Advanced Statistical Modeling

Part 2. Nonparametric Modeling

Session 3:

Nonparametric regression model III

Pedro Delicado

Departament d'Estadística i Investigació Operativa

Universitat Politècnica de Catalunya

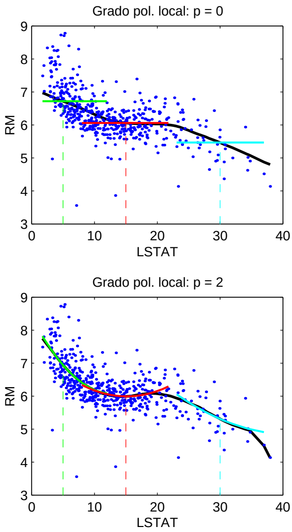
Choosing the degree q of the local polynomial

- The effect on the final estimation of the choice of the local polynomial degree, is much less important than the effect of the bandwidth choice.
- The larger is q the better are the asymptotic properties (in bias) but in practice it is recommended to use $q = s + 1$, where s is the order of the derivative of $m(x)$ that is estimated.
- When estimating $m(t)$, it is preferable to use the odd degree $q = 2k + 1$ than the preceding even degree $2k$.
- Among other advantages of local polynomials with odd degree, they are able to automatically adapt to the boundary of the explanatory variable support (when it is not the whole real line).

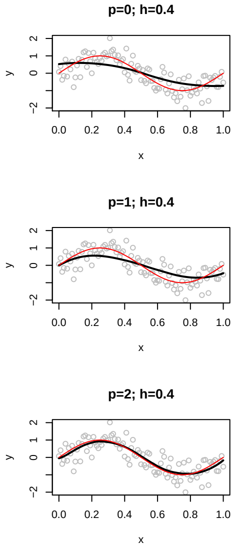
- To decide if it is worth fitting a local cubic model ($p = 3$) instead of just fitting a local linear model ($p = 1$), we must take into account the asymptotic expression of the local linear estimator bias:

$$\text{Bias}(\hat{m}_1(t)) = \frac{m''(t)}{2} h^2 \mu_2(K) + o(h^2).$$

- Bias is high for t in intervals where the function $m(t)$ has high curvature: large values of $|m''(t)|$.
- Therefore, if we suspect that the regression function $m(t)$ could be very bumpy it would be better to use $p = 3$ instead of $p = 1$.


Grado pol. local: p = 0
Grado pol. local: p = 1
Grado pol. local: p = 2
Grado pol. local: p = 3
RM
LSTAT
0
10
20
30
40
3
4
5
6
7
8
9

