Aspects of more-dimensional non-parametric regression

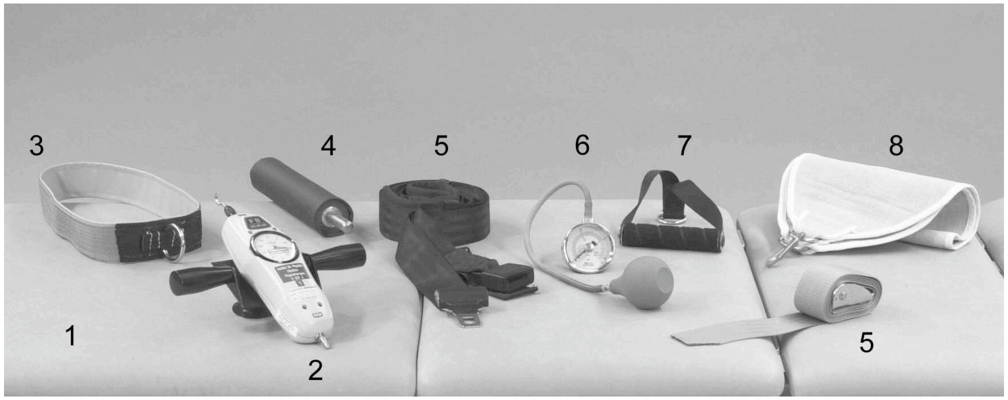
Burkhardt Seifert

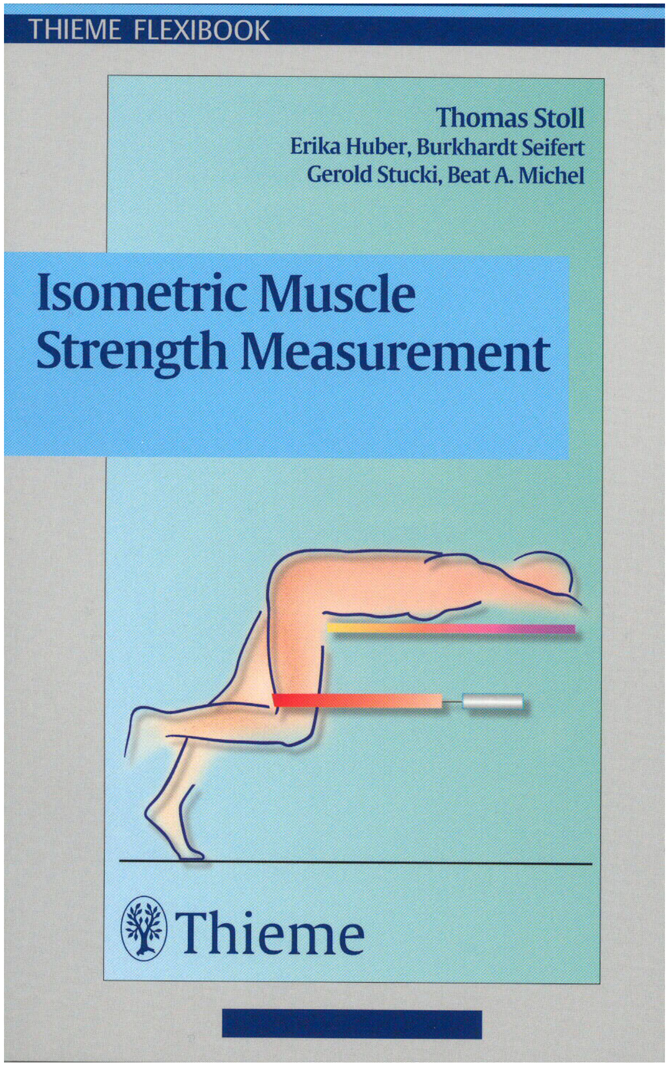
Biostatistics Unit, ISPM

University Zürich

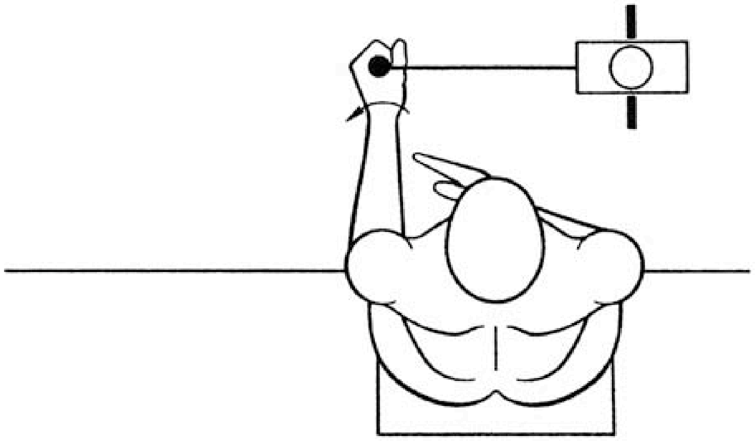
- additive regression
- penalty on departure from additivity
- local additive regression

Gender-specific and age-dependent normative percentile curves for isometric muscle strength

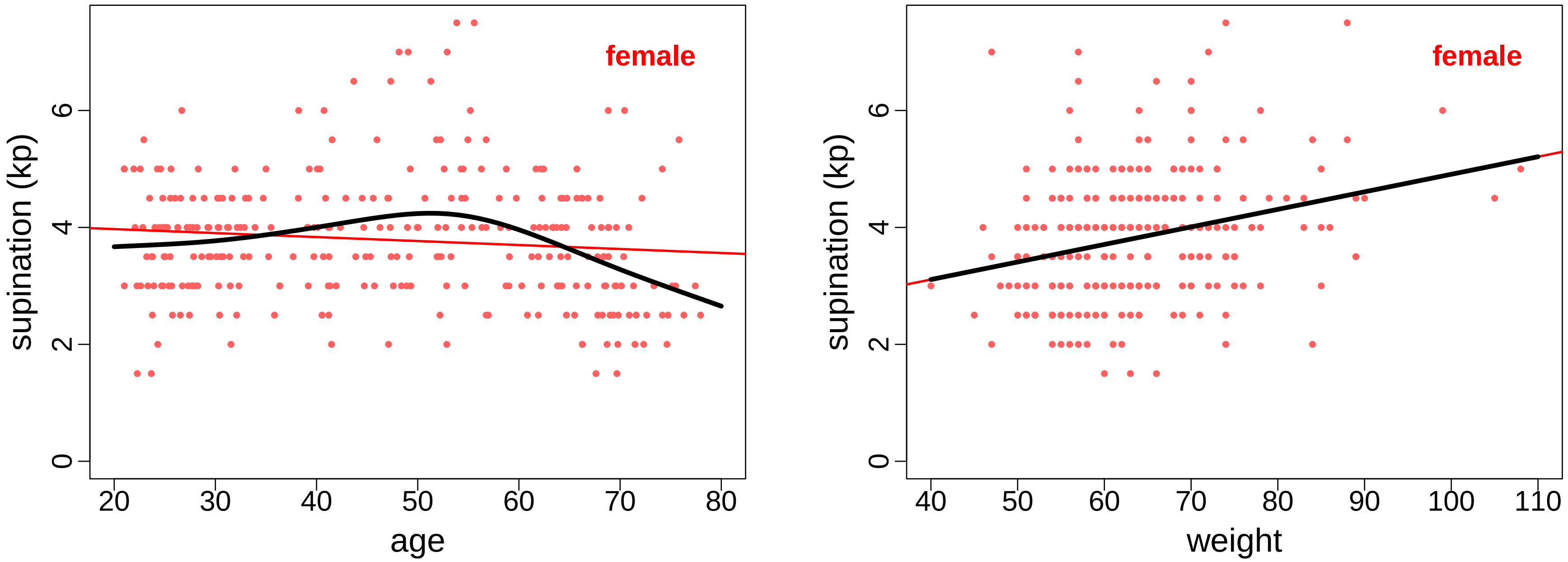




Supination left:



Automatic snmoothing in R

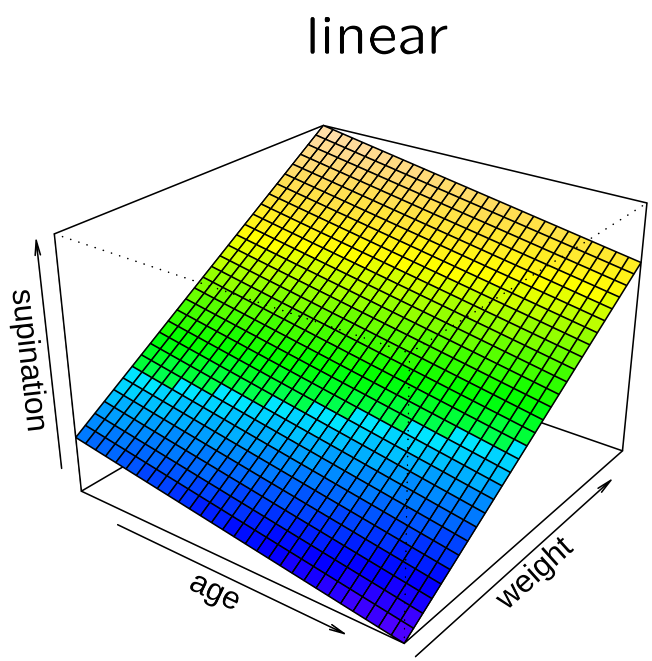


- $n = 287$

- **Interpretation:**
  - reduction of strength after age 55 by 30%
  - strength increases linearly with weight

R package mgcv: regression splines with smoothness penalty and GCV parameter selection

Multiple linear regression



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

y = supination (women) $R^2_{\text{adj}} = 0.099$

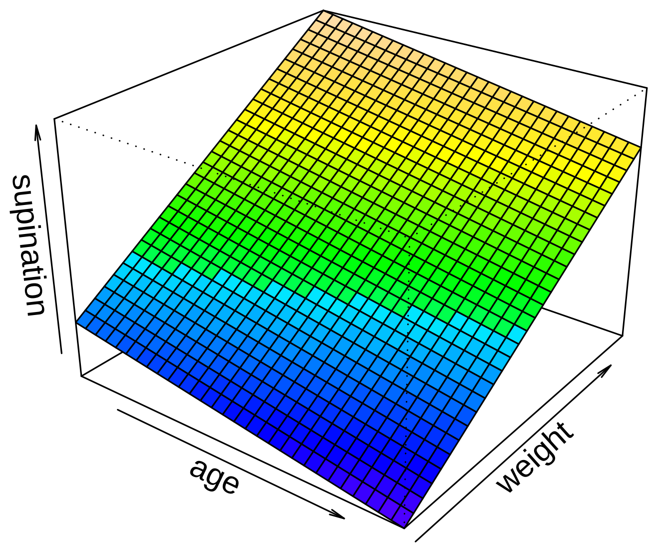
x_1 = age $p = 0.003$

x_2 = body weight $p < 0.0001$

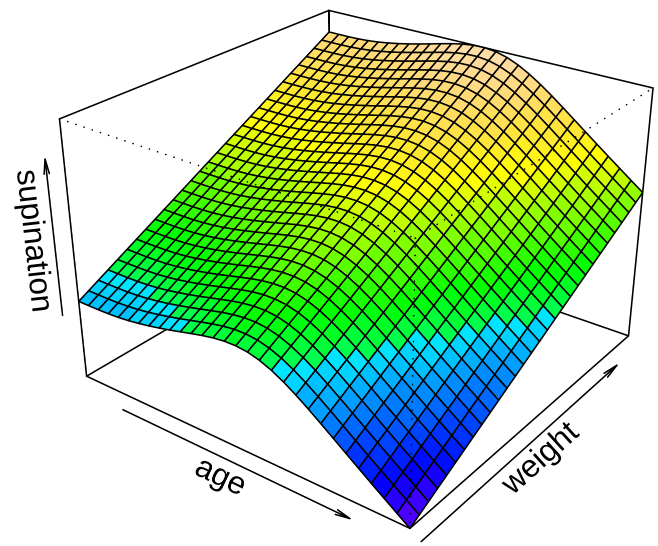
gam in R package mgcv

Additive regression

linear



additive



$$y = \overline{y} + s_1(x_1) + s_2(x_2)$$

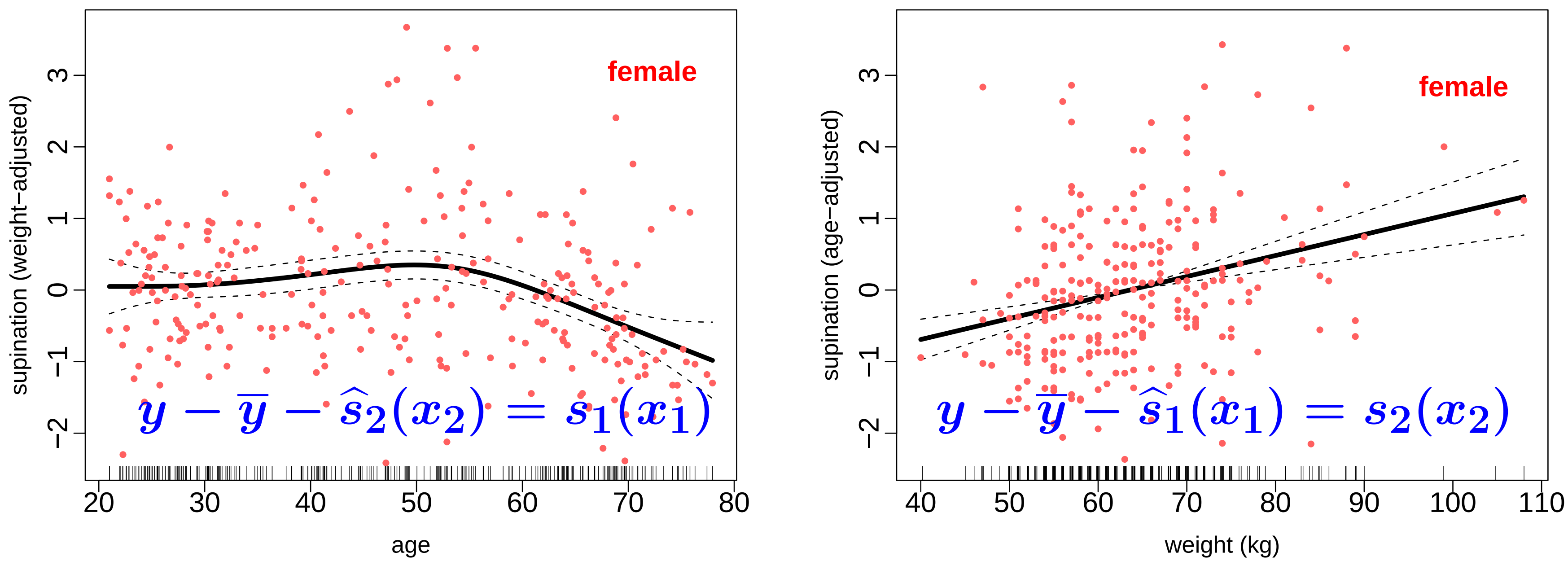
y = supination (women) $R^2_{\text{adj}} = 0.159$

