Nonparametric Methods in Economics and Finance Lecture 3

Oliver Linton

May 10, 2006

Additive Models

- The curse of dimensionality means that it is hard to get good estimates of a high dimensional regression surface. The term curse of dimensionality actually comes from Bellman in the context of computing high dimensional dynamic programs but has been also by many different authors in this nonparametric statistical context.

- One way out of this is to assume a compromise between fully nonparametric specification and a fully parametric specification. A leading example if this is additive regression.

- According to Luce and Tukey (1964), additivity is basic to science. It is certainly hard to think of functions that are not additive in some sense, i.e., after transformations or relabelling of variables.

Theorem 0.1 (Kolmogorov (1957)) *There exist* d *constants* $\lambda_\alpha > 0,\ \alpha = 1, \ldots, d,\ \sum_{\alpha=1}^{d} \lambda_\alpha \leq 1,$ *and* $2d + 1$ *continuous strictly increasing functions* $\phi_k,$ $k = 1, \ldots, 2d + 1,$ *which map* $[0, 1]^d$ *to* $[0, 1]^d$ *and have the propery that for each continuous function* m *from* $[0, 1]^d$ *to* R

$$m(x_1, \ldots, x_d) = \sum_{k=1}^{2d+1} g\left(\sum_{\alpha=1}^{d} \phi_k(x_\alpha)\right)$$

for some function g *continuous on* $[0, 1]^d$.

- This simplifying structure is present in many models of economic behavior starting with Leontieff (1947); see Deaton and Muellbauer (1980) for examples.

- Additivity is also widely used in parametric and semiparametric models of economic data.

- A function $m(x)$ is additively separable if

$$m(x) = \sum_{\alpha=1}^{d} m_\alpha(x_\alpha)$$

  for some functions m_α.

- Estimation in these models was first discussed by Stone (1985,1986) who showed that the optimal rate for estimating $m(\cdot)$ is the one-dimensional rate of convergence e.g., $n^{2/5}$ for twice continuously differentiable functions. Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1991).

Model and Notation

- We suppose that one observes i.i.d. observations (X_i, Y_i) for $i = 1, \ldots, n$, where the response Y_i is real valued and where the covariates $X_i = (X_{1i}, \ldots, X_{di})$ take values in R^d. Define the regression function $m(x) = E(Y|X = x)$. Then the additive model can be written as

$$Y_i = c + m_1(X_{1i}) + \ldots + m_d(X_{di}) + \varepsilon_i,$$

where the error variables ε_i satisfy

$$E(\varepsilon_i|X_i) = 0 a.s.$$

We shall maintain throughout that

$$\text{var}(\varepsilon_i|X_i) = \sigma^2(X_i) < \infty a.s.$$

- The functions $m_1, \ldots, m_d$ and the constant c are unknown and have to be estimated by the data.

- For identifiability we make the additional assumption that

$$\int m_\alpha(x_\alpha) dQ_\alpha(x_\alpha) = 0$$

for $\alpha = 1, \ldots, d$, where Q_α is a signed measure. For example, Q_α could be the marginal distribution of X_α.

- Without this assumption replacing e.g. $m_\alpha(x_\alpha)$ by $m_\alpha(x_\alpha)+c_\alpha$, $\alpha = 1, \ldots d$, such that $\sum_{\alpha=1}^d c_\alpha = 0$ would not change the sum $\sum_{\alpha=1}^d m_\alpha(x_\alpha)$. So the model remains unchanged although the functions m_α changed. Such arbitrariness is eliminated by our assumption.

- It follows that $c = \int m(x) dQ(x)$, where Q is any measure for which Q_α are the marginals.

- We assume one further condition on the covariate distribution for identification of m_α. We suppose that for $\alpha = 1, \ldots, d$,

$$\sum_{\alpha=1}^{d} f_\alpha(X_\alpha) = 0 \text{ a.s.} \implies f_\alpha \equiv 0 \text{ a.s.},$$

- This rules out what is called 'concurvity' in Hastie and Tibshirani (1991). It is a generalization of the usual full rank condition on the cross product matrix in linear regression. It rules out not just linear functional relationships but also any non-linear functional relationships.

- Alternatively, one can normalize by taking

$$m_\alpha(x_{\alpha 0}) = 0,$$

say, for some $x_{\alpha 0}$ for each α.

