1}^{p} x_{ijr}\beta_r,\ j = 1, \ldots, k,\ \pi_{i0} = 1 - \sum_{j=1}^{k} \pi_{ij}$. Then the GLM with the multinomial distribution is fit. The AMSC for the logistic linear classification is obtained as follows:

$$AMSC = n^{-1}\left(-\sum_{i=1}^{n} \log p(\mathbf{y}_i|\hat{\boldsymbol{\mu}}_i) + E\sum_{i=1}^{n}\sum_{j=1}^{k}\left((y_{ij} - \mu_{ij})\sum_{r=1}^{p} x_{ijr}\beta_r\right)\right) = n^{-1}\left(-\sum_{i=1}^{n} \log p(\mathbf{y}_i|\hat{\boldsymbol{\mu}}_i) + \sum_{r=1}^{p} \frac{\partial}{\partial \beta_r} E\left(\hat{\beta}_r\right)\right).$$

5. AMSC for Probability Density Estimation

In this section, we apply the AMSC for the density estimation problem. Suppose that $Y_1, Y_2, \ldots, Y_n \overset{\text{iid}}{\sim} f,\ f : [0, 1] \to \mathbf{R}_+$. In this section, the histogram and the logspline estimators are investigated.

Histogram density estimation

Given mesh $\mathbf{b} = (b_0, \ldots, b_K)$ for $0 \leq b_0 < \cdots < b_{K+1} \leq 1$ where $(K + 1)$ is the number of the bins and $B_j = [b_j, b_{j+1}),\ j = 0, \ldots, K$, and $\pi_j = \int_{B_j} f$, the histogram and the AMSC for the histogram are defined.

$$\hat{f}(y, \mathbf{b}) = \frac{1}{n}\sum_{j=0}^{K} \frac{n_j}{(b_{j+1} - b_j)} I_{[b_j, b_{j+1})}(y) = \frac{n_j}{n(b_{j+1} - b_j)}, \quad y \in B_j,\ j = 0, \ldots, K.$$

$$AMSC = n \log n - \sum_{j=0}^{K} n_j \log(n_j(b_{j+1} - b_j)) + \sum_{j=1}^{K} \mathrm{E}(\hat{\eta}_j(Y_j - \mu_j)).$$

Logspline density estimation

Let $s(x)$ be a spline function of order m $(m \geq 1)$ with a knot sequence $\mathbf{t} = (t_1, \ldots, t_k)$ for $-\infty < t_0 < t_1 \leq \cdots \leq t_k < t_{k+1} < \infty$, including k_j $(k_l \leq m)$ repeated knots at each location t_j. A spline function $s(x)$ can be represented by the normalized B-spline basis $\{B_l(x; t), l = 1, \cdots, m + k\}$, $s(x) = \sum_{l=1}^{k+m} \theta_l B_l(x; \mathbf{t})$, where $B_l(x; \mathbf{t}) = (t_l - t_{l-m})[t_{l-m}, \cdots, t_l](\cdot - x)_+^{m-1}$ and $[t_{l-m}, \cdots, t_l](\cdot)$ is the m-th order divided difference of $(\cdot)$. Then the logsplie density estimator and its AMSC are defined as follows:

$$f(y; \boldsymbol{\theta}) = \exp\left(\sum_{j=1}^{k} \theta_j B_j(y) - c(\boldsymbol{\theta})\right) = \exp\left(s(y) - c(\boldsymbol{\theta})\right), = c(\boldsymbol{\theta}) = \log\left(\int \exp\left(\sum_{j=1}^{k} \theta_j B_j(y)\right) dy\right),\ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k).$$

$$AMSC = -\sum_{i=1}^{n}\left(\sum_{j=1}^{k} \hat{\theta}_j B_j(y_i) - c(\boldsymbol{\theta})\right) + \sum_{j=1}^{k} E(\hat{\theta}_j(B_j(Y) - E(B_j(Y)))).$$

The performance of the proposed models was demonstrated by the simulation study.