Nieparametryczna regresja logistyczna, wielowymiarowe splajny, wygładzanie falkowe

(podrozdziały 5.6, 5.7, 5.9)

Patrycja Hęćka

Spis treści

1. Nieparametryczna regresja logistyczna

Splajny mogą być również używane w sytuacji, gdy zmienna odpowiedzi jest jakościowa.
Zajmiemy się problemem wygładzania splajnów w kontekście regresji logistycznej.

Model:

$$\log \frac{\Pr(Y=1|X=x)}{\Pr(Y=0|X=x)} = f(x)$$

$$\Pr(Y=1|X=x) = \frac{e^{f(x)}}{1+e^{f(x)}}$$

Konstruujemy kryterium log-wiarogodności:

$$\begin{aligned} \ell(f;\lambda) &= \sum_{i=1}^{N} [y_i \log p(x_i) + (1-y_i)\log(1-p(x_i))] - \frac{1}{2}\lambda \int \{f''(t)\}^2 dt \\ &= \sum_{i=1}^{N} \left[y_i f(x_i) - \log(1+e^{f(x_i)}) \right] - \frac{1}{2}\lambda \int \{f''(t)\}^2 dt, \qquad (5.30) \end{aligned}$$

gdzie p(x)=P(Y=1|x), a λ to dopasowany parametr wygładzający.

Optymalna funkcja f to skończenie-wymiarowy naturalny splajn z węzłami (knots) w unikatowych wartościach x. Zatem możemy przedstawić funkcję f jako: $f(x) = \sum_{j=1}^{N} N_j(x)\theta_j$,

gdzie $N_j(x)$ to N-wymiarowy zestaw funkcji bazowych, służący reprezentacji rodziny splajnów naturalnych.

Teraz zajmiemy się estymacją współczynników θ_j. Po wstawieniu funkcji f(x) do kryterium log-wiarogodności, liczymy pierwszą i drugą pochodną:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \mathbf{N}^T(\mathbf{y}-\mathbf{p}) - \lambda\mathbf{\Omega}\theta,$$
$$\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T} = -\mathbf{N}^T\mathbf{W}\mathbf{N} - \lambda\mathbf{\Omega},$$

- p – wektor z elementami p(x_i)
- W – macierz diagonalna z wagami $\mathrm{p}(\mathrm{x_i})\left(1-\mathrm{p}(\mathrm{x_i})\right)$
- $\{N\}_{ij} = N_j(x_i), \{\Omega_N\}_{jk} = \int N_j''(t)N_k''(t)\,\mathrm{d}t$
- Pierwsza pochodna jest nieliniowa, dlatego użyjemy iteracyjnego algorytmu Newtona-Raphsona (4.4.1)

$$\theta^{new} = \theta^{old} - \left(\frac{\partial^2 \ell(\theta^{old})}{\partial \theta \partial \theta^T}\right)^{-1} \frac{\partial \ell(\theta^{old})}{\partial \theta}$$

- Otrzymujemy:

$$\begin{aligned} \theta^{\text{new}} &= (\mathbf{N}^T\mathbf{W}\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}^T\mathbf{W}\left(\mathbf{N}\theta^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})\right) \\ &= (\mathbf{N}^T\mathbf{W}\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}^T\mathbf{W}\mathbf{z}. \end{aligned}$$

- W terminach dopasowanych wartości:

$$\begin{aligned} \mathbf{f}^{\text{new}} &= \mathbf{N}(\mathbf{N}^T\mathbf{W}\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}^T\mathbf{W}\left(\mathbf{f}^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})\right) \\ &= \mathbf{S}_{\lambda,w}\mathbf{z}. \end{aligned}$$

- $S_{\lambda,w}$ jest nazywana macierzą wygładzenia.
- Aktualizacje dopasowują wygładzony splajn do zmiennej odpowiedzi z.
- W drugim wyniku możemy zastępować $S_{\lambda,\omega}$ dowolnym nieparametrycznym operatorem regresji w celu otrzymania różnych nieparametrycznych modeli regresji logistycznej.
- Możemy uogólniać procedurę na wyższe wymiary.

2. Wielowymiarowe splajny

- Rozważmy $X \in R^2$,
- Wtedy funkcje bazowe $h_{1k}(X_1), k = 1, \dots, M_1$ będą reprezentowały funkcje dla współrzędnej X_1, a zbiór M_2 funkcji bazowych $h_{2k}(X_2)$ dla współrzędnej X_2
- Wówczas następująca $M_1 \times M_2$ -wymiarowa baza

$$g_{jk}(X) = h_{1j}(X_1)h_{2k}(X_2),\ j = 1, \dots, M_1,\ k = 1, \dots, M_2$$

może być użyta do reprezentacji dwuwymiarowej funkcji:

$$g(X) = \sum_{j=1}^{M_1} \sum_{k=1}^{M_2} \theta_{jk} g_{jk}(X)$$

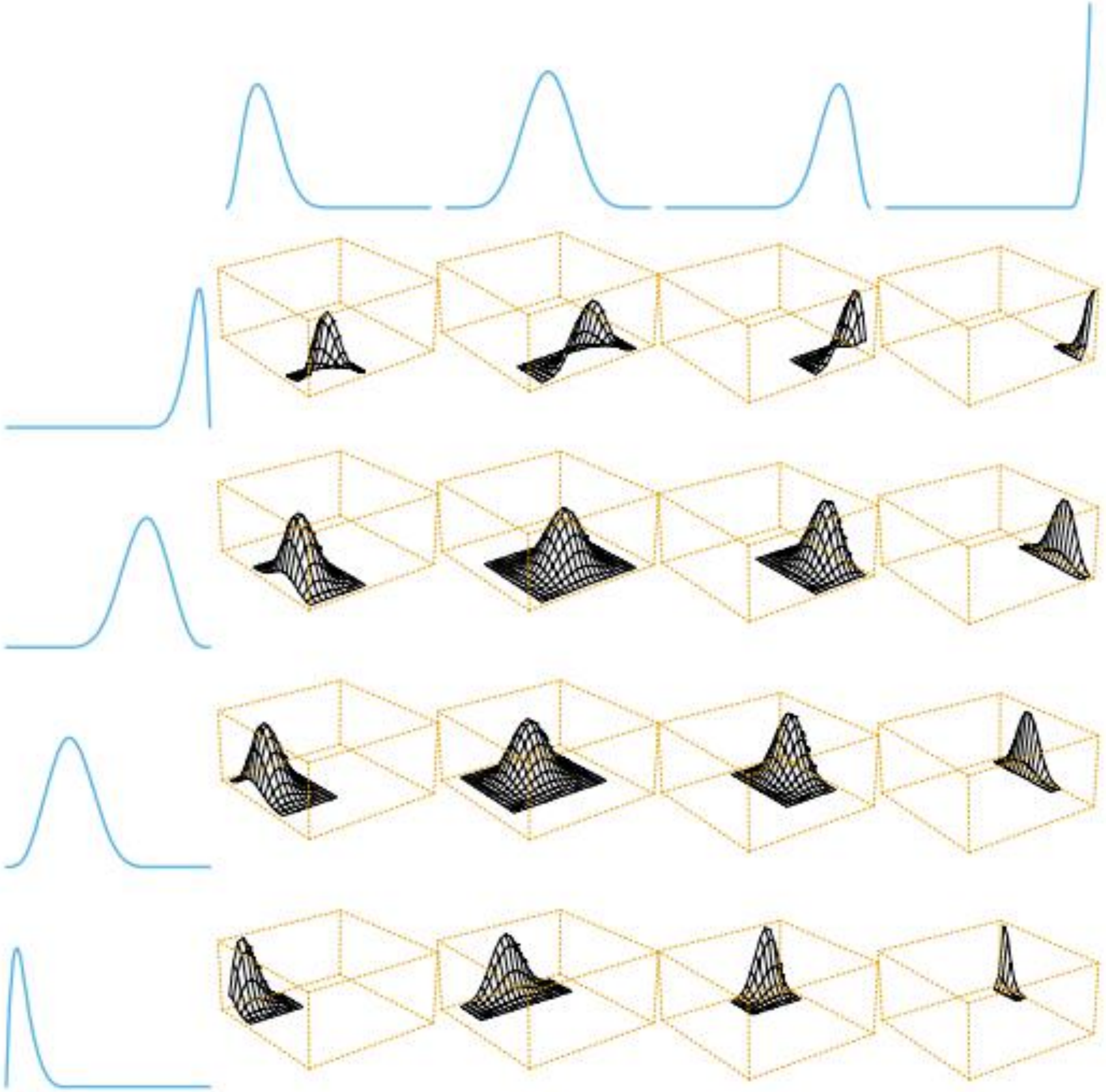
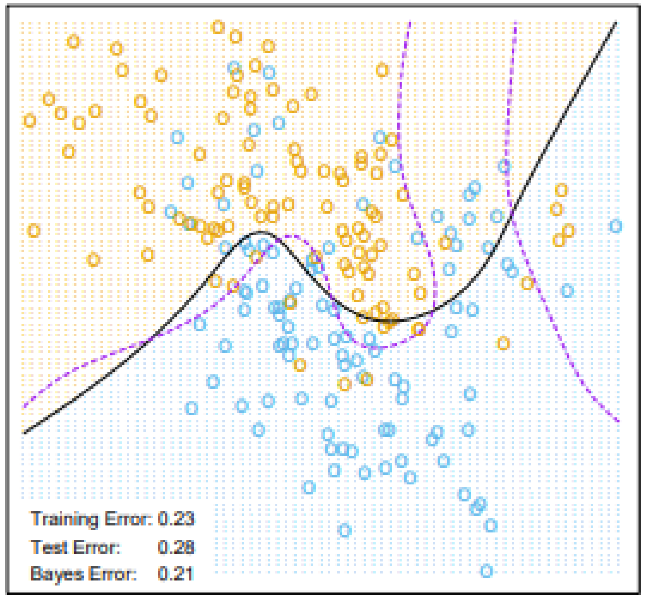


FIGURE 5.10. *A tensor product basis of B-splines, showing some selected pairs. Each two-dimensional function is the tensor product of the corresponding one dimensional marginals.*