- This might be convenient in some cases as particular values of the covariate might have special meaning, like zero input should produce zero output. We shall not pursue this further.

- Write for each α, $x = (x_\alpha, x_{-\alpha})$, $X = (X_\alpha, X_{-\alpha})$, and $X_i = (X_{\alpha i}, X_{-\alpha i})$. We shall suppose for simplicity that X are absolutely continuous with respect to Lebesgue measure on some set $\mathcal{X}$ (usually a compact subset of R^d) and have a density function $f(x)$, which has marginals $f_\alpha(x_\alpha)$ and $f_{-\alpha}(x_{-\alpha})$ for all α.

Marginal Integration

- This method is due to Linton and Nielsen (1995), who called it Marginal Integration, to Newey (1994), who called it Partial Mean, and to Tjøstheim and Auestad (1994), who called it Projection.

- Define

$$g_\alpha(x_\alpha) = \int m(x) dQ_{-\alpha}(x_{-\alpha}),$$

where $Q_{-\alpha}(x_{-\alpha})$ is a $d-1$ dimensional probability measure.

- It follows that

$$\begin{aligned} g_\alpha(x_\alpha) &= c + m_\alpha(x_\alpha) + \sum_{\gamma \neq \alpha}^{d} \int m_\gamma(x_\gamma) dQ_{-\alpha}(x_{-\alpha}) \\ &\equiv m_\alpha(x_\alpha) + \mu_\alpha \end{aligned}$$

so that $g_\alpha(x_\alpha)$ is equal to $m_\alpha(x_\alpha)$ upto an additive constant.

- The constants μ_α are determined by c and by the choice of measures $Q_{-\alpha}$. Since we have assumed that m_α are mean zero with respect to Q_α,

$$m_\alpha(x_\alpha) = g_\alpha(x_\alpha) - \int g_\alpha(x_\alpha) dQ_\alpha(x_\alpha).$$

- We use these relations to generate estimators. In practice we have to replace m by an unrestricted nonparametric regression estimator $\widehat{m}(x)$ (and perhaps the $Q_{-\alpha}, Q_\alpha$ by estimates $\widehat{Q}_{-\alpha}, \widehat{Q}_\alpha$ when they are unknown) and approximate the integral by some method. We then let

$$\begin{aligned} \widetilde{g}_\alpha(x_\alpha) &= \int \widehat{m}(x) d\widehat{Q}_{-\alpha}(x_{-\alpha}) \\ \widetilde{m}_\alpha(x_\alpha) &= \widetilde{g}_\alpha(x_\alpha) - \int \widetilde{g}_\alpha(x_\alpha) d\widehat{Q}_\alpha(x_\alpha) \end{aligned}$$

$$\widetilde{m}(x) = \widehat{c} + \sum_{\alpha=1}^{d} \widetilde{m}_\alpha(x_\alpha), \quad \widehat{c} = \int \widehat{m}(x) d\widehat{Q}(x).$$

- There are many choices for $\widehat{m}$ here. For example, the multivariate local constant estimator. Alternatively, one can use local polynomial estimators.

There are several common choices of weighting measure $Q_{-\alpha}$:

1. $\widehat{Q}_{-\alpha}$ is the empirical distribution $\widehat{F}_{-\alpha}$ of $X_{-\alpha i}$

2. $\widehat{Q}_{-\alpha}$ is the integral of a kernel estimate $\widehat{f}_{-\alpha}$ of the density of $X_{-\alpha i}$

3. $Q_{-\alpha}$ is the integral of some fixed density $q_{-\alpha}$ defined on a subset of the support of $X_{-\alpha i}$

- The common implementation of the integration method is computationally demanding. This is based on taking the empirical distribution of the covariates $X_{-\alpha}$, and involves computing

$$\widetilde{g}_\alpha(x_\alpha) = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(x_\alpha, X_{-\alpha i})$$

for each point of interest x_α.

- If we compute $\widetilde{m}_\alpha(x_\alpha)$ at each sample observation $X_{\alpha i}$ in effect one needs to compute $\widehat{m}(X_{\alpha j}, X_{-\alpha i})$ for $i, j = 1, \ldots, n$, i.e., one has to compute n different $n \times n$ smoothing matrices. We next discuss an alternative approach.

- Linton and Nielsen (1995) and Fan, Mammen, and Härdle (1998) consider the choice of optimal weighting.

