Semiparametric Modeling for Longitudinal Data Analysis

Jianqing Fan

Chinese University of Hong Kong

joint work with Runze Li

http://www.stat.unc.edu/faculty/fan.html

October 4, 2002

Longitudinal Data

Data: From individual treatment effects over time.

Nature: Highly unbalanced, observing at irregular time points.

Time: i^{th} individual observed at time t_{ij}, $j = 1, \cdots, J_i$.

Collected data: covariate vector $\mathbf{x}_i(t)$ along with its associated response y_i at time t_{ij}:

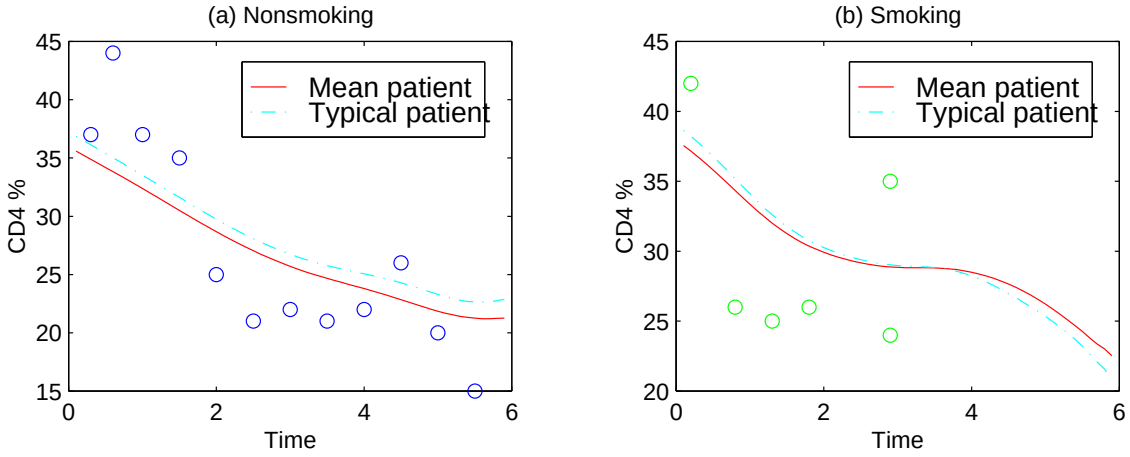
$$\{(t_{ij}, \mathbf{x}_i(t_{ij}), y(t_{ij})), j = 1, \cdots, J_i\}$$

Notation: $\mathbf{y}_i = (y_i(t_{i1}), \cdots, y_i(t_{iJ_i}))^T$, $\mathbf{X}_i = (\mathbf{x}_i(t_{i1}), \cdots, \mathbf{x}_i(t_{iJ_i}))^T$.

Example 1 — CD4 data

Measurements (from multi-center AIDS cohort study): CD4-cell percentages and other important covariates for 283 homosexual men infected during 84-91 were measured over a period of time in order to monitor AIDS progression (Kaslow *et al.* 1987).

Covariates: $X_1(t) =$ Smoking, $X_2(t) =$ age, $X_3(t) =$ preCD4.



Missing data: Very heavy; fewer data points at the end

Question: — How to model $\mathbf{X}(t)$ and $Y(t)$?

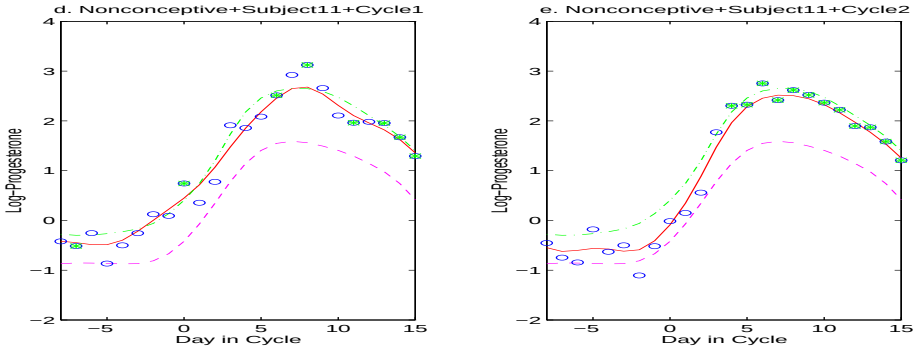
— How to predict individual's progression?

Example 2 — Hormone Data

Measurements: urinary metabolite progesterone curves measured over 21 conceptive and 70 nonconceptive women menstrual cycles.

Nesting:

conceptive		nonconceptive	
subject	subject	subject	subject
cycle	cycle	cycle	cycle



Alignment: aligned and truncated around the day of ovulation.

Data: $Y_{ijk}(t)$ with missing values in some cycles.

Question: How to analyze nested functional data?

