Advanced Statistical Modeling

Part 2. Nonparametric Modeling

Session 7: Generalized additive models and Semiparametric models

Pedro Delicado

Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

Multiple nonparametric regression

- The extension of the nonparametric regression model to the case in which there are p explanatory variables is straightforward:

$$y_i = m(x_{i1}, \dots, x_{ip}) + \varepsilon_i,$$

  with $E(\varepsilon_i) = 0$ and $V(\varepsilon_i) = \sigma^2$, for $i = 1, \dots, n$.
- The regression function m indicates how y varies depending on the p-dimensional explanatory variable $\mathbf{x} = (x_1, \dots, x_p)$.
- To define the local polynomial estimator of the regression function $m(\mathbf{x})$ we need, first, to define the weights w_i for each observation, and secondly, to specify which explanatory variables are included at each local linear regression model.

Defining the weights w_i

- When estimating $m(\mathbf{t})$, with $\mathbf{t} = (t_1, \dots, t_p)$, data $(y_i; x_{i1}, \dots, x_{ip})$ with $\mathbf{x_i} = (x_{i1}, \dots, x_{ip})$ closer to $\mathbf{t} = (t_1, \dots, t_p)$ should have greater weight than those data that are further.
- Now distances between $\mathbf{t}$ and $\mathbf{x_i}$ are measured in a p-dimensional space and there are many sensible ways to define distances in such space.
- A way to define weights w_i with good performance in practice is

$$w_i = w(\mathbf{t}, \mathbf{x}_i) \propto \prod_{j=1}^{p} K\left(\frac{x_{ij} - t_j}{h_j}\right),$$

where K is a univariate kernel function, h_j is a smoothing parameter well suited for variable j-th, and symbol $\propto$ means proportionality.

Explanatory variables at each local linear model

- To fit a degree q polynomial depending on p-variables, all possible terms with the form

$$\beta_{s_1 \cdots s_p} \prod_{j=1}^{p} (x_{ij} - t_j)^{s_j},$$

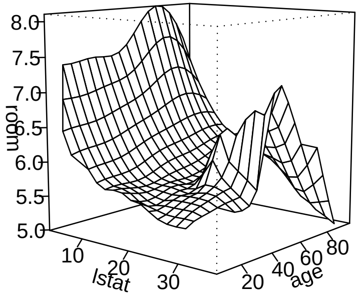
with degree $\sum_{j=1}^{p} s_j \leq q$, must be included.
- The estimate of $m(\mathbf{t})$ will be the intercept of the local polynomial fitted around point $\mathbf{t}$: $\hat{m}(\mathbf{t}) = \hat{m}(t_1, \ldots, t_p) = \hat{\beta}_{0\ldots0}$.
- Example. Two explanatory variables, local polynomial with degree 2,

$$\beta_{00} + \beta_{10}(x_{i1} - t_1) + \beta_{01}(x_{i2} - t_2) + \beta_{11}(x_{i1} - t_1)(x_{i2} - t_2) + \beta_{20}(x_{i1} - t_1)^2 + \beta_{02}(x_{i2} - t_2)^2.$$

The estimate of m at $\mathbf{t} = (t_1, t_2)$ will be $\hat{\beta}_{00}$.

Example: Boston housing data, bivariate regression.

- Nonparametric fit of `ROOM` as a function of `LSTAT` and `AGE`.
- `AGE`: For each neighborhood, the percentage of houses built before 1940.
- A kernel product of two univariate Gaussian kernels is used.
- Smoothing parameters: $h_1 = 2.5$ for `LSTAT`, and $h_2 = 10$ for `AGE`.



Practice:

Bivariate non-parametric regression with library **sm**

The curse of dimensionality

- A general problem in multivariate nonparametric estimation, and in particular in nonparametric regression with multiple regressors, is the phenomenon known as curse of dimensionality:

  In high dimensional spaces the neighborhood of any point $\mathbf{t}$ contains virtually no observational data.

- To construct a ball centered at a point $\mathbf{x_0} \in \mathbb{R}^d$ that contains say the 25% of the observed points, the ball must be so large that we can hardly say that it represents a neighborhood of $\mathbf{x_0}$.
- Let $X \sim U([-1,1]^d)$.

Dimension p	1	2	5	10
$P(X \in B_p(\mathbf{0_p}, 1))$	1	0.79	0.16	0.0025

- One way to overcome this problem is to work with extremely large sample sizes.

  Attention: For some problems, to have $842,000$ data in dimension 10 is really like to have 4 data in dimension 1.
- Therefore it is recommended not to go beyond 3 or 4 dimensions.
- For an explanatory variable in $\mathbb{R}^p$, it can be proved that local linear regression has $\text{AMSE}_0 = O(n^{-4/(4+p)})$.
- The higher the dimension p of explanatory variable, the lower the precision with which the regression function is estimated.
- There exist proposals alternative to local polynomial regression that overcome the curse of dimensionality: Additive models and Projection pursuit are two of them.