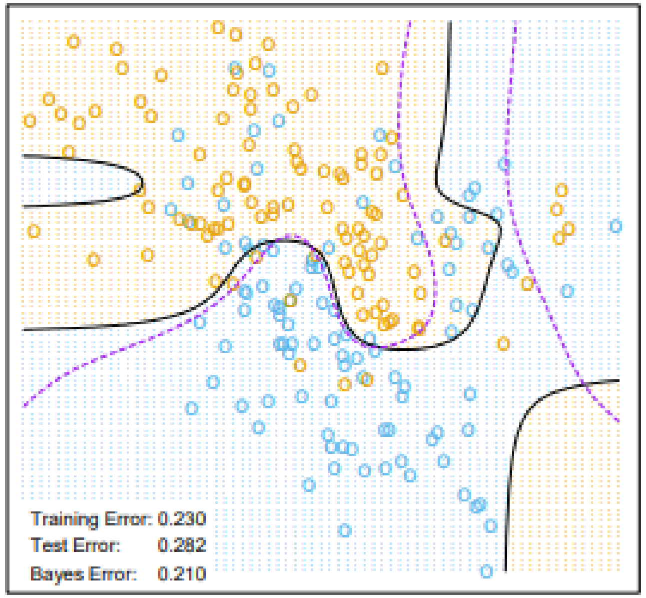
- Figura ilustruje bazę dla iloczynu, używając metody B-splines.
- Współczynniki są dopasowywane poprzez metodę najmniejszych kwadratów.
- Możemy uogólnić metodę do d wymiarów. Wtedy wymiar bazy rośnie wykładniczo-szybko.
- W rozdziale 9 zostanie omówiona procedura „MARS”, która służy do wybrania tylko tych elementów bazy iloczynowej, które są konieczne do zaprezentowania funkcji.

Additive Natural Cubic Splines - 4 df each



Granica decyzyjna dla addytywnego logistycznego modelu regresji.

Natural Cubic Splines - Tensor Product - 4 df each



Granica decyzyjna dla iloczynu naturalnych funkcji bazowych.

Fioletowa, przerywana linia to Bayesowska granica decyzyjna.

Uogólnienie na wyższe wymiary

- Załóżmy, że mamy pary y_i, x_i, gdzie $x_i \in R^d$;
- Szukamy d-wymiarowej funkcji regresji f(x)
- Problem:

$$\min_f \sum_{i=1}^{N} \{y_i - f(x_i)\}^2 + \lambda J[f]$$

Gdzie J to kara na funkcję f. Np. w R^2:

$$J[f] = \int\int_{\mathbb{R}^2} \left[\left(\frac{\partial^2 f(x)}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 f(x)}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 f(x)}{\partial x_2^2}\right)^2\right] dx_1 dx_2$$

- Optymalizacja problemu z taką karą prowadzi do dwuwymiarowej płaszczyzny, znanej jako „thin-plate spline”, która ma następujące właściwości:
- Gdy $\lambda \rightarrow 0$, rozwiązanie zbliża się do funkcji interpolującej,
- Gdy $\lambda \rightarrow \infty$, rozwiązanie zbliża się do płaszczyzny najmniejszych kwadratów,
- Dla pośrednich wartości λ, rozwiązanie może być reprezentowane jako liniowa rozbudowa funkcji bazowych. Wtedy współczynniki są otrzymywane poprzez uogólnioną regresję grzbietową.

- Rozwiązanie ma formę:

$$f(x) = \beta_0 + \beta^T x + \sum_{j=1}^{N} \alpha_j h_j(x)$$

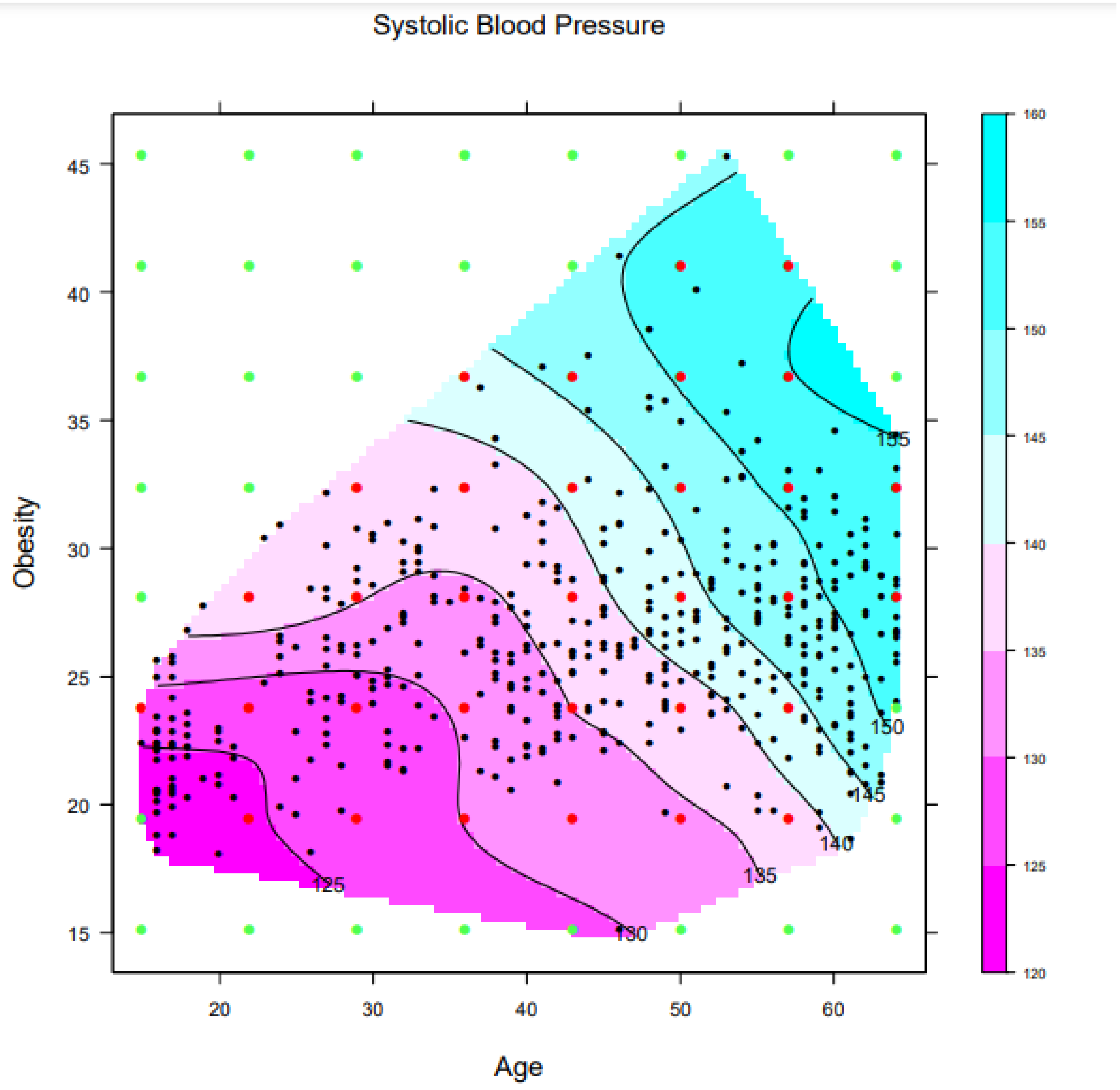
gdzie $h_j(x) = \|x - x_j\|^2 \log\|x - x_j\|$.

- Współczynniki znajdujemy poprzez metodę najmniejszych kwadratów.

„Thin-plate splines”

Zdefiniujemy problem ogólniej dla wymiaru d.

- Jest wiele podejść hybrydowych, popularnych w praktyce przez prostotę obliczeniową.
- Złożoność obliczeniowa wynosi $O(N^3)$.
- Możemy używać mniejszej liczby węzłów niż zalecana przez rozwiązanie.
- W praktyce zazwyczaj pracuje się z kratą węzłów pokrywającą dziedzinę.
- Użycie K węzłów redukuje obliczenia do $O(NK^2 + K^3)$.



„Thin-plate spline” dopasowany do danych o chorobie serca. Zmienną odpowiedzi jest ciśnienie skurczowe, modelowane jako funkcja wieku i otyłości.

- Ogólniej można przedstawić f∈ R^d jako rozszerzenie dowolnego, dużego zbioru funkcji bazowych, a złożoność kontrolować poprzez zastosowanie kary - J(f).
- Np. możemy stworzyć bazę, tworząc iloczyny wszystkich par jednowymiarowych funkcji bazowych, używając jednowymiarowych b-splajnów jako składników.
- Wraz ze wzrostem wymiaru rosłaby jednak wykładniczo liczba funkcji bazowych. Należałoby wtedy zmniejszyć liczbę funkcji na każdą współrzędną.

3. Modele addytywne - przypomnienie

- Modele regresji odgrywają ważną rolę w analizie danych, predykcji i klasyfikacji,
- Prosty, tradycyjny model liniowy czasami zawodzi, gdyż w rzeczywistości efekty często nie są liniowe,
- Zobaczymy teraz bardziej elastyczne metody statystyczne, które mogą służyć do identyfikacji i charakteryzowania efektów regresji nieliniowej. Metody te nazywane są uogólnionymi modelami addytywnymi (GAM).

- Uogólniony model addytywny jest postaci:

$$\mathrm{E}(Y|X_1, X_2, \ldots., X_d) = \alpha + f_1(X_1) + \cdots + f_d(X_d),$$

- gdzie f_j są nieokreślonymi funkcjami gładkimi,
- $X_1, X_2, \ldots., X_d$ to predyktory.

- Dla dwuklasowej klasyfikacji możemy użyć modelu regresji logistycznej. Oznaczając poprzez $\mu(X) = P(Y = 1|X)$ otrzymamy:

$$\log \frac{\mu(X)}{1 - \mu(X)} = \alpha + \beta_1 X_1 + \ldots + \beta_p X_p$$

- W addytywnym, modelu logistycznym zamieniamy każdy składnik liniowy na bardziej funkcjonalną formę:

$$\log \frac{\mu(X)}{1 - \mu(X)} = \alpha + f_1(X_1) + \ldots + f_p(X_p)$$

gdzie ponownie każda funkcja f jest nieokreśloną funkcją gładką.

