Nonparametric Convex Imputation and Its Application

Jian-Hui Ning
Department of Mathematics and Statistics, Central China Normal University, CHINA –jhning@mail.ccnu.edu.cn

Michelle Liou
Institute of Statistical Science, Academia Sinica, TAIWAN – mliou@stat.sinica.edu.tw

Philip E. Cheng*
Institute of Statistical Science, Academia Sinica, TAIWAN – pcheng@stat.sinica.edu.tw

Abstract

Methods of regression function imputation with incomplete data have exhibited various applications. Nearest neighbor regression and kernel regression have been used for imputing missing data in survey sampling and missing treatment assignments since 1970s. Asymptotic normality for estimating the mean of missing data was given for kernel imputation in 1986, and for k-nearest neighbor imputation in 2012. In this study, a novel convex mixture of these two regression imputation methods is constructed for extracting the advantage of both methods so as to offset irregular distribution conditions in the data. A naive mixture and its modified version are found to yield improved and stable performance for estimating the mean of an incomplete variable and for predicting discrete types with or without supervised learning under general distribution conditions. A simulation study of general missing data and an empirical study of two data sets in the UCI Machine Learning Repository are examined for useful applications.

Keywords: classification, convex mixture, kernel imputation, *k*-nearest neighbor imputation.

1. Introduction

The topic of nonparametric estimation for the mean of an incomplete response variable has also been studied in the literature under the MAR condition. Without assuming parametric models for the regression function or the missing data pattern, nonparametric regression approaches to estimating the mean of a response variable have been examined since 1980s. Under the MAR assumption and a few regularity conditions on the joint distribution of the variables, asymptotic normality for the kernel regression (KR) imputation was initially examined by Cheng and Wei (1986). Suppose that a random sample with incomplete responses and fully observed covariates arise from a double sampling design, denoted by

$$(X_i, Y_i, \delta_i),\ i = 1, \cdots, n. \tag{1.1}$$

All the covariates X_i are observed, and $\delta_i = 1$ if Y_i is observed, $\delta_i = 0$ otherwise. The parameter of interest is the mean of Y, $\mu = EY$, which is estimated under MAR, that is, missing Y depends solely on the covariate X

$$P(\delta = 1 | X, Y) = P(\delta = 1 | X) \equiv p(X). \tag{1.2}$$

Approximating the same asymptotic normality, two asymptotically equivalent KR imputation estimators for μ (Cheng and Wei, 1986; Cheng, 1994) were defined as

$$\tilde{\mu}_{KR} = \frac{1}{n}\sum_{i=1}^{n}\hat{m}_{KR}(X_i), \tag{1.3}$$

and $$\hat{\mu}_{KR} = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i Y_i + (1-\delta_i)\hat{m}_{KR}(X_i)\}; \tag{1.4}$$

where $$\hat{m}_{KR}(X_i) = \sum_{j=1}^{n} W_h(X_i, X_j)\delta_j Y_j \Big/ \sum_{j=1}^{n} W_h(X_i, X_j)\delta_j, \tag{1.5}$$

$W_h(u,x) = h^{-1}W((u-x)/h)$, and W is a symmetric probability density function.

The method of *k*-NN imputation was used to predict the Iris species, it could be trained to be a close alternative to the method of support vector machine (Ning and Cheng, 2012). For a finite positive integer *k*, the *k*-NN imputation estimator for the mean is defined to be

$$\hat{\mu}_{kNN} = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i Y_i + (1-\delta_i)\hat{m}_{kNN}(X_i)\}. \tag{1.6}$$

Here, the kernel imputation estimates $\hat{m}_{KR}(X_i)$ of (1.5) are replaced by the nearest neighbor estimates $\hat{m}_{kNN}(X_i) = \sum_{j=1}^{k} Y_{i(j)}/k$, using the k nearest complete paired-data units $\{(X_{i(j)}, Y_{i(j)}) : \delta_{i(j)} = 1, j = 1, \cdots, k\}$, where $X_{i(j)}$ denotes the jth nearest neighbor to X_i among the fully observed pairs. Thus, the fixed kernel bandwidth h is replaced by a random distance from X_i to its kth nearest neighbor $X_{i(k)}$ with $\delta_{i(k)} = 1$, where the Euclidean or the Mahalanobis distance is commonly used.

