## Mη παραμετρική παλινδρόμηση ΙV: Πολλαπλή μη παραμετρική παλινδρόμηση, GAMs και Regression Trees

* Μέχρι τωρα υποθέταμε ότι οι ανεξάρτητες μεταβλητές . Συχνά όμως είναι διαθέσιμες  ανεξάρτητες μεταβλητές και αρα υποθέτουμε τωρα .
* Θέλουμε πάλι να εκτιμήσουμε την , όπου η  δίνει την αναμενόμενη τιμή των εξαρτημένων :

 με  (ή αν  τυχαίες μεταβλητές:)

* Οι μέθοδοι της απλής παλινδόμηση επεκτείνονται στη πολλαπλή. Ωστόσο **βρισκόμαστε αντιμετωποι με την ανάγκη για πολύ περισσότερες (εκθετικά περισότερες) παρατηρήσεις**:

**«η κατάρα των πολλών διαστάσεων» (curse of diamensionality):**

Έστω  ομοιόμορφη στο  και μια περιοχή ακτίνας  ενός σημείου . Πόσες παρατηρήσεις είναι αναγκαίες ώστε να πέσουν περίπου 10 στη δοσμένη περιοχή



Επίσης, ενώ για  και κάτω από την υπόθεση ότι η  έχει ολοκληρώσιμες δεύτερες παραγώγους παίρνουμε  ως βέλτιστη ταχύτητα του μέσου τετραγωνικού σφάλματος, για  η ταχύτητα αυτή γίνεται  .

Έτσι αν για  επιτυγχάνουμε μια δοσμένη ακρίβεια στην εκίμηση της , για **να πετύχουμε την ίδια ακρίβεια σε  θα πρέπει να έχουμε δείγμα μεγέθους**  όπου . Δηλαδή εκθετικά μεγαλύτερο.

Αν έχουμε «αρκετές» διαθέσιμες παρατηρήσεις (π.χ. ) οι μέθοδοι που γνωρίσαμε επεκτείνονται ως εξής:

### Τοπική παλινδρόμηση και splines

#### Τοπική παλινδρόμηση

* Εστω ένας πυρήνας  σε  μεταβλητές.
* Π.χ. αν τυποποιήσουμε όλες τις ανεξάρτητες μεταβλητές θα μπορούσαμε να πάρουμε

, όπου .

* Τότε εκτιμούμε τοπικά (για κάθε ) τις παραμέτρους  που ελαχιστοποιούν την  με  και θέτουμε  (το intercept).
* Παίρνουμε ,

όπου  και .

#### Splines

Tο ανάλογο των splines σε περισσότερες διαστάσεις λέγεται **thin plate splines** που για παράδειγμα στο  βρίσκεται με ελαχιστοποίηση της



, όπου







### Γενικευμένα Αθροιστικά Μοντέλα (Generalized Additive Models, GAM)

* Και στις δύο παρπάνω περιπτώσεις (τοπική παλινδρόμηση και splines) όσο μεγαλώνει το , εκτός από το πρόβλημα της επάρκειας δεδομένων, **δυσκολεύει** και ο υπολογιστικός φόρτος, αλλά κυρίως και **η εποπτεία μας στην ερμηνεία των αποτελεσμάτων**:
  + ήδη για  η είναι μια επιφάνεια στο . Αυτήν ίσως οριακά να μπορούμε να τη παραστήσουμε και να τη κατανοήσουμε.
  + Για μεγαλύτερα  είναι πολύ δύσκολο.
  + Γι’ αυτό είναι προτιμότερο για μεγαλύτερα  (ισως ήδη και για ) να εισάγουμε κάποια δομή στο μοντέλο μας (κάνοντας ένα βήμα «πίσω» προς τη παραμετρική στατιστική), απαιτώντας π.χ. να αθροίζονται οι επιδράσεις τους: κάνουμε δηλάδη την υπόθεση του **αθροιστικού μοντέλου**



 ,

όπου οι  είναι «ομαλές» συναρτήσεις των 



* Εισάγοντας επιπλέον υποθέσεις για την κατανομή των λαθών και τη συνάρτηση σύνδεσμο, μπορούμε να φθάσουμε και στο **Γενικευμένο Αθροιστικό Μοντέλο (GAM)** (αντίστοιχο της τοπικής πιθανοφάνειας).







