Hierarchical Local Regression

Payman Sadegh and Henrik Öjelund

Department of Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark.

Fax: +45 45 881397

Phone: +45 45 253419

e-mail: ps@imm.dtu.dk

ABSTRACT This paper presents a novel semi-parametric approach for local function approximation under limited global information about the underlying data generating structure. The related technique, named hierarchical local regression (HLR), is based on an aggregate of local models with random mean parameters. The structure is hierarchical in that the local random mean parameters assume probability distributions that are defined by common globally parameterized mean-variance functions. While the global mean function expresses global information about the expected response, the variance function quantifies the uncertainty associated with the global information. On the one hand, this formulation accounts for the parsimonious nature of the global mean function. On the other hand, the information provided by estimated global parameters is combined with locally weighted data to achieve robust local adaptation in data sparse regions which occur frequently in high-dimensional situations (curse of dimensionality). We suggest a criterion for estimation of the parameters and derive an empirical Bayes prediction formula. We present two numerical studies to illustrate different aspects of the method. One example involves prediction of power production in windmill farms based on real data.

Keywords: Hierarchical local regression, Function approximation, Curse of dimensionality, Empirical Bayes, BLUP, Global information.

1 Introduction

Data-based prediction (simulation) of the response of a given system is a central problem in statistics, system theory, and artificial intelligence. A fully parametric approach utilizes for the observed data a family of probability distributions indexed by a parameter. A parameter estimate is a value that minimizes an appropriately defined "discrepancy measure" between the predicted and observed responses. In many situations of practical interest, it may be difficult to construct global models that accommodate the trade-off between model parsimony and unbiasedness of the resulting estimates without extensive knowledge of the underlying data generating structure. For instance, a global regression function that could provide good trend description of data may perform poorly in adapting to localized function variations. Simple increase in the number of regression function parameters may not remedy the situation as over-parameterization leads to over-fitting and random error in estimated parameter values. A nonparametric approach based on local fit, on the other hand, aims at predicting the response at a point by utilizing statistical properties of the nearby data. Examples of such approach are local regression in statistics (Cleveland & Delvin, 1988) and locally weighted learning in artificial intelligence (Atkeson, Moore, & Schaal, 1997). Local models with small number of parameters are usually sufficient to describe function variations in small neighborhoods. However, they generally fail to provide good estimates in high dimensional situations, since as the dimensions increase the bandwidth in each coordinate should be selected very large in order to accommodate only a moderate percentage of data points. It is easy to verify that a bandwidth covering a certain fraction of each coordinate, say $0 < \alpha < 1$, on a p-dimensional cube covers only α^p of the cubic volume and this quickly tends to zero as p increases. Techniques such as projection pursuit regression (Friedman & Stuetzle, 1981) attempt to alleviate the situation. However, no prior information is exploited in these approaches. The problem of high-dimensionality, sometimes referred to as the curse of dimensionality, is a major limitation in many areas related to regression analysis.

The method presented here extends the ideas of hierarchical modeling (Bryk & Raudenbush, 1992) to local regression, hence the technique is named hierarchical local regression (HLR). In this approach, expected response parameters are assumed to be random effects whose probability distributions are defined by a common mean-variance function with unknown global parameters. An important part of HLR modeling involves estimation of the parameters of the mean-variance function. Note that in contrast to traditional regression analysis where parameters of the response distributions are estimated, HLR provides a technique for estimation of the expected response distributions. The globally parameterized mean-variance functions respectively express information about global behavior of the expected response and the associated uncertainty. In this spirit, the global mean function only attempts to capture basic trends in data rather than provide detailed description. Depending on the particular problem of interest, structures such as those derived from low order polynomials and splines, various types of qualitative information (periodicity, symmetry, etc.), and basic laws of physics may be regarded as suitable candidates for incorporation into the global mean function. HLR provides detailed description of the response by relying on assumptions analogous to those related to local regression, hence introducing only minimal *a priori* smoothness restrictions. HLR is distinguished from local regression by the fact that the information embedded in the estimated global parameter values is combined with locally weighted data to obtain robust local adaptation in data sparse regions which occur frequently in high-dimensional situations. In this way, HLR enjoys the benefits of both global and local approaches. In analogy with hierarchical models, the combination of globally estimated parameters with locally weighted data has the consequence that local regression estimates are pulled towards the estimated global mean function which may be thought of as an *attraction surface* (in contrast to regression surface) for prediction of function values. Another major difference between HLR and local regression is in the way the optimization problem for estimation of the parameters is set up. A local regression estimate at a point may be regarded as the solution to an optimization criterion that uses locally weighted data around the neighborhood of interest. As we move from one neighborhood to another, a new criterion is obtained, not only because the data weighting changes, but also because a new set of local optimization parameters is involved.

Estimates at various neighborhoods may then be thought of as solutions to disjoint problems of this type. Missing in such formulation is a method of incorporating global information in form of imposing certain functional relations among the local optimization parameters. The HLR global parameter estimate, on the other hand, is the solution to a single criterion, computed as the sum of local log-likelihoods of data in different neighborhoods. The estimated values of the global parameters determine the common mean-variance function for the local random effects, or equivalently, their unconditional (prior) probability distributions. Empirical Bayes predictions (Maritz & Lwin, 1989) of the function values, also known as BLUP estimates (Robinson, 1991), are computed as the posterior mode of the local effects using the aforementioned estimated prior distributions and locally weighted data. The result is a prediction formula with the appealing property that local regression estimates are attracted towards the (estimated) global mean function, where the magnitude of attraction is a decreasing function of the number of available data points in the neighborhood of interest and the ratio of estimated values of the global variance function to the measurement noise variance at the location of interest. Finally, attention is restricted to normally distributed variables throughout the paper. Even though the normal assumptions facilitate presentation and interpretability of the results, the technique is quite general and may be almost directly applied to other distribution assumptions.

The rest of this paper is organized as follows. Section 2 presents the problem formulation, assumptions, and the method. In Section 3, the theory is applied to modeling of a linear time varying dynamic system using simulated data, and prediction of power production in windmill farms using real data. Section 4 offers concluding remarks.

2 Motivation and Formulation

Let $\boldsymbol{x} \in \mathbb{R}^n$, $y \in \mathbb{R}$, and $\phi \in \mathbb{R}$ generically denote the explanatory variable, the observed response, and the expected response, respectively. We only consider fixed design (non-random explanatory variables) situations. For random design cases, the following discussions hold if all distributions are conditioned on the explanatory variables. Our primary goal is to predict the response at arbitrary values of the explanatory variable or equivalently to estimate a value for the expected response ϕ based on observed values of $\boldsymbol{x}$ and y. In most applications of practical interest, some information about the global dependency of y upon $\boldsymbol{x}$ is given which may translate into a parametric structure $f(\boldsymbol{x};\boldsymbol{\theta})$ for the expected response $\phi(\boldsymbol{x})$ where $\boldsymbol{\theta}$ is the parameter. Relying totally on such parametric models may lead to highly biased estimates if they are only crude approximations to the "true" mean structure. Relying on purely data driven approaches such as local regression, on the other hand, ignores the existing information which might turn indispensable in predicting the response in sparse data regions. The hierarchical structure we propose leads to predictions of the form