Instrumental Variables

- Letting $\eta_i = \sum_{\gamma \neq \alpha} m_\gamma \left(X_{\gamma i} \right) + \varepsilon_i$, we rewrite the model as

$$Y_i = g_\alpha \left(X_{\alpha i} \right) + \eta_i = c + m_\alpha \left(X_{\alpha i} \right) + \eta_i,$$

which is a classical example of "omitted variable" regression.

- That is, although this appears to take the form of a univariate nonparametric regression model, smoothing Y on X_α will incur a bias due to the omitted variable η, because η contains $X_{-\alpha}$, which in general depends on X_α.

- One solution to this is suggested by the classical econometric notion of instrumental variable. That is, we look for an instrument Z such that

$$E \left(Z | X_\alpha \right) \neq 0 \quad ; \quad E \left(Z \eta | X_\alpha \right) = 0$$

with probability one.

- If such a random variable exists,

$$E\left(ZY|X_\alpha\right) = E\left(Z|X_\alpha\right) m_\alpha\left(X_\alpha\right)$$

so that

$$g_\alpha\left(x_\alpha\right) = \frac{E(ZY|X_\alpha = x_\alpha)}{E(Z|X_\alpha = x_\alpha)}.$$

- This suggests that we estimate the function $m_\alpha\left(\cdot\right)$ by nonparametric smoothing of ZY on X_α and Z on X_α.

- In parametric models the choice of instrument is usually not obvious and requires some caution. However, our additive model has a natural class of instruments $-$ $f_{-\alpha}\left(X_{-\alpha}\right)/f\left(X\right)$ times any measurable function of X_α will do.

- Suppose that we take

$$Z(X) = \frac{f_\alpha(X_\alpha) f_{-\alpha}\left(X_{-\alpha}\right)}{f\left(X\right)}.$$

- We have

$$\begin{aligned} & E\left(Z\eta|X_\alpha\right) \\ = \; & E\left(Z\left(\sum_{\gamma\neq\alpha} m_\gamma(X_\gamma)\right)|X_\alpha\right) \\ = \; & \int \frac{f_\alpha(X_\alpha)f_{-\alpha}\left(X_{-\alpha}\right)}{f\left(X\right)}\left(\sum_{\gamma\neq\alpha} m_\gamma(X_\gamma)\right)\frac{f(X)}{f_\alpha\left(X_\alpha\right)}dX \\ = \; & \sum_{\gamma\neq\alpha}\int m_\gamma(X_\gamma)f_{-\alpha}\left(X_{-\alpha}\right)dX_{-\alpha} \\ = \; & 0. \end{aligned}$$

- Furthermore, $E\left(Z|X_\alpha\right) = 1$ so that

$$g_\alpha\left(x_\alpha\right) = E(ZY|X_\alpha = x_\alpha).$$

- Of course, the distribution of the covariates is rarely known *a priori.* In practice, we have to rely on estimates of these quantities.

- Let $\widehat{f}(\cdot)$, $\widehat{f}_{\alpha}(\cdot)$, and $\widehat{f}_{-\alpha}(\cdot)$ be kernel estimates of the densities $f(\cdot)$, $f_{\alpha}(\cdot)$, and $f_{-\alpha}(\cdot)$, respectively. Then, the feasible procedure is defined by replacing the instrumental variable Z_i by

$$\widehat{Z}_i = \widehat{f}_{\alpha}(X_{\alpha i})\,\widehat{f}_{-\alpha}(X_{-\alpha i})\,/\widehat{f}(X_i)$$

and computing an internally normalized one dimensional smooth of $\widehat{Z}_i Y_i$ on $X_{\alpha i}$.

- Thus

$$\begin{aligned}
&\widetilde{g}_{\alpha}(x_{\alpha}) \\
&= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_{\alpha}-X_{\alpha i}}{h}\right)\frac{\widehat{Z}_i Y_i}{\widehat{f}_{\alpha}(X_{\alpha i})} \\
&= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_{\alpha}-X_{\alpha i}}{h}\right)\frac{\widehat{f}_{-\alpha}(X_{-\alpha i})}{\widehat{f}(X_i)}Y_i
\end{aligned}$$

as our estimate of $g_{\alpha}(x_{\alpha}) = c + m_{\alpha}(x_{\alpha})$.

