

# Nonparametric and Semiparametric Regression for Longitudinal/Clustered Data and High Dimensional Data

Xihong Lin

Department of Biostatistics

Harvard School of Public Health

## Outline

- Motivation and background
- Nonparametric/semiparametric regression for clustered/longitudinal data
  - Nonparametric Regression: Kernels and smoothing splines and their relationship
  - Semiparametric Regression: Profile kernel/spline methods
- Nonparametric/semiparametric regression for high-dimensional data
  - Kernel Machines (KMs)
  - Relationship between KMs and mixed models
- Simulation studies and data example
- Conclusions

## Introduction

- Nonparametric and semiparametric regression methods are well understood for independent and low dimensional data.
- These techniques are not well understood for clustered/longitudinal data and high-dimensional covariate data (e.g., microarrays, SNPs, and proteomics).
- Main interests: Develop nonparametric and semiparametric regression models for
  - clustered/longitudinal data
  - high-dimensional genetic pathway data.

## Part I(a) Nonparametric Regression For Longitudinal/clustered Data: Introduction

- Clustered Data: longitudinal data, familial data, etc.

Subject	Time		
	1	...	m
1	×	...	×
⋮		...	
n	×	...	×

- Feature of clustered data: **Within-cluster correlation**.
- Parametric regression for clustered data is well developed: GEEs and GLMMs, and efficiency is improved by accounting for correlation.
- For independent data, **splines=kernels** asymptotically and both are local.
- **Main question:** **Are these still true for nonparametric regression in clustered/longitudinal data?**

## Nonparametric Regression for Clustered Data

$i$ =cluster  $i$  ( $i = 1, \dots, n$ )       $j$ =observation  $j$  ( $j = 1, \dots, m_i < \infty$ )  
 $Y_{ij}$ =outcome       $T_{ij}$ =covariate (e.g., time)

### Model:

$$Y_{ij} = \theta(T_{ij}) + e_{ij}, \quad \mathbf{e}_i \sim N(0, \Sigma)$$

### Goals:

- (1) Are the most efficient kernel and spline estimators obtained by accounting for the within-cluster correlation?
- (2) Splines=Kernels asymptotically?

## Conventional Kernel GEEs

- **Key idea:** Extension of local polynomials by introducing kernel weights in GEEs.
- **Kernel GEEs:**

$$\sum_{i=1}^n \tilde{\mathbf{T}}_i(t)^T \mathbf{K}_{ih}^{1/2}(t) \mathbf{V}_i^{-1}(t) \mathbf{K}_{ih}^{1/2}(t) \{ \mathbf{Y}_i - [\beta_0 \mathbf{1} + \beta_1 (\mathbf{T}_i - t)] \} = 0$$

where  $K(\cdot)$  = kernel function,  $h$  = bandwidth  
 $\mathbf{V}_i$  = working covariance matrix

- **Conventional Kernel GEE estimator:**  $\hat{\theta}(t) = \hat{\beta}_0$ .
- **Key Finding:** The most efficient  $\hat{\theta}(t)$  requires ignoring the correlation ( $\mathbf{V} = \mathbf{I}$ ), i.e., the kernel GEE fails.

## Smoothing Spline

- A smoothing spline minimizes

$$n^{-1} \sum_{i=1}^n \{\mathbf{Y}_i - \boldsymbol{\theta}(\mathbf{T}_i)\}^T \mathbf{V}^{-1} \{\mathbf{Y}_i - \boldsymbol{\theta}(\mathbf{T}_i)\} + \lambda \int \{\theta''(t)\}^2 dt$$

- The smoothing spline estimator is

$$\hat{\boldsymbol{\theta}}_S(\mathbf{T}) = (\tilde{\mathbf{V}}^{-1} + n\lambda\Psi)^{-1} \tilde{\mathbf{V}}^{-1} \mathbf{Y},$$

- One can calculate  $\hat{\boldsymbol{\theta}}_S(\mathbf{T})$  using the BLUP by fitting a mixed model.

$$\mathbf{Y} = \delta_0 \mathbf{1} + \delta_1 \mathbf{T} + \mathbf{B}\mathbf{a} + \boldsymbol{\epsilon},$$

where  $\mathbf{a} \sim N(0, \tau\mathbf{I})$ ,  $\tau = 1/\lambda$ ,  $\boldsymbol{\epsilon} \sim N(0, \mathbf{V})$ , and  $\boldsymbol{\epsilon}_i = \mathbf{Z}_i^T \mathbf{b}_i + e_i$ .

$$\implies \hat{\boldsymbol{\theta}}_S(\mathbf{T}) = \hat{\delta}_0 \mathbf{1} + \hat{\delta}_1 \mathbf{T} + \mathbf{B}\hat{\mathbf{a}},$$

## Relationship Between Smoothing Splines and GEE Kernels

- The conventional kernel GEE estimator is **local**, however splines are **not local**.
- The most efficient GEE kernel requires **ignoring correlation**, however, the most efficient spline estimator requires **accounting for correlation**.