$$\hat{\phi}(\boldsymbol{x}) = \frac{\left(\sum_i w(\boldsymbol{x}_i - \boldsymbol{x})\right)^{-1} \sum_i w(\boldsymbol{x}_i - \boldsymbol{x}) y_i + \left(\sum_i w(\boldsymbol{x}_i - \boldsymbol{x})\right)^{-1} \eta^{-2}(\boldsymbol{x})\sigma^2(\boldsymbol{x}) f(\boldsymbol{x};\boldsymbol{\theta})}{1 + \left(\sum_i w(\boldsymbol{x}_i - \boldsymbol{x})\right)^{-1} \eta^{-2}(\boldsymbol{x})\sigma^2(\boldsymbol{x})} \tag{2.1}$$

where $\boldsymbol{x}_i$ and y_i, $i = 1, \cdots, N$, are observed pairwise data, $\{w(\boldsymbol{x}_i - \boldsymbol{x})\}$ is a positive weight sequence that downplays or totally eliminates the contribution of observations whose corresponding value of the explanatory variable is "far" from the point of interest $\boldsymbol{x}$, $\sigma^2(\boldsymbol{x})$ is the noise variance at $\boldsymbol{x}$, and finally $\eta^2(\boldsymbol{x})$ is a global variance function that reflects the "quality" of the parametric model $f(\cdot;\boldsymbol{\theta})$ at $\boldsymbol{x}$. Small values of $\sigma^2(\boldsymbol{x})\eta^{-2}(\boldsymbol{x}) \left(\sum_i w(\boldsymbol{x}_i - \boldsymbol{x})\right)^{-1}$ in the equation (2.1) lead to the kernel estimator $\sum_i w(\boldsymbol{x}_i - \boldsymbol{x}) y_i / \sum_i w(\boldsymbol{x}_i - \boldsymbol{x})$ (Nadaraya, 1964; Watson, 1964; Gasser & Müller, 1979). Large values of this ratio, on the other hand, simply render $f(\boldsymbol{x};\boldsymbol{\theta})$ as the predicted response. The prediction formula (2.1) implies that $f(\boldsymbol{x};\boldsymbol{\theta})$ is an attraction surface for $\hat{\phi}(\boldsymbol{x})$ where $\sigma^2(\boldsymbol{x})\eta^{-2}(\boldsymbol{x}) \left(\sum_i w(\boldsymbol{x}_i - \boldsymbol{x})\right)^{-1}$ determines the magnitude of attraction towards the surface. Since $\boldsymbol{\theta}$, $\sigma(\boldsymbol{x})$, and $\eta(\boldsymbol{x})$ are unknown, they should be first estimated from the data. We present local and global assumptions that lead to derivation of a suitable objective function for calculation of $\boldsymbol{\theta}$, $\sigma(\boldsymbol{x})$, and $\eta(\boldsymbol{x})$ and thereafter study the response prediction problem.

Local assumptions: Around a point $\boldsymbol{x}$ the local assumption is stated through the model

$$\text{for all data points } (\boldsymbol{x}_i, y_i) \text{ such that } i \in I(\boldsymbol{x}):\; y_i|\phi(\boldsymbol{x}) \sim \mathcal{N}\left(\phi(\boldsymbol{x}), \frac{\sigma^2(\boldsymbol{x})}{w(\boldsymbol{x}_i - \boldsymbol{x})}\right) \tag{2.2}$$

where $I(\boldsymbol{x})$ denotes the set of data indices i for which $w(\boldsymbol{x}_i - \boldsymbol{x}) > 0$ and $\sigma^2(\boldsymbol{x})$ is a variance term that reflects measurement noise variance at $\boldsymbol{x}$. We only consider weight sequences that are computed from product kernels (Härdle, 1990) with finite support ($w(\boldsymbol{x}_i - \boldsymbol{x}) = 0$ for $\boldsymbol{x}_i$ outside the support of the kernel around $\boldsymbol{x}$). We remind that product kernel functions in multivariate cases are computed as the product of one-dimensional kernel functions that correspond to each explanatory variable. The variance term $\sigma^2(\boldsymbol{x})$ may be assumed to be an unknown global constant or parameterized as a function of $\boldsymbol{x}$ with unknown global parameters. In both cases, for simplicity we denote the parameters of $\sigma^2(\boldsymbol{x})$ by $\boldsymbol{\sigma}$. We return to the problem of estimating the parameters of the measurement noise model later when estimation of global parameters is discussed. Using (2.2) to obtain a maximum likelihood estimate for the mean leads to a kernel estimator (Nadaraya, 1964; Watson, 1964; Gasser & Müller, 1979). It is easy to see that this estimate does not change if the weights are scaled. Such scaling will however change the estimated values of the global parameters and predicted values of the response for HLR. It is therefore important to scale the weights properly. The scaling is chosen such that for infinitely many uniformly distributed explanatory variables within a finite support kernel, a constant function, and constant measurement noise variance, the mean of the maximum likelihood estimate for the measurement noise variance, parameterized as a constant, is equal to the true noise variance. This holds if $(\int w(\boldsymbol{x}' - \boldsymbol{x}) d\boldsymbol{x}')(\int d\boldsymbol{x}')^{-1} = 1$ where both integrations are taken over $\boldsymbol{x}' : w(\boldsymbol{x}' - \boldsymbol{x}) > 0$.

We continue by forming a set S of sampled values of the explanatory variable and constructing local models around $\boldsymbol{x}_j \in S$. Later in this section, we connect the local models around $\boldsymbol{x}_j \in S$ through a hierarchical structure. The set of sampled values S may not necessarily coincide with the set of available values of the explanatory variable in the dataset. Instead, selection of S is reserved as a design factor for optimal estimation of global parameters. Even though we do not treat this topic formally in the present article, we provide numerical support for relative merit of several sampling schemes in the examples of Section 3.

Now, for each $\boldsymbol{x}_j \in S$ let $\mathcal{Y}^{(\boldsymbol{x}_j)}$ denote $\{y_i | i \in I(\boldsymbol{x}_j)\}$. Given (2.2) and assuming independent measurements, the likelihood of observing $\mathcal{Y}^{(\boldsymbol{x}_j)}$ is given by

$$p(\mathcal{Y}^{(\boldsymbol{x}_j)}|\phi_j) = \frac{1}{\sqrt{(2\pi)^{n_j}\sigma_j^{n_j}}} \exp\left(-\frac{1}{2\sigma_j^2} \sum_{i \in I(\boldsymbol{x}_j)} w_{ij}(y_i - \phi_j)^2\right) \tag{2.3}$$

where $\phi_j = \phi(\boldsymbol{x}_j)$, $w_{ij} = w(\boldsymbol{x}_i - \boldsymbol{x}_j)$, $\sigma_j = \sigma(\boldsymbol{x}_j)$, and n_j is the number of elements of $I(\boldsymbol{x}_j)$.

Remark 1 The smoothness assumption underlying (2.2) is that the function may be sufficiently described as a local constant in small neighborhoods. Extension to local polynomials of higher order, say order m, is possible through

$$\forall i \in I(\boldsymbol{x}): \; y_i|\phi(\boldsymbol{x}), \cdots, \phi^{(m)}(\boldsymbol{x}) \sim \mathcal{N}\left(\phi(\boldsymbol{x}) + \cdots + \phi^{(m)}(\boldsymbol{x})(\boldsymbol{x}_i - \boldsymbol{x})^m, \frac{\sigma^2(\boldsymbol{x})}{w(\boldsymbol{x}_i - \boldsymbol{x})}\right) \tag{2.4}$$

where $\phi^{(k)}(\boldsymbol{x})$ is proportional to the kth derivative of the expected response at $\boldsymbol{x}$. If (2.4) is considered, all discussions in the paper may be applied, keeping in mind that the local parameter at $\boldsymbol{x}$ is $(\phi(\boldsymbol{x}), \cdots, \phi^{(m)}(\boldsymbol{x}))^\top$ rather than $\phi(\boldsymbol{x})$. Similarly, other appropriate forms of local models may be employed within the analysis.