- The main advantage that the local instrumental variable method has is in terms of the computational cost. There is a convenient matrix formula

for the IV estimator in the bivariate case

$$\widetilde{g}_1 = W_1(y. * (W_2 i)./n(W_1. * W_2)i),$$

where W_1 and W_2 are one-dimensional smoothing matrices, i is the n vector of ones, and $.*$ and $./$ denote element by element multiplication and division respectively.

- The marginal integration method needs n^2 regression smoothings evaluated at the pairs $(X_{\alpha i}, X_{-\alpha j})$, for $i, j = 1, \ldots, n$, while the backfitting method requires nr operations-where r is the number of iterations to achieve convergence. The instrumental variable procedure takes at most $2n$ operations of kernel smoothings in a preliminary step for estimating the instrumental variable, and another n operations for the regressions. Kim, Linton, and Hengartner (1999)].

- It has several interpretations in addition to the above instrumental variable estimate. First, as a version of the one-dimensional regression smoother but adjusting internally by a conditional density estimate

$$\widehat{f}_{\alpha|-\alpha}(X_{\alpha i}|X_{-\alpha i}) = \frac{\widehat{f}(X_{\alpha i}, X_{-\alpha i})}{\widehat{f}_{-\alpha}(X_{-\alpha i})},$$

instead of by a marginal density estimate.

- Second, one can think of it as a one-dimensional standard Nadaraya-Watson (externalized) regression smoother of the adjusted data $\widehat{Y}_i$ on $X_{\alpha i}$, where

$$\widehat{Y}_i = \widehat{f}_\alpha(x_\alpha)\widehat{f}_{-\alpha}(X_{-\alpha i})Y_i \Big/ \widehat{f}(X_{\alpha i}, X_{-\alpha i})\,.$$

- Finally, note that $\widetilde{g}_\alpha(X_{\alpha i})$ can be interpreted as a marginal integration estimator in which the pilot estimator is a fully internalized smoother [see

Jones, Davies and Park (1994)] and the integrating measure is the empirical covariate one

$$\widehat{m}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \frac{Y_i}{\widehat{f}(X_i)},$$

rather than the Nadaraya-Watson:

- by interchanging the orders of summation, we obtain

$$\begin{aligned}
& \widetilde{g}_\alpha(X_{\alpha i}) \\
= \; & \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{X_{\alpha i} - X_{\alpha j}}{h}\right) \frac{\widehat{f}_{-\alpha}(X_{-\alpha j})}{\widehat{f}(X_j)} Y_j \\
= \; & \frac{1}{nh} \sum_{j=1}^{n} \frac{K\left(\frac{X_{\alpha i} - X_{\alpha j}}{h}\right) Y_j}{\widehat{f}(X_j)} \left\{ \frac{1}{nh^{d-1}} \sum_{k=1}^{n} K\left(\frac{X_{-\alpha k} - \right.\right. \\
= \; & \frac{1}{n^2 h^d} \sum_{k=1}^{n} \sum_{j=1}^{n} \frac{K\left(\frac{X_{\alpha i} - X_{\alpha j}}{h}\right) K\left(\frac{X_{-\alpha k} - X_{-\alpha j}}{h}\right) Y_j}{\widehat{f}(X_j)} \\
= \; & \frac{1}{n} \sum_{k=1}^{n} \left\{ \frac{1}{nh^d} \sum_{j=1}^{n} K\left(\frac{X_{\alpha i} - X_{\alpha j}}{h}\right) K\left(\frac{X_{-\alpha k} - X_{-}\right.\right.
\end{aligned}$$

$$= \frac{1}{n} \sum_{k=1}^{n} \widehat{m}(X_{\alpha i}, X_{-\alpha k}),$$

where $\widehat{m}(X_{\alpha i}, X_{-\alpha k})$ is an internally normalized pilot smoother.

Backfitting

- Consider the population problem of finding functions $m(x) = \sum_{\alpha=1}^{d} m_\alpha(x_\alpha)$ to minimize the least squares criterion

$$Q = E\left[\{Y - m(X)\}^2\right],$$

where $E(Y^2) < \infty$.