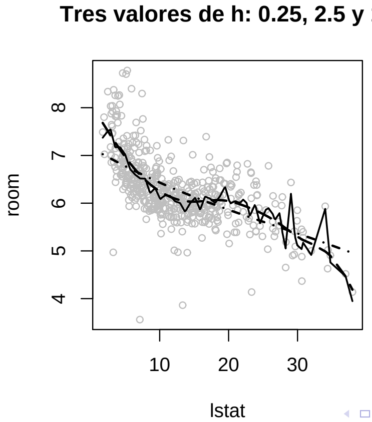
Effect of degree p on a single sample



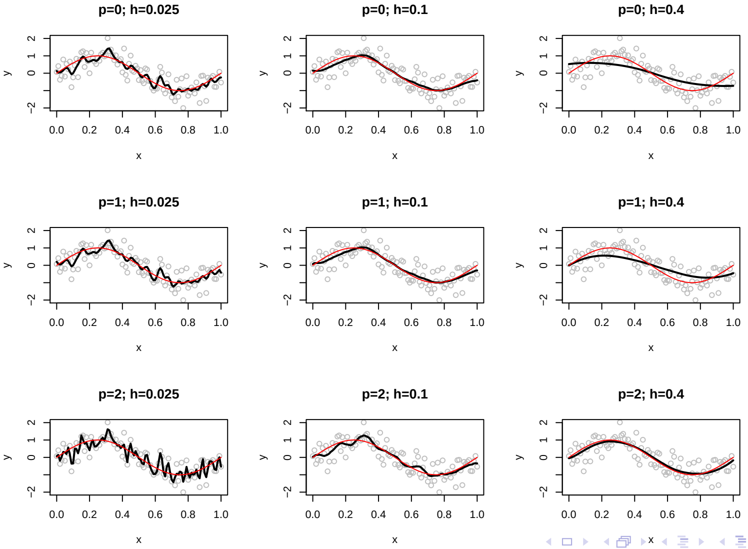
Bandwidth choice

The choice of smoothing parameter h is of crucial importance in the appearance and properties of the regression function estimator.

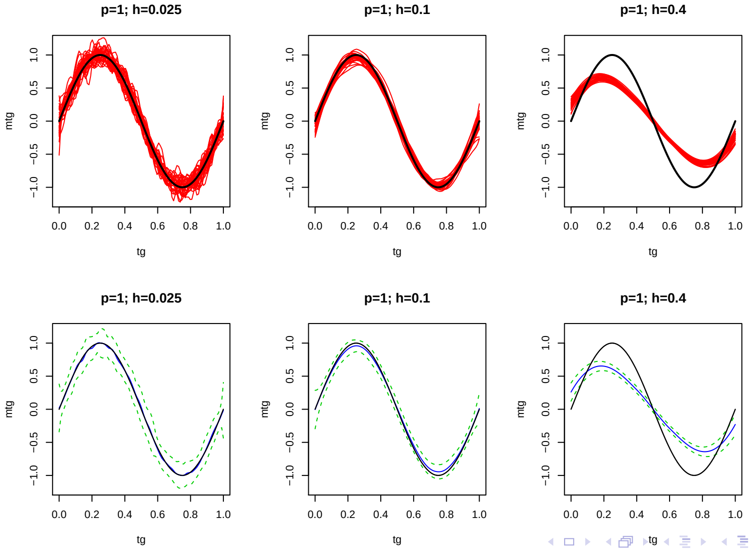
Example: Boston housing data. Local linear fit with Gaussian kernel.



Effect of bandwidth *h* and degree *p* on a single sample



Effect of bandwidth h on many samples



- The smoothing parameter h controls the balance between fitting well the observed data and the ability to predict future observations.
- Small values of h give great flexibility to the estimator and allow it to approach all the observed data (when h tends to 0 the estimator tends to interpolate the data), but the prediction errors will be high. There is overfitting.
- If h is too large, there is underfitting, as may occur with global parametric models. In this case both, the errors in the observed sample as well as the prediction errors in independent data, will be high.
- The bandwidth also controls the bias-variance trade-off.
- For h small the estimator is highly variable (applied to different samples from the same model gives very different results) and has small bias (the average of the estimators obtained for different samples is approximately the true regression function). If h is large the opposite happens.

Choosing the degree of the local polynomial

Choosing the smoothing parameter

- Global measures of fitting quality
- Bandwidth choice
- Variable bandwidth

Bandwidth choice: According to which criterion?

Several criteria are sensible. They represent global measures of fitting quality, or they are related with prediction error for new observations.

- Integrated Mean Squared Error (IMSE). A first global measure of the error made when using the nonparametric estimator $\hat{m}(t)$, $t \in [a, b]$, as an estimation of function $m(t)$, $t \in [a, b]$:

$$\text{IMSE}(\hat{m}) = \int_a^b E_{\mathbf{Z}}\left((\hat{m}(t) - m(t))^2\right) f(t)dt = \int_a^b \text{MSE}(\hat{m}(t))f(t)dt,$$