Additive models

- Additive models: nonparametric regression models that are less flexible than the multiple nonparametric regression model that we have seen before.
- They can be estimated with good practical results even when the number of explanatory variables is high: They are able to overcome the curse of dimensionality problem.
- Moreover the estimation results are more easily interpreted than in the case of the multiple nonparametric regression model.

- Additive model:

$$y_i = \alpha + \sum_{j=1}^{p} g_j(x_{ij}) + \varepsilon_i,$$

$$E(\varepsilon_i) = 0,\ V(\varepsilon_i) = \sigma^2,\ i = 1, \ldots, n.\ E(g_j(X_j)) = 0,\ j = 1, \ldots, p\ .$$

- Functions g_j must be estimated nonparametricaly because no parametric model is specified for them.
- The main assumption in this model is that the nonparametric univariate functions g_j are combined additively to produce the nonparametric p-dimensional regression function.
- The additive model is halfway between the multiple linear regression model (which additively combines linear transformations of the explanatory variables: $\beta_j x_{ij}$) and the multiple nonparametric regression model.

- Multiple linear regression model:

$$y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i.$$

- Additive model:

$$y_i = \alpha + \sum_{j=1}^{p} g_j(x_{ij}) + \varepsilon_i.$$

- Multiple nonparametric regression model:

$$y_i = m(x_{i1}, \ldots, x_{ip}) + \varepsilon_i.$$

Estimating the additive model

- Additive model: $y_i = \alpha + \sum_{j=1}^{p} g_j(x_{ij}) + \varepsilon_i$.
- Observe that $E(y_i) = \alpha$, because $E(\varepsilon_i) = 0$ and $E(g_j(X_j)) = 0$.
- Assume for a moment that the parameter α and all functions g_j, except g_k, were known.
- In this case the unknown function g_k could be estimated by using any nonparametric univariate smoother (i.e., a local linear fit).
- It would be enough to apply the smoother to data $(x_{ik}, y_i^{(k)})$, where

$$y_i^{(k)} = y_i - \alpha - \sum_{j=1, j \neq k}^{p} g_j(x_{ij}).$$

- This reasoning leads to propose the algorithm known as backfitting to estimate the additive model.

Backfitting algorithm

- Estimate α by $\hat{\alpha} = (1/n)\sum_{i=1}^{n} y_i$.
- Take an arbitrary function $\hat{g}_k = g_k^0$ as initial estimate of function g_k, for $k = 1, \ldots, p$ (for instance, $g_k^0(x_{ik}) = \hat{\beta}_k x_{ik}$, where the coefficients $\hat{\beta}_k$ are the multiple linear regression estimated coefficients).
- REPEAT

  FOR EACH $k = 1, \ldots, p$,

  estimate g_k by a univariate smoothing of the data $(x_{ik}, y_i^{(k)})$, where

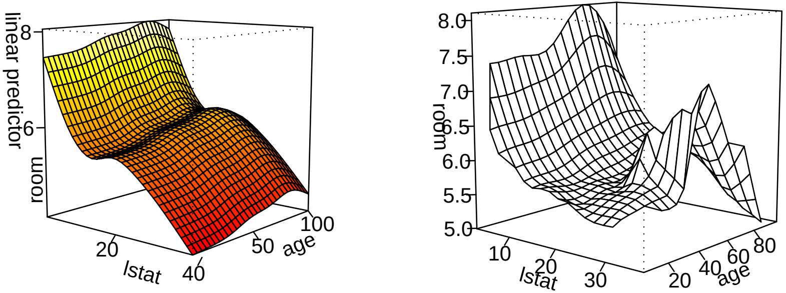
$$y_i^{(k)} = y_i - \hat{\alpha} - \sum_{j=1, j \neq k}^{p} \hat{g}_j(x_{ij}).$$

  END FOR

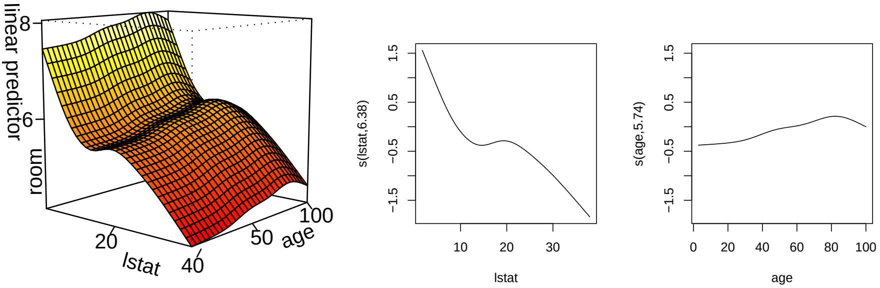
  UNTIL convergence.