- This can be characterized as a projection problem in Hilbert space. Let $\mathcal{H}$ be the Hilbert space of square integrable functions of X, then the regression function m is the function in $\mathcal{H}$ that minimizes Q. Define the subspace of additive functions

$$\mathcal{H}_{add} = \oplus_{\alpha=1}^{d} \mathcal{H}_\alpha \subset \mathcal{H},$$

that is the subspace of random variables $\sum_{\alpha=1}^{d} m_\alpha(X_\alpha)$ for square integrable m_α.

- Then the function that minimizes Q over $\mathcal{H}_{add}$, denoted $m^*(x) = \sum_{\alpha=1}^{d} m_\alpha(x_\alpha)$, satisfies the set of equations:

$$\begin{aligned} m_1(x_1) &= E(Y|X_1 = x_1) - m_0 - E[m_2(X_2)|X_1 = x \\ &\quad \ldots - E\left[m_d(X_d)|X_1 = x_1\right], \\ \vdots &= \vdots \\ m_d(x_d) &= E(Y|X_d = x_d) - m_0 - E\left[m_1(X_1)|X_d = x\right. \\ &\quad \ldots - E\left[m_{d-1}(X_{d-1})|X_d = x_d\right]. \end{aligned}$$

or more compactly

$$P_\alpha \{Y - m^*(X)\} = 0, \alpha = 1, \ldots, d,$$

where

$$P_\alpha(\cdot) = E\left(\cdot|X_\alpha\right)$$

is the projection operator on the subspace $\mathcal{H}_\alpha$.

- We can represent these first order conditions as in Hastie and Tibshirani (1990):

$$\begin{pmatrix} I & P_1 & \cdots & P_1 \\ P_2 & I & \cdots & P_2 \\ \vdots & & \ddots & \vdots \\ P_d & \cdots & P_d & I \end{pmatrix} \begin{pmatrix} m_1^* \\ m_2^* \\ \vdots \\ m_d^* \end{pmatrix}$$

$$= \begin{pmatrix} P_1 Y \\ P_2 Y \\ \vdots \\ P_d Y \end{pmatrix}.$$

- The sample analogue of the projection operator P_α is the sample smoothing matrix W_α (the n by n smoother matrix used in computing $\widehat{E}(\cdot|X_\alpha)$). Therefore, we have the corresponding sample first order condition

$$\begin{pmatrix} I & W_1 & \cdots & W_1 \\ W_2 & I & \cdots & W_2 \\ \vdots & & \ddots & \vdots \\ W_d & \cdots & W_d & I \end{pmatrix} \begin{pmatrix} \widetilde{m}_1 \\ \widetilde{m}_2 \\ \vdots \\ \widetilde{m}_d \end{pmatrix}$$

$$= \begin{pmatrix} W_1 y \\ W_2 y \\ \vdots \\ W_d y \end{pmatrix},$$

where $y = (Y_1, \ldots, Y_n)^\top$ and $\widetilde{m}_\alpha = (\widetilde{m}_\alpha(X_{\alpha 1}), \ldots, \widetilde{m}_\alpha$

- This is a linear system of large dimensions nd equations in nd unknowns

$$W\widetilde{m} = s$$

- The estimator $\widetilde{m}$ can then be defined through

$$\widetilde{m} = W^{-1}s$$

when this inverse exists. However, in practice the inversion of W is quite difficult when n is large. Opsomer and Ruppert (1997) recommended re-centering the smoothers so that we replace W_α by $W_\alpha^* = (I - ii^\top/n)W_\alpha$.

- In the bivariate case there is a simple solution

$$\begin{aligned}\widetilde{m}_1 &= \left\{I - (I - W_1^* W_2^*)^{-1}(I - W_1^*)\right\} y \\ \widetilde{m}_2 &= \left\{I - (I - W_2^* W_1^*)^{-1}(I - W_2^*)\right\} y\end{aligned}$$

provided the inverses exist. These only involve inverting $n \times n$ matrices.

- In practice, the backfitting (Gauss-Seidel) algorithm is often used instead. This is as follows

1. For each $\alpha = 1, \ldots, d$ compute $\widetilde{m}_\alpha^{[0]} = W_\alpha y$

2. For each $\alpha = 1, \ldots, d$ and $r = 1, 2, \ldots$

$$\widetilde{m}_\alpha^{[r]} = W_\alpha \left\{ y - \sum_{\gamma<\alpha} \widetilde{m}_\gamma^{[r-1]} - \sum_{\gamma>\alpha} \widetilde{m}_\gamma^{[r-1]} \right\}$$

3. Repeat until some convergence criterion is satisfied like the sum of squared residuals.

- SPLUS/R etc.