Statistical models (I)

Classical Linear Models: [Diggle, Liang and Zeger (1994) and Hand and Crowder (1996)]

$$\mathbf{y(t)} = \beta_0 + \beta_1 \mathbf{x_1(t)} + \cdots + \beta_p \mathbf{x_p(t)} + \varepsilon(\mathbf{t})$$

Coefficients are independent of time t

Method: Weighted LS using a **working** covariance matrix W_i:

$$\sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta})^T W_i (\mathbf{y}_i - \boldsymbol{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta})$$

A semiparametric model: **allowing baseline to be time-dependent**

$$y(t) = \beta_0(\mathbf{t}) + \beta_1 x_1(t) + \cdots + \beta_p x_p(t) + \varepsilon(t)$$

See Moyeed and Diggle (1994), Zeger and Diggle (1994), Martinussen and Scheike (1999), Lin and Ying (2001).

Statistical models (II)

Functional linear model: (**Varying-coefficient model**)

$$Y(t) = \beta_0(t) + \beta_1(t)X_1(t) + \cdots + \beta_p(t)X_p(t) + \varepsilon(t)$$

Hoover *et al* (1997), Brumback and Rice (1998) and Fan and Zhang (2000), Chiang, Rice and Wu (2001), Chiang, Wu and Zhou (2002).

Comparisons:

— Linear model is too restrictive, can not capture time effect.

— Functional linear model is very flexible and understandable, but its coefficients can not be well estimated (collinearity), and are **not** as interpretable.

— The semiparametric model falls between these two. The $\boldsymbol{\beta}$ admits similar interpretation. It captures some time effect but **not as flexible** as the FL model.

Lin and Ying's approach

Model: $y(t) = \alpha(t) + \boldsymbol{\beta}^T \mathbf{x}(t) + \varepsilon(t)$.

Counting process: $N_i(t) \equiv \sum_{j=1}^{J_i} I(t_{ij} \leq t)$

Assumption's:

(a) **Observed time**: $N_i(t) = N_i^*(t \wedge c_i)$ (somewhat artificial)

(b) **Noninform. censoring**: $E\{y_i(t)|\mathbf{x}_i(t), c_i \geq t\} = E\{y_i(t)|\mathbf{x}_i(t)\}$.

(c) **Observations**: The trajectories $[\mathbf{x}(t), y(t)]$ are fully observable until the censoring time c_i. This is unrealistic, and can be implemented by **linear interpolation** or **constant interpolation**

(d) **Working independence**: W_i is diagonal.

LY method — independence

Assumption: Observation times are independent of $\mathbf{x}(t)$.

Direct approach: Regarding $\alpha(t)$ as a parameter, minimize

$$\sum_{i=1}^{n} \int_0^{+\infty} w(t)\{y_i(t) - \alpha(t) - \boldsymbol{\beta}^T \mathbf{x}_i(t)\}^2 \, dN_i(t) = \sum_i \left\{ \sum_{j=1}^{J_i} w(t_{ij})\{y_i(t_{ij}) - \cdots \right.$$

Baseline est.: $\hat{\alpha}(t;\boldsymbol{\beta}) = \bar{y}(t) - \boldsymbol{\beta}^T \bar{\mathbf{x}}(t)$ with $\xi_i(t) = I(c_i \geq t)$

$$\bar{\mathbf{x}}(t) = \sum_{i=1}^{n} \xi_i(t)\mathbf{x}_i(t) / \sum_{i=1}^{n} \xi_i(t), \quad \bar{y}(t) = \sum_{i=1}^{n} \xi_i(t) y_i(t) / \sum_{i=1}^{n} \xi_i(t)$$

Parameter estimate: Minimize

$$\sum_{i=1}^{n} \int_0^{+\infty} w(t)[\{y_i(t) - \bar{y}(t)\} - \boldsymbol{\beta}^T \{\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\}]^2 \, dN_i(t).$$

LY method — dependence

Intensity: $E\{dN_i^*(t)|\mathbf{x}_i(t), y_i(t), c_i \geq t\} = \lambda(t)\exp\{\boldsymbol{\gamma}'\mathbf{x}_i(t)\}$ (**Cox model**). $\boldsymbol{\gamma} = 0$ corresponds to the independence case.