Global assumptions: As stated earlier, local assumptions stated so far are analogous to the general framework for local regression. We depart from this framework by introducing information about the global behavior of the response. Regarding the local parameters $\{\phi_j\}$ as random variables, we use the global information to derive a parametric class of probability distributions for these variables. The (global) parameters involved are unknown and should be estimated from the data. This is quite analogous to the approach followed in hierarchical modeling (Bryk & Raudenbush, 1992). More specifically, we assume

$$\phi(\boldsymbol{x}) \sim \mathcal{N}\left(f(\boldsymbol{x}; \boldsymbol{\theta}), \eta^2(\boldsymbol{x})\right), \tag{2.5}$$

where $f(\boldsymbol{x}; \boldsymbol{\theta})$ and $\eta^2(\boldsymbol{x})$ are common mean-variance functions for local effects. The variance function $\eta^2(\boldsymbol{x})$ may be assumed to be an unknown global constant or parameterized as a function of $\boldsymbol{x}$ with unknown global parameters. In both cases, for simplicity we denote the variance parameters of (2.5) by $\boldsymbol{\eta}$. From (2.3) and (2.5), we compute the (marginal) likelihood of $\mathcal{Y}^{(\boldsymbol{x}_j)}$ as a function of the global parameters by integrating out the local effect in (2.3). Denoting this likelihood by $p(\mathcal{Y}^{(\boldsymbol{x}_j)}; \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta})$ to emphasize the dependency upon the global parameters, we have

$$p(\mathcal{Y}^{(\boldsymbol{x}_j)}; \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta}) = \int_{\phi_j} p(\mathcal{Y}^{(\boldsymbol{x}_j)}|\phi_j)p(\phi_j)d\phi_j \tag{2.6}$$

where $p(\phi_j)$, as given by (2.5), is the density function of a Gaussian distribution with mean $f_j = f(\boldsymbol{x}_j; \boldsymbol{\theta})$ and variance $\eta_j = \eta(\boldsymbol{x}_j)$.

Remark 2 HLR is able to incorporate global information in a variety of forms. To illustrate this point by an example, let us consider a function of the scalar variable x which is "almost" symmetric around the midpoint of some interval. Defining the equidistant sampling $S = \{x_0, \cdots, x_{2T}\}$ where x_0 and x_{2T} coincide with the end points of the interval and $\boldsymbol{\theta} = (\theta_0, \cdots, \theta_T)^\top$, the global mean function may be written as $f_j = \theta_j$ for $j = 0, \cdots, T$, and $f_j = \theta_{2T-j}$ for $j = T+1, ..., 2T$ (the global variance function may be assumed constant). In the simple "almost" symmetric case here, HLR essentially uses dense data regions at one half of the range of x to infer upon function values at data sparse half of the range.

Estimation of the global parameters: While separate optimization of local likelihoods (2.3) with respect to ϕ_j for varying j yields local estimates for the local parameters, the approach fails for estimation of the global parameters since optimizing (2.6) for $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\eta}$ yields *local estimates* for the same *global parameters* as j varies. The situation may be viewed as a multi-objective optimization where varying local criteria $\log p(\mathcal{Y}^{(\boldsymbol{x}_j)}; \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta})$ compete for the same global resources $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\eta}$. The local criteria may be aggregated into a single global objective function by

$$J(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta}) = \sum_{j:\boldsymbol{x}_j \in S} L_j(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta}) \tag{2.7}$$

where $L_j(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta}) = \log p(\mathcal{Y}^{(\boldsymbol{x}_j)}; \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta})$. The objective function in (2.7) is proportional to the logarithm of the total likelihood of data, if $\forall \boldsymbol{x}_j, \boldsymbol{x}_{j'} \in S$, $j \neq j'$: $I(\boldsymbol{x}_j) \cap I(\boldsymbol{x}_{j'}) = \emptyset$, i.e. in case the supports of the kernels centered around $\boldsymbol{x}_j \in S$ are non-overlapping. In the general situation of overlapping kernel supports, the objective function (2.7) may be regarded as the decomposition

$$J(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta}) = \sum_{m=1}^{M} \sum_{j:\boldsymbol{x}_j \in S_m} L_j(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta}) \tag{2.8}$$

where each subset S_m consists of kernel centers with non-overlapping supports and $S = \bigcup_m S_m$. Each term $\sum_{j:\boldsymbol{x}_j \in S_m} L_j(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta})$ in the decomposition (2.8) is the log-likelihood of data given the local models centered around $\boldsymbol{x}_j \in S_m$. Optimizing $\sum_{j:\boldsymbol{x}_j \in S_m} L_j(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta})$ with respect to $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, $\boldsymbol{\eta}$ yields various estimates for the same global parameters as m varies. By integrating these criteria into the global objective function (2.8) a global estimate of the global parameters is obtained. This heuristically justifies the use of $J(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta})$ as an appropriate objective function to estimate the global parameters from.

Asymptotic behavior: Consider a situation where a set of kernels cover the whole dataset of N independent observations. In case the observations within kernels are independently obtained, N coincides with the sum of number of observations within the kernels, and under general regularity conditions the law of large numbers for weighted sums implies convergence of $N^{-1}J(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta})$ to its asymptotic limit $\sum_{j:\boldsymbol{x}_j \in S} E\{L_j(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\eta})\}$ as $N \to \infty$. This can be shown following arguments similar to the proof of Proposition (1) below. The independence for example holds for kernels with non-overlapping supports as discussed above. The proposition below shows similar convergence results for a more general case.

Proposition 1 Let the true data generating mechanism be such that each data point is Gaussian with bounded deterministic mean and variance. Deleting all kernels that contain no observation, consider the decomposition (2.8) where for each m the kernels corresponding to S_m cover all the observations. Denote the number of elements of S_m by K_m and let $N \to \infty$ and $K_m \to \infty$ for all $m = 1, \cdots, M$ (note that the average number of observations per kernel for each m cannot tend to zero, or equivalently $K_m/N < \infty$ for all $m = 1, \cdots, M$, since no empty kernel exists in the set). Also, assume that for all values of $\boldsymbol{\sigma}$ and $\boldsymbol{\eta}$, the functions

$\eta^2(\boldsymbol{x})$, $\sigma^2(\boldsymbol{x})$, $\eta^{-2}(\boldsymbol{x})$, and $\sigma^{-2}(\boldsymbol{x})$ are bounded for all $\boldsymbol{x}$. Then

$$N^{-1}J(\boldsymbol{\theta},\boldsymbol{\sigma},\boldsymbol{\eta}) \xrightarrow{\text{a.s.}} \sum_{j:\boldsymbol{x}_j\in S} E\{L_j(\boldsymbol{\theta},\boldsymbol{\sigma},\boldsymbol{\eta})\}$$

if M is smaller than a variable of order $N^{\frac{1}{2}-\frac{1}{\alpha_0}}$ for some $\alpha_0 \geq 1$.

PROOF: See the appendix.