- Each step involves just one dimensional smoothing. The estimators are linear in y. There are some problems with this algorithm.

- A sufficient condition ($d = 2$) for convergence of Backfitting is if either $\|W_1 W_2\| < 1$ or if both W_1 and W_2 are symmetric e.g. cubic splines. approach is to iteratively solve empirical versions of the above equations, see Breiman and Friedman (1985), Buja, Hastie an Tibshirani (1989), and Hastie and Tibshirani (1991). Hastie and Tibshirani (1990).

- These estimators are computed at each observation point and so are quite computationally demanding as writ when n or d is large.

Smooth Backfitting

- Mammen, Linton, and Nielsen (1998) define backfitting estimates $\widetilde{m}_\alpha$ as the minimizers of the following empirical norm

$$\begin{aligned} & \|\widehat{m} - \overline{m}\|^2_{\widehat{f}} \\ = & \int [\widehat{m}(x) - \mu - \bar{m}_1(x_1) - \ldots - \bar{m}_d(x_d)]^2 \, \widehat{f}(x) dx, \end{aligned}$$

where the minimization runs over all functions $\overline{m}(x) = \mu + \sum_\alpha \bar{m}_\alpha(x_\alpha)$, with

$$\int \bar{m}_\alpha(x_\alpha) \widehat{f}_\alpha(x_\alpha) dx_\alpha = 0.$$

Here, $\widehat{f}(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i)$ is the density estimator with marginals $\widehat{f}_\alpha(x_\alpha) = \int \widehat{f}(x) dx_{-\alpha}$ [this is the one-dimensional kernel density estimate $\widehat{f}_\alpha(x_\alpha) = n^{-1} \sum_{i=1}^n K_h(x_\alpha - X_{\alpha i})$.], while $\widehat{m}(x)$ is the unrestricted Nadaraya-Watson estimator

$$\widehat{m}(x) = \frac{\sum_{i=1}^n K_h(x - X_i) Y_i}{\sum_{i=1}^n K_h(x - X_i)}$$

- A minimizer exists if the density estimate $\widehat{f}$ is non-negative. Equation means that

$$\widetilde{m}(x) = \widetilde{m}_0 + \widetilde{m}_1(x_1) + \ldots + \widetilde{m}_d(x_d)$$

is the projection in the space $\mathbf{L}_2(\widehat{f})$ of $\widehat{m}$ onto the subspace of additive functions

$$\{m \in \mathbf{L}_2(\widehat{f}) : m(x) = m_0 + m_1(x_1) + \ldots + m_d(x_d)\}.$$

This is a central point of our thesis. For projection operators backfitting is well understood (method of alternating projections, see below). Therefore, this interpretation will enable us to understand convergence of the backfitting algorithm and the asymptotics of $\widetilde{m}_\alpha$. Not every backfitting algorithm based on iterative smoothing can be interpreted as an alternating projection method.

- The solution is characterized by the following system of equations ($\alpha = 1, \ldots, d$):

$$\begin{aligned}\widetilde{m}_\alpha(x_\alpha) &= \int \widehat{m}(x)\frac{\widehat{f}(x)}{\widehat{f}_\alpha(x_\alpha)}dx_{-\alpha} \\ &\quad - \sum_{\gamma\neq\alpha}\int \widetilde{m}_\gamma(x_\gamma)\frac{\widehat{f}(x)}{\widehat{f}_\alpha(x_\alpha)}dx_{-\alpha} - \widetilde{m}_0\end{aligned}$$

$$0 = \int \widetilde{m}_\alpha(x_\alpha)\widehat{f}_\alpha(x_\alpha)dx_\alpha.$$

- Straightforward algebra gives

$$\begin{aligned}\int \widehat{m}(x)\frac{\widehat{f}(x)}{\widehat{f}_\alpha(x_\alpha)}dx_{-\alpha} &= \frac{n^{-1}\sum_{i=1}^n K_h(x_\alpha - X_{\alpha i})Y_i}{\widehat{f}_\alpha(x_\alpha)} \\ &\equiv \widehat{m}_\alpha(x_\alpha),\end{aligned}$$

because of

$$\int \prod_{\ell\neq\alpha} K_h(x_\ell - X_{\ell i})dx_{-\alpha} = 1,$$

where $\widehat{m}_\alpha(x_\alpha)$ is exactly the corresponding univariate Nadaraya-Watson estimator.