Baseline estimation: $\hat{\alpha}(t;\boldsymbol{\beta}) = \bar{y}(t;\boldsymbol{\gamma}) - \boldsymbol{\beta}^T\bar{\mathbf{x}}(t;\boldsymbol{\gamma})$ with

$$\bar{\mathbf{x}}(t,\boldsymbol{\gamma}) = \frac{\sum_{i=1}^n \xi_i(t)\exp\{\boldsymbol{\gamma}^T\mathbf{x}_i(t)\}\mathbf{x}_i(t)}{\sum_{i=1}^n \xi_i(t)\exp\{\boldsymbol{\gamma}^T\mathbf{x}_i(t)\}},$$

Parameter estimation: Minimize

$$\sum_{\mathbf{i=1}}^{\mathbf{n}} \int_{\mathbf{0}}^{+\infty} \mathbf{w(t)}[\{\mathbf{y_i(t)} - \bar{\mathbf{y}}\mathbf{(t;\boldsymbol{\gamma})}\} - \boldsymbol{\beta}^{\mathbf{T}}\{\mathbf{x_i(t;\boldsymbol{\gamma})} - \bar{\mathbf{x}}\mathbf{(t;\boldsymbol{\gamma})}\}]^{\mathbf{2}}\,\mathbf{dN_i(t)}.$$

Estimation of nuisance parameter: Solving the partial likelihood equation (see Pepe and Cai, 1993):

$$\sum_{\mathbf{i=1}}^{\mathbf{n}} \int_{\mathbf{0}}^{+\infty} \{\mathbf{x_i(t)} - \bar{\mathbf{x}}\mathbf{(t,\boldsymbol{\gamma})}\}\,\mathbf{dN_i(t)} = \mathbf{0}.$$

Remarks

— LY method is **simple** and does not involve **any smoothing**.

— No smoothness is used in estimation. Hence, the efficiency can be **significantly improved**.

— Interpolation can create large biases, making **procedures inconsistency**.

— Counting process formulation and its nuisance parameters estimation are **cumbersome and artificial**.

— The difference based estimator overcome all disadvantages and does not involve any smoothing. It is useful for smoothing in the profile least-squares estimator.

Difference based method

Idea: From Fan and Huang (2001), Yatchew (1997). Order

$$\{(t_{ij}, \mathbf{x}(t_{ij})^T, \mathbf{y}(t_{ij})), j = 1, \cdots, J_i, i = 1, \cdots, n\}$$

according to $\{t_{ij}\}$ as $\{(t_i, \mathbf{x}_i^T, \mathbf{y}_i), i = 1, \cdots, n^*\}$, with $n^* = \sum_{i=1}^{n} J_i$.

Marginal model: $y_i = \alpha(t_i) + \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i$.

Differencing: The nuisance function eliminated by differencing:

$$y_{i+1} - y_i = \alpha(t_{i+1}) - \alpha(t_i) + \boldsymbol{\beta}^T(\mathbf{x}_{i+1} - \mathbf{x}_i) + e_i \approx \boldsymbol{\beta}^T(\mathbf{x}_{i+1} - \mathbf{x}_i) + e_i.$$

Parameter estimation: Applied weighted LS to

$$y_{i+1} - y_i = \alpha_0 + \alpha_1(t_{i+1} - t_i) + \boldsymbol{\beta}^T(\mathbf{x}_{i+1} - \mathbf{x}_i) + e_i.$$

Limited lost of efficiency: among working independence estimators. Pretend independence. The data $\{y_{2i+1} - y_{2i}\}$ are indep. Lose only $\{y_{2i+1} + y_{2i}\}$, containing less information about $\boldsymbol{\beta}$.

Profile least-squares (I)

Nonparametric regression: Letting $y^*(t) \equiv y(t) - \boldsymbol{\beta}^T \mathbf{x}(t)$, we have

$$y^*(t) = \alpha(t) + \varepsilon(t).$$

Local linear fit: For each given t_0, approximate

$$\alpha(t) \approx \alpha(t_0) + \alpha'(t_0)(t - t_0) \equiv a + b(t - t_0).$$

Minimize with respect to a and b

$$\sum_{i=1}^{n} \sum_{j=1}^{J_i} \{y_i^*(t_{ij}) - a - b(t_{ij} - t_0)\}^2 w(t_{ij}) K_h(t_{ij} - t_0),$$

where $K_h(\cdot) = h^{-1} K(\cdot/h)$, K is a kernel and h is a bandwidth, resulting in $\hat{\alpha}(t_0; \boldsymbol{\beta}) = \hat{a}$.

Notation. $\mathbf{y} = (\mathbf{y}_1^T, \cdots, \mathbf{y}_n^T)$, $\mathbf{X} = (\mathbf{X}_1^T, \cdots, \mathbf{X}_n^T)^T$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_n^T)^T$.

Profile least-squares (II)

Marginal model: $\mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Local linear estimator: linear in response $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. Hence, $\hat{\boldsymbol{\alpha}} = \mathbf{S}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, where $\mathbf{S}$ is a smoothing matrix.