Example: Boston housing data, additive model.

`ROOM` as an additive function of `LSTAT` and `AGE`.



The additive model can not pick up the local maximum (located around `LSTAT` = 35, tt AGE = 50) because it is more rigid than the nonparametric regression model.



- Axis labels show the equivalent number of parameters of both univariate estimates.
- If we cut the three-dimensional graph with cuts parallel to the plane ($\mathtt{LSTAT}$, $\mathtt{ROOM}$), the profiles that are obtained are copies of $g_{\mathtt{LSTAT}}(\cdot)$. Analogous results if the cuts are parallel to the plane ($\mathtt{AGE}$, $\mathtt{ROOM}$).
- The surface can be obtained by shifting the function $g_{\mathtt{LSTAT}}$ over the function $g_{\mathtt{AGE}}(\cdot)$ (or vice versa) and adding the mean of $\mathtt{ROOM}$.

Practice:

Additive models

Projection pursuit

- The *projection pursuit* nonparametric regression model is:

$$y_i = \alpha + \sum_{j=1}^{M} g_j(\alpha_j^T \mathbf{x_i}) + \varepsilon_i,$$

  with $E(\varepsilon_i) = 0$ and $V(\varepsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$, and each α_j is a unit norm vector in $\mathbb{R}^p$.
- Moreover it is assumed that $E(g_j(\alpha_j^T \mathbf{x})) = 0$ for all $j = 1, \ldots, M$.
- Each $z_j = \alpha_j^T \mathbf{x}$ is the projection of vector $\mathbf{x}$ in the direction of vector α_j.
- The model can be written as an additive model with explanatory variables $z_1, \ldots, z_M$.
- The model look for directions α_j that maximize the explained variance of y_i.
- This is the reason for naming it *projection pursuit*.

Projection pursuit fitting algorithm

Step 1. Let $j = 1$, $\hat{\alpha} = \bar{y}_n$ and $\hat{\varepsilon}_i = y_i - \hat{\alpha}$.

Step 2. Find the direction $\hat{\alpha}_j$ minimizing

$$RSS(\alpha_j) = \sum_{i=1}^{n} (\hat{\varepsilon}_i - \hat{g}(\alpha_j^T \mathbf{x_i}))^2,$$

where $\hat{g}$ is a nonparametric estimator for the regression of $\hat{\varepsilon}_i$ as a function of $\alpha_j^T \mathbf{x_i}$. Let $\hat{g}_j$ be the function $\hat{g}$ corresponding to the optimal value $\hat{\alpha}_j$.

Step 3. Update the residuals, $\hat{\varepsilon}_i = \hat{\varepsilon}_i - \hat{g}_j(\hat{\alpha}_j^T \mathbf{x_i})$, and do $j = j + 1$.

Step 4. Return to Step 2 if the stopping rules are not fulfilled:

(a) Stop if $j = M$.

(b) Stop if $RSE(\hat{\alpha}_j)/\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 < \delta$.

The values M and/or δ can be chosen by cross-validation.

Practice:

Projection pursuit regression

Generalized nonparametric multiple regression model

- The r.v. $(Y; \mathbf{X})$, with $\mathbf{X} = (X_1, \ldots, X_p)$, has distribution such that
$$(Y|\mathbf{X} = (x_1, \ldots, x_p)) \sim f(y; m(x_1, \ldots, x_p), \psi)$$
where $m(x_1, \ldots, x_p) = E(Y|\mathbf{X} = (x_1, \ldots, x_p))$ is a smooth function of $(x_1, \ldots, x_p)$, possibly subject to certain constrains (non-negativity or boundedness, for instance), and ψ represents other parameters (variance, for instance) not depending on $(x_1, \ldots, x_p)$.
- There exists an invertible link function $g(\cdot)$ such that
$$\theta(x_1, \ldots, x_p) = g(m(x_1, \ldots, x_p)), \quad m(x_1, \ldots, x_p) = g^{-1}(\theta(x_1, \ldots, x_p))$$
where $\theta(x_1, \ldots, x_p)$ is a smooth function of $(x_1, \ldots, x_p)$ free of constrains ($\theta(x_1, \ldots, x_p)$ can take any real value).
- Alternatively, $(Y|\mathbf{X} = (x_1, \ldots, x_p)) \sim f_2(y; \theta(x_1, \ldots, x_p), \psi) = f(y; g^{-1}(\theta(x_1, \ldots, x_p), \psi))$.