$s(x_1)$ = s(age) $p < 0.0001$

$s(x_2)$ = s(weight) $p < 0.0001$

gam in R package mgcv

Additive regression

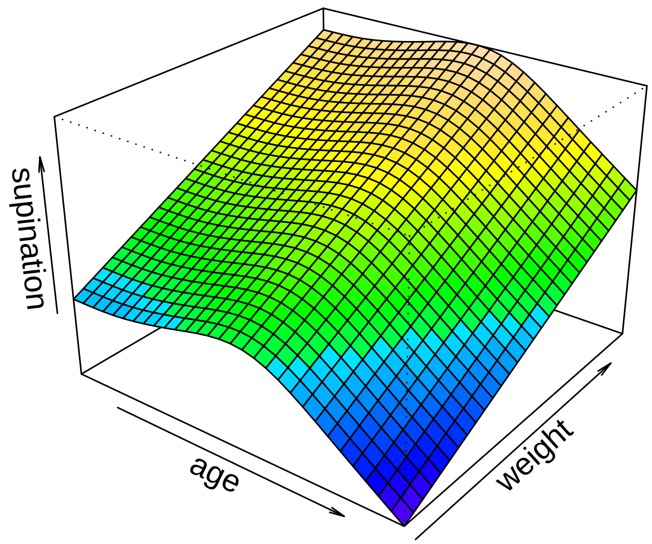


estimation:

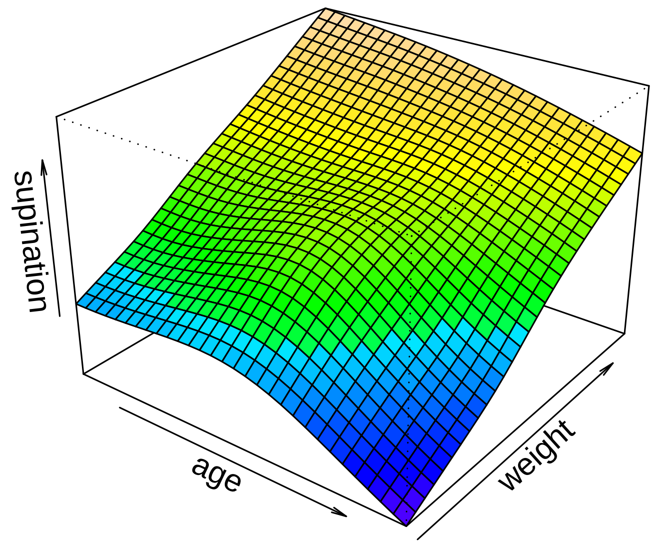
- backfitting (Breiman & Friedman 1985, JASA)
- marginal integration (Linton & Nielsen 1995, Biometrika)
- smooth backfitting (Mammen, Linton & Nielsen 1999, Annals Statistics)
- regression splines w. smoothness penalty (Wood 2000–08, JASA, JRSS B)

Nonparametric multiple regression

additive



full nonparametric

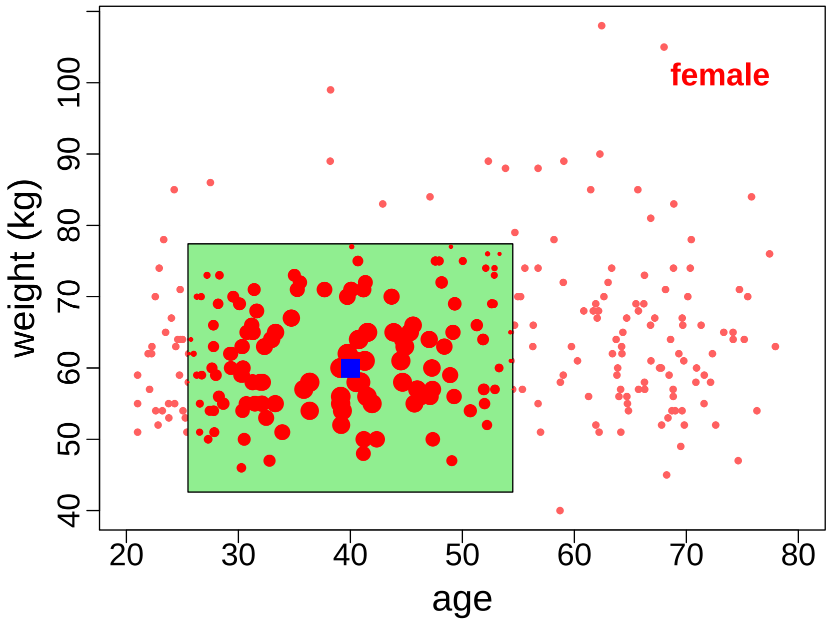


$$y = \overline{y} + s(x_1, x_2)$$

$R^2_{\text{adj}} = 0.161$

gam in R package mgcv

Local linear estimator



$$y_i = r(\mathbf{x}_i) + \varepsilon_i$$

$$\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})$$

estimation in $\mathbf{u} = (u_1, \ldots, u_d)$

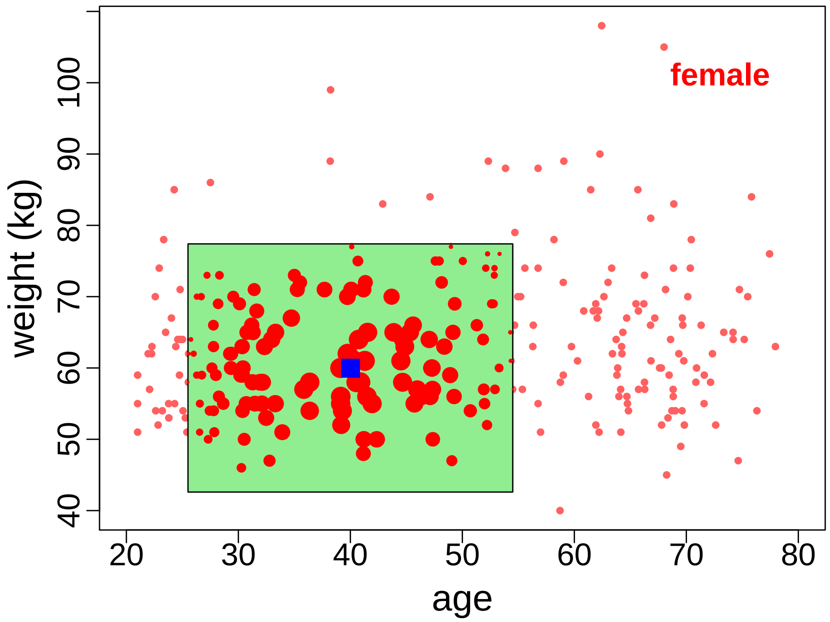
bandwidth $\mathbf{h} = (h_1, \ldots, h_d)$

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \beta^0(\mathbf{u}) - \sum_{k=1}^{d}\beta^k(\mathbf{u})\frac{x_{i,k}-u_k}{h_k}\right]^2 K_h(\mathbf{x}_i, \mathbf{u})$$

$\longrightarrow$ Min

- $\widehat{r}_{\mathsf{ll}}(\mathbf{u}) = \widehat{\beta}^0(\mathbf{u})$

Local linear estimator



$$y_i = r(\mathbf{x}_i) + \varepsilon_i$$

$$\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})$$

estimation in $\mathbf{u} = (u_1, \ldots, u_d)$

bandwidth $\mathbf{h} = (h_1, \ldots, h_d)$

$$\int \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \beta^0(\mathbf{u}) - \sum_{k=1}^{d} \beta^k(\mathbf{u}) \frac{x_{i,k} - u_k}{h_k} \right]^2 K_h(\mathbf{x}_i, \mathbf{u})\, d\mathbf{u}$$

$\longrightarrow$ Min

- $\widehat{r}_{||} = \widehat{\beta}^0 \quad \longleftarrow$ **function**

Local linear estimator is projection

- Mammen, Marron, Turlach, Wand (1997) $\longrightarrow$ (2001) **Statist. Science:**

$$\|\underline{r}_y - \underline{r}\|_*^2 = \int \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \beta^0(\mathbf{u}) - \sum_{k=1}^{d} \beta^k(\mathbf{u}) \frac{x_{i,k} - u_k}{h_k} \right]^2 K_h(\mathbf{x}_i, \mathbf{u})\, d\mathbf{u}$$