Prediction: Predictions of function values may be easily obtained once the global parameters are calculated from (2.7). We insert estimated values of $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, $\boldsymbol{\eta}$ in $p\left(\phi(\boldsymbol{x})\right)$, and $p\left(\mathcal{Y}^{(\boldsymbol{x})}|\phi(\boldsymbol{x})\right)$, and compute the posterior mode of $\phi(\boldsymbol{x})$,

$$\hat{\phi}(\boldsymbol{x}) = \arg\max_{\phi(\boldsymbol{x})} p\left(\mathcal{Y}^{(\boldsymbol{x})}|\phi(\boldsymbol{x})\right) p\left(\phi(\boldsymbol{x})\right), \tag{2.9}$$

at a point $\boldsymbol{x}$ where the prediction is desired. The posterior mode $\hat{\phi}(\boldsymbol{x})$ is the predicted value, which in the Gaussian case coincides with posterior mean. The prediction formula (2.1) is the closed form solution to (2.9) for the locally constant model (2.2). More general local models, such as local polynomials of arbitrary orders (see Remark 1), may be treated similarly to derive suitable prediction formulas.

Finally, for the locally constant model (2.2) we compute the variance of the posterior distribution of $\phi(\boldsymbol{x})$ which yields

$$\mathrm{var}(\hat{\phi}(\boldsymbol{x})) = \frac{\sigma^2(\boldsymbol{x})\eta^2(\boldsymbol{x})}{\sum_{i\in I(\boldsymbol{x})} w(\boldsymbol{x}_i-\boldsymbol{x}) + \sigma^2(\boldsymbol{x})}.$$

This equation is useful for calculation of confidence bounds for the predicted function values.

3 Applications

3.1 Modeling of a linear time-varying system

In this numerical study, we consider modeling of a linear time-varying dynamic system based on simulated data. We pursue several objectives in this example. First, it is well-known that local regression performs poorly in high dimensional situations as a result of increasing data sparsity. This example concerns a 4 dimensional case with moderate number of observations for which local regression fails due to lack of sufficient data points in several neighborhoods. Second, we study the ability of HLR to adapt to localized variations as well as the ability to provide good overall description of the trend in data as compared to a fully parametric approach where the comparison is performed for different number of observations. Third, we study the effect of kernel placement schemes on the performance of HLR.

Let us consider the following data generating mechanism,

$$y_t = a_{1,t}\, y_{t-1} + a_{2,t}\, y_{t-2} + b_{0,t}\, u_t + b_{1,t}\, u_{t-1} + \epsilon_t \tag{3.10}$$

where $u_t \sim N(0,1)$ is the measured input and $\epsilon_t \sim \mathrm{N}(0, 0.04)$ is the (measurement) noise term. Moreover, $\{\epsilon_t\}$ and $\{u_t\}$ are i.i.d. and mutually independent. The time-dependency of the parameters is described by

$$\begin{array}{rcl} a_{1,t} & = & 0.4 \sin\left(\frac{\mathrm{y_{t-1}+y_{t-2}}}{2}\right), \\ a_{2,t} & = & 0.4 \cos\left(\frac{\mathrm{y_{t-1}+y_{t-2}}}{2}\right), \\ b_{0,t} & = & 1.2 + 0.4 \cos\left(\frac{\mathrm{y_{t-1}+y_{t-2}}}{2} + \mathrm{u_t} + \mathrm{u_{t-1}}\right), \\ b_{1,t} & = & -0.2 + 0.4 \sin\left(\frac{\mathrm{y_{t-1}+y_{t-2}}}{2} + \mathrm{u_t} + \mathrm{u_{t-1}}\right), \end{array} \tag{3.11}$$

and histograms over parameter variations are shown in Figure 1. Let $\boldsymbol{x}(t)$ denote the vector

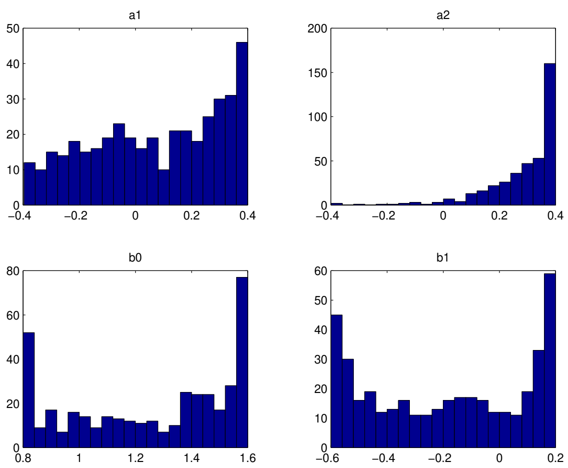

Figure 1: Histograms of the parameter values of the training data.

$(y_{t-1}\ y_{t-2}\ u_t\ u_{t-1})^\top$, and assume a linear mean and constant variance function for local effects, i.e. $\phi(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{x}^\top\boldsymbol{\theta}, \eta^2)$ where $\boldsymbol{\theta}$ is an unknown global parameter vector and η^2 is an unknown global constant. For the local model (2.2), we select $\sigma^2(\boldsymbol{x})$ to be a global constant and denote it by σ^2. We generate data for parameter estimation by simulating 400 time steps of the recursive equation (3.10) and optimize the criterion (2.7) to calculate global parameter estimates. The same dataset is later used as the set of pairwise data points $\boldsymbol{x}_i$, y_i, $i = 1, \cdots, N$, for calculations of the equation (2.1) and for local regression calculations. We repeat the procedure of estimating the global parameters for three different kernel placement schemes viz. uniformly distributed sets of disjoint kernels, multivariate normally distributed kernels, and kernels centered at each observation. In all cases, triangular product kernels each having hyper-rectangular support have been used.

Uniformly distributed sets of disjoint kernels: First fix a collection of disjoint kernels that cover the entire range of the explanatory variables. We can obtain different collections of this type by setting various position parameters, and simultaneously shifting all the kernel centers by the same position parameter. Starting from a fixed set of disjoint kernels, we generate various sets by sampling the position parameters from a uniform distribution. Deleting the kernels that contain no data, the set S is taken as the center positions of the union of all the generated kernels. Addition of a new set of disjoint kernels obviously changes the value of the estimated parameter. We select the number of sets of disjoint kernels (M) such that further addition of a set of kernels does not cause a significant change in the estimated values of the global parameters. In this example, 200 sets of disjoint kernels, each containing an average of 28 kernels is selected.

Multivariate normally distributed kernels: The center positions of the kernels are sampled from a multivariate normal distribution whose mean and covariance are the sample mean and covariance estimates for the sequence of explanatory variables in the dataset. We select 7000 normally sampled kernels. Similar to previous kernel placement scheme, the criterion for selection of the number of kernels is the convergence of estimated values of the global parameters.

Kernels centered at observations: In this kernel placement scheme, the set S is selected as the set of available explanatory variable values in the dataset.

For validation and selection of bandwidth, we generate new data from the data generating mechanism (3.10) by simulating 500 time steps with $\epsilon_t = 0$. Using the formula (2.1) recursively, where at stage t of the recursion the explanatory variable $\boldsymbol{x}$ is set to $\boldsymbol{x}(t) = (\hat{\phi}(\boldsymbol{x}(t-1)) \ \hat{\phi}(\boldsymbol{x}(t-2)) \ u_t \ u_{t-1})^\top$, we generate a sequence of 500 simulated values. Initial conditions in both of the 500 step simulations are selected to be zero. The mean squared error (sim^2) is used to assess the performance of the estimated model where the error sequence is defined as the difference between output sequences of the two 500 step simulations. The optimal bandwidth for each case is selected as the bandwidth minimizing sim^2.