Generalized additive models (GAM)

- The nonparametric estimation of θ and m by maximum local likelihood also suffers the effects of the curse of dimensionality.
- A possible solution. The Generalized additive model:

$$(Y|X = (x_1, \ldots, x_p)) \sim f_2(y|\theta(x_1, \ldots, x_p), \psi), \ \theta(x_1, \ldots, x_p) = \alpha + \sum_{j=1}^{p} g_j(x_j).$$

- If the restriction that the functions g_j are linear is added, we obtain the Generalized linear model.
- Then the Generalized additive model is halfway between the Generalized nonparametric multiple regression model and the Generalized linear model.

- Generalized Linear Model:

$$(Y|X=(x_1,\ldots,x_p)) \sim f_2(y|\theta(x_1,\ldots,x_p),\psi),\ \theta(x_1,\ldots,x_p)=\alpha+\sum_{j=1}^{p}\beta_j x_j.$$

- Generalized Additive Model:

$$(Y|X=(x_1,\ldots,x_p)) \sim f_2(y|\theta(x_1,\ldots,x_p),\psi),\ \theta(x_1,\ldots,x_p)=\alpha+\sum_{j=1}^{p}g_j(x_j).$$

- Generalized nonparametric multiple regression model:

$$(Y|X=(x_1,\ldots,x_p)) \sim f_2(y|\theta(x_1,\ldots,x_p),\psi),\ \theta(x_1,\ldots,x_p) \text{ arbitrary.}$$

GAM Estimation

- The estimation of a Generalized additive model combines the backfitting algorithm (used to fit additive models) with the IRWLS algorithm (used to maximize the likelihood in GLM).
- In the IRWLS algorithm, each multiple linear regression fit by WLS is replaced by the fitting of an additive model using backfitting.
- This way the model finally fitted is a GAM instead of a GLM.

Local scoring algorithm for logistic GAM

(Source: Algorithm 9.2 in Hastie, Tibshirani, and Friedman 2001)

1. Compute starting values: $\hat{\alpha} = \log[\bar{y}/(1-\bar{y})]$, where $\bar{y}$ is the sample proportion of ones, and set $\hat{f}_j = 0$ for all j.
2. Define $\hat{\eta}_i = \hat{\alpha} + \sum_j \hat{f}_j(x_{ij})$ and $\hat{p}_i = \exp(\hat{\eta}_i)/[1+\exp(\hat{\eta}_i)]$.
   Iterate:
   (a) Construct the working target variable $z_i = \hat{\eta}_i + \frac{(y_i-\hat{p}_i)}{\hat{p}_i(1-\hat{p}_i)}$.
   (b) Construct weights $w_i = \hat{p}_i(1-\hat{p}_i)$.
   (c) Fit an additive model to the targets z_i with explanatory variables x_{ij} and weights w_i, using a weighted backfitting algorithm. This gives new estimates $\hat{\alpha}$, $\hat{f}_j$, $\forall j$.
3. Repeat step 2 until the change in the functions falls below a prespecified threshold.

Semiparametric models

- Sometimes some of the explanatory variables involved in the definition of a generalized additive model (or an additive model) affect the response variable linearly.
- If this is known in advance, the GAM model can be reformulated allowing some of the functions $g_j(x_j)$ to be linear: $g_j(x_j) = \beta_j x_j$.
- Other possible modifications of the GAM model are:
  - Nonparametrically estimating the combined effect of two (or more) explanatory variables. This includes, for example, replacing $g_j(x_j) + g_h(x_h)$ by $g_{j,h}(x_j, x_h)$.
  - Estimating the effect of a variable x_j differently at each of the classes determined by another categorical variable x_h. These effects could be estimated linearly or nonparametrically.

Semiparametric models. References

- The models obtained by incorporating these modifications to the GAM model are known as Semiparametric models.
- These models can be fitted using the function `gam` in the R package `mgcv`, developed by S. Wood (see Wood 2006).
- The book Ruppert, Wand, and Carroll (2003) is entirely dedicated to semiparametric models.
- These authors have developed in parallel the R package `SemiPar`, that allows fitting these models.
- Specifically, function `spm` has similarities with function `gam` in `mgcv`, but it incorporates additional options.

Practice:

Generalized additive models and semiparametric models

Bowman, A. W. and A. Azzalini (1997).
Applied Smoothing Techniques for Data Analysis.
Oxford: Oxford University Press.

Hastie, T., R. Tibshirani, and J. Friedman (2001).
The Elements of Statistical Learning. Data Mining, Inference, and Prediction.
Springer.

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

Ruppert, D., M. P. Wand, and R. J. Carroll (2003).
Semiparametric Regression.
Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

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

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