Widzimy, że addytywność zostaje zachowana, a postać nieparametryczna funkcji f sprawia, że model jest bardziej elastyczny.

- Uogólniając warunkowa średnia $\mu(X)$ jest związana z funkcją addytywną predyktorów poprzez funkcję łączącą g:

$$g[\mu(X)] = \alpha + f_1(X_1) + \ldots + f_p(X_p)$$

- Przykładami klasycznych funkcji łączących są:
- g(μ)= μ – funkcja identycznościowa, używana w liniowych i addytywnych modelach;
- g(μ)=logit(μ) lub g(μ)=probit(μ) – używane w modelowaniu prawdopodobieństw dwumianowych;
- g(μ)=log(μ) dla log-liniowych i log-addytywnych modeli.

GAM z naturalnymi splajnami

- Jeśli zamodelujemy każdą funkcję f_j jako naturalny splajn, możemy dopasować model używając metody najmniejszych kwadratów lub metody największej wiarogodności.
- Dla naturalnych sześciennych splajnów (cubic splines) mamy następujący uogólniony addytywny model:

$$g(\mu) = \sum_{j=1}^{p} \sum_{k=1}^{K(j)} \beta_{jk} N_k(X_j) + \epsilon,$$

gdzie K(j) to liczba węzłów dla zmiennej j.

Modele addytywne z gładkimi splajnami

- Istnieje kara J(f), która gwarantuje rozwiązanie postaci:

$f(X) = \alpha + f_1(X_1) + \cdots + f_d(X_d)$, gdzie każda z funkcji f jest jednowymiarowym splajnem.

- W tym przypadku wystarczy nałożyć karę na każdą z funkcji składowych:

$$\begin{aligned} J[f] &= J(f_1 + f_2 + \cdots + f_d) \\ &= \sum_{j=1}^{d} \int f_j''(t_j)^2 dt_j. \end{aligned}$$

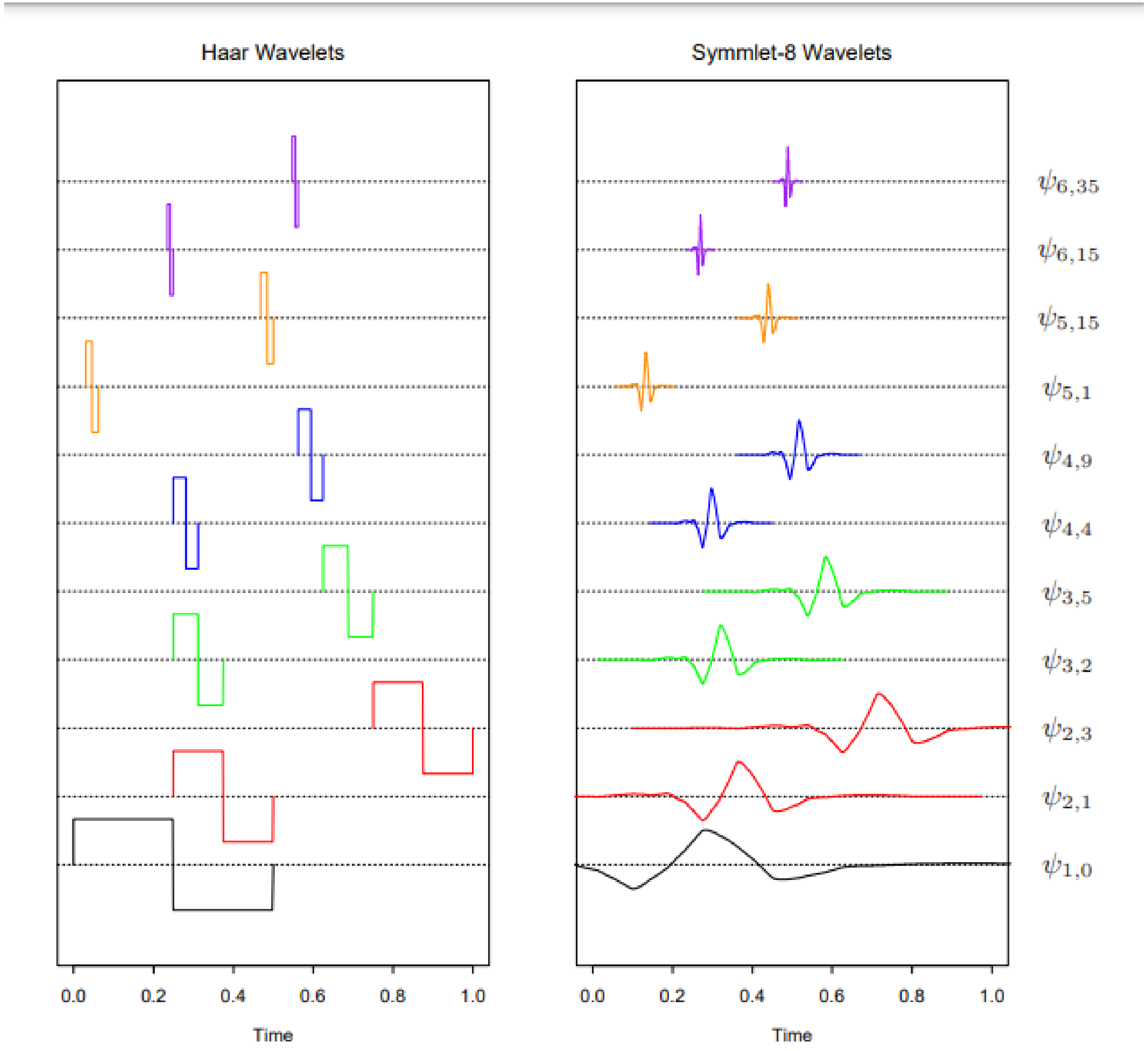
- Możemy określić dla tego problemu:

$$SS(\alpha, f_1, \ldots, f_p) = \sum_{i=1}^{N} \left(y_i - \alpha - \sum_{j=1}^{p} f_j(x_{ij}) \right)^2 + \sum_{j=1}^{p} \lambda_j \int f_j''(t_j)^2 dt_j,$$

Minimalizując SS rzeczywiście otrzymujemy, że każda z funkcji f jest splajnem sześciennym (cubic spline) z węzłami w unikatowych wartościach x_{ij}.

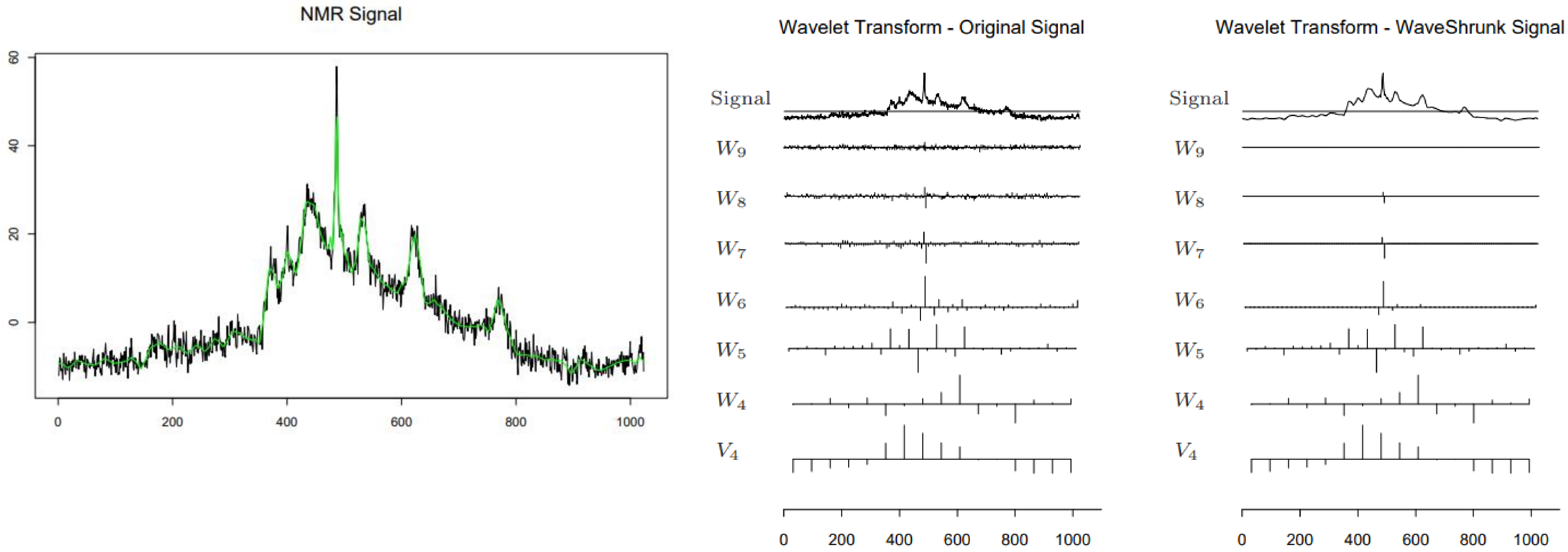
4. Wygładzanie falkowe (wavelet smoothing)

- Falki używają kompletnie ortonormalnych baz do reprezentacji funkcji. Współczynniki kurczy się i wybiera w taki sposób, aby tworzyły one rzadką reprezentację.
- Tak jak gładką funkcję można reprezentować za pomocą kilku bazowych splajnów, w większości gładka funkcja z kilkoma wypukłościami może być reprezentowana poprzez kilka funkcji bazowych z wypukłościami.
- Bazy falkowe są bardzo popularne w przetwarzaniu i kompresji sygnałów, ponieważ są one w stanie reprezentować oba te aspekty – zarówno gładkie jak i lokalnie wyboiste funkcje, w sposób efektywny.



Wybrane falki w różnych translacjach
i dylatacjach stosowanych dla falek
Haar i symmlet.
Oś pionowa wskazuje skalę falek.