The results are summarized in Table 3.1 and Figure 2. In Table 3.1, values of the estimated variance parameters, optimized sim^2 values, and the corresponding optimal bandwidths are shown for the three kernel placement schemes and different number of observations. For comparison, the table also shows the results obtained from estimating a linear ARX model. Figure 2 compares the linear (ARX) and the HLR estimated models where uniform kernel placement scheme is used for HLR. It is found that even though the linear ARX model is able to capture the main dynamics of the system, it is clearly outperformed by HLR regardless of the kernel placement scheme. The superior performance of HLR becomes more pronounced as the number of observations increases. Notice that HLR uses no information about the dynamical relations (3.11) nor does it attempt to model the dynamics of parameter variations. Furthermore, we apply local regression with the same bandwidth as the one used for HLR. The simulated values are obtained recursively in exactly the same manner as the one for HLR except for that all simulated values (including those inserted in $\boldsymbol{x}(t)$ for the recursive calculations) are now local regression estimates rather than HLR estimates. Local regression fails to simulate the behavior of the system as in all our simulations (even when the number of observations is 400) the recursions hit a point around which the kernel contains no data points.

Finally, our experiment with this example shows that kernel placement scheme influences the performance of the estimated model and is an important design factor. Furthermore,

Model type	$\hat{\sigma}^2$	$\hat{\eta}^2$	b	sim^2
Linear 400 obs.	0.3412	–	–	**0.40**
Linear 200 obs.	0.4038	–	–	**0.41**
Linear 100 obs.	0.4076	–	–	**0.43**
Linear 50 obs.	0.5759	–	–	**0.46**
HLR, On Obser. 400 obs.	0.33	0.33	1.6	**0.21**
HLR, On Obser. 200 obs.	0.46	0.40	2.0	**0.22**
HLR, On Obser. 100 obs.	0.60	0.27	2.1	**0.26**
HLR, On Obser. 50 obs.	0.71	0.54	2.5	**0.28**
HLR, Uniformly 400 obs.	0.15	0.46	1.3	**0.22**
HLR, Uniformly 200 obs.	0.24	0.59	1.8	**0.22**
HLR, Uniformly 100 obs.	0.31	0.52	2.1	**0.28**
HLR, Uniformly 50 obs.	0.63	0.61	3.0	**0.29**
HLR, Normally 400 obs.	0.23	0.21	1.4	**0.23**
HLR, Normally 200 obs.	0.32	0.19	1.7	**0.25**
HLR, Normally 100 obs.	0.46	0.20	2.0	**0.28**
HLR, Normally 50 obs.	0.83	0.18	2.6	**0.28**

Table 1: Validation results using three kernel placement schemes of HLR and linear ARX model for different number of observations. The estimated values of σ^2 and η^2, and the bandwidth are denoted by $\hat{\sigma}^2$, $\hat{\eta}^2$, and b respectively.

we find that the number of kernels required for the convergence of the global parameter (i.e. a situation where addition of new kernels does not cause a significant change in the estimated value of the global parameter) is smaller for uniformly distributed sets than for normally distributed kernels.

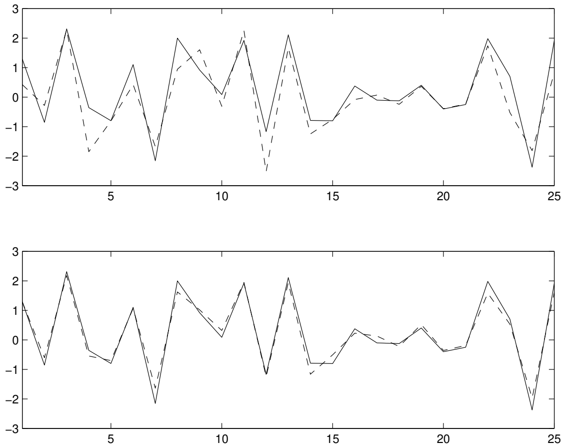

Figure 2: The solid curves show a simulation of the true function with no added noise. The dashed lines show the results of simulations with a linear (upper figure) and HLR (lower figure) models.

3.2 Modeling of power production in windmill farms

The second study concerns prediction of the output power in windmill farms as a function of observed meteorological data. The objective of this example is to apply the technique to a real modeling problem where a parsimonious physical model will be employed as the global mean function. Furthermore, this example illustrates a situation where noise and model quality variances are best modelled as functions of $\boldsymbol{x}$ rather than constants.

In-depth study of the windmill data used in this example can be found in Nielsen, Madsen, Nielsen, & Tøfting (1999). Sampled observations of the generated power, wind speed, and wind direction were taken every 30 minutes for a period of one year. We split the dataset into the data collected during odd weeks and the data collected during even weeks and use these two datasets for training (estimation) and testing (validation) respectively. The explanatory variables are wind speed and wind direction. Figure 3 shows a polar plot of the explanatory variables for the training set. Notice that in some wind directions, no high wind speed value has been observed during the year. For instance at 120 degrees the maximum observed wind speed is less than 10 m/s. In Figure 4 the training data is used to plot the generated power as a function of the wind speed for the wind direction 330±15 degrees and as a function of wind direction for the wind speed 7.25±0.25 m/s. It can be seen that the produced power approximately follows a s-shaped function of the wind speed and deviation from the s-shaped function varies with wind speed. The plot further indicates the dependency of the generated power upon wind direction. Based on these plots, we see that the unknown data generating structure is clearly non-linear, the noise intensity is not constant, and no data exists in certain regions. Since we are interested in predicting the generated power as a function of meteorological forecasts, which may assume values outside the range of available data, we should devise a way of making predictions even in sparse or no data regions together with an uncertainty assessment for the predicted values.

A simple model commonly used for prediction of produced power employs the so-called Gompertz function. Denoting the wind speed by v, the Gompertz function computes the

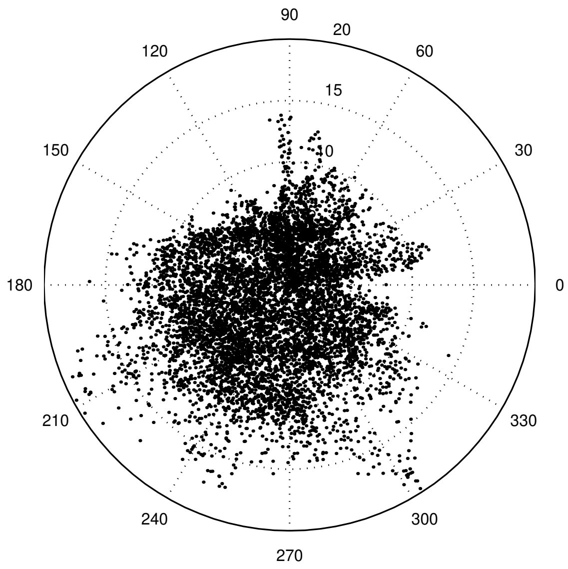

Figure 3: Wind direction and wind speed at which data are obtained.