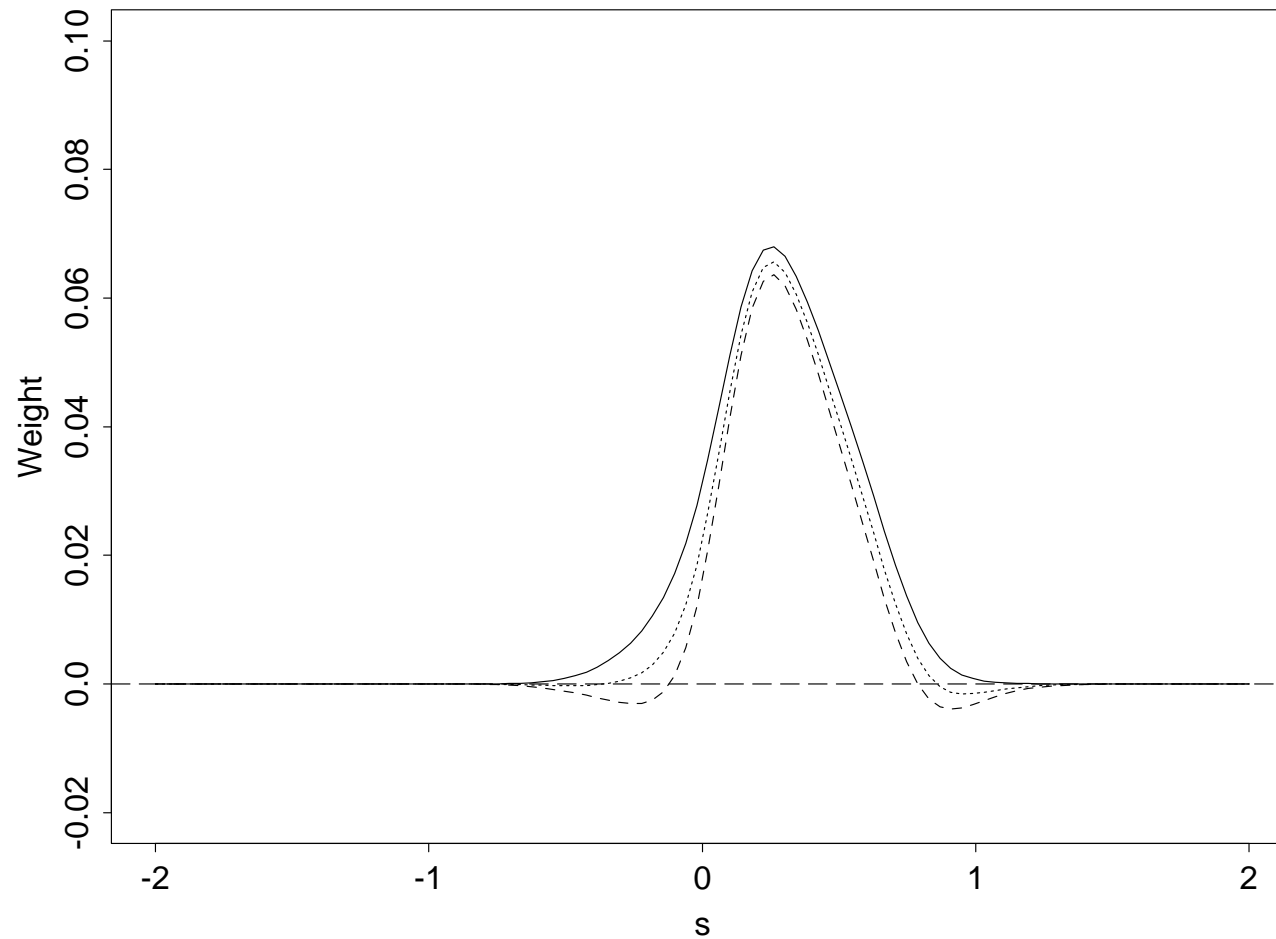
### Efficiency

spline  $>$  spline = GEE kernel  $>$  GEE kernel  
 (true) (work indep) (work indep) (true)

- **Main message:** For clustered data, splines are advantageous over **conventional** kernels (kernel GEEs).