- Wygładzanie falkowe dopasowuje współczynniki poprzez metodę najmniejszych kwadratów, a następnie odrzuca zbyt małe współczynniki.
- Ponieważ jest wiele funkcji bazowych w każdej skali, można używać dowolnych funkcji bazowych gdziekolwiek byłyby potrzebne i odrzucić te, które są niepotrzebne w celu osiągnięcia odpowiedniego czasu i lokalizacji częstotliwości.
- Falki Haar są łatwe do zrozumienia, ale niewystarczająco gładkie dla większości celów.
- Falki symmlet mają te same właściwości ortonormalne i są gładsze.



Rysunek przedstawia sygnał NMR (jądrowego rezonansu magnetycznego), który
składa się z gładkich składników, kolców oraz delikatnego szumu.
Przedstawiono transformację falkową, wykorzystującą bazę symmlet. Współczynniki
falkowe są ułożone w wierszach – od współczynnika z najmniejszą skalą na dole, do tego z największą do góry.
Ostatni obrazek pokazuje współczynniki falkowe po nałożeniu ograniczeń (po zmniejszeniu).
Zauważmy, że wiele mniejszych współczynników zostało wyzerowanych.
Zielona linia na pierwszym obrazku przedstawia wygładzoną wersję oryginalnego sygnału.

Konstrukcja baz falkowych

- Bazy falkowe są generowane poprzez translacje i dylatacje pojedynczych funkcji ϕ(x) (zwanych „ojcami”).
- Czerwone linie to funkcje skalujące

Haar i symmlet, a zielone to tzw. „matki”.

- Konstrukcja bazy Haar:

Jeśli ϕ(x) = I(x ∈ [0, 1]), to $\phi_{0,k}$(x) = ϕ(x−k) generuje ortonormalną bazę

dla funkcji ze skokami w liczbach całkowitych k.

Nazywamy to referencyjną bazą V_0.

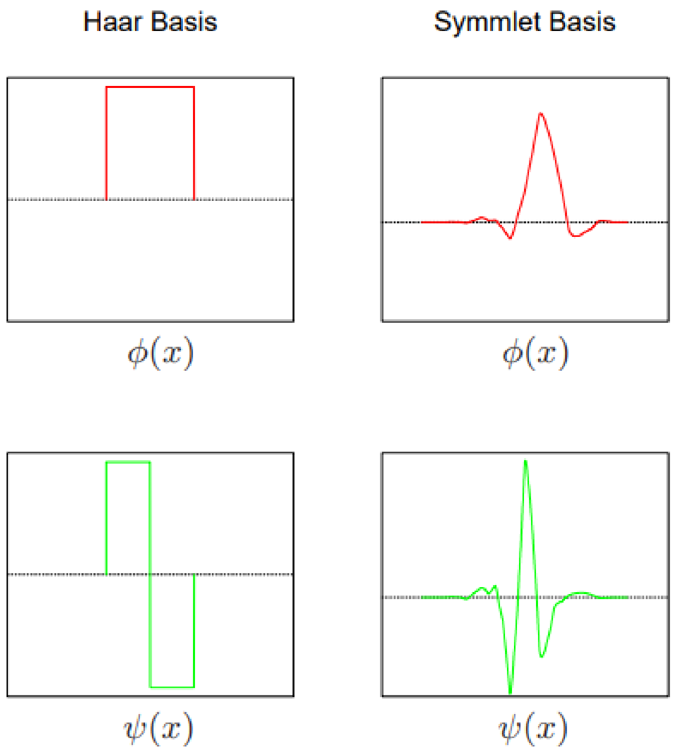
Dylatacja $\phi_{1,k}$(x) = $\sqrt{2}$ϕ(2x−k) tworzy ortonormalną bazę dla

przestrzeni $V_1 \supset V_0$ funkcji kawałkami ciągłej na

przedziale długości ½. Ogólniej mamy

$\cdots \supset V_1 \supset V_0 \supset V_{-1} \supset \cdots$, gdzie dla każdego V_j:

$$\phi_{j,k} = 2^{j/2}\phi\left(2^j x - k\right)$$



- Reprezentujemy funkcję w przestrzeni V_{j+1} poprzez składnik z V_j plus składnik z dopełnienia ortogonalnego przestrzeni V_j: W_j, co zapisujemy jako V_{j+1} = $V_j \oplus W_j$.
- Funkcja ψ(x−k) generowana przez matkę-falkę ψ(x) = ϕ(2x)−ϕ(2x−1) tworzy ortonormalną bazę dla przestrzeni W_0 w rodzinie Haar.
  Zaś $\psi_{j,k} = 2^{j/2}\psi\left(2^j x - k\right)$ tworzy bazę dla W_j.
- V_{j+1} = $V_j \oplus W_j$ = $V_{j-1} \oplus W_{j-1} \oplus W_j$, stąd:

V_j = $V_0 \oplus W_0 \oplus W_1 \cdots \oplus W_{j-1}$.

Dopóki te przestrzenie są ortogonalne, wszystkie funkcje bazowe są ortonormalne.