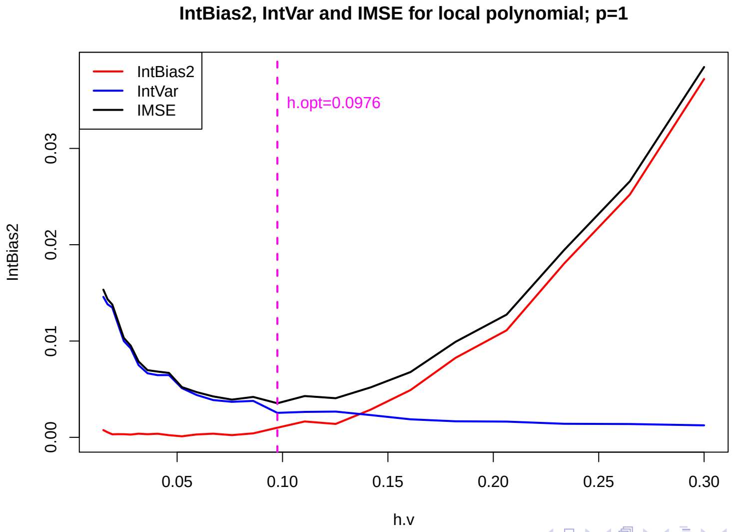
  where $\mathbf{Z} = \{(x_i, Y_i) : i = 1, \ldots, n\}$ is the sample used to compute $\hat{m}$.

- Mean Integrated Squared Error (MISE). It coincides with the IMSE:

$$\text{MISE}(\hat{m}) = E_{\mathbf{Z}}\left(\int_a^b (\hat{m}(t) - m(t))^2 f(t)dt\right) \overset{\text{Fubini's Theorem}}{=}$$

$$\int_a^b E_{\mathbf{Z}}\left((\hat{m}(t) - m(t))^2\right) f(t)dt = \text{IMSE}(\hat{m}).$$

Integrated Variance, integrated squared Bias and IMSE as a function of h



- Asymptotic Mean Integrated Squared Error (AMISE). Is the integrated value of the main part in the asymptotic expression of $\text{MSE}(\hat{m}(t))$, with respect to the density $f(t)$ of the explanatory variable.
- Example: For the local linear estimator with constant variance, it is

$$\text{AMISE}(\hat{m}) = \int_a^b \text{AMSE}(t) f(t) dt =$$

$$\int_a^b \frac{(m''(t))^2}{4} h^4 \left(\int_{-1}^1 u^2 K(u) du \right)^2 f(t) dt + \int_a^b \frac{\sigma^2}{nh} \int_{-1}^1 K^2(u) du dt =$$

$$\frac{h^4 \mu_2^2(K)}{4} \int_a^b (m''(t))^2 f(t) dt + \frac{R(K)\sigma^2}{nh}(b-a).$$

- Consider again MISE:

$$\text{MISE}(\hat{m}) = E_{\mathbf{Z}}\left(\int_a^b (\hat{m}(t) - m(t))^2 f(t)dt\right) = E_{\mathbf{Z}}\left[E_T\{(\hat{m}(T) - m(T))^2 \mid \mathbf{Z}\}\right],$$

  where $\mathbf{Z} = \{(x_i, Y_i) : i = 1, \ldots, n\}$ is the sample used to compute $\hat{m}$, and T is a random variable independent from $\mathbf{Z}$, with the same distribution that generates the independent variable values x_i, $i = 1, \ldots, n$.

- Average Squared Error (ASE). In the definition of MISE, the expectation with respect to $\mathbf{Z}$ is eliminated, and the expectation with respect to T is replaced by the average over the observed regressor values x_i, $i = 1, \ldots, n$, that are distributed as T:

$$\text{ASE}(\hat{m}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{m}(x_i) - m(x_i))^2.$$

  Usually the ASE is unknown (as well as MISE and AMISE) because it depends on the unknown function m.