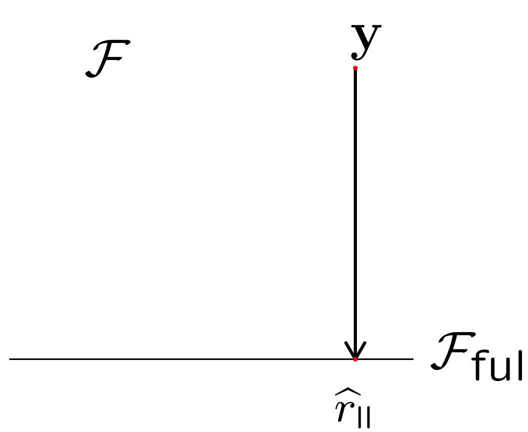
$$\xrightarrow{\mathcal{F}_{\text{full}}} \text{Min}$$

- $\|\cdot\|_*^2$ in $n(d+1)$–dimensional Hilbert space $\mathcal{F}$

- $\widehat{r}_{\text{ll}} = \widehat{\beta}_0$ is projection $\mathcal{F} \longrightarrow \mathcal{F}_{\text{full}}$

- $\widehat{r}_{\text{ll}}$ is optimal and design–adaptive

- but: "curse of dimensionality"



Additive regression

- Mammen, Linton & Nielsen (1999) **Annals Statistics:**

  Smooth backfitting estimator (**SBE**): $\|\underline{r}_y - \underline{r}\|_*^2 \quad \xrightarrow{\mathcal{F}_{\text{add}}} \quad$ Min

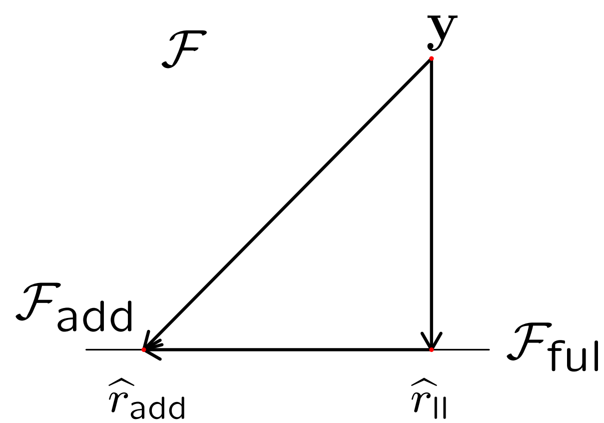
- $\mathcal{F}_{\text{add}}$ = subspace of additive functions in $\mathcal{F}$:

  $\beta^0(\mathbf{u}) = \beta^0_0 + \beta^0_1(u_1) + \ldots + \beta^0_d(u_d), \quad \beta^k(\mathbf{u}) = \beta^k(u_k)$

- $\hat{r}_{\text{add}} = \hat{\beta}^0$ **(SBE) is projection** $\mathcal{F} \longrightarrow \mathcal{F}_{\text{add}}$

- and: projection $\mathcal{F} \longrightarrow \mathcal{F}_{\text{full}} \longrightarrow \mathcal{F}_{\text{add}}$

- $\hat{r}_{\text{add}}$ is “oracle” optimal and design–adaptive



Penalty on departure from additivity

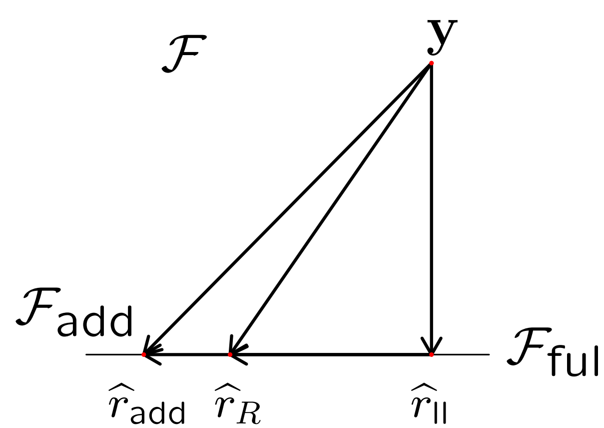
- Studer, Seifert & Gasser (2005) **Annals Statistics:**

$$\|\underline{r}_y - \underline{r}\|_R^2 = \|\underline{r}_y - \underline{r}\|_*^2 + R\|(\mathcal{I} - \mathcal{P}_{\text{add}})\underline{r}\|_2^2 \quad \xrightarrow{\mathcal{F}_{\text{full}}} \quad \text{Min}$$

- $\widehat{r}_{\text{R}}$ is projection $\mathcal{F} \longrightarrow \mathcal{F}_{\text{full}}$

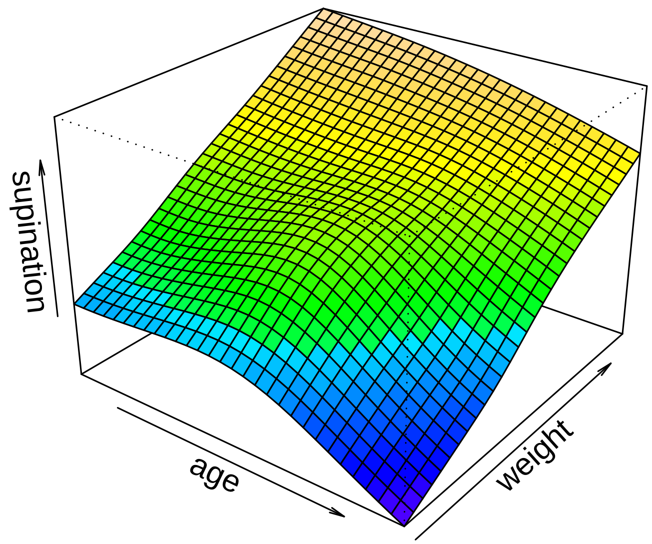
- $\widehat{r}_{\text{add}}$ is projection of $\widehat{r}_{\text{R}}$: $\mathcal{F}_{\text{full}} \longrightarrow \mathcal{F}_{\text{add}}$

- $\widehat{r}_R$ is design–adaptive compromise between $\widehat{r}_{\text{ll}}$ and $\widehat{r}_{\text{add}}$