- Furthermore,

$$\widetilde{m}_0 = \int \widehat{m}(x)\widehat{f}(x)dx,$$

and because of $\int \prod_{\ell=1}^{d} K_h(x_\ell - X_{\ell i})dx_{-\alpha} = 1$, we find, as in Hastie and Tibshirani (1991), that

$$\widetilde{m}_0 = n^{-1}\sum_{i=1}^{n} Y_i.$$

Therefore, $\widetilde{m}_0$ is a n-consistent estimate of the population mean and the randomness from this estimation is of smaller order and can be effectively ignored.

- Note also that

$$\widetilde{m}_0 = \int \widetilde{m}_\alpha(x_\alpha)\widehat{f}_\alpha(x_\alpha)\ dx_\alpha \quad \alpha = 1, \ldots, d.$$

- We therefore define a backfitting estimator $\widetilde{m}_\alpha(x_\alpha)$, $\alpha = 1, \ldots, d$, as a solution to the system of equations $[\alpha = 1, \ldots, d]$

$$\begin{aligned} & \widetilde{m}_\alpha(x_\alpha) \\ = \; & \widehat{m}_\alpha(x_\alpha) - \sum_{\gamma \neq \alpha} \int \widetilde{m}_\gamma(x_\gamma) \frac{\widehat{f}(x)}{\widehat{f}_\alpha(x_\alpha)} dx_{-\alpha} - \widetilde{m}_0, \end{aligned}$$

$$0 = \int \widetilde{m}_\alpha(x_\alpha) \widehat{f}_\alpha(x_\alpha) dx_\alpha.$$

with $\widetilde{m}_0 = n^{-1} \sum_{i=1}^n Y_i$

- Upto now we have assumed that multivariate estimates of the density and of the regression function exist for all x. This assumption is not reasonable for large dimensions d (or at least such estimates can perform very poorly). Furthermore, this assumption is not necessary.

- Note that can be rewritten as

$$\begin{aligned} & \widetilde{m}_\alpha(x_\alpha) \\ = \; & \widehat{m}_\alpha(x_\alpha) - \sum_{\gamma \neq \alpha} \int \widetilde{m}_\gamma(x_\gamma) \frac{\widehat{f}_{\alpha,\gamma}(x_\alpha, x_\gamma)}{\widehat{f}_\alpha(x_\alpha)} dx_\gamma - \widetilde{m}_0, \end{aligned}$$

  where

$$\widehat{f}_{\alpha,\gamma}(x_\alpha, x_\gamma) = n^{-1} \sum_{i=1}^{n} K_h(x_\alpha - X_{\alpha i}) K_h(x_\gamma - X_{\gamma i})$$

  is the two-dimensional marginal of the full dimensional kernel density estimate $\widehat{f}(x)$. In this equation only one and two dimensional marginals of $\widehat{f}$ are used.

- The integrals are computed numerically.

- The estimator can be computed on a grid of points in the covariate support $I_1 \times \cdots \times I_d$ so it does not need residuals as in the standard backfitting approach.

- This estimator has been called smooth backfitting by Nielsen and Sperlich (2005).

- In practice, our backfitting algorithm works as follows. One starts with an arbitrary initial guess $\widetilde{m}_\alpha^{[0]}$ for $\widetilde{m}_\alpha$; for example, $\widetilde{m}_\alpha^{[0]} = \widehat{m}_\alpha$ or $\widetilde{m}_\alpha^{[0]}$ is the marginal integration estimator of Linton and Nielsen (1995). In the α-th step of the r-th iteration cycle one puts

$$\begin{aligned} & \widetilde{m}_\alpha^{[r]}(x_\alpha) \\ = \; & \widehat{m}_\alpha(x_\alpha) - \sum_{\gamma<\alpha} \int \widetilde{m}_\gamma^{[r]}(x_\gamma) \frac{\widehat{f}_{\alpha,\gamma}(x_\alpha, x_\gamma)}{\widehat{f}_\alpha(x_\alpha)} dx_\gamma \\ & - \sum_{\gamma>\alpha} \int \widetilde{m}_\gamma^{[r-1]}(x_\gamma) \frac{\widehat{f}_{\alpha,\gamma}(x_\alpha, x_\gamma)}{\widehat{f}_\alpha(x_\alpha)} dx_\gamma - \widetilde{m}_0, \end{aligned}$$

and the process is iterated until a desired convergence criterion is satisfied.

- Upto now we have implicitly assumed that the support of X is unbounded or at least that the density approaches zero at the boundary suitably fast.

- We now consider a generalization of the method which takes care of the boundary effects that are present when the densities have compact support. We suppose that

$$K_h(u,v) = \mathbf{1}(u,v \in [0,1]) \frac{K_h(u-v)}{\int_0^1 K_h(w-v)\, dw}$$

with, again, $K_h(u) = h^{-1}K(h^{-1}u)$. We will suppose that the kernel K has compact support $[-C_1, C_1]$, see B1. For this reason we get that

$$K_h(u,v) = K_h(u-v)$$

for $v \in [C_1h, 1-C_1h]$ or for $u \in [2C_1h, 1-2C_1h]$.

- So $K_h(u,v)$ differs from $K_h(u-v)$ only on the boundary.

- The norming gives that

$$\int_0^1 K_h(u,v)\ du = 1.$$

Therefore we have that:

$$\begin{aligned}\int_0^1 \widehat{f}_{\alpha,\gamma}(x_\alpha, x_\gamma)\ dx_\gamma &= \widehat{f}_\alpha(x_\alpha) \\ \int_0^1 \widehat{f}_\alpha(x_\alpha)\ dx_\alpha &= 1.\end{aligned}$$

Interpretation

- The backfitting method has a nice population interpretation as the projection of the given function $m(x)$ on to the space of additive functions where the norm is in terms of the expectation. This interpretation says what is happening when the additive model is not true, and it also turns out to be key in establishing efficiency properties. Here we present a similar interpretation of marginal integration.

- Let

$$\begin{aligned} & \pi_I(m)(x) \\ = & \int m(x) dQ_{-\alpha}(x_{-\alpha}) dQ_\alpha(x_\alpha) \\ & + \sum_{\alpha=1}^{d} \Big[\int m(x) dQ_{-\alpha}(x_{-\alpha}) \\ & - \int m(x) dQ_{-\alpha}(x_{-\alpha}) dQ_\alpha(x_\alpha) \Big] \end{aligned}$$

  be the integration 'map', that takes a function $m(x)$ in the space of additive functions.

- It is easy to see that π_I is a linear idempotent map from $\mathcal{H}$ into itself,

$$\pi_I^2 = \pi_I$$

$$\pi_I(am + bm') = a\pi_I(m) + b\pi_I(m')$$

and moreover

$$\pi_I(m) = m \text{ if } m \in \mathcal{H}_{add}.$$

- However, π_I is not self-adjoint, i.e., it is not an orthogonal projection with respect to the norm induced by expectation with respect to the joint distribution of the covariates.

- However, if we change the definition of norm on the space $\mathcal{H}$ we can find an interpretation of π_I as an orthogonal projection. Specifically, if distance is calculated by the product measure

$$\otimes_{\alpha=1}^{d} Q_\alpha,$$

then π_I is self-adjoint, and hence an orthogonal projection, Nielsen and Linton (1998). In other words we can consider $\pi_I(m)$ as the solution to the minimization problem

$$\int \left\{ m(x) - c - \sum_{\alpha=1}^{d} m_\alpha(x_\alpha) \right\}^2 d\mu(x),$$

where $\mu = \otimes_{\alpha=1}^{d} Q_\alpha$.

- Thus far, the choice between integration and backfitting is reminiscent of the choice between ordinary least squares and generalized least squares in regression. The latter estimator finds the closest linear approximation to the regression function in the covariance matrix norm, while the former method finds the closest linear approximation in the unweighted Euclidean norm, see Drygas (1970). Although generalized least squares is more efficient in the Gauss-Markov sense when the linear structure is true, the efficiency gain may not be huge and the estimator can be harder to compute.

- Finally, we point out that under quite reasonable conditions, the solutions to the minimization are continuous (in the supremum norm) in the weighting function μ, so that small changes in weighting produce small differences in the fitted functions. In fact, there is a non-infinitesimal bound available for general weight functions, see Cleveland (1971).