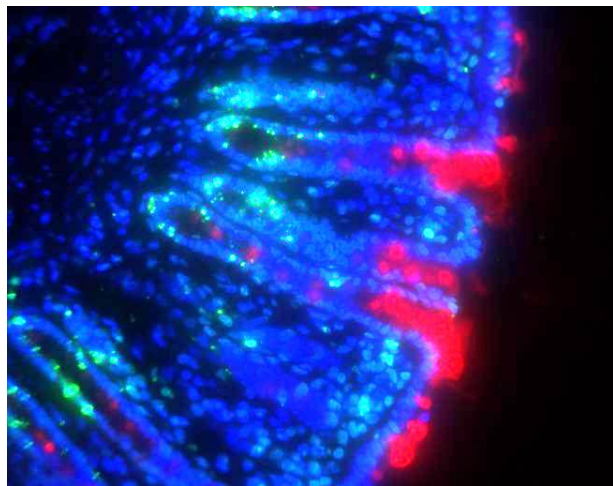
# Non/Semiparametric Regression and Clustered/Longitudinal Data

Raymond J. Carroll

Texas A&M University

http://stat.tamu.edu/~carroll

carroll@stat.tamu.edu

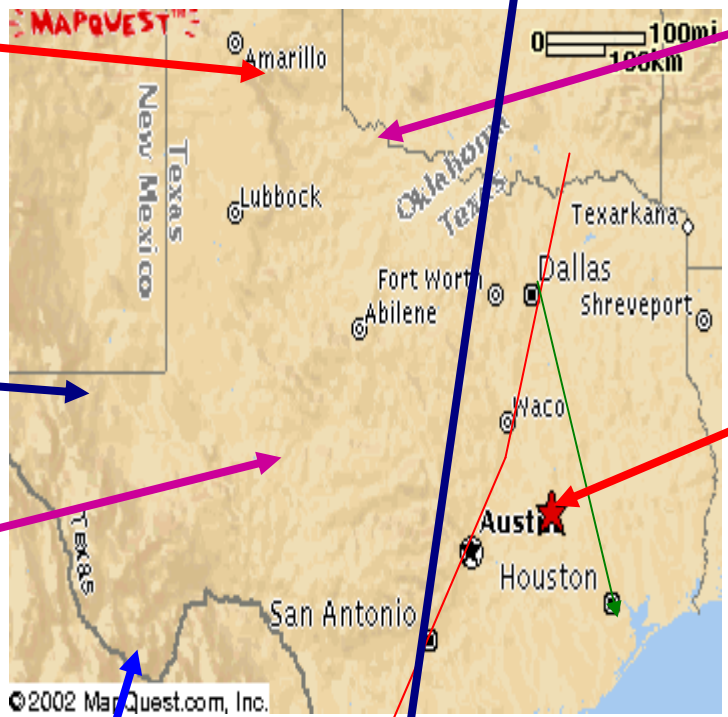


**Palo Duro Canyon, the Grand Canyon of Texas**

**West Texas** ☺

**East Texas** ☹

Wichita Falls, my hometown




**Guadalupe Mountains National Park**

**College Station, home of Texas A&M University**

Midland

**I-45**

**Big Bend National Park**

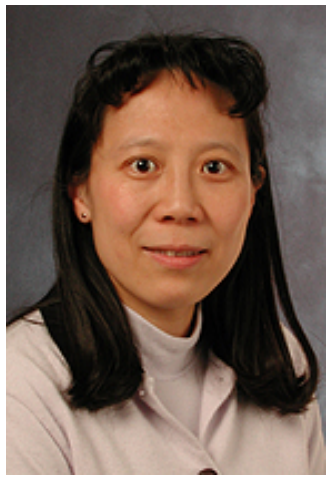
**I-35**

# Outline

- **Longitudinal models**:
  - Panel data
  - Nonparametric models
  - Partially linear semiparametric models
- **Theme**:
  - How are splines and kernels related in this problem?
- **Correlated data**:
  - Efficient methods of estimation

# Acknowledgments

Naisyin Wang
Texas A&M



Xihong Lin
University of Michigan

Alan Welsh
Australian National University

# Panel Data (for simplicity)

- i = 1,…,n clusters/individuals
- j = 1,…,m observations per cluster

| Subject | Wave 1 | Wave 2 | … | Wave m |
|---|---|---|---|---|
| 1 | X | X | | X |
| 2 | X | X | | X |
| … | | | | X |
| n | X | X | | X |

# Panel Data (for simplicity)

- i = 1,...,n clusters/individuals
- j = 1,...,m observations per cluster
- **Important points**:
  - The cluster size **m** is meant to be fixed
  - This is not a **multiple time series** problem where the cluster size increases to infinity
  - Some comments on the **single time series** problem are given near the end of the talk

# The Marginal Parametric Model

- Y = Response
- X,Z = time-varying covariates

$$Y_{ij} = Z_{ij}\beta + X_{ij}\Theta + \varepsilon_{ij}$$
$$\mathrm{cov}(\varepsilon_{ij}) = \Sigma$$

- **General Result**: We can improve efficiency for $(\beta,\Theta)$ by accounting for correlation: GLS

# The Marginal Semiparametric Model

- Y = Response
- X,Z = time-varying covariates

$$Y_{ij} = Z_{ij}\beta + \Theta(X_{ij}) + \varepsilon_{ij}$$
$$\mathrm{cov}(\varepsilon_{ij}) = \Sigma$$

- **Question**: can we improve efficiency for $\beta$ by accounting for correlation?

# The Marginal Nonparametric Model

- Y = Response
- X = time-varying covariate

$$Y_{ij} = \Theta(X_{ij}) + \varepsilon_{ij}$$

$$\Theta(\bullet) = \text{unknown function}$$

$$\text{cov}(\varepsilon_{ij}) = \Sigma$$

- **Question**: can we improve efficiency by accounting for correlation? (GLS)

# Independent Data

- **Splines** (smoothing, P-splines, etc.) with penalty parameter = $\lambda$

$$\textbf{minimize} \sum_{i=1}^{n} \left\{\underline{Y}_i - \underline{\Theta}(\underline{X}_i)\right\}^T \left\{\underline{Y}_i - \underline{\Theta}(\underline{X}_i)\right\} + \lambda \int \left\{\Theta''(t)\right\}^2 dt$$

- Ridge regression fit
- Some bias, smaller variance
- $\lambda = 0$ is over-parameterized least squares
- $\lambda = \infty$ is a polynomial regression