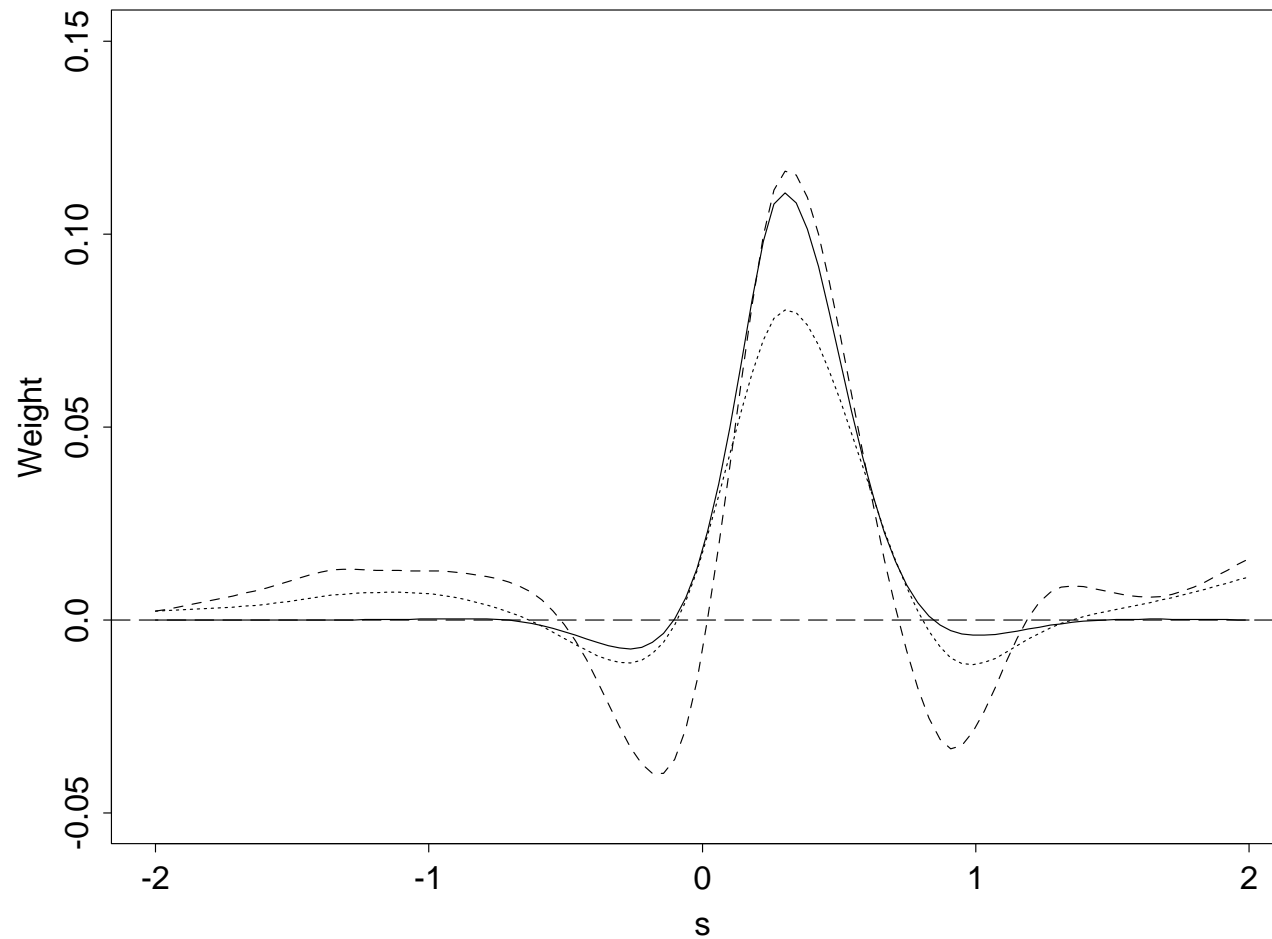
## GEE Kernel Weight for $\hat{\theta}_K(t = 0.25)$

Solid ( $\rho = 0$ ); Dotted ( $\rho = 0.5$ ); Dashed ( $\rho$  unstructured).



## Smoothing Spline Weight for $\hat{\theta}_K(t = 0.25)$

Solid ( $\rho = 0$ ); Dotted ( $\rho = 0.5$ ); Dashed ( $\rho$  unstructured).



## Simulation Results

- Number of clusters  $n=100$  with cluster size  $m = 3$ .
- Generated  $T_{ij}$  from  $U(-2, 2)$ .
- True correlation structure:  
1=AR(1), 2=exchangeable, 3=unstructured.
- $\theta(t)$ =bi-modal

### MSE Efficiency of Splines vs Kernel GEEs (200 runs)

Corr	True Cov	Independence
1	1.43	1.04
2	1.43	1.04
3	2.54	1.04

## Equivalent Kernels of Splines

- **Dilemma:**

Splines  $\neq$  kernels and kernels perform worse, i.e., Silverman's results do not apply for clustered data.

- **Questions:**

(1) Is there an equivalent kernel of a spline for clustered data?

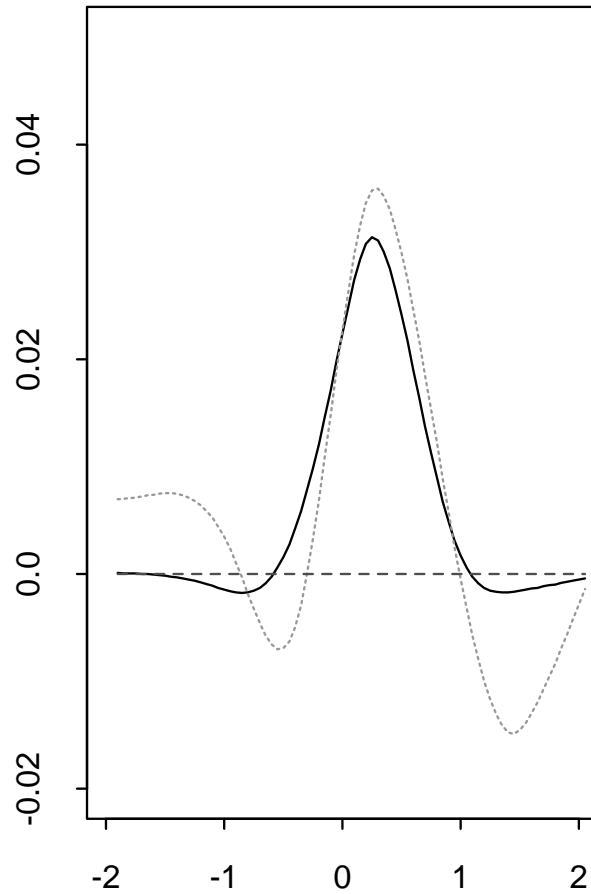
(2) Since splines are not local, can a nonlocal nonparametric estimator still be consistent? i.e., bias  $\rightarrow \infty$ ?