- Residual Sum of Squares (RSS). It is an attempt to obtain a feasible version of ASE.
- In the expression of ASE, the unknown values $m(x_i)$ are replaced by the observed $y_i = m(x_i) + \varepsilon_i$:

$$\text{RSS}(\hat{m}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{m}(x_i))^2.$$

  It is also known as error in the training sample.
- Observe that in the definition of RSS, the data are used twice: first they are used to compute $\hat{m}$, and then they are used to evaluate if $\hat{m}$ is a good estimator of m.
- This is the reason why RSS is an optimistic version of ASE or MISE: remember that in MISE, the data $\mathbf{Z}$ used to compute $\hat{m}$ and those used to evaluate it, T, are independent.
- This independence is no longer maintained when using RSS.
- More important: The random quantities y_i replacing $m(x_i)$ in RSS also depend on $\mathbf{Z}$.

- Predictive Mean Square Error (PMSE). It is the expected squared error made when predicting

$$Y = m(t) + \varepsilon$$

by $\hat{m}(t)$, where t is an observation of the random variable T, distributed as the observed explanatory variable, when T and ε are independent from the sample $\mathbf{Z}$ used to compute $\hat{m}$. Then

$$\mathrm{PMSE}(\hat{m}) = E_{\mathbf{Z},T,\varepsilon}\left[(Y - \hat{m}(T))^2\right] = E_{\mathbf{Z},T,\varepsilon}\left[(\hat{m}(T) - m(T) - \varepsilon)^2\right] =$$

$$E_{\mathbf{Z},T}\left[(\hat{m}(T) - m(T))^2\right] + E_{\varepsilon}(\varepsilon^2) = \mathrm{MISE}(\hat{m}) + \sigma^2.$$

- Observe that MISE and PMSE are equivalent criteria for evaluating a nonparametric estimator.

- Predictive Average Square Error (PASE), also known as average squared prediction error in sample.
- Suppose we draw a new observation $Y_i^* = m(x_i) + \varepsilon_i^*$ at each observed x_i and previously we have made the prediction of Y_i^* by $\hat{m}(x_i)$. Then

$$\text{PASE}(\hat{m}) = \frac{1}{n}\sum_{i=1}^{n} E_{Y_i^*}(Y_i^* - \hat{m}(x_i))^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{m}(x_i) - m(x_i))^2 + \sigma^2 = \text{ASE}(\hat{m}) + \sigma^2.$$

- Observe that ASE and PASE are equivalent criteria for evaluating a nonparametric estimator.

- We have seen several global measures indicating whether a nonparametric estimator $\hat{m}$ is good or not for estimating an unknown regression function m.
- It is equivalent measuring closeness between $\hat{m}$ and m or prediction errors.
- These measures could be used in the model selection process.
- Bandwidth choice is in fact a model selection process.
- Unfortunately these measures are unfeasible because they depend on unknown functions or quantities.
- The only exception is RSS, that is optimistically biased.
- We will see now how to obtain feasible versions of these criteria.

Prediction error in a validation set

- When the amount of available data is high (as it usually happens in data mining or in Big Data problems) the sample is randomly divided in three sets:
  - The training set: it is used to fit the model.
  - The validation set: it is used to compute feasible versions of the above criteria for model selection.
  - The test set: it is used to evaluate the generalization (or prediction) error of the final chosen model.
- Assuming that at least a validation set has been preserved, an estimation of PMSE is the Predictive Mean Squared Error in the validation set:

$$\mathrm{PMSE}_{\mathrm{Val}}(h) = \frac{1}{n_V} \sum_{i=1}^{n_V} (y_i^V - \hat{m}(x_i^V))^2,$$

  where (x_i^V, y_i^V), $i = 1, \ldots, n_V$, is the validation set and $\hat{m}(x)$ is the estimator computed with bandwidth h using the training set.
- $h_{\mathrm{Val}} = \arg\min_h \mathrm{PMSE}_{\mathrm{Val}}(h)$.