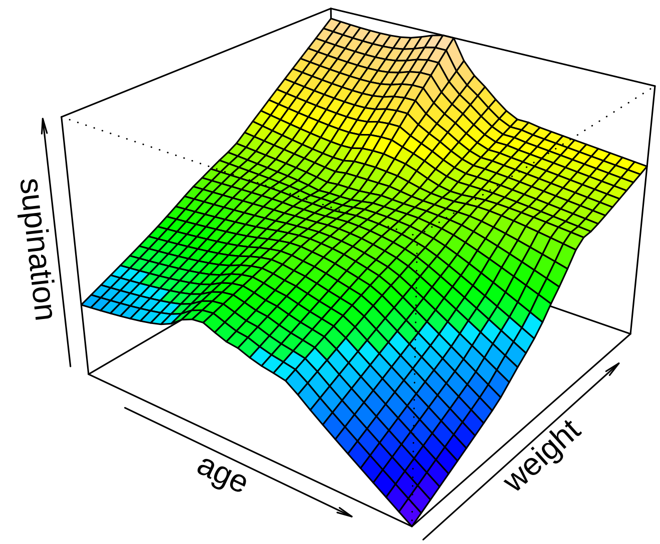


Penalty on departure from additivity

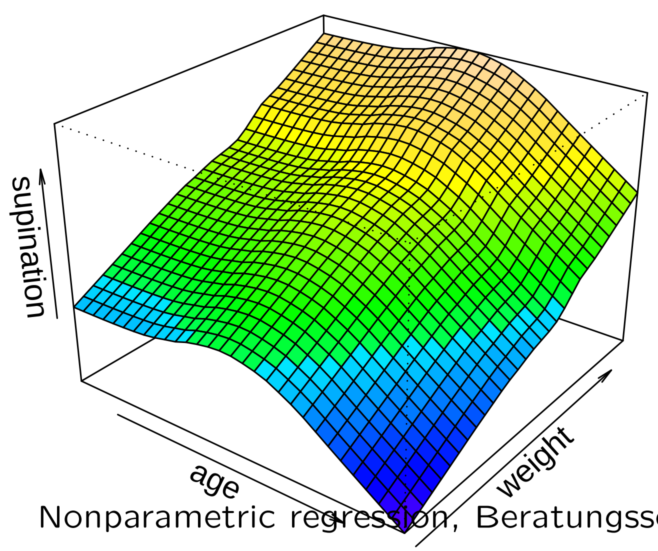
spline with smoothness penalty



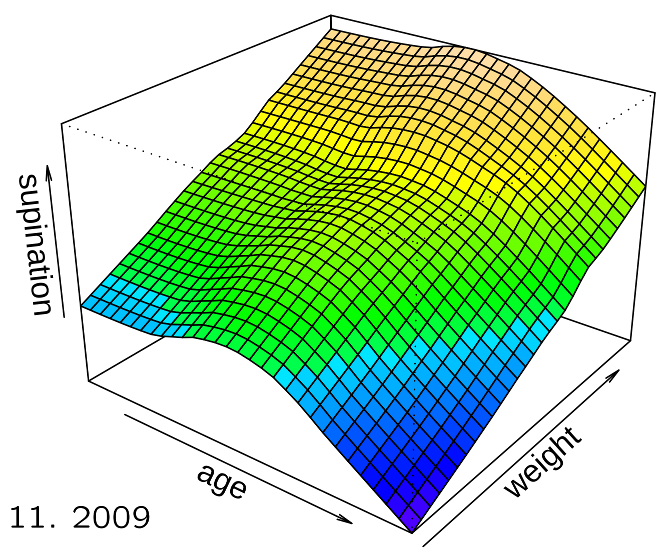
local linear



SBE additive

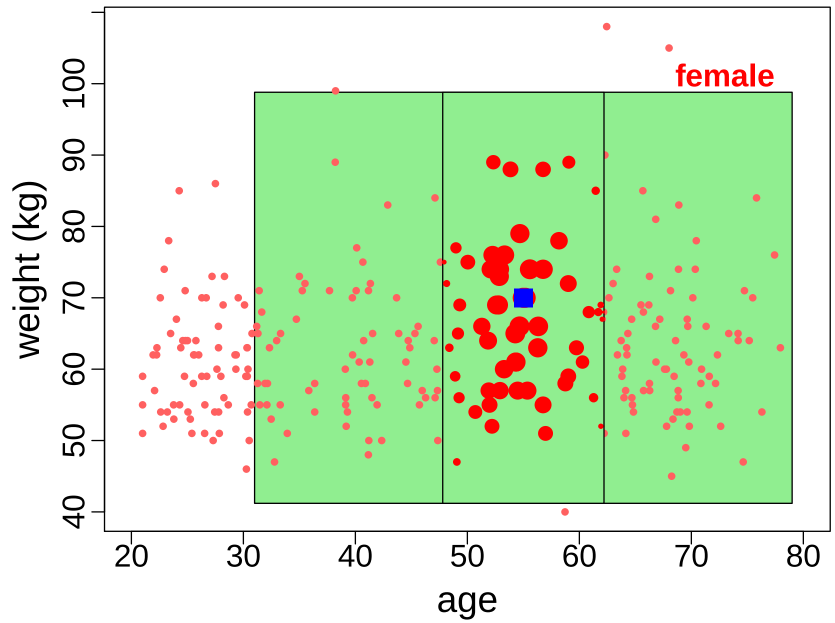


penalty on non–additivity

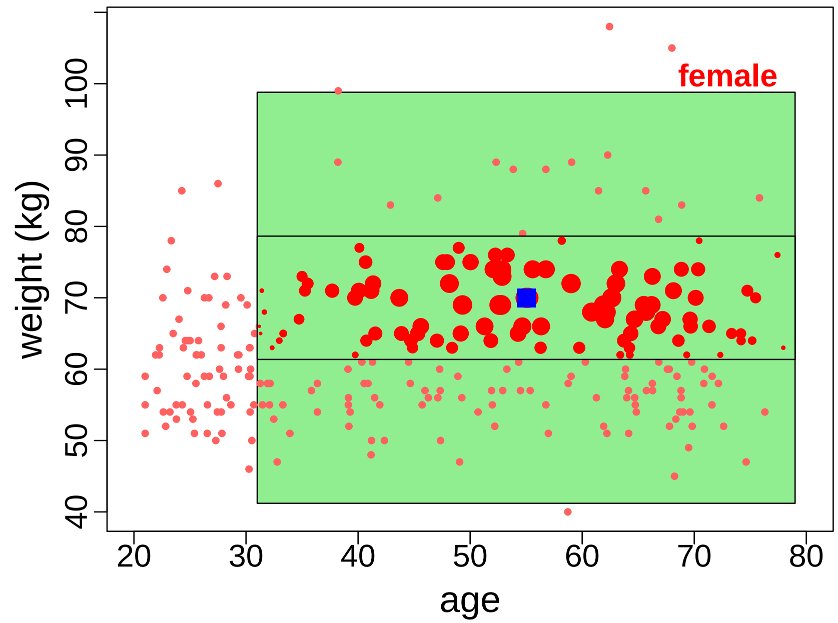


Local additive regression

- Park & Seifert (to appear) **JRSS B**



$$y - \overline{y} - \widehat{s}_2(x_2) = s_1(x_1)$$



$$y - \overline{y} - \widehat{s}_1(x_1) = s_2(x_2)$$

Idea:

- more data used than for local linear $\rightarrow$ reduced variance
- smaller bandwidth compared to local linear $\rightarrow$ reduced bias

Main result

- w = diameter of region of data used: $\mathbf{x}_i \in [\mathbf{u} \pm \mathbf{w}]$, $\mathbf{x}_i \in R^d$
- h = internal bandwidth of additive estimator used locally
- $b_{jk}(\mathbf{x}) = (x_j - u_j)(x_k - u_k)$ bilinear terms

Main theorem: Assume the additive estimator used locally is linear in Y, oracle optimal and Lipschitz continuous in $f(x) \rightarrow$ uniform. Then,

$$V\Big[\widehat{r}_{ladd}(\mathbf{u})\Big] = O\Big((nw^{d-1}h)^{-1}\Big)$$

$$B^2[\widehat{r}_{ladd}(\mathbf{u})] = \max\{O(h^4), O(w^8), w^4 O(|\widehat{b}_{jk}(\mathbf{u})|^2)\}$$

Local additive estimation using SBE

- bias of bilinear terms is negligible if $\widehat{r}_{ladd}$ is based on SBE

- MSE–optimal w and h ($d \leq 8$):

$$w \sim n^{-1/(9+d)}, \quad h \sim n^{-2/(9+d)},$$

- **intuition confirmed:**

$$\boldsymbol{w > h_{ll} > h_{ladd} > h_{add}}$$

$$\longrightarrow \quad MSE[\widehat{r}_{ladd}(\mathbf{u})] \sim n^{-8/(9+d)} = n^{-4/\left(4+\frac{d+1}{2}\right)}.$$

$$MSE[\widehat{r}_{ll}(\mathbf{u})] \sim n^{-4/(4+d)} \longrightarrow \tilde{d} = (d+1)/2$$

