Sampling-based Gaussian Mixture Regression for Big Data

JooChul Lee, Elizabeth D. Schifano, and Haiying Wang
Department of Statistics, University of Connecticut

Introduction

• This paper proposes a nonuniform subsampling method for finite mixtures of regression models to reduce large data computational tasks.
A general estimator based on a subsample is investigated, and its asymptotic normality is established. We assign optimal subsampling probabilities to data points that minimize the asymptotic mean squared errors of the general estimator and linearly transformed estimators.
Since the proposed probabilities depend on unknown parameters, an implementable algorithm is developed.
We first approximate the optimal subsampling probabilities using a pilot sample. After that, we select a subsample using the approximated subsampling probabilities and compute estimates using the subsample. We present a real data example using appliance energy data.

Mixture of Gaussian Regressions

We review a finite mixture of Gaussian linear regressions. Suppose that y is a response and $\mathbf{x}$ is a d dimensional covariate with the first entry being one. The conditional density function of y given $\mathbf{x}$ is

$$f(y|\mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^{J} p_j f_j(y|\mathbf{x};\boldsymbol{\beta}_j,\sigma_j), \qquad (1)$$

where J is a given number of components, p_j's are the component weights satisfying $p_j > 0$ for each j and $\sum_{j=1}^{J} p_j = 1$, $f_j(y|\mathbf{x};\boldsymbol{\beta}_j,\sigma_j)$ is the density of a normal distribution with mean $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j$ and variance σ_j^2, $\boldsymbol{\beta}_j$ is a $d\times 1$ vector of unknown regression coefficients including an intercept, and $\boldsymbol{\theta} = (\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_J,\sigma_1,...,\sigma_J,p_1,...,p_{J-1})$. The maximum likelihood estimator (MLE), $\hat{\boldsymbol{\theta}}$, for the unknown parameter $\boldsymbol{\theta}$ is the maximizer of the following log-likelihood,

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{J} p_j f_j(y_i|\mathbf{x}_i;\boldsymbol{\beta}_j,\sigma_j)\right). \qquad (2)$$

Estimation and Optimal Subsampling Strategy

Denote the full data as $\mathcal{D}_n = \{(\mathbf{x}_i, y_i) : i = 1,...,n\}$. Let $\{\pi_i\}_{i=1}^n$ be the subsampling probabilities assigned to all observations satisfying $\sum_{i=1}^n \pi_i = 1$.
Consider a random subsample of size r selected from the full data $\mathcal{D}_n$ based on the subsampling probabilities.
Then, the subsampling estimator $\tilde{\boldsymbol{\theta}}$ can be obtained by maximizing the target function $\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{r} \frac{1}{\pi_i^*} \log\left(\sum_{j=1}^{J} p_j f_j(y_i^*|\mathbf{x}_i^*;\boldsymbol{\beta}_j,\sigma_j)\right)$, where $\mathbf{x}_i^*$'s, y_i^*'s and π_i^*'s, are covariates, responses, and subsampling probabilities in the subsample, respectively. The EM algorithm can be applied to optimize the target function. The details of the algorithm are presented in Algorithm 1.

Algorithm 1 EM Algorithm for the target function

Estimates can be obtained by maximizing the sampled complete-data target function, $\ell_c^*(\boldsymbol{\theta}) = \sum_{i=1}^{r} \frac{1}{\pi_i^*} \sum_{j=1}^{J} z_{ij}^* \log\left\{p_j f_j(y_i^*|x_i^*)\right\}$,
where z_{ij}^* is equal to one if y_i^* belongs to the jth component and zero otherwise.
E-step : Given the current estimate $\boldsymbol{\theta}^{(s)}$,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \sum_{i=1}^{r} \frac{1}{\pi_i^*} \sum_{j=1}^{J} \tau_{ij}^{*(s)} \log p_j f_j(y_i^*|\mathbf{x}_i^*;\boldsymbol{\beta}_j,\sigma_j), \text{ where}$$

$$\tau_{ij}^{*(s)} = p_j^{(s)} f_j(y_i^*|\mathbf{x}_i^*;\boldsymbol{\beta}_j^{(s)},\sigma_j^{(s)}) / \sum_{k=1}^{J} p_k^{(s)} f_k(y_i^*|\mathbf{x}_i^*;\boldsymbol{\beta}_k^{(s)},\sigma_k^{(s)}).$$