(3) Asymptotic properties of splines?

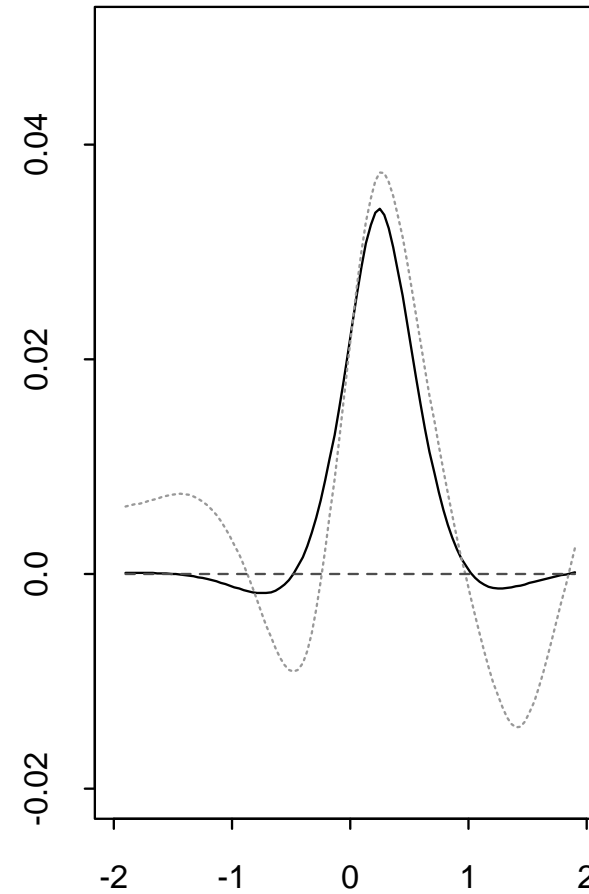
## Equivalent Kernels of Splines

- **Seemingly Unrelated (SUR) Kernel (Wang, 2003):** An iterative kernel estimating equation.
  - The most efficient SUR kernel estimator is obtained by accounting for the within-cluster correlation.
- **Key results (Lin, et al, 2004):**
  - (1) Splines = SUR Kernel asymptotically
  - (2) Both are non-local and consistent and most efficient if accounting for correlation.

## SUR Kernel Weight



## Smoothing Spline Weight



Solid (Working independence) and Dotted (True Covariance)

## Intuition of Spline $\stackrel{asy}{=} \text{SUR Kernel}$ Using Iterative Pseudo-obs

- At  $(l + 1)$ th iteration, define pseudo-observations as

$$Y_{ij}^{(l+1)} = Y_{ij} + (v^{jj})^{-1} \sum_{j=1}^m \sum_{k \neq j} v^{jk} \left\{ Y_{ik} - \hat{\theta}^{(l)}(T_{ik}) \right\}.$$

- Fit  $Y_{ij}^{(l+1)} = \theta(T_{ij}) + e_{ij}$ , where  $e_{ij} \stackrel{iid}{\sim} N(0, v^{jj})$  using standard kernel  $\hat{\theta}_K^{(l+1)}(t)$  or spline  $\hat{\theta}_S^{(l+1)}(t)$ .

- At convergence,

If spline:  $\hat{\theta}_S^{(l+1)}(\mathbf{T}) \rightarrow \hat{\theta}_S(\mathbf{T}) = (\tilde{\mathbf{V}}^{-1} + n\lambda\Psi)^{-1}\tilde{\mathbf{V}}^{-1}\mathbf{Y}$

If kernel:  $\hat{\theta}_K^{(l+1)}(t) \rightarrow \text{SUR Kernel.}$

- The smoothing spline is consistent ( $\text{bias}\{\hat{\theta}_S(t)\} = O\{h^4(t)\}$ ) and its variance  $(\{nh(t)\}^{-1}C)$  is minimized when  $\mathbf{V} = \Sigma$ .

## Simulation Study

- Setting: same as before ( $n = 100$  and  $m = 3$ ).
- Goal: Compare the final sample MSE efficiency of SUR kernels and splines assuming true covariance.

### MSE Efficiency of Splines vs SUR Kernels Assuming True Covariance

Corr	n=50	n=100
1	1.09	1.07
2	1.08	1.06
3	1.25	1.13

## Longitudinal progesterone data

- 32 women collected urine samples on alternative days during a menstrual cycle (11-28 obs/woman)
- Goal: Time profile of reproductive hormone progesterone and the effects of the covariates age and BMI.

- Semiparametric model:

$$Y_{ij}(t) = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \theta(t) + b_i + U_{ij}(t) + e_{ij},$$

where  $\mathbf{X} = (\text{AGE}, \text{BMI})$ ,  $b_i \sim N(0, \theta)$ ,  $U_{ij}(t) \sim \text{NOU}(\xi)$ ,  $e_{ij} \sim N(0, \sigma^2)$

- Fit the linear mixed model using PROC MIXED

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{T}\boldsymbol{\delta} + \mathbf{B}\mathbf{a} + \mathbf{Z}\mathbf{b} + \mathbf{U} + \mathbf{e}$$

where  $\mathbf{a} \sim N(0, \tau\mathbf{I})$ ,  $\mathbf{b} \sim N(\theta\mathbf{I})$ ,  $\mathbf{U} \sim \text{NOU}(\xi)$ ,  $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$ .