Another notable nonparametric method for estimating the mean is the classical Horvitz and Thompson inverse-weighting estimation. Under MAR, the basic form is

$$\hat{\mu}_{HT} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Y_i}{w_i}, \tag{1.7}$$

where for $i = 1, \ldots, n$, $w_i = \hat{p}(X_i) = \sum_{j=1}^{n}\delta_j W_h(X_j, X_i) \Big/ \sum_{j=1}^{n} W_h(X_j, X_i)$ are locally weighted kernel estimates of the missing pattern function $p(x)$, which is an analog of the kernel regression estimator (1.5). Like estimator (1.7), an IPW type regression imputation estimator for the mean can be derived from the KR imputation estimator:

$$\hat{\mu}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_{KR}(X_i) + \frac{\delta_i\{Y_i - \hat{m}_{KR}(X_i)\}}{w_i}\right], \tag{1.8}$$

where $\hat{m}_{KR}$ and w_i are given by (1.5) and (1.7), respectively (Ning and Cheng, 2012). **Lemma 1.** The estimators $\hat{\mu}_{KR}$, $\hat{\mu}_{HT}$ and $\hat{\mu}_{IPW}$ asymptotically approximate the same normal distribution $N(\mu,\ \sigma_{KR}^2)$ under a set of regularity conditions. The common asymptotic variance is

$$\sigma_{HT}^2 = \sigma_{IPW}^2 = \sigma_{KR}^2 = Var(Y) + E\left[\frac{\{1-P(X)\}\sigma^2(X)}{p(X)}\right]. \tag{1.9}$$

2. Convex mixture regression imputation

In this section, a new imputation method using a convex mixture of the *k*-NN (**1**.6) and the KR (1.4) imputation estimator for the mean is introduced. This is defined as

$$\hat{\mu}_{CM} = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i Y_i + (1-\delta_i)\hat{m}_{CM}(X_i)\}, \tag{2.1}$$

where
$$\hat{m}_{CM}(X_i) = w_i \cdot \hat{m}_{KR}(X_i) + \{1-w_i\}\cdot \hat{m}_{kNN}(X_i). \tag{2.2}$$

Theorem 2.1 Under regularity conditions on missing pattern and regression function. The convex imputation estimator (2.1) approximates the normal distribution

$$\sqrt{n}(\hat{\mu}_{CM} - \mu) \to N(0,\ \sigma_{CM}^2), \quad \text{where}$$

$$\sigma_{CM}^2 = Var[m(X)] + E\left[\frac{\sigma^2(X)}{p(X)}\right] + E\left[\sigma^2(X)\{1-p(X)\}^3\left\{2p(X)-1+\frac{1}{k}\right\}\right], \tag{2.3}$$

and the sum of the first two summands in (2.3) is equal to σ_{KR}^2 of (1.9).

3. Conclusion

For estimating the mean, the CM estimator and a variant version are found to yield better and more stable performance than the existing ones under irregular conditions. In case of prediction for the Iris data or the wine data, the CM imputation also yields satisfactory performance. We remark that the same idea of convex mixture imputation, combining a kernel smoothing estimator and a *k*-NN estimator, can also be applied to estimating other population parameters. For example, nonparametric estimation of a regression function is a typical case. It is anticipated that this study will induce further research toward improvement in various applications of nonparametric regression.

References

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random, J. Amer. Stat. Assoc., **89**, 81-87.

Cheng, P.E. and Wei, L.J. (1986). Nonparametric inference under ignorable missing data process and treatment assignment, International Statistical Symposium, Taipei, **1**, 97-112.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems, **47**, 547-553.

Ning, J.H. and Cheng, P.E. (2012). A comparison study of nonparametric imputation methods, Statistics & Computing, **22**, 273-285.