M-step : Updates the estimate $\boldsymbol{\theta}^{(s+1)}$. For $j = 1,...,J$,

$$\hat{p}_j^{(s+1)} = \left(\sum_{i=1}^{r} \frac{1}{\pi_i^*}\right)^{-1} \sum_{i=1}^{n} \frac{\tau_{ij}^{*(s)}}{\pi_i^*},$$

$$\hat{\boldsymbol{\beta}}_j^{(s+1)} = \left(\sum_{i=1}^{r} \frac{\tau_{ij}^{*(s)} \mathbf{x}_i^* \mathbf{x}_i^{*\mathrm{T}}}{\pi_i^*}\right)^{-1} \sum_{i=1}^{n} \frac{\tau_{ij}^{*(s)} y_i^* \mathbf{x}_i^*}{\pi_i^*},$$

$$\hat{\sigma}_j^{2(s+1)} = \left(\sum_{i=1}^{r} \frac{\tau_{ij}^{*(s)}}{\pi_i^*}\right)^{-1} \sum_{i=1}^{n} \frac{\tau_{ij}^{*(s)} (y_i^* - \mathbf{x}_i^{*\mathrm{T}} \hat{\boldsymbol{\beta}}_j^{(s)})^2}{\pi_i^*}.$$

Repeat until convergence.

Asymptotic Result

Theorem 1 *Under some assumptions, as $r, n \to \infty$, if $r/n = o(1)$,*

$$\sqrt{r}\mathbf{V}^{-1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_t) \longrightarrow N(\mathbf{0}, \mathbf{I}) \quad \textit{in distribution},$$

where $\mathbf{V} = \mathbf{M}_t^{-1}\mathbf{V}_\pi\mathbf{M}_t^{-1}$, $\mathbf{M}_t = \frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}$, *and* $\mathbf{V}_\pi = \sum_{i=1}^{n} \frac{\dot{\ell}_i(\boldsymbol{\theta}_t)\dot{\ell}_i(\boldsymbol{\theta}_t)^{\mathrm{T}}}{n^2\pi_i}$.

A-optimal subsampling probability

We specify the optimal subsampling probabilities based on the result in Theorem 1.

Theorem 2 *The optimal subsampling probabilities that minimize $tr(\mathbf{V})$ are*

$$\pi_i^{\mathbf{V}} = \frac{\left\|\mathbf{M}_t^{-1}\dot{\ell}_i(\boldsymbol{\theta}_t)\right\|}{\sum_{k=1}^{n}\left\|\mathbf{M}_t^{-1}\dot{\ell}_i(\boldsymbol{\theta}_t)\right\|}, \quad i = 1,...,n. \qquad (3)$$

L-optimal subsampling probability

We assign the optimal subsampling probabilities by minimizing the trace of $\mathbf{V}_\pi$ which is equivalent to minimizing the asymptotic MSE of $\mathbf{M}\tilde{\boldsymbol{\theta}}$. In addition to that, we also focus on the asymptotic MSE of the coefficient estimator $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1,...,\tilde{\boldsymbol{\beta}}_J)$. Denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_J)$, $\boldsymbol{\theta}_{-\boldsymbol{\beta}} = (\sigma_1,...,\sigma_J,p_1,...,p_{J-1})$, $\mathbf{M}_{t,11} = \frac{1}{n}\frac{\partial^2\ell(\boldsymbol{\theta}_t)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}$, $\mathbf{M}_{t,12} = \frac{1}{n}\frac{\partial^2\ell(\boldsymbol{\theta}_t)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}_{-\boldsymbol{\beta}}^{\mathrm{T}}}$, and $\mathbf{M}_{t,22} = \frac{1}{n}\frac{\partial^2\ell(\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_{-\boldsymbol{\beta}}\partial\boldsymbol{\theta}_{-\boldsymbol{\beta}}^{\mathrm{T}}}$.