- AGE: 0.92 (1.92); BMI: -2.91 (2.37).

Figure 1

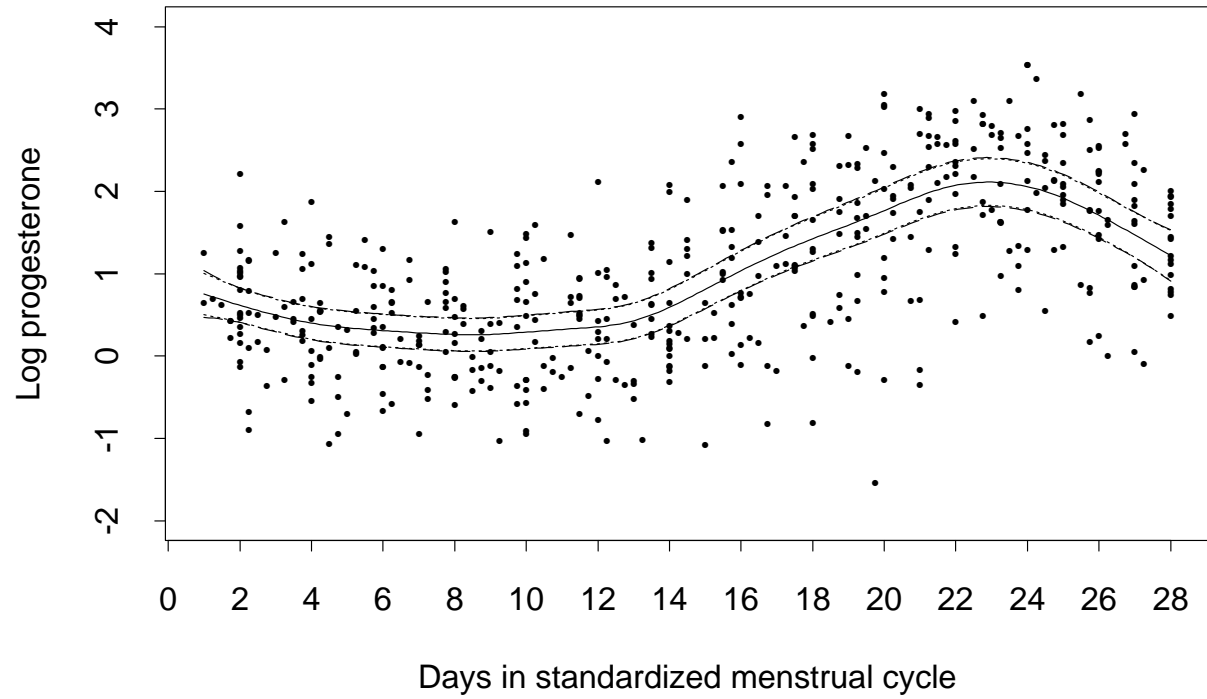
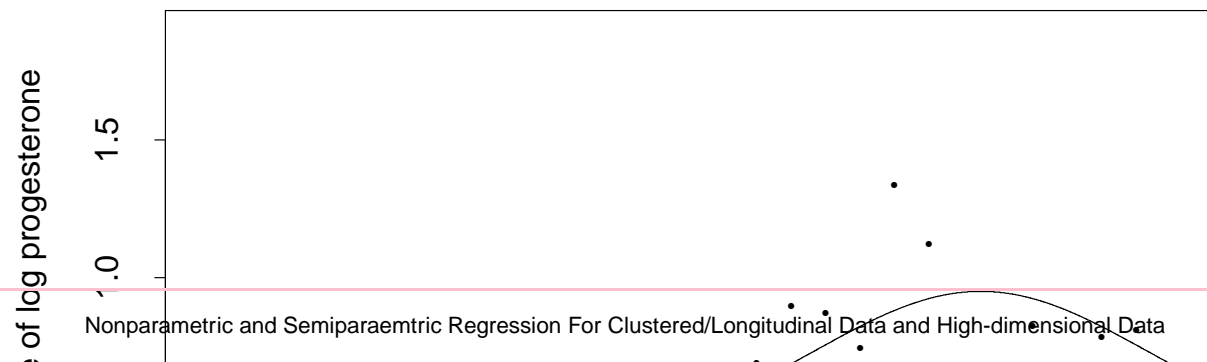


Figure 2



## Conclusions of Part I

- For clustered data using marginal models, splines and kernels are asymptotically consistent for any working covariance matrix.
- Splines and SUR kernels are asymptotically equivalent, and they effectively account for correlation and are more efficient than conventional local-polynomial kernels.
- Unlike independent data, non-local nonparametric methods, such as splines, can still be consistent for clustered data.
- Profile-spline(kernel) methods yield the semiparametric efficient estimator in semiparametric models for clustered data.

## Part II: Semiparametric Regression for High-dimensional data

### – Kernel Machines and Mixed Models

## Motivation and Background

- High-dimensional data are arising more frequently, e.g, microarrays, SNPs and proteomics.
- Kernel machines (KMs), such as support vector machines (SVMs), are powerful for analyzing high-dimensional data in machine learning.
- The most familiar version of KMs is SVM classification.
- Our focus: Kernel machine regression.
- Typical form of the data:

	Response	Covariates			Genes/Proteins		
Sample	$Y$	$X_1$	$\dots$	$X_q$	$T_1$	$\dots$	$T_p$
1	×	×	$\dots$	×	×	$\dots$	×
$\dots$							
n	×	×	$\dots$	×	×	$\dots$	×

## Prostate Cancer Genetic Pathway Data

- Outcome ( $Y_i$ )=pre-surgery  $\ln(\text{PSA}+1)$ : 59 prostate cancer patients.
- Objective: To assess the effects of covariates ( $\mathbf{X}_i$ ) (age and Gleason score) and a pathway with genes ( $T_{i1}, \dots, T_{ip}$ ) on PSA.