power production through the relation

$$G(v;\boldsymbol{\theta}) = \theta_1 \exp(\theta_2 \exp(\theta_3 v)) \tag{3.12}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^\top$ is a parameter vector. Even though the dependency of generated power on wind direction, as established in detailed analysis of Nielsen et al. (1999), is disregarded by the Gompertz function, there is strong reason to choose (3.12) as the global mean function of HLR. This is due to the simplicity of the Gompertz function as well as its ability to explain 90% of the power variations ($R^2 = 0.9$) if used as the regression function of a fully parametric approach. Defining the explanatory variable as $\boldsymbol{x} = (v, \gamma)^\top$, where γ denotes wind direction, we assume the distributions

$$\phi(\boldsymbol{x}) \sim \mathcal{N}\left(G(v;\boldsymbol{\theta}), \eta^2(v)\right) \tag{3.13}$$

for the random effects of the locally constant model (2.2). Further examination of the data in Figure 4 indicates strong dependence of σ upon v. Hence, we consider the following piecewise

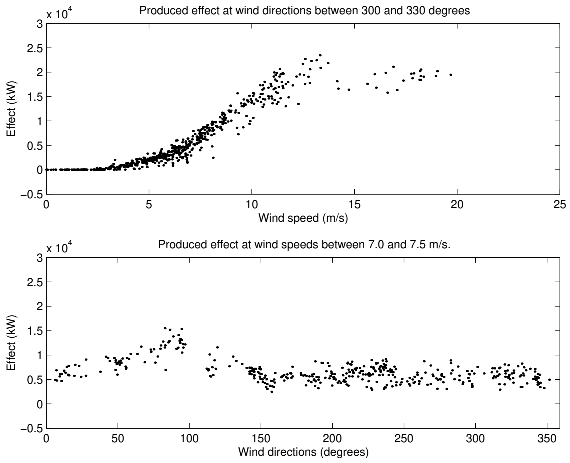

Figure 4: The produced effect as function of the wind speed at 315±15 degrees, and the wind direction at 7.25±0.25 m/s.

constant functional relation,

$$\sigma^2(v) = \begin{cases} \sigma_a^2 & & v < 3 \\ \sigma_b^2 & 3 \leq & v < 6 \\ \sigma_c^2 & 6 \leq & v < 9 \\ \sigma_d^2 & 9 \leq & v < 12 \\ \sigma_e^2 & & v \geq 12 \end{cases} \tag{3.14}$$

where σ_a, σ_b, σ_c, σ_d, and σ_e are unknown parameters. We parameterize η in exactly the same way.

Similar to the previous example, we select triangular product kernels with hyper rectangular support and spread them randomly using the uniform distribution scheme (see the previous example). We select two different bandwidths $h_\gamma = 45$ degrees and $h_v = 1.5$ m/s in the γ and v directions respectively, and find that 20 sets of disjoint kernels, each containing 97 kernels in average, are required to obtain stable estimates of the global parameters for the training set. We apply the HLR prediction formula for the test dataset and achieve an R^2 value of 94%. This value is noticeably larger than 90% found for the simple Gompertz function model. In Figure 5, the validation data is plotted together with the predicted values $\pm 2\times$ the standard errors calculated as $\sqrt{\mathrm{var}(\hat{\phi}(\boldsymbol{x})) + \hat{\sigma}(\boldsymbol{x})^2}$. Notice that the confidence interval increases as the data become increasingly sparse at around 20 m/s. Further beyond this point, predictions are completely calculated on the basis of the global mean function. Local

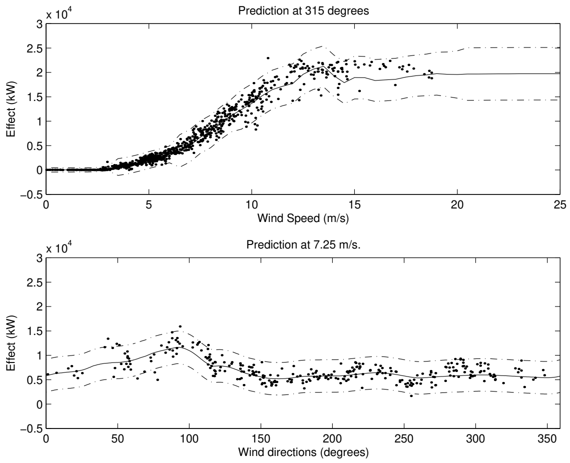

Figure 5: Empirical Bayes prediction with two times the standard errors bands. The shown data is the validation data for wind direction 315±15 degrees and wind speed 7.25±0.25 m/s.

regression will fail at wind speed values around and beyond 20 m/s.

4 Conclusion

We have presented HLR as a new technique for function approximation under incomplete global information about the underlying data generating structure. The technique combines local approximation schemes with global probabilistic information about the function. While function approximation in dense data regions relies heavily on local properties of the data, the parameters of the global distributions (the global parameters) essentially provide information about the function in sparse data regions using the whole dataset. We have derived a cost function for estimation of the global parameters and examined the asymptotic properties of the cost as the number of observations tend to infinity. Numerical studies with both simulated and real data indicate the superiority of the technique under the realistic situation that only limited information about the function (e.g. trend information) is available.

There is a variety of directions for future research in this area. Assuming availability of on-line controlled experiments, HLR may be incorporated within a sequential experiment design/parameter estimation scheme as follows. Given an initial dataset and starting with a collection of non-overlapping kernels, a maximum likelihood estimate for the global parameter vector may be calculated, which together with an appropriately defined experimental design cost function may determine the optimal placement for a new kernel. *Independent* measurements within the optimally found kernel may be added to the dataset to re-calculate maximum likelihood estimates and repeat the whole procedure several times. Another extension is to regard the variance component as a random variable. This might ease the modeling efforts in situations like the second numerical study of the paper where the data clearly indicate dependency of the variance upon the explanatory variables. HLR may also find application in model validation area through testing the significance of the variance term that reflects the quality of the global mean function (attraction surface). Another interesting direction is applying bootstrapping techniques to generate data for calculation of the cost when the data are collected off-line. The cost for estimation of the global parameter is proportional to the log-likelihood of data only if the data within kernel supports are collected independently. One may start with placing a set of non-overlapping kernels throughout the space of the explanatory variable and estimating a value for the global parameter vector. This estimate may then be used to simulate *independent* series of data within highly overlapping kernels (whose locations are for example calculated according to some appropriately defined experimental design cost) as they are added to the set and re-estimate the parameters. This procedure may be repeated several times. Other interesting research directions include quantification of bias in the estimated HLR parameter values (as compared to global regression models), use of low-order splines as global mean function structures, various applications in predictive control, and finally comparative studies between HLR and "black-box" function approximation approaches such as those based on artificial neural network training.

Acknowledgement: The authors would like to thank Professor Poul Thyregod and Professor Henrik Madsen for their comments.

References

Atkeson, C. G., Moore, A. W., & Schaal, S. (1997). Locally weighted learning. *Artificial Intelligence Review*, **11**, 11–73.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical Linear Models.* Newbury Park, CA: Sage Publications, Inc.

Chow, Y. S. & Lai, T. L. (1973). Limiting behavior of weighted sums of independent random variables. *The Annals of Probability*, **1**(5), 810–824.

Cleveland, W. & Delvin, S. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of American Statistical Association*, **83**, 596–610.

Friedman, J. H. & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, **76**(376), 817–823.

Gasser, T. & Müller, H. G. (1979). Kernel estimation of regression functions. In Gasser, T. & rosenblatt, M. (Eds.), *Smoothing Techniques for Curve Estimation*, pp. 23–68. Springer-Verlag, Heidelberg.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press.

Maritz, J. S. & Lwin, T. (1989). *Empirical Bayes.* Monographs on Statistics and Applied Probability. Thomson Science and Professional.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, **9**, 141–142.