Tutorial example

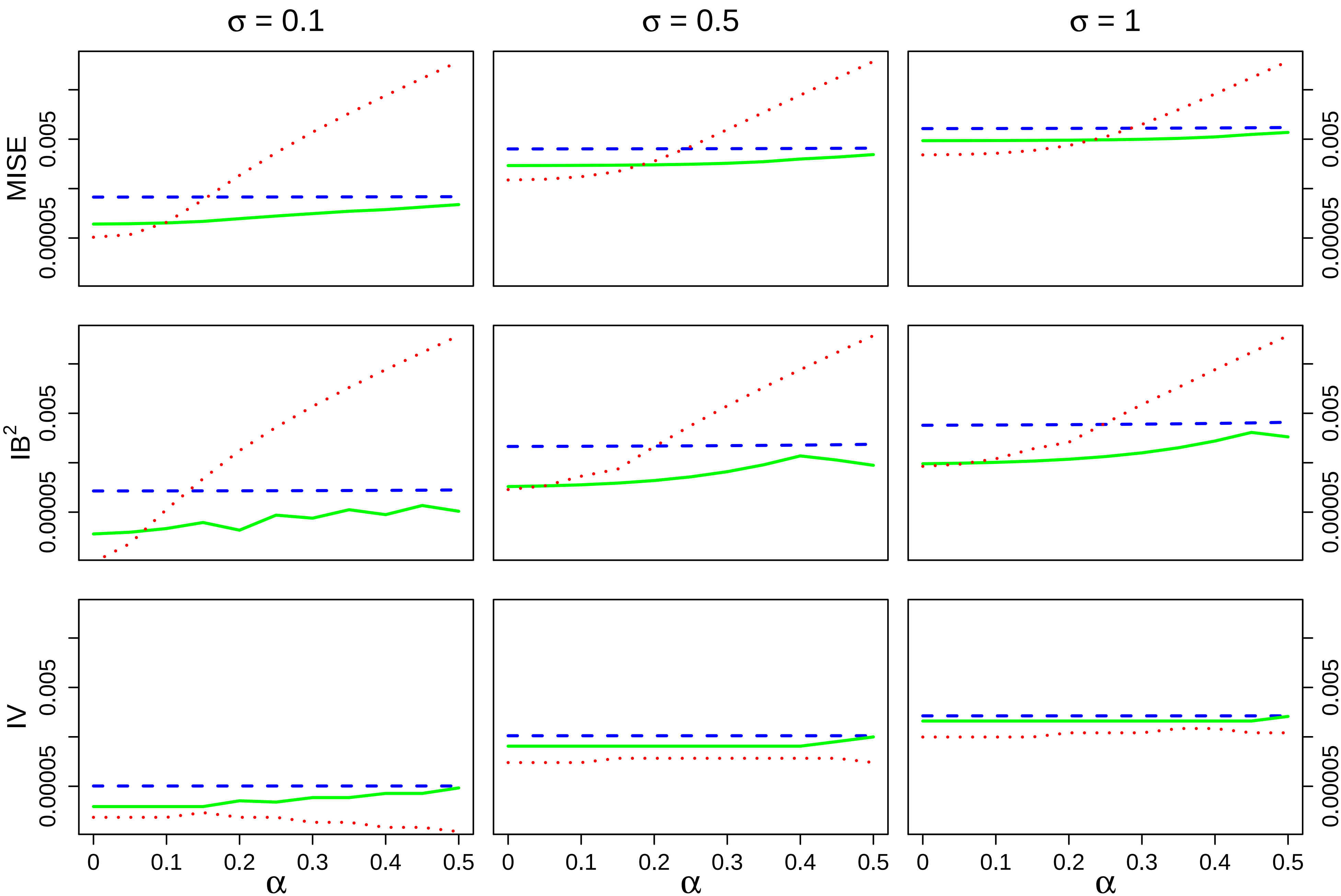
- $d = 2$, $n = 400$

- random uniform design on $[-1, 1]^2$

- $r(\mathbf{x}) = x_1^2 + x_2^2 + \frac{\alpha}{1-\alpha} x_1 x_2$

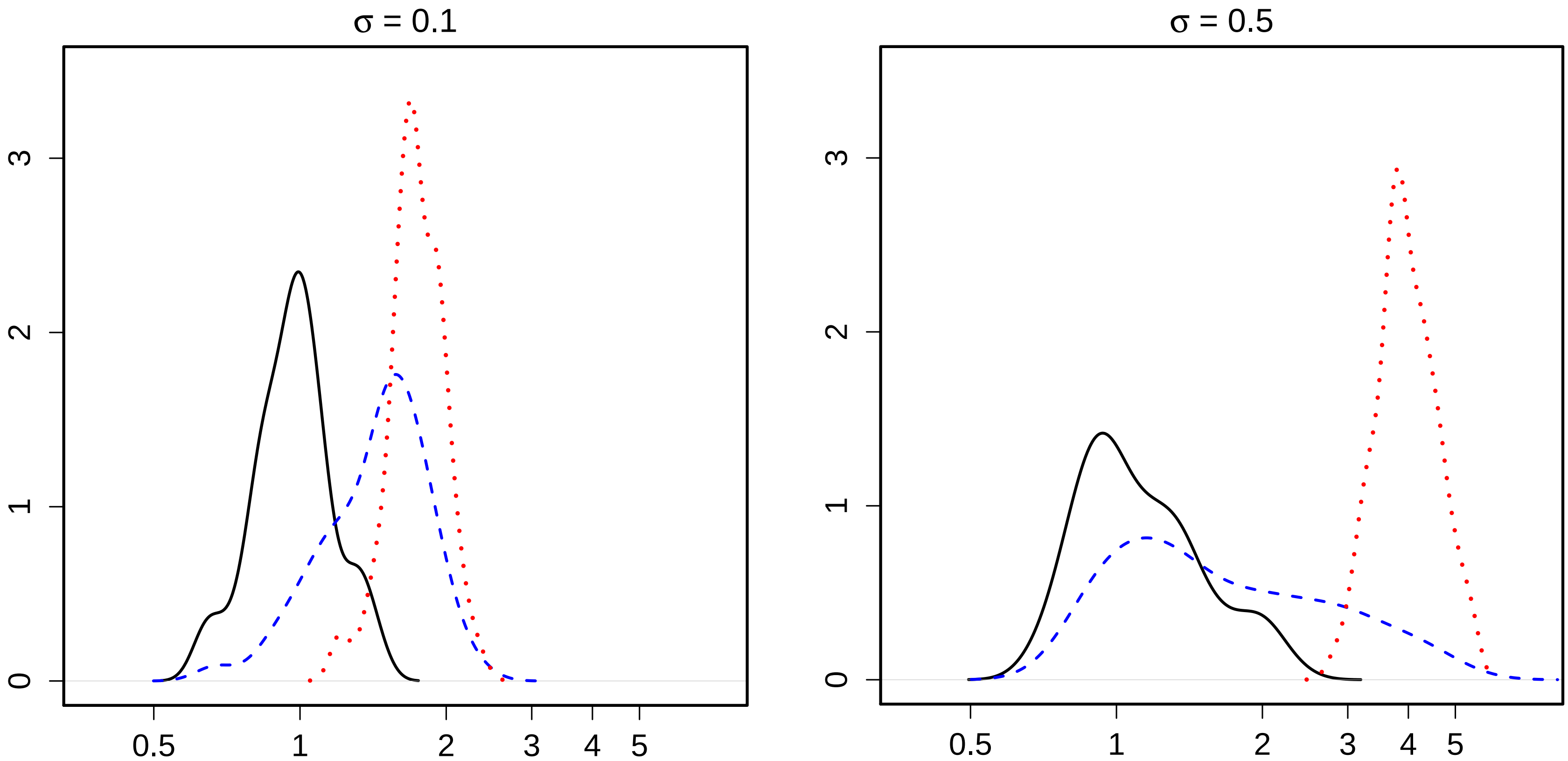
- residuals $\mathcal{N}(0, \sigma^2)$