- Model:

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \theta(T_{i1}, \dots, T_{ip}) + e_i,$$

where  $e_i \sim N(0, \sigma^2)$  and  $\theta(\cdot)$  allows interactions among genes.

- Limitations of individual gene analysis:

- (1) Cellular processes often affect sets of genes;
- (2) Individual highly ranked genes can be poorly annotated and are often not reproducible from studies to studies;
- (3) Knowledge-based Studies on gene sets, e.g. genetic and metabolic pathways, are more biologically interpretable and reproducible.

## Semiparametric Model for High-Dimensional Genetic Pathway Data

- Model covariate effects parametrically and gene effects within a pathway nonparametrically.

$$Y_i = \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta} + \theta(\mathbf{T}_{i1}, \dots, \mathbf{T}_{ip}) + e_i$$

where  $\theta(\mathbf{T}_i)$  = unknown smooth function in **functional** (feature) space  $\mathcal{H}$  and  $\dim(\mathbf{T}_i) = p$  (might be high).

- Estimate  $\boldsymbol{\beta}$  and  $\theta(\cdot)$  by minimizing the penalized RSS:

$$\sum_{i=1}^n \{Y_i - (\beta_0 + \mathbf{X}_i^T \boldsymbol{\beta} + \theta(\mathbf{T}_i))\}^2 + \lambda \|\theta\|_{\mathcal{H}}^2$$

## Primal and Dual Formulations

- Write  $\theta(\mathbf{T}) = \sum_{j=1}^J \omega_j \phi_j(\mathbf{T}) \triangleq \phi(\mathbf{T})^T \boldsymbol{\omega}$ , where  $\{\phi_i(\mathbf{T})\}_{i=1}^{\infty}$  is an orthonormal basis of  $\mathcal{H}$  and  $\|\theta\|_{\mathcal{H}}^2 = \|\boldsymbol{\omega}\|^2$ .
- Difficulties with the primal formulation: (1) Need to specify basis  $\{\phi_j(\mathbf{T})\}_{j=1}^J$ ; (2)  $\boldsymbol{\omega}$  and  $\phi$ 's are high dimensional.
- Introduce the Lagrangian multiplier (dual parameters)  $\boldsymbol{\gamma}$

$$\mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\beta}, \mathbf{e}; \boldsymbol{\gamma}) = \frac{1}{2} \sum_{i=1}^n e_i^2 + \frac{1}{2} \lambda \|\boldsymbol{\omega}\|^2 - \sum_{i=1}^n \gamma_i \{Y_i - \beta_0 - \mathbf{X}_i^T \boldsymbol{\beta} - \phi(\mathbf{T}_i)^T \boldsymbol{\omega} - e_i\}$$

- The dual formulation is obtained by removing  $\boldsymbol{\omega}$  (high dim) and writing  $\mathcal{L}(\cdot)$  as a function of dual parameters  $(\boldsymbol{\gamma}, \boldsymbol{\beta})$  alone.

$$\begin{cases} Y_i - \mathbf{X}_i^T \boldsymbol{\beta} - \frac{1}{\lambda} \sum_{i'=1}^n \gamma_{i'} \phi(\mathbf{T}_i)^T \phi(\mathbf{T}_{i'}) - \gamma_i = 0 \\ \sum_{i=1}^n \gamma_i \mathbf{X}_i = 0 \end{cases}$$

## Dual Formulation Continues

- Estimation in the dual formulation is low dimensional.
- The estimator  $\hat{\theta}(\mathbf{T}) = \lambda^{-1} \sum_{i=1}^n \hat{\gamma}_i \phi(\mathbf{T})^T \phi(\mathbf{T}_i)$ .
- Computation of  $\hat{\gamma}$  and  $\hat{\theta}(\mathbf{T})$  hence only requires evaluating the kernel function

$$k(\mathbf{T}, \mathbf{T}') = \langle \phi(\mathbf{T}), \phi(\mathbf{T}') \rangle = \phi(\mathbf{T})^T \phi(\mathbf{T}').$$

- If  $k(\mathbf{T}, \mathbf{T}')$  is specified, no need to explicitly know the basis  $\{\phi_j(\mathbf{T})\}_{j=1}^{\infty}$ .
- Given  $k(\cdot, \cdot)$ , the functional space  $\theta(\cdot) \in \mathcal{H}_k$  is called the **Reproducing Kernel Hilbert Space**.

## Two most popular kernel functions

- Gaussian kernel (We are currently using):

$$k(\mathbf{T}_i, \mathbf{T}_{i'}) = \exp\left(-\frac{\|\mathbf{T}_i - \mathbf{T}_{i'}\|^2}{\rho}\right)$$

Functional space: radius basis

- $d^{\text{th}}$  degree polynomial kernel:

$$k(\mathbf{T}_i, \mathbf{T}_{i'}) = (\langle \mathbf{T}_i, \mathbf{T}_{i'} \rangle + c)^d$$

Functional space:  $d$ th polynomial basis (up to  $d$ -way interactions).