Leave-one-out cross-validation

- When the sample size does not allow us to set a validation set aside, leave-one-out cross-validation is an attractive alternative.
- Remove the observation (x_i, y_i) from the sample and fit the nonparametric regression using the other $(n-1)$ data. Let $\hat{m}_{(i)}(x)$ be the resulting estimator.
- Now use $\hat{m}_{(i)}(x_i)$ to predict y_i.
- Repeat for $i = 1, \ldots, n$.
- For any bandwidth candidate value h, compute

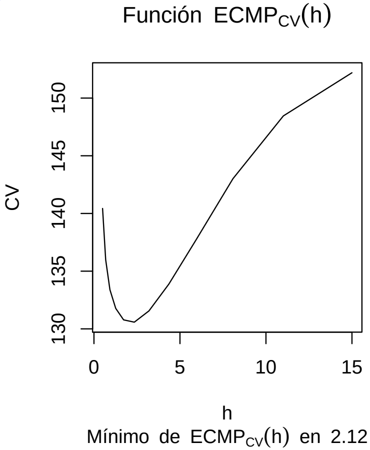
$$\text{PMSE}_{\text{CV}}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{m}_{(i)}(x_i))^2.$$

- $h_{\text{CV}} = \arg\min_h \text{PMSE}_{\text{CV}}(h)$.
- $\text{PMSE}_{\text{CV}}(h)$ is an approximately unbiased estimator of $\text{PMSE}(h)$, but has a considerable variance.

- The variance can be reduced doing K-fold cross-validation: The sample is randomly divided in K subsets, each of them is removed by turns from the sample, the model is estimated with the other $(K-1)$ subsamples and the removed subsample is used to compute prediction errors.
- N-fold cross-validation is leave-one-out cross-validation.
- K-fold cross-validation has lower variance than leave-one-out cross-validation but larger bias.
- General recommendation: Use 5-fold or 10-fold cross-validation.

Example. Leave-one-out cross-validation

$\text{PMSE}_{\text{CV}}(h)$ as a function of h in the example of local linear regression of `ROOM` against `LSTAT`.



Efficient computation of PMSE_{CV}

- Consider a linear smoother: $\hat{y}_i = \sum_{j=1}^{n} w(x_i, x_j) y_j$.
- In matrix formulation: $\hat{Y} = SY$, with S the smoothing matrix.
- In these cases PMSE_{CV} can be calculated avoiding the computational cost of fitting n different nonparametric regressions.
- It can be proved that

$$\text{PMSE}_{\text{CV}}(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - s_{ii}} \right)^2 .$$

See Wood (2006), pages 169-170.

Generalized cross-validation

- For linear smoothers a modification can be done in the measure of PMSE_{CV}.
- It is known as generalized cross-validation (GCV).
- It consists in replacing in the expression of $\text{PMSE}_{\text{CV}}(h)$ the values s_{ii}, coming from the diagonal of S, by their average value:

$$\text{PMSE}_{\text{GCV}}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-\nu/n}\right)^2,$$

  $\nu = \text{Trace}(S) = \sum_{i=1}^{n} s_{ii}$ is the effective number of parameters.
- $h_{\text{GCV}} = \arg\min_h \text{PMSE}_{\text{GCV}}(h)$.
- Manipulating the expression of $\text{PMSE}_{\text{GCV}}(h)$ it follows that

$$\text{PMSE}_{\text{GCV}}(h) = \frac{n\hat{\sigma}_\varepsilon^2}{n-\nu},$$

  where $\hat{\sigma}_\varepsilon^2 = \frac{1}{n-\nu}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ estimates the residual variance.

Plug-in bandwidth choice in the local linear estimator

- We have obtained before that for the local linear fit

$$\text{AMISE}(\hat{m}) = \frac{h^4 \mu_2^2(K)}{4} \int_a^b (m''(x))^2 f(x) dx + (b-a) \frac{R(K)\sigma^2}{nh}.$$

- The value of h minimizing this expression is

$$h_0 = \left(\frac{R(K)\sigma^2}{\mu_2^2(K) \int_a^b (m''(x))^2 f(x) dx} \right)^{1/5} n^{-1/5}.$$

- Some quantities there are unknown: the expected value of $(m''(X))^2$ and σ^2.
- h_{PI}: Replacing the unknowns by estimations.

