Multi-output Nonparametric Regression

José M. Matías

Dpto. de Estadística, Universidad de Vigo, 36310 Vigo, Spain
jmmatias@uvigo.es

Abstract. Several non-parametric regression methods with various dependent variables that are possibly related are explored. The techniques which produce the best results in the simulations are those which incorporate the observations of the other response variables in the estimator. Compared to analogous single-response techniques, this approach results in a significant reduction in the quadratic error in the response.

1 Introduction

Let the problem be the non-parametric estimation of the (vectorial) regression function $\mathbf{m}(\mathbf{x}) = \mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x}) \in \mathbb{R}^c$ from n observations $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n$ with $\mathbf{X}_i \in \mathbb{R}^d$, $\mathbf{Y}_i \in \mathbb{R}^c$ according to the model:

$$\mathbf{Y}_i = \mathbf{m}(\mathbf{X}_i) + \varepsilon(\mathbf{X}_i),\ i = 1, ..., n \tag{1}$$

where $\varepsilon = \varepsilon(\mathbf{x})$ is random noise such that $\mathbb{E}[\varepsilon(\mathbf{x})] = \mathbf{0}_{c\times 1}$ and $\mathrm{Cov}(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}')) = C(\mathbf{x}, \mathbf{x}')$ is the $c \times c$ matrix, which is not necessarily diagonal or symmetrical, with elements $C_{jl}(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}(\varepsilon_j(\mathbf{x}), \varepsilon_l(\mathbf{x}'))$.

In what follows, the above approach may be generalized to the case in which each variable Y_j, $j \in \{1, ..., c\}$ may be observed at different points. However, for the sake of notational simplicity, we will maintain the above formulation.

The motivation for this work is the fact that, when estimating each component m_j, $j \in \{1, ..., c\}$ in $\mathbf{m}$, from a limited set of data, there are in theory no logical reasons for ignoring the information on the Y_j variable that may be held by the observations Y_{li}, $i = 1, ..., n$ of the remaining co-responses Y_l, $l \in \{1, ..., c\}$ with $l \neq j$, as is the normal practice in non-parametric regression. On the contrary, it could be claimed that, for example, in linear multivariate regression the joint estimation of the parameters produces the same results as an estimation of each response separately. Nonetheless, this property would not be necessarily true in more complex models or when using non-parametric techniques, when a limited quantity of data is available.

Despite the many areas in which multi-output non-parametric regression techniques may be useful (finance, engineering, biology, etc.), to our knowledge no formal proposals of this kind have been described in the literature, nor studies that evaluate the possible benefits of this approach.

With a view to exploring the advantages of incorporating information from all the co-responses in the estimation of each component of $\mathbf{m}$, in this work we construct different non-parametric alternatives and evaluate their behavior in simulated tests.

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 288–292, 2005.

2 Non-parametric Alternatives for Multi-output Regression

Given that it supposes no restriction on our discussion, for the sake of notational simplicity we will assume just two responses ($c = 2$) denoted as $\mathsf{Y}, \mathsf{Z} \in \mathbb{R}$.

The multi-output non-parametric regression models which are postulated and subsequently evaluated in our simulations are described below (we shall only specify the estimator for $m_{\mathsf{Y}}(\mathbf{x})$, as that for $m_{\mathsf{Z}}(\mathbf{x})$ is completely analogous).

The first group of models have a **double non-parametric regression** structure in the form described as follows. Given that the regression function $m_{\mathsf{Y}}(\mathbf{x})$ can be written as two expectations:

$$m_{\mathsf{Y}}(\mathbf{x}) = \mathbb{E}_{\mathsf{Y}|\mathbf{x}}(\mathsf{Y}) = \mathbb{E}_{\mathsf{Z}|\mathbf{x}}[\mathbb{E}_{\mathsf{Y}|\mathsf{Z},\mathbf{x}}(\mathsf{Y})]$$

we construct the following estimator:

$$\hat{m}_{\mathsf{Y}}(\mathbf{x}) = \widehat{\mathbb{E}}_{\mathsf{Z}|\mathbf{x}}[\widehat{\mathbb{E}}_{\mathsf{Y}|\mathsf{Z},\mathbf{x}}(\mathsf{Y})]$$

where the symbol $\widehat{\mathbb{E}}$ represents a non-parametric estimator of the corresponding expectation. In our simulations, we test the following particular cases:

1. The double Nadaraya-Watson estimator (DNW):

$$\hat{m}_{\mathsf{Y}}(\mathbf{x}) = \frac{\frac{1}{n^2}\sum_{i,j=1}^{n} k_{H_1}^{(1)}(\mathbf{X}_i - \mathbf{x})\left[k_{H_2}^{(2)}(\mathbf{V}_j - \mathbf{v}_i)\mathsf{Y}_j\right]}{\frac{1}{n^2}\sum_{i,j=1}^{n} k_{H_1}^{(1)}(\mathbf{X}_i - \mathbf{x})k_{H_2}^{(2)}(\mathbf{V}_j - \mathbf{v}_i)}$$

   where $\mathbf{v}_i = (\mathbf{x}^t \ z_i)^t$ and $k_{H_j}^{(j)}(\mathbf{u}) = \frac{1}{|H_j|}k^{(j)}(H_j^{-1}\mathbf{u})$, $j = 1, 2$, where $k^{(1)}$ and $k^{(2)}$ are admisible kernels.
2. The double local linear regression (DLLR), obtaining by successively resolving:

$$\min_{\alpha_{\mathbf{v}},\beta_{\mathbf{v}}} \sum_{j=1}^{n} \left[\mathsf{Y}_j - \alpha_{\mathbf{v}} - \beta_{\mathbf{v}}^t(\mathbf{V}_j - \mathbf{v}_i)\right]^2 k_{H_2}^{(2)}(\mathbf{V}_j - \mathbf{v}_i)$$

$$\min_{\alpha_{\mathbf{x}},\beta_{\mathbf{x}}} \sum_{i=1}^{n} \left[\psi_i - \alpha_{\mathbf{x}} - \beta_{\mathbf{x}}^t(\mathbf{X}_i - \mathbf{x})\right]^2 k_{H_1}^{(1)}(\mathbf{X}_i - \mathbf{x})$$

   where $\psi_i = \psi(\mathsf{Z}_i, \mathbf{X}_i) = \widehat{\mathbb{E}}_{\mathsf{Y}|\mathsf{Z}_i,\mathbf{X}_i}(\mathsf{Y}) = \hat{\alpha}_{\mathbf{V}_i}$. The final estimator is $\hat{m}_{\mathsf{Y}}(\mathbf{x}) = \hat{\alpha}_{\mathbf{x}}$, the estimator of the constant term of the last regression.
3. The double empirical regression (DER), resulting from ([3], [2]) a calculation of the respective conditional expectations using the following estimators of the density functions. For the first regression $\hat{f}(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{U}}(\mathbf{u}; \mathbf{U}_i)$, where:

$$f_{\mathbf{U}}(\mathbf{u}; \mathbf{U}_i) = \frac{1}{2\pi\sigma^2 \left|S_{\mathbf{U}}\right|^{1/2}} \exp\{-\frac{1}{2\sigma^2/n}(\mathbf{u} - \mathbf{U}_i)^t S_{\mathbf{U}}^{-1}(\mathbf{u} - \mathbf{U}_i)\}$$

   with $\mathbf{u} = (\mathbf{x}^t \ y \ z)^t$ and $S_{\mathbf{U}}$ as the empirical variance-covariance matrix for the variable $\mathbf{U} = (\mathbf{X}^t \ \mathsf{Y} \ \mathsf{Z})^t$.