Adaptacyjne filtrowanie falkowe

- Zajmiemy się przypadkiem jednowymiarowym z N=2^j dyskretnymi punktami.
- y – wektor odpowiedzi, W – NxN ortonormalna baza falkowa
- Wtedy $y^* = W^T y$ nazywamy falkową transformacją y.
- Popularną metodą, której można użyć do adaptacyjnego filtrowania falkowego jest SURE (Stein Unbiased Risk Estimation).
- Zaczynamy z kryterium: $\min_{\boldsymbol{\theta}} ||\mathbf{y} - \mathbf{W}\boldsymbol{\theta}||_2^2 + 2\lambda||\boldsymbol{\theta}||_1$

Ponieważ W jest ortonormalna, rozwiązaniem jest: $\hat{\theta}_j = \mathrm{sign}(y_j^*)(|y_j^*| - \lambda)_+$

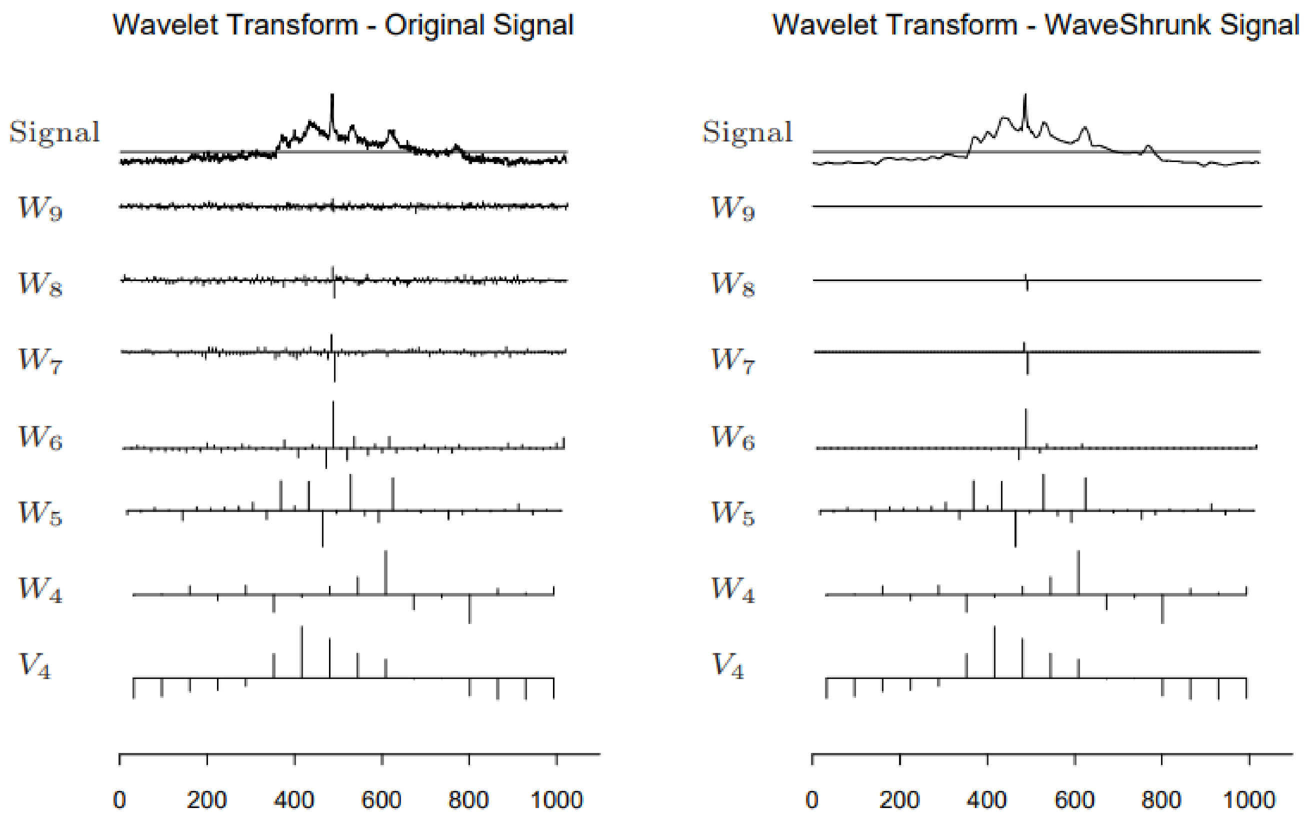
- Jak wybrać λ ?

W jest ortonormalną transformacją, jeśli elementy y są białym szumem.

Dodatkowo jeśli $Z_1, \ldots, Z_N$ to biały szum, to oczekiwane maksimum | Z_j| to σ $\sqrt{2\log n}$, gdzie σ to estymator odchylenia standardowego szumu. .

Stąd wszystkie współczynniki poniżej σ $\sqrt{2\log n}$ są uważane za szum i przyrównywane do 0, więc wybieramy λ = σ $\sqrt{2\log n}$.

• Ucinanie zastosowane
przy sygnale NMR
zostało uzyskane poprzez SURE.



Bibliografia

Trevor Hastie, Robert Tibshirani, Jerome Friedman - The Elements of Statistical Learning Data Mining, Inference, and Prediction, Springer Series in Statistics, Second Edition, February 2009, pages: 161-181

DZIĘKUJĘ ZA UWAGĘ!