Nielsen, T. S., Madsen, H., Nielsen, H. A., & Tøfting, J. (1999). Using meteorological forecasts in on-line predictions of wind power. Technical Report, ELSAM, Fredericia, Denmark.

Robinson, G. K. (1991). That blup is a good thing: the estimation of random effect. *Statistical Science*, **6**(1), 15–51.

Watson, G. S. (1964). Smooth regression analysis. *Sankyā A*, **26**, 359–372.

Appendix: proof of Proposition 1

Let us examine each term in (2.8) more closely. We assume that the data are generated by an arbitrary mechanism $y_i = \bar{\mu}_i + \bar{\sigma}_i\epsilon_i$ where $\bar{\mu}_i$ and $\bar{\sigma}_i$ are unknown bounded constants and $\{\epsilon_i\}$ are i.i.d. zero-mean Gaussian with $\mathrm{var}(\epsilon_i) = 1$. Straightforward calculations yield

$$\begin{aligned}\frac{L_j(\boldsymbol{\theta},\boldsymbol{\sigma},\boldsymbol{\eta})}{N} &= \\ &-\frac{n_j}{N}\log\sigma_j - \frac{\log\eta_j}{N} - \frac{1}{2N}\log(\frac{1}{\eta_j^2}+\frac{1}{\sigma_j^2}\sum_i w_{ij}) - \frac{1}{2N\sigma_j^2}\sum_i w_{ij}y_i^2 - \frac{1}{2}\frac{f_j^2}{N\eta_j^2} \\ &+\frac{1}{2N}(\frac{1}{\eta_j^2}+\frac{1}{\sigma_j^2}\sum_i w_{ij})^{-1}(\frac{f_j}{\eta_j^2}+\frac{1}{\sigma_j^2}\sum_i w_{ij}y_i)^2 + \text{constant}. \qquad \text{(A.1)}\end{aligned}$$

Now consider $N^{-1}\sum\limits_j L_j(\boldsymbol{\theta},\boldsymbol{\sigma},\boldsymbol{\eta})$ for $j : \boldsymbol{x}_j \in S_m$. We study the limiting behavior of the sums

$$N^{-1}\sum_{j:\boldsymbol{x}_j\in S_m}\sum_{i\in I(\boldsymbol{x}_j)}\sigma_j^{-2}w_{ij}y_i^2, \qquad \text{(A.2)}$$

$$N^{-1}\sum_{j:\boldsymbol{x}_j\in S_m}\left\{(\eta_j^{-2}+\sigma_j^{-2}\sum_i w_{ij})^{-1}(\frac{f_j}{\eta_j^2}+\frac{1}{\sigma_j^2}\sum_i w_{ij}y_i)^2\right\}. \qquad \text{(A.3)}$$

To do so, we apply the results of Chow & Lai (1973) to give strong convergence of weighted sums of interest under squared summability of the weights. In all summations below, i and j are such that $\boldsymbol{x}_j \in S_m$ and $i \in I(\boldsymbol{x}_j)$. For the summation (A.2)

$$\begin{aligned}\sum_j\sum_i \sigma_j^{-2}w_{ij}y_i^2 &= \sum_j\sum_i \sigma_j^{-2}w_{ij}\bar{\sigma}_i^2\frac{(y_i-\bar{\mu}_i+\bar{\mu}_i)^2}{\bar{\sigma}_i^2} \\ &= \sum_j\sum_i \sigma_j^{-2}w_{ij}\bar{\sigma}_i^2(\frac{y_i-\bar{\mu}_i}{\bar{\sigma}_i})^2 + 2\sum_j\sum_i \sigma_j^{-2}w_{ij}\bar{\mu}_i\bar{\sigma}_i\frac{y_i-\bar{\mu}_i}{\bar{\sigma}_i} + \sum_j\sum_i \sigma_j^{-2}w_{ij}\bar{\mu}_i^2.\end{aligned}$$

We first study the squared summability of the weights in the above sums. Noting that $0 \le w_{ij} \le w_{\max}$ and assuming that $\bar{\mu}_i$ and $\bar{\sigma}_i$ are bounded for all i then obviously

$$\sum_j\sum_i\left(\frac{w_{ij}\bar{\sigma}_i^2}{\sigma_j^2\sqrt{N}}\right)^2 \le w_{\max}^2 \mathrm{sup}_i\bar{\sigma}_i^4 \mathrm{sup}_j\sigma_j^{-4}\frac{1}{N}N < \infty,$$

$$\sum_j\sum_i\left(\frac{w_{ij}\bar{\mu}_i\bar{\sigma}_i}{\sigma_j^2\sqrt{N}}\right)^2 \le w_{\max}^2 \mathrm{sup}_i\bar{\sigma}_i^2 \mathrm{sup}_j\sigma_j^{-4}\mathrm{sup}_i\bar{\mu}_i^2\frac{1}{N}N < \infty.$$

In addition to the squared summability of the weights as shown above, it holds that

$$\frac{y_i - \bar{\mu}_i}{\bar{\sigma}_i} \sim \mathcal{N}(0,1)\text{ , hence } E\left\{\left(\frac{y_i - \bar{\mu}_i}{\bar{\sigma}_i}\right)^2\right\} = 1, E\left\{\frac{y_i - \bar{\mu}_i}{\bar{\sigma}_i}\right\} = 0,\ E\left\{|\frac{y_i - \bar{\mu}_i}{\bar{\sigma}_i}|^\alpha\right\} < \infty$$

for all $\alpha \geq 1$. Notice further that each S_m is constructed to contain centers of kernels with non-overlapping support. This implies that the sequence $\{y_i\}$ in the summation $\sum_j \sum_i \sigma_j^{-2} w_{ij} y_i^2$ is independent. Now apply the strong convergence of weighted sums to conclude that with probability one

$$N^{-\frac{1}{2}} \sum_j \sum_i \sigma_j^{-2} w_{ij} y_i^2 =$$
$$N^{-\frac{1}{2}} \sum_j \sigma_j^{-2} \left[\sum_i w_{ij} \bar{\sigma}_i^2 E\left\{\left(\frac{y_i - \bar{\mu}_i}{\bar{\sigma}_i}\right)^2\right\} + 2 \sum_i w_{ij} \bar{\mu}_i \bar{\sigma}_i E\left\{\frac{y_i - \bar{\mu}_i}{\bar{\sigma}_i}\right\} + \sum_i w_{ij} \bar{\mu}_i^2\right] + o(N^{\frac{1}{\alpha}})$$

for all $\alpha \geq 1$ or equivalently

$$N^{-1} \sum_j \sum_i \sigma_j^{-2} w_{ij} y_i^2 = N^{-1} \left\{\sum_j \sum_i \sigma_j^{-2} w_{ij} \bar{\sigma}_i^2 + \sum_j \sum_i \sigma_j^{-2} w_{ij} \bar{\mu}_i^2\right\} + \frac{o(N^{\frac{1}{\alpha}})}{N^{\frac{1}{2}}} \tag{A.4}$$

almost everywhere. As for the sum (A.3), first denote

$$\kappa_j = \eta_j^{-2} f_j + \sigma_j^{-2} \sum_i w_{ij} \bar{\mu}_i,\ \zeta_j = \sigma_j^{-2} \sqrt{\sum_i w_{ij}^2 \bar{\sigma}_i^2}.$$