## Dual Formulation Continues

- Matrix notation of the dual problem:

$$\begin{bmatrix} \mathbf{0} & \mathbf{X}^T \\ \mathbf{X} & \tau\mathbf{K} + \sigma^2\mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{y} \end{bmatrix}$$

where the  $n \times n$  matrix  $\mathbf{K} = \mathbf{K}(\rho) = \{k(\mathbf{T}_i, \mathbf{T}_{i'})\}$ , where  $\tau = 1/\lambda$ .

- Given  $(\tau, \rho, \sigma^2)$ , we have

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}^T(\tau\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{X}]^{-1}\mathbf{X}^T(\tau\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{Y}$$

$$\hat{\boldsymbol{\gamma}} = (\tau\mathbf{K} + \sigma^2\mathbf{I})^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$\hat{\boldsymbol{\theta}} = \tau\mathbf{K}\hat{\boldsymbol{\gamma}} = \tau\mathbf{K}(\tau\mathbf{K} + \sigma^2\mathbf{I})^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

- No obvious way to jointly estimate  $(\tau, \rho, \sigma^2)$ .

## Connection of LS Kernel Machines and Linear Mixed Models

- Key Message: LS KM (semi)non-parametric regression can be fitted using linear mixed models by PROC MIXED.
- The forms of the KM estimators  $\hat{\beta}$  and  $\hat{\theta}$  are identical to the BLUP estimators under the linear mixed effects model

$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{X}\beta + \mathbf{b} + \mathbf{e} \quad (1)$$

where  $\mathbf{b}(n \times 1) \sim N\{\mathbf{0}, \tau \mathbf{K}(\rho)\}$  and  $\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ .

- The LS KM estimator  $\hat{\theta}(\mathbf{T}) = \{\hat{\theta}(\mathbf{T}_1), \dots, \hat{\theta}(\mathbf{T}_n)\}^T$  is the BLUP:

$$\hat{\theta} = \hat{\mathbf{b}}$$

- Unified estimation of  $(\tau, \rho, \sigma^2)$ : Treat  $\tau, \rho, \sigma^2$  as variance components and estimate them simultaneously using REML in the linear mixed model (1).

## Test for the Nonparametric Function

- Hypothesis of interest:

$$H_0 : \theta(\mathbf{T}) = 0 \quad vs \quad H_1 : \theta(\mathbf{T}) \in \mathcal{H}_k.$$

- This hypothesis is equivalent to variance component testing:

$$H_0 : \tau = 0 \quad vs \quad H_1 : \tau > 0.$$

- Difficulties:

Standard theory fails, since  $H_0$  is on the boundary of the parameter space and  $\mathbf{K}$  is not block diagonal and involves an unknown scale parameter  $\rho$  for the Gaussian kernel, which is unestimable under  $H_0$ .

- Current approach: Fix  $\rho$  and calculate score statistic for  $\tau$  for a range of  $\rho$  values.

## Variable/Model Selection Using the LS Kernel Machine

- Recall semiparametric model:  $Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \theta(\mathbf{T}_i) + e_i$ .
- Estimate  $\theta(\mathbf{T})$  using kernel machine  $\hat{\theta}(\mathbf{T}) = \sum_{i=1}^n \hat{\gamma}_i k(\mathbf{T}, \mathbf{T}_i)$ .
- Variable/model selection is equivalent to kernel selection under the semiparametric model.
- Examples:
  - Variable selection in linear regression  $\theta(\mathbf{T}) = T_1 \omega_1 + \dots + T_{10} \omega_{10}$ :  
 $H_0: \omega_6 = \dots = \omega_{10} = 0$  is equivalent to  $H_0: k(\cdot, \cdot) = 1$ st-degree polynomial kernel using  $T_1, \dots, T_5$  vs using all  $T_1, \dots, T_{10}$ .
  - Linear vs quadratic regression of  $\theta(\mathbf{T})$ :  $k(\cdot, \cdot)$  is a 1st-degree ( $H_0$ ) vs 2nd-degree ( $H_1$ ) polynomial kernel.
  - Linear vs smooth  $\theta(\mathbf{T})$ :  $k(\cdot, \cdot)$  is a 1st-degree polynomial ( $H_0$ ) vs Gaussian or smoothing spline kernel ( $H_1$ ).

## Variable/model Selection Using Kernel Machine AIC/BIC

- For any given kernel, the predicted value of  $Y$  can be written as  $\hat{\mathbf{Y}} = \mathbf{A}\mathbf{Y}$ , where

$$\mathbf{A} = (\mathbf{I} + \lambda^{-1}\mathbf{K})^{-1} \left[ \lambda^{-1}\mathbf{K} + \mathbf{X} \{ \mathbf{X}^T (\mathbf{I} + \lambda^{-1}\mathbf{K})^{-1} \mathbf{X} \}^{-1} \mathbf{X}^T (\mathbf{I} + \lambda^{-1}\mathbf{K})^{-1} \right].$$

- Kernel machine Degree-of-Freedom (KM\_DF):  $r = \text{tr}(\mathbf{A})$ .
- Kernel Machine AIC and BIC:

$$\text{KM\_AIC} = n \log \{ (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}) \} + 2r,$$

$$\text{KM\_BIC} = n \log \{ (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}) \} + r \log(n).$$