Synthetic model: $(\mathbf{I} - \mathbf{S})\mathbf{y} = (\mathbf{I} - \mathbf{S})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

PLS: $\hat{\boldsymbol{\beta}} = \{\mathbf{X}^T(\mathbf{I} - \mathbf{S})^T\mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{X}\}^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{S})^T\mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{y}$

Estimated covariance matrix: $\text{cov}\{\hat{\boldsymbol{\beta}}|t_{ij}, \mathbf{x}_i(t_{ij})\} = \mathbf{D}^{-1}\mathbf{V}\mathbf{D}^{-1}$, where $\mathbf{D} = \mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{X}$ and $\mathbf{V} = \text{cov}\{\mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}\boldsymbol{\varepsilon}\}$.

The matrix $\mathbf{V}$ can be estimated as $\hat{\mathbf{V}} = \mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}\mathbf{C}\mathbf{W}^T(\mathbf{I}-\mathbf{S})\mathbf{X}$ with $\mathbf{C} = \text{diag}\{\hat{\boldsymbol{\varepsilon}}_1\hat{\boldsymbol{\varepsilon}}_1^T, \cdots, \hat{\boldsymbol{\varepsilon}}_n\hat{\boldsymbol{\varepsilon}}_n^T\}$.

Implementations

Sparsity: The function $\alpha(\cdot)$ can not be estimated well at **both tails**. We only estimate function at inter 90% of the data.

Bandwidth selection: Use DBE to get an estimate $\hat{\boldsymbol{\beta}}_{DBE}$. Letting $y^*(t) = y(t) - \hat{\boldsymbol{\beta}}_{DBE}^T \mathbf{x}(t)$, we have $y^*(t) \approx \alpha(t) + \varepsilon(t)$. Apply a bandwidth selection method to get an $\hat{h}$. Using this $\hat{h}$ to obtain the profile least-squares estimator and the estimator of $\boldsymbol{\alpha}(\cdot)$.

Methods: Cross-validation, pre-asymptotic substitution method (Fan and Gijbels, 1995), asymptotic substitution method (Ruppert *et al*, 1995), empirical bias method (Ruppert, 1997), among others.

Simulation Models

Model: $y(t) = \alpha(t) + \boldsymbol{\beta}^T \mathbf{x}(t) + \varepsilon(t)$, where $\alpha(t) = \tau\sqrt{t/\tau}$ or $\tau \sin(2\pi t/\tau)$, $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$

$\varepsilon(t)$ is Gaussian proc. with cov. $E\{\varepsilon(s)\varepsilon(t)\} = \exp(-2|t-s|)$

$\mathbf{x} \sim N(0, \Sigma)$ with $\text{cov}(x_i, x_j) = 0.5^{|i-j|}$.

Design of observation time

Case I : LY design — **independence**. The process $N^*(t)$ was generated from a random effect **Poisson process** with intensity rate η, where $\eta \sim$ Gamma(1, 0.5). The censoring $c \sim$ unif$(0, \tau)$ with $\tau = 4$ or 20, resulting in, on average, 3 or 11 observations per subject.

Case II : LY design — **dependence**. The intensity function is of form $\eta \exp(0.5x_1)$, instead of η.

case III : Observation time to be $t_{ij} = 0, 1, \cdots, [c_i]$, with $c_i \sim \text{unif}(0, \tau)$, where $\tau = 10$ or 20. On average, 6 or 11 obs. per subject.

case IV Huang, Hu and Zhou (2002) design: Each subject has a set of 'scheduled' time {0,1,3,...,29}, and each scheduled time has a probability of being skipped 60%. For non-skipped time, a uniform [-1, 1] is added to each scheduled time.

Objective: **Performance comparison** and **accuracy of the Sandwich formula**.

Performance measure: MSE $= E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$.