**GAM library**

gam.fit <- gam(chd~ lo(sbp,span=0.3,degree=1)+

lo(tobacco,span=0.3,degree=1)+

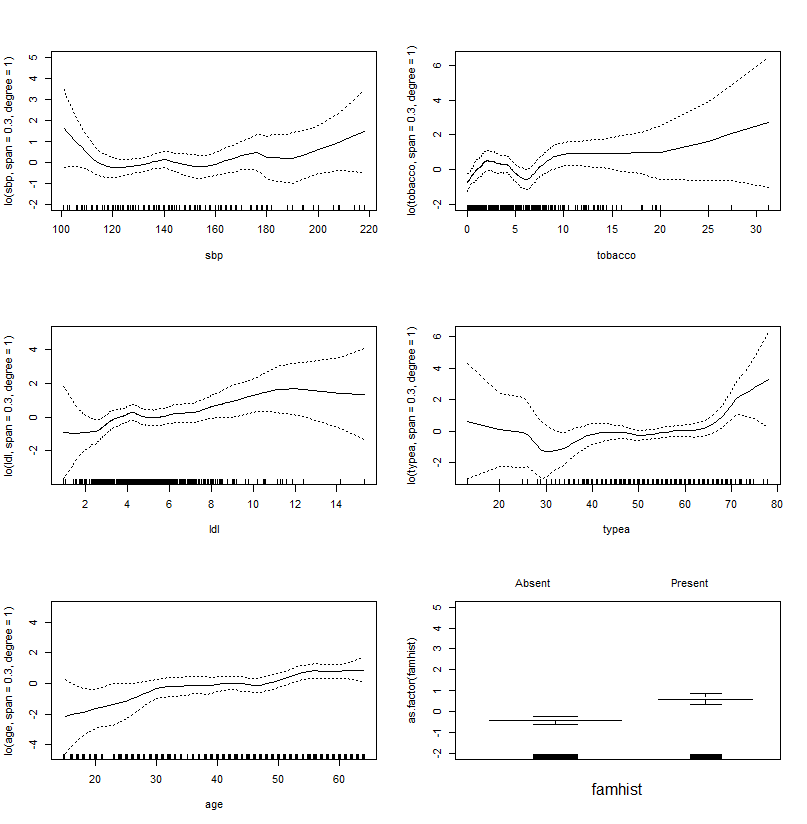
lo(ldl,span=0.3,degree=1)+

lo(typea,span=0.3,degree=1)+

lo(age,span=0.3,degree=1)+as.factor(famhist),

data=heart, family=binomial(link=logit))





**MGCV library**

gam.fit <- gam(chd~ s(sbp)+

s(tobacco)+

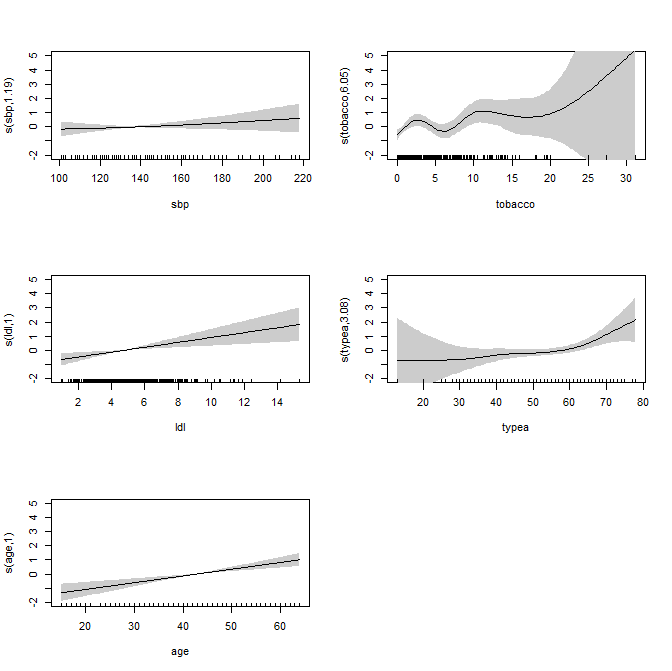
s(ldl)+

s(typea)+

s(age) +

as.factor(famhist),

data=heart, family=binomial(link=logit))



#### Αλγόριθμος για εκτίμηση GAM

Ένας αλγόριθμος για την εκτίμηση των  είναι ο λεγόμενος **“backfitting” algorithm**

1. Αρχικά θέτουμε  και 
2. Διαδοχικά για κάθε  και μέχρι να επιτευχθεί σύγκλιση
   1. Υπολόγισε  , 
   2. Εφάρμοσε κάποια μέθοδο τοπικής εξομάλυνσης στα  ώστε να πάρουμε 
   3. Θέτουμε 

* Μπορεί στο 2b άλλες μεταβλητές να εκτιμώνται με τοπική παλινδρόμηση και άλλες με splines. Επίσης μπορεί κάποιοι όροι να εμφανίζονται παραμετρικά  και άλλοι μη-παραμετρικά
* Οι βαθμοί ελευθερίας σε κάθε  είναι  όπου  είναι ο πίνακας εξομάλυνσης στον υπολογισμό του 
* Καλό είναι όλες οι ανεξάρτητες μεταβλητές να τυποποιούνται, ώστε να μετρώνται σε συγκρίσιμες μεταξυ τους μονάδες