## Simulation Study

- Sample size  $n=150$  and  $p=5$  active genes.
- Model:  $y = x_1 + \theta(T_1, \dots, T_5) + e$ , where  $e \sim N(0, 1)$  and  $\theta(\mathbf{T}) = 10 \cos(T_1) - 15T_2^2 + 10 \exp(-T_3)T_4 - 8 \sin(T_5) \cos(T_3) + 20T_1T_5$ .
- Use the 5 right genes and 5 junk genes and fit  $y = x_1 + \theta(T_1, \dots, T_{10}) + e$  using the LS kernel machine via the linear mixed model.
- Goal 1: Study bias and SEs of the point estimates (1000 runs)
- Goal 2: Study size and power of the VC score test of  $H_0 : \theta(\mathbf{T}) = 0$ , where  $\theta_*(\mathbf{T}) = \alpha\theta_0(\mathbf{T})$  (2000 runs (size) and 1000 runs (power)).
- Set  $n = 50$  and  $\alpha = 0, 0.2, 0.4, 0.6, 0.8, 1.0$ .

### Simulation results of point estimates

Parameter Estimates				Regression of $\theta$ on $\hat{\theta}$		
$\hat{\beta}_1$	$\hat{\sigma}^2$	$\hat{\tau}$	$\hat{\rho}$	int	slope	$R^2$
1.09	0.85	575.81	36.72	0.12	1.01	0.98

### Simulation Results for the VC Score Test for $H_0 : \theta(T) = 0$

Scale	Size	Power				
		$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1.0$
$\rho$	$\alpha = 0$					
1	0.057	0.081	0.212	0.400	0.619	0.770
25	0.053	0.164	0.427	0.673	0.855	0.943
100	0.044	0.139	0.378	0.609	0.747	0.879

## PSA Data Example

- $n = 59$  prostate cancer patients.
- **Phasphatase/Kinase**: Potentially prostate-cancer related genetic pathway with 5 gene expressions (normalized to have mean 0 var 1).
- **Objective**: To assess the effects of age, Gleason score and the phasphatase pathway on  $\log(PSA + 1)$ .
- The semiparametric model:  $y = \beta_0 + \mathbf{x}^T \boldsymbol{\beta} + \theta(T_1, \dots, T_5) + e$ .
- Fit the model using the KM with the Gaussian kernel via the linear mixed model.

## Analysis Results of Semiparametric Models

Covariate	Estimate	S.E.	P-value
Intercept	-1.44	0.99	0.15
Gleason	0.38	0.12	0.002
Age	0.02	0.01	0.15

$\tau$	$\rho$	$\sigma^2$
0.06	2.32	0.43

Score test for Gene Effect  $H_0 : \theta(\mathbf{T}) = 0$

$\rho$	$\mathcal{S}$	df	P-value
1.0	1711.52	1588.19	0.016
3.0	140.50	104.44	0.011
10.0	39.39	19.37	0.004
30.0	27.12	8.45	0.001
60.0	30.53	6.12	0.0001

## Model/Variable Selection for the PSA Data

- Consider all subsets of genes (31 models).
- Using the Gaussian kernel based KM AIC/BIC.
- KM AIC ranged from 187.06 to 298.89 and KM BIC ranged from 196.34 to 421.47.
- The model with the best AIC/BIC is the gene set *FGF2, IGFBP1* (AIC=187.06, BIC=196.34).

## Conclusions

- We develop a semiparametric regression model where clinical covariates are modeled parametrically and high-dimensional gene expressions are modeled nonparametrically using the LS kernel machine.
- The LS KM can be fit using a linear mixed model in a unified framework, where the regression coefficients and the nonparametric function can be estimated by the BLUPs and the smoothing parameters and the kernel scale parameter can be estimated using REML.
- A variance component score test is developed for testing the nonparametric function.

- Our simulation studies show the proposed methods work well in finite samples.
- The close connection between KMs and mixed models open doors to familiar Bayesian techniques in high-dimensional data problems.
- It provides an flexible framework for modeling covariate interactions and performing model/variable selection.
- Extensions in progress:
  - Semiparametric additive model for multiple pathways:

$$y_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + \theta_1(\mathbf{T}_{1i}) + \theta_2(\mathbf{T}_{2i}) + e_i$$

- Generalized semiparametric model for categorical responses and the connection of KMs and Generalized Linear Mixed Models.
- Survival analysis and the connection of KMs and frailty models.

## Acknowledgements of Collaborators

Non-(semi-)parametric regression in longitudinal data:

Raymond J Carroll

Peter Hall

Jonathan Raz

Naisyin Wang

Alan Welsh

Daowen Zhang

Kernel Machines and Mixed Models

Debashis Ghosh

Dawei Liu

## References the talk is based on

Zhang, Lin, Raz and Sowers (1998, JASA)

Lin and Zhang (1999, JRSSB)

Lin and Carroll (2000, 2001, JASA; 2001, Biometrika)

Welsh, Lin and Carroll (2002, JASA)

Lin, Wang, Welsh and Carroll (2004, Biometrika)

Carroll, Hall, Apanasovich, and Lin (2004, Stat Sinica)

Lin and Carroll (2005, JRSSB),

Wang, Carroll and Lin (2004, JASA)

Liu, Lin and Ghosh (2005)