Numerical Results — performance comparison

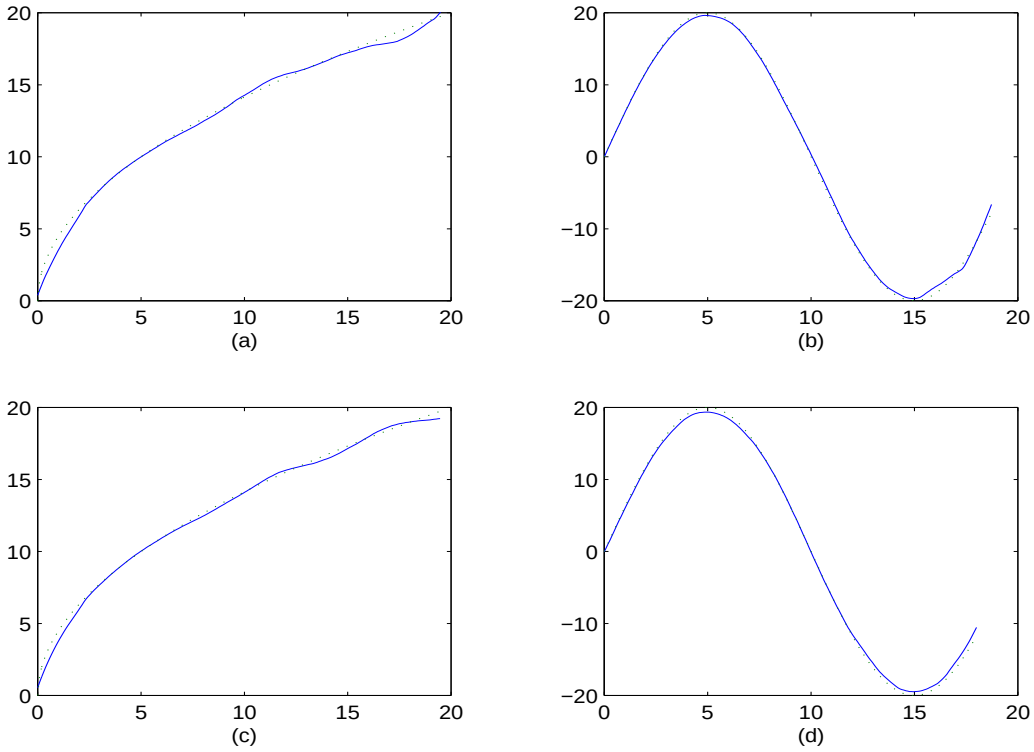
Table 1: Relative Efficiency — Ratios of MSEs with respect to LY estimator

Case I		$n = 50$		$n = 75$	
$\alpha(t)$	τ	DBE	Profile LSE	DBE	Profile LSE
$\tau\sqrt{t/\tau}$	4	0.8481	0.6592	0.8407	0.6661
$\tau\sqrt{t/\tau}$	20	**0.3450**	**0.2962**	0.3963	0.3246
$\tau\sin(2\pi t/\tau)$	4	0.6632	0.5065	0.6756	0.5377
$\tau\sin(2\pi t/\tau)$	20	**0.2798**	**0.2324**	0.3001	0.2359
Case II					
$\tau\sqrt{t/\tau}$	4	0.7868	0.6500	0.7209	0.6004
$\tau\sqrt{t/\tau}$	20	**0.2518**	**0.2138**	0.2438	0.2015
$\tau\sin(2\pi t/\tau)$	4	0.5627	0.4641	0.5409	0.4502
$\tau\sin(2\pi t/\tau)$	20	**0.1705**	**0.1395**	0.1623	0.1280
Case III					
$\tau\sqrt{t/\tau}$	10	**1.0785**	0.7040	1.1299	0.7348
$\tau\sqrt{t/\tau}$	20	**0.7748**	**0.5006**	0.8316	0.5424
$\tau\sin(2\pi t/\tau)$	10	**1.2188**	0.6818	1.2560	0.7347
$\tau\sin(2\pi t/\tau)$	20	**0.9868**	**0.5007**	1.0721	0.5973
Case IV					
$30\sqrt{t/30}$		0.9666	0.6842	1.0501	0.7086
$30\sin(2\pi t/30)$		**0.1434**	**0.0953**	0.1360	0.0869

Number of simulations: 400.

Remark: The deteriorated performance in case 3 is due to wide gap between observation times.

Numerical results— baseline estimation



Remarks. Typical estimated baseline curves with $n = 50$ and $\tau = 20$. Right panel $\alpha(t) = \tau\sqrt{t/\tau}$ for case I and case II, respectively. Left panel for $\alpha(t) = \tau \sin(t/\tau)$ for Cases I and II, respectively.

Numerical results — accuracy of SE

Table 2: Stds and SEs of Profile LSE for Case I with $\alpha(t) = \tau\sqrt{t/\tau}$

(n, τ)	β_1		β_2		β_5	
	SD	se (SD(se))	SD	se (SD(se))	SD	se (SD(se))
(50, 4)	**0.1512**	**0.1377**(0.0327)	0.1683	0.1579(0.0369)	0.1664	0.1543(0.0381)
(75, 4)	**0.1200**	**0.1148**(0.0211)	0.1262	0.1273(0.0243)	0.1274	0.1287(0.0240)
(50, 20)	**0.0854**	**0.0820**(0.0182)	0.1004	0.0910(0.0211)	0.1012	0.0933(0.0203)
(75, 20)	**0.0651**	**0.0675**(0.0130)	0.0718	0.0748(0.0144)	0.0708	0.0749(0.0149)

Note. — The relative Monte Carlo error is of size $1/\sqrt{800}$ for the sample SD. It is **negligible**.

— The discrepancy is less than half of SD(se).

— Results for other cases are similar.

Asymptotic results

Formulation: The asymptotic result depends on the formulation.

Follow Lin and Ying's formulation for comparison.