- MISE–optimal h and w


σ = 0.1
σ = 0.5
σ = 1
MISE
IB2
IV
0.00005
0.005
0
0.1
0.2
0.3
0.4
0.5
α

Data driven parameter selection

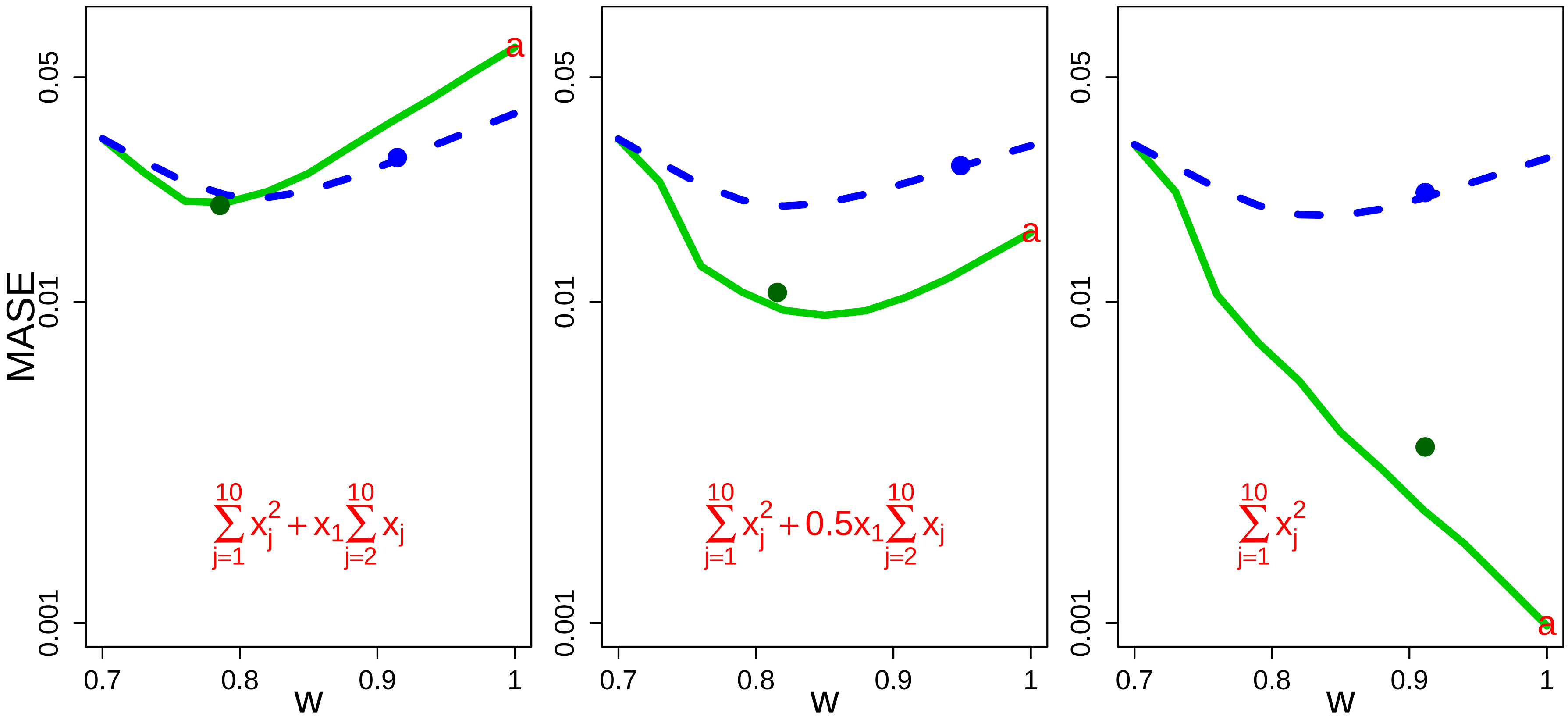
- $\alpha = 0.3$



PLS, AIC, AIC_C vs ASE–optimal number of parameters (tr(H))