Now

$$N^{-\frac{1}{2}} \sum_j \left\{(\eta_j^{-2} + \sigma_j^{-2} \sum_i w_{ij})^{-1} (\eta_j^{-2} f_j + \sigma_j^{-2} \sum_i w_{ij} y_i)^2\right\} =$$
$$N^{-\frac{1}{2}} \sum_j \left\{(\eta_j^{-2} + \sigma_j^{-2} \sum_i w_{ij})^{-1} \zeta_j^2 \left(\frac{\eta_j^{-2} f_j + \sigma_j^{-2} \sum_i w_{ij} y_i - \kappa_j}{\zeta_j}\right)^2\right\} +$$
$$N^{-\frac{1}{2}} \sum_j 2 \left\{(\eta_j^{-2} + \sigma_j^{-2} \sum_i w_{ij})^{-1} \zeta_j \kappa_j \left(\frac{\eta_j^{-2} f_j + \sigma_j^{-2} \sum_i w_{ij} y_i - \kappa_j}{\zeta_j}\right)\right\} +$$
$$N^{-\frac{1}{2}} \left\{\sum_j (\eta_j^{-2} + \sigma_j^{-2} \sum_i w_{ij})^{-1} \kappa_j^2\right\}.$$

Further

$$\frac{\eta_j^{-2} f_j + \sigma_j^{-2} \sum_i w_{ij} y_i - \kappa_j}{\zeta_j} \sim \mathcal{N}(0,1) \text{ hence}$$

$$E\left\{\left(\frac{\eta_j^{-2} f_j + \sigma_j^{-2} \sum_i w_{ij} y_i - \kappa_j}{\zeta_j}\right)^2\right\} = 1, E\left\{\frac{\eta_j^{-2} f_j + \sigma_j^{-2} \sum_i w_{ij} y_i - \kappa_j}{\zeta_j}\right\} = 0,$$

$$E\left\{|\frac{\eta_j^{-2} f_j + \sigma_j^{-2} \sum_i w_{ij} y_i - \kappa_j}{\zeta_j}|^\alpha\right\} < \infty,$$

for all $\alpha \geq 1$. Again notice that S_m contains centers of kernels with non-overlapping support which implies that

$$\left\{\frac{\eta_j^{-2} f_j + \sigma_j^{-2} \sum_i w_{ij} y_i - \kappa_j}{\zeta_j}\right\}$$

is i.i.d. Hence as $K_m \to \infty$

$$N^{-\frac{1}{2}} \sum_j \left\{(\eta_j^{-2} + \sigma_j^{-2} \sum_i w_{ij})^{-1} (\eta_j^{-2} f_j + \sigma_j^{-2} \sum_i w_{ij} y_i)^2\right\} =$$

$$N^{-\frac{1}{2}} \sum_j \left\{(\eta_j^{-2} + \sigma_j^{-2} \sum_i w_{ij})^{-1} \zeta_j^2\right\} + N^{-\frac{1}{2}} \left\{\sum_j (\eta_j^{-2} + \sigma_j^{-2} \sum_i w_{ij})^{-1} \kappa_j^2\right\} +$$

$$o(K_m^{\frac{1}{\alpha}}), \quad \text{a.s.}$$

given squared summability of the weight sequences

$$\left\{N^{-\frac{1}{2}} (\eta_j^{-2} + \sigma_j^{-2} \sum_i w_{ij})^{-1} \zeta_j^2\right\} \ , \ \left\{N^{-\frac{1}{2}} (\eta_j^{-2} + \sigma_j^{-2} \sum_i w_{ij})^{-1} \kappa_j \zeta_j\right\}.$$

But notice that since $0 \leq w_{ij} \leq w_{\max}$ then

$$\zeta_j^2 = \sigma_j^{-4} (\sum_i w_{ij}^2 \bar{\sigma}_i^2) = w_{\max}^2 \sigma_j^{-4} (\sum_i \frac{w_{ij}^2}{w_{\max}^2} \bar{\sigma}_i^2) \leq w_{\max}^2 \sigma_j^{-4} (\sum_i \frac{w_{ij}}{w_{\max}} \bar{\sigma}_i^2) = w_{\max} \sigma_j^{-4} (\sum_i w_{ij} \bar{\sigma}_i^2),$$

and since σ_j, η_j, σ_j^{-1}, and η_j^{-1} are bounded, then

$$\left(N^{-\frac{1}{2}} (\eta_j^{-2} + \sigma_j^{-2} \sum_i w_{ij})^{-1} \zeta_j^2\right)^2 \leq k_1 N^{-1}$$

for a constant k_1 and

$$\sum_j \left(N^{-\frac{1}{2}} (\eta_j^{-2} + \sigma_j^{-2} \sum_i w_{ij})^{-1} \zeta_j^2\right)^2 \leq k_1 \frac{K_m}{N} < \infty.$$

Also

$$\zeta_j = \sigma_j^{-2}\sqrt{\sum_i w_{ij}^2 \bar{\sigma}_i^2} \le \sigma_j^{-2}(\sum_i w_{ij}\bar{\sigma}_i)$$

and therefore

$$\left(N^{-\frac{1}{2}}(\eta_j^{-2} + \sigma_j^{-2}\sum_i w_{ij})^{-1}\zeta_j\kappa_j\right)^2 \le k_2 N^{-1}$$

for some constant k_2 implying that

$$\sum_j \left(N^{-\frac{1}{2}}(\eta_j^{-2} + \sigma_j^{-2}\sum_i w_{ij})^{-1}\zeta_j\kappa_j\right)^2 \le k_2\frac{K_m}{N} < \infty.$$

We have therefore established the convergence

$$\begin{aligned}
&N^{-1}\sum_j \left\{(\eta_j^{-2} + \sigma_j^{-2}\sum_i w_{ij})^{-1}(\eta_j^{-2}f_j + \sigma_j^{-2}\sum_i w_{ij}y_i)^2\right\} = \\
&N^{-1}\sum_j \left\{(\eta_j^{-2} + \sigma_j^{-2}\sum_i w_{ij})^{-1}\zeta_j^2\right\} + N^{-1}\left\{\sum_j(\eta_j^{-2} + \sigma_j^{-2}\sum_i w_{ij})^{-1}\kappa_j^2\right\} + \\
&N^{-\frac{1}{2}}o(N^{\frac{1}{\alpha}}), \text{ a.s.}
\end{aligned} \tag{A.5}$$

where we have replaced $o(K_m^{\frac{1}{\alpha}})$ by $o(N^{\frac{1}{\alpha}})$ noting that $K_m^{-\frac{1}{\alpha}}o(K_m^{\frac{1}{\alpha}}) \to 0$ implies $N^{-\frac{1}{\alpha}}o(K_m^{\frac{1}{\alpha}}) \to 0$ as $K_m \to \infty$ since $K_m/N < \infty$. Now (A.4) and (A.5) prove

$$N^{-1}\sum_{j:\boldsymbol{x}_j\in S_m} L_j(\boldsymbol{\theta},\boldsymbol{\sigma},\boldsymbol{\eta}) = \sum_{j:\boldsymbol{x}_j\in S_m} E\{L_j(\boldsymbol{\theta},\boldsymbol{\sigma},\boldsymbol{\eta})\} + N^{-\frac{1}{2}}o(N^{\frac{1}{\alpha}})$$

with probability one for all $\alpha \ge 1$. Summing the above equations over m and noting M is smaller than a variable of order $N^{\frac{1}{2}-\frac{1}{\alpha_0}}$ for some $\alpha_0 \ge 1$ yields the desired convergence.