Notation. Let $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ and

$$\mathbf{A} = E\int_0^\infty \{\mathbf{x}(t) - E\mathbf{x}(t)\}^{\otimes 2} w(t) dN(t)$$

$$\mathbf{B} = E\left\{\int_0^\infty \{\mathbf{x}(t) - E\mathbf{x}(t)\}\varepsilon(t) w(t) dN(t)\right\}^{\otimes 2}.$$

Results: If $h_n = bn^{-a}$, for $1/8 < a < 1/2$, then as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} N(0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}).$$

Consistency of Sandwich Formula: $\mathbf{D}^{-1}\hat{\mathbf{V}}\mathbf{D}^{-1}$ consistently estimates $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$.

Variable selection

Background: To reduce modeling biases, large parametric models are introduced.

Questions: How to

— **automatically select significant variables?**

— **construct confidence intervals?**

— **verify properties of data-driven procedures?**

Challenge: Not yet studied neither in semiparametric models nor longitudinal data.

Traditional Approaches: stepwise procedure, best subset, $\cdots$

Drawbacks: — hard to establish sampling properties;

— expensive in computation

A viable solution

Setting: i.i.d samples first (see Fan and Li, 2001; 2002).

— **Penalized likelihood**: $\ell(\mathbf{X}\boldsymbol{\beta}, Y) - \sum_j p_\lambda(|\beta_j|)$ where e.g. $p_\lambda(|\beta_j|) = \lambda|\beta_j|$.

— **Sandwich formula** based on the penalized likelihood.

— **Oracle property**: Suppose that $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1^T\boldsymbol{\beta}_1 + \mathbf{X}_2^T\boldsymbol{\beta}_2$ with $\boldsymbol{\beta}_2 = 0$. The oracle estimator: $\hat{\boldsymbol{\beta}}_2 = 0$ and $\hat{\boldsymbol{\beta}}_1$ being the MLE of submodel.

— What kind of $p_\lambda(\cdot)$?

Penalty functions

Assumption: Normal likelihood with **orthonormal** $\mathbf{X}$. Then, PLK is equivalent to

$$\|y - \mathbf{x}\boldsymbol{\beta}\|^2 + \sum_j p_\lambda(|\beta_j|) = c + \sum_j \{(z_i - \beta_j)^2 + p_\lambda(|\beta_j|)\},$$

where $\mathbf{z} = \mathbf{x}^T\mathbf{y}$ and $c = \|\mathbf{y} - \mathbf{xz}\|^2$.

Componentwise minimization: $(z_i - \theta)^2 + p_\lambda(|\theta|)$.

L_2 penalty: $p_\lambda(|\theta|) = \lambda|\theta|^2 \Longrightarrow$ **ridge regression**

Entropy penalty: $p_\lambda(|\theta|) = \lambda I(|\theta| \neq 0) \Longrightarrow$ **Best subset**

Hard-thresholding penalty: $p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda)$

$\Longrightarrow$ **Best subset**.

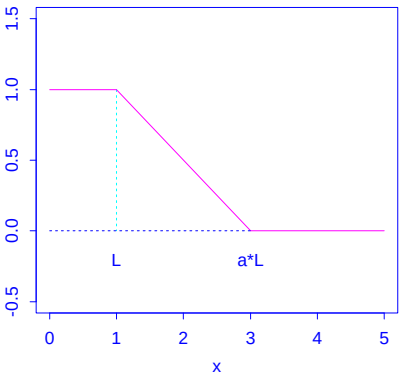
L_1-penalty: $p_\lambda(|\theta|) = \lambda|\theta| \Longrightarrow$ **soft-thresholding** (Dohono and Johnstone, 1994) and LASSO (Tibshirani 1996, 97, Knight and Fu, 2000).

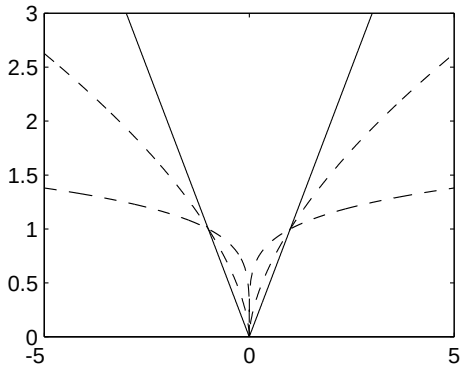
Other penalty functions

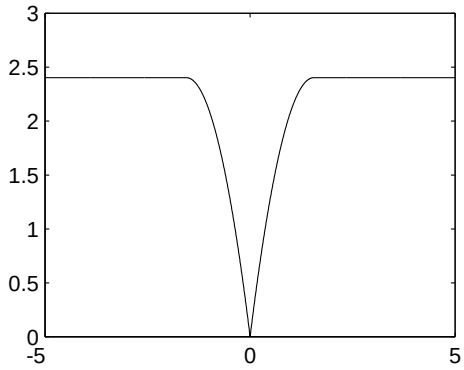
SCAD: $p'_\lambda(\theta) = I(\theta \le \lambda) + \frac{(a\lambda-\theta)_+}{(a-1)\lambda} I(\theta > \lambda)$,