Data driven parameter selection, $d = 10$

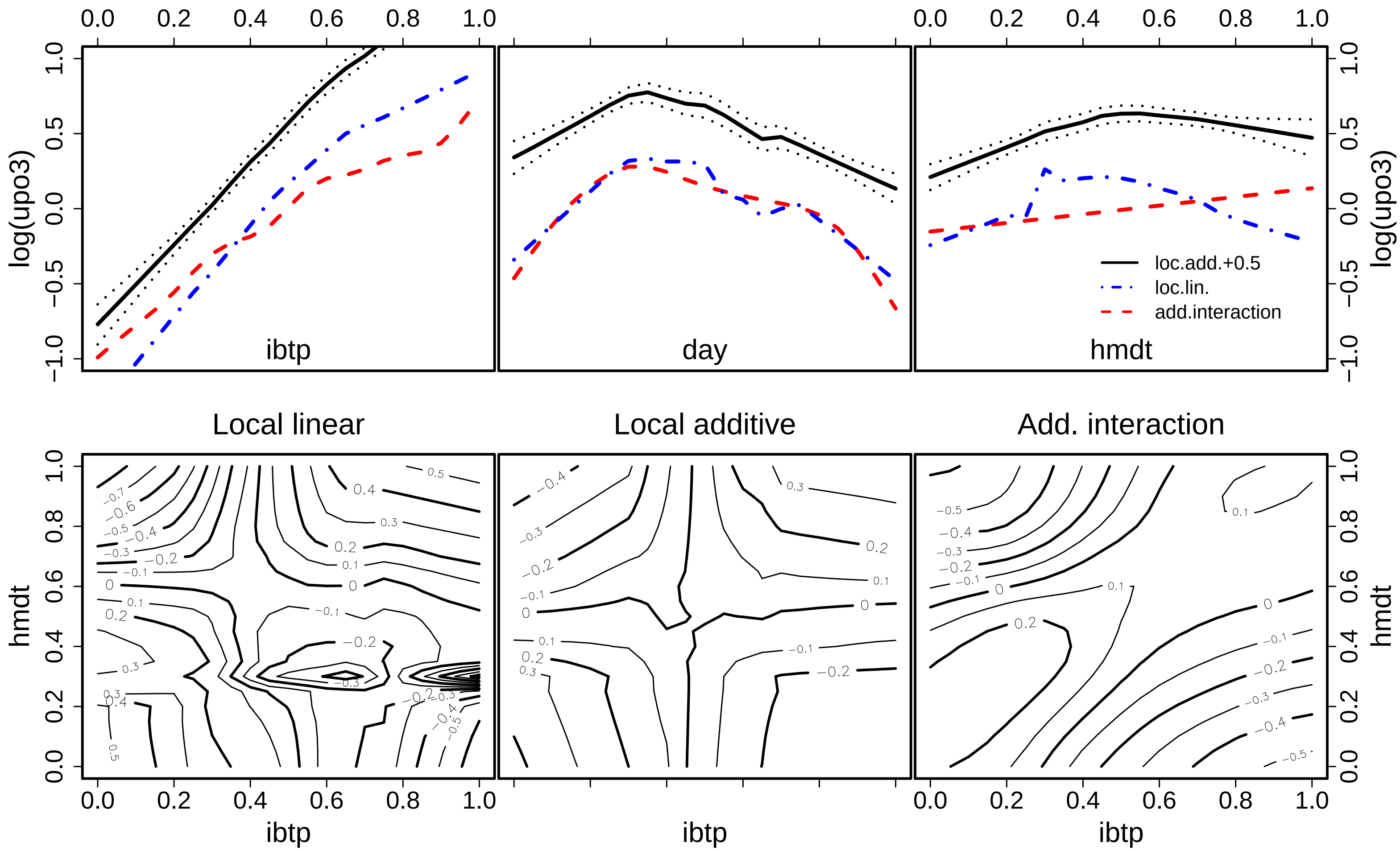
$\sigma = 0.2$



GCV: • = local linear, • = local additive, a = additive

gam in R package mgcv

Ozone concentration in Los Angeles basin

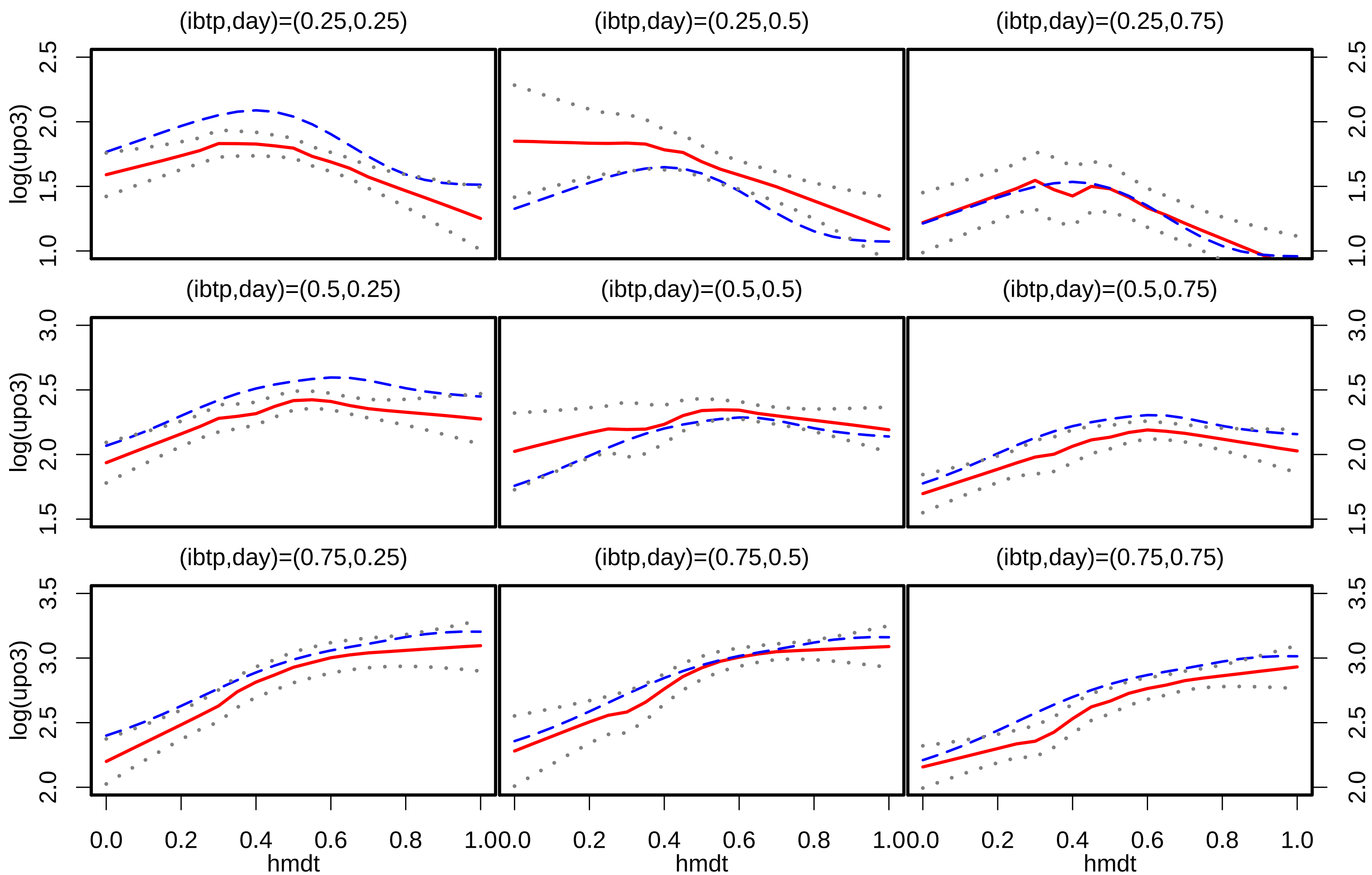


y = log upland ozone concentration (upo3)

$\mathbf{x}$ = humidity (hmdt), inversion base temperature (ibtp), day (day)

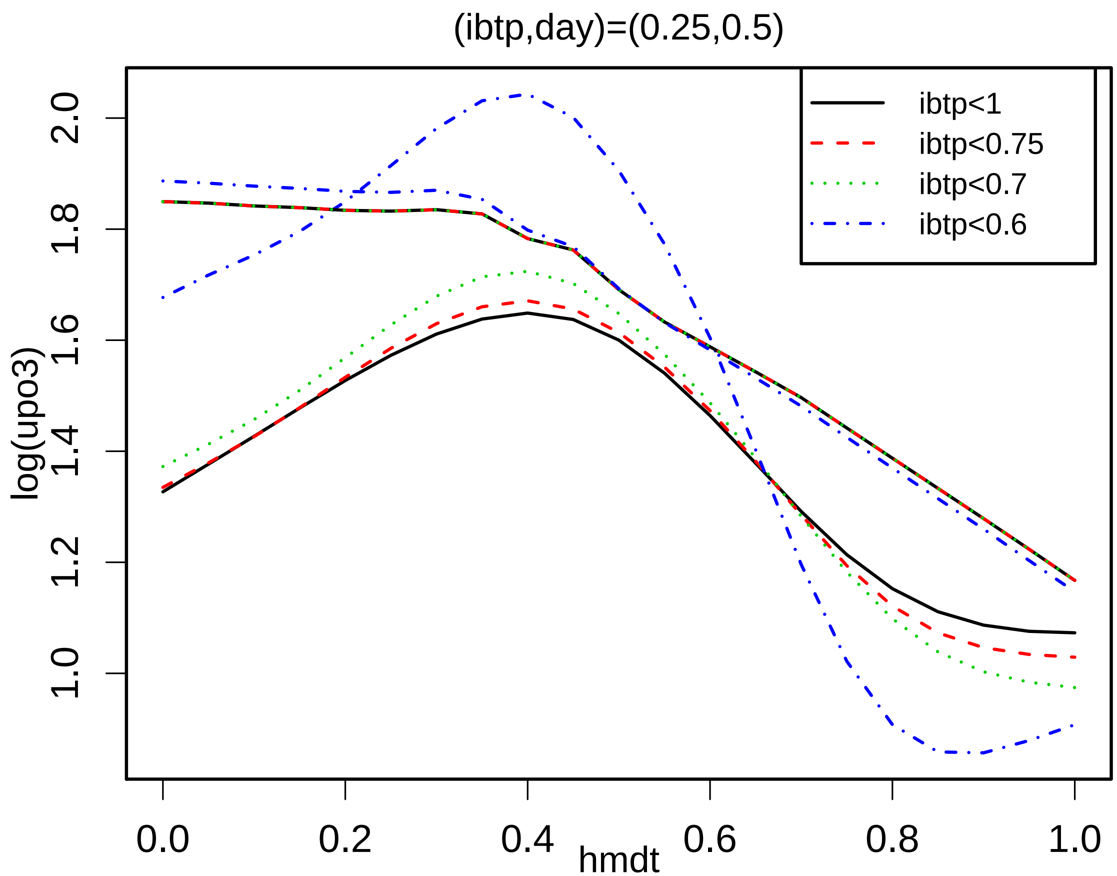
R package gss (Hastie, Tibshirani 1990)

Estimates at slices log(upo3) vs hmdt | ibtp,day



- local additive with 95% CI
- additive with interaction (gam)

Comparison of estimates for reduced sample



- gam estimator for small hmdt mainly depends on data for large hmdt

Conclusions

Non-parametric smoothing methods are

- increasingly implemented in statistical packages,

- useful tools in the evaluation of relations without restricting constraints.

- similarities are larger than differences

- choice of smoothing parameters is central