Theorem 3 *The optimal subsampling probabilities that minimize $tr(\mathbf{V}_\pi)$ are*
$\pi_i^{\mathbf{V}_\pi} = \left\|\dot{\ell}_i(\boldsymbol{\theta}_t)\right\| / \sum_{k=1}^{n}\left\|\dot{\ell}_k(\boldsymbol{\theta}_t)\right\|$, $i = 1,...,n$, *and the optimal subsampling probabilities that minimize the asymptotic MSE of $\tilde{\boldsymbol{\beta}}$ are*

$$\pi_i^{\mathbf{V}_\beta} = \frac{\left\|(\mathbf{M}_{\boldsymbol{\beta}}^{inv}, \mathbf{M}_{\boldsymbol{\theta}_{-\boldsymbol{\beta}}}^{inv})\dot{\ell}_i(\boldsymbol{\theta}_t)\right\|}{\sum_{k=1}^{n}\left\|(\mathbf{M}_{\boldsymbol{\beta}}^{inv}, \mathbf{M}_{\boldsymbol{\theta}_{-\boldsymbol{\beta}}}^{inv})\dot{\ell}_k(\boldsymbol{\theta}_t)\right\|}, \quad i = 1,...,n,$$

where $\mathbf{M}_{\boldsymbol{\beta}}^{inv} = (\mathbf{M}_{t,11} - \mathbf{M}_{t,12}\mathbf{M}_{t,22}^{-1}\mathbf{M}_{t,12}^{\mathrm{T}})^{-1}$ *and* $\mathbf{M}_{\boldsymbol{\theta}_{-\boldsymbol{\beta}}}^{inv} = -\mathbf{M}_{\boldsymbol{\beta}}^{inv}\mathbf{M}_{t,12}\mathbf{M}_{t,22}^{-1}$,

Real Data Analysis

• Appliance energy data[a]
• Appliances energy consumption (response), and three humidities in different areas (covariates: kitchen area (H-Kit), living room area (H-Liv), and laundry area (H-Lau)).
• Full data size is $n = 19,735$

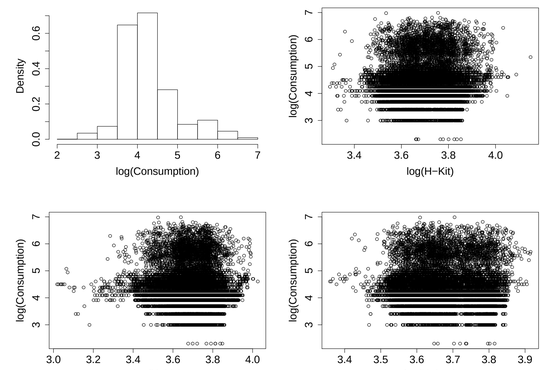

Figure 1: Histogram of log-transformed appliances energy consumption (top-left) and scatter plots between appliances energy consumption and humidity at different areas (top-right, bottom-left, bottom-right).

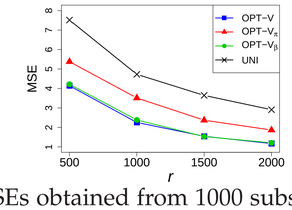

Figure 2: MSEs obtained from 1000 subsamples. OPT-$\mathbf{V}$, OPT-$\mathbf{V}_\pi$, and OPT-$\mathbf{V}_\beta$ use $\pi_i^{\mathbf{V}}$, $\pi_i^{\mathbf{V}_\pi}$, and $\pi_i^{\mathbf{V}_\beta}$, respectively. UNI uses uniform subsampling probabilities.

[a]The data is available at the UCI Machine Learning repository https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction

Reference

[1] HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018.