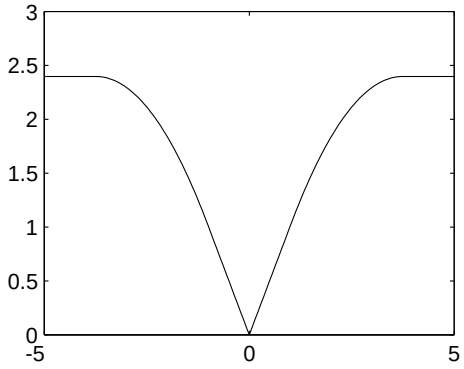
Transformed L_1: $p_\lambda = \lambda a|x|/(1+a|x|)$

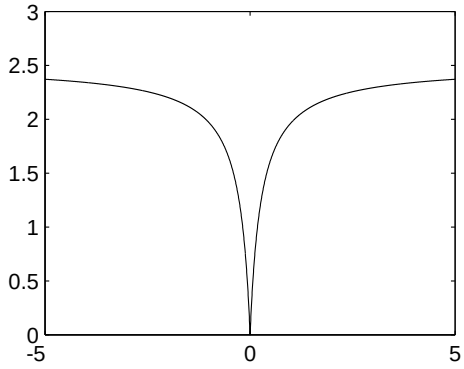
L_p-penalty: $p_\lambda = \lambda|x|^p$











Desired Properties

— **continuity**: to avoid instability in model prediction

$\Longleftrightarrow \text{argmin}_\theta\{|\theta| + p'_\lambda(|\theta|)\} = 0$

— **sparsity**: to reduce model complexity $\Longleftarrow p'_\lambda(0+) > 0$

— **Unbiasedness**: to avoid unnecessary modeling bias $\Longleftrightarrow$

$\lim_{|\theta|\to\infty} p_\lambda(|\theta|) = 0$

Method	Best subset	Ridge	LASSO	SCAD
Continuity		x	x	x
Sparsity	x		x	x
Unbiasedness	x			x

Variable Selection for Semiparametric Model

Weighted profile least-squares: Minimize

$$\ell(\boldsymbol{\beta}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{I} - \mathbf{S})^T\mathbf{W}(\mathbf{I} - \mathbf{S})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Variable selection: Penalized WPLS $\mathcal{L}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|)$

It can be obtained from the penalized likelihood via profiling.

Computation: Use the iterated ridge regression (modified Newton-Raphson method)

Covariance matrix can be estimated from a modified sandwich formula.

Choice of regularization parameter can be obtained by a form of GCV.

Sampling Properties

Notation: $a_n = \max_j\{|p'_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}$

and $b_n = \max_j\{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}$. Note that $a_n = b_n = 0$ for SCAD and hard-thresholding penalty and $a_n = \lambda_n$, $b_n = 0$ for L_1.

Rate of convergence: If $a_n \to 0$ and $b_n \to 0$ then there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $\mathcal{L}(\boldsymbol{\beta})$ such that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_P(n^{-1/2} + a_n)$.

Significant Variables: WOLG $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}, \mathbf{0})$.

Oracle properties: If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, then the above root-n consistent estimator satisfies

— (**Sparsity**) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;

— (**Normality**) $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \to N_s(\mathbf{0}, \mathbf{A}_{11}^{-1}\mathbf{B}_{11}\mathbf{A}_{11}^{-1})$

Remark. No oracle property of L_1-penalty.

Evaluation of the Procedure

Criterion — Prediction error. Letting $\{\tilde{\mathbf{x}}(t), \tilde{y}(t), \tilde{N}(t)\}$ be a new observation,

$$\mathrm{PE}(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = E \int_0^\infty \{\tilde{y}(t) - \hat{\alpha}(t) - \hat{\boldsymbol{\beta}}^T \tilde{\mathbf{x}}(t)\}^2 \, d\tilde{N}(t).$$

$$\mathrm{PE} = \text{Noise error} + \textbf{error of } \hat{\alpha} + \textbf{error of } \hat{\boldsymbol{\beta}} + \ldots$$

Generalize MSE: The effectiveness of $\hat{\boldsymbol{\beta}}$ is assessed via

$$\textbf{GMSE} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \left\{ \int_0^\infty E\tilde{\mathbf{x}}(t)^{\otimes 2} \exp\{\boldsymbol{\gamma}^T \tilde{\mathbf{x}}(t)\} \xi(t) \, d\Lambda(t) \right\} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

The blacket term can be evaluated by the Monte Carlo method.