Estimating $E[(m''(X))^2]$ and σ^2.

- In order to estimate $\int_a^b (m''(x))^2 f(x)dx = E[(m''(X))^2]$ a local cubic polynomial regression can be fitted, using weights $w(x_i, t) = K((x_i - t)/g)$, where the bandwidth g must be chosen.
- Once $m''(t)$ has been estimated for $t = x_1, \ldots, x_n$, $E[(m''(X))^2]$ is estimated as $\frac{1}{n}\sum_{i=1}^n (\hat{m}_g''(x_i))^2$.
- The optimal value of g for estimating the second derivative of $m(x)$ is

$$g_0 = C_2(K)\left(\frac{\sigma^2}{|\int_a^b m''(x)m^{(iv)}(x)f(x)dx|}\right)^{1/7} n^{-1/7}.$$

- At this point the estimation of $m''(x)$ and $m^{(iv)}(x)$ is done dividing the range of the explanatory variable in subintervals (4, for instance) and fitting a degree 4 polynomial at each subinterval.
- This last step also provides another estimation of σ^2.

Asymptotic behavior of bandwidth selectors of h

- We have seen three bandwidth selectors that do not require a validation set: h_{CV}, h_{GCV} and h_{PI}.
- The three methods provide bandwidths that converge to the value h_0 minimizing the AMISE when n goes to infinity, but their rates of convergence are different:

$$\frac{h_{\mathrm{CV}}}{h_0} - 1 = O_p(n^{-1/10}), \ \frac{h_{\mathrm{GCV}}}{h_0} - 1 = O_p(n^{-1/10}), \ \frac{h_{\mathrm{PI}}}{h_0} - 1 = O_p(n^{-2/7}).$$

Variable bandwidth

- The expression of the bandwidth h_{AMSE} minimizing the asymptotic mean square error, AMSE, of $\hat{m}(t)$ as an estimator of $m(t)$ is

$$h_{\text{AMSE}}(t) = \left(\frac{R(K)\sigma^2(t)}{\mu_2^2(K) f(t) (m''(t))^2} \right)^{1/5} n^{-1/5}.$$

- This expression suggests that sometimes it could be better to use different bandwidth at different points t.
- Variable bandwidth. The bandwidth depends on the point t where the function is being estimated: $h(t)$.

When is it recommended to use a variable bandwidth?

- When the density of the explanatory variable varies considerably along the support of the explanatory variable (in areas with much data the bandwidth can be smaller than in areas where there are few observations).
- When the residual variance is a function of the explanatory variable (in areas with great residual variability it is recommended to use large values of the window).
- When the curvature of the regression function is different in different parts of the support of the explanatory variable (in areas where curvature is larger, smaller values of h should be used).

How to define a variable bandwidth in practice?

- The most common way to include a variable bandwidth is to fix the proportion s of data points to be used in the estimation of each value $m(t)$ and define $h(t)$ such that the number of data (x_i, y_i) with x_i belonging the interval $(t - h(t), t + h(t))$ is sn. The ratio s is called *span*.
- If a local polynomial of degree $q = 0$ is fitted (Nadaraya-Watson estimator) using the uniform kernel and choosing $s = k/n$, the resulting estimator is known as the *k-nearest neighbours* estimator. The choice of s (or $k = sn$) can be done by cross-validation or using a validation set.

Practice:

Bandwidth choice

Fan, J. and I. Gijbels (1996).
Local polynomial modelling and its applications.
London: Chapman & Hall.

Hastie, T. J. and R. J. Tibshirani (1990).
Generalized additive models.
Monographs on Statistics and Applied Probability. London: Chapman and Hall Ltd.

Wand, M. P. and M. C. Jones (1995).
Kernel smoothing.
London: Chapman and Hall.

Wasserman, L. (2006).
All of Nonparametric Statistics.
New York: Springer.

Wood, S. (2006).
Generalized Additive Models: An Introduction with R.
Chapman and Hall/CRC.