When $\mathbf{x}(t)$ is Gaussian, it has an analytic form.

Effectiveness of PPLS

Table 3: Comparison of Variable Selection Procedures: Ratio of PPLS versus PLS based on 400 simulations

	$\alpha(t) = \tau\sqrt{t/\tau}$			$\alpha(t) = \tau\sin(2\pi t/\tau)$		
Method	RGMSE	Zero Coefficient		RGMSE	Zero Coefficient	
Case I: $n = 50, \tau = 20$						
Method	mean (std)	C	I	mean (std)	C	I
L_1	**0.3936**(0.2966)	4.9950	0	0.3923(0.2863)	4.9900	0
SCAD	**0.3549**(0.2453)	4.9950	0	0.3533(0.2453)	4.9925	0
Oracle	**0.3502**(0.2412)	5.0000	0	0.3480(0.2425)	5.0000	0
Case II: $n = 75, \tau = 4$						
L_1	**0.5772**(0.2614)	4.3325	0	0.5733(0.2648)	4.3500	0
SCAD	**0.5127**(0.2101)	4.4275	0	0.5115(0.2107)	4.4250	0
Oracle	**0.3939**(0.2326)	5.0000	0	0.3915(0.2318)	5.0000	0
Case III: $n = 50, \tau = 20$						
L_1	**0.3975**(0.2843)	4.9950	0	0.4002(0.2860)	4.9975	0
SCAD	**0.3450**(0.2278)	4.9975	0	0.3460(0.2279)	4.9975	0
Oracle	**0.3438**(0.2269)	5.0000	0	0.3450(0.2271)	5.0000	0
Case IV: $n = 50$, $\tau = 30$						
L_1	**0.4091** (0.2716)	4.9975	0	0.4074 (0.2717)	5.0000	0
SCAD	**0.3554** (0.2210)	5.0000	0	0.3546 (0.2205)	5.0000	0
Oracle	**0.3549** (0.2200)	5.0000	0	0.3542 (0.2199)	5.0000	0

Table 4: Accuracy of Standard Errors formula

	β_1		β_2		β_5	
	std	se (std(se))	std	se (std(se))	std	se (std(se))
L_1	**0.0823**	**0.0798** (0.0176)	0.0826	0.0775 (0.0177)	0.0735	0.0702 (0.0166)
SCAD	**0.0810**	**0.0808** (0.0180)	0.0821	0.0793 (0.0187)	0.0738	0.0708 (0.0169)
Oracle	**0.0808**	**0.0808** (0.0181)	0.0810	0.0794 (0.0188)	0.0737	0.0709 (0.0169)

Based on case I design with $n = 50$, $\alpha(t) = \tau\sqrt{t/\tau}$ with $\tau = 20$.

Application to the CD4 Data

Previous Analysis: Wu and Chiang (2000) and Fan and Zhang (2000) used the model

$$y(t) = \beta_0(t) + \beta_1(t)\text{Smoking} + \beta_2(t)\text{Age}(t) + \beta_3(t)\text{PreCD4}(t) + \varepsilon(t).$$

Results of Huang, Wu and Zhou (2002) suggests that functions $\beta_1(\cdot)$, $\beta_3(\cdot)$, and $\beta_3(\cdot)$ are constant.

Our Analysis: Involve interactions

$$\begin{aligned} y(t) \;=\; & \alpha(t) + \beta_1 x_1 + \beta_2 x_2(t) + \beta_3 x_3(t) + \beta_4 x_2^2(t) + \beta_5 x_3^2(t) \\ & + \beta_6 x_1 x_2(t) + \beta_7 x_1 x_3(t) + \beta_8 x_2(t) x_3(t) + \varepsilon(t). \end{aligned}$$

x_2 and x_3 are standardized age and PreCD4, resp.

Results

Table 5: Estimated Coefficients

Variable	Profile LS $\hat{\beta}(\text{se}(\hat{\beta}))$	L_1 $\hat{\beta}(\text{se}(\hat{\beta}))$	SCAD $\hat{\beta}(\text{se}(\hat{\beta}))$
Smoking	0.5333(1.0972)	0(0)	0(0)
Age	-0.1010(0.9167)	0(0)	0(0)
PreCD4	**2.8252(0.8244)**	**3.0932(0.5500)**	**3.1993(0.5699)**
Age^2	0.1171(0.4558)	0(0)	0(0)
PreCD4^2	-0.0333(0.3269)	0(0)	0(0)
Smoking*Age	**-1.7084(1.1192)**	**-0.9684(0.4904)**	**-1.0581(0.5221)**
Smoking*PreCD4	1.3277(1.3125)	0(0)	0(0)
Age*PreCD4	-0.1360(0.5413)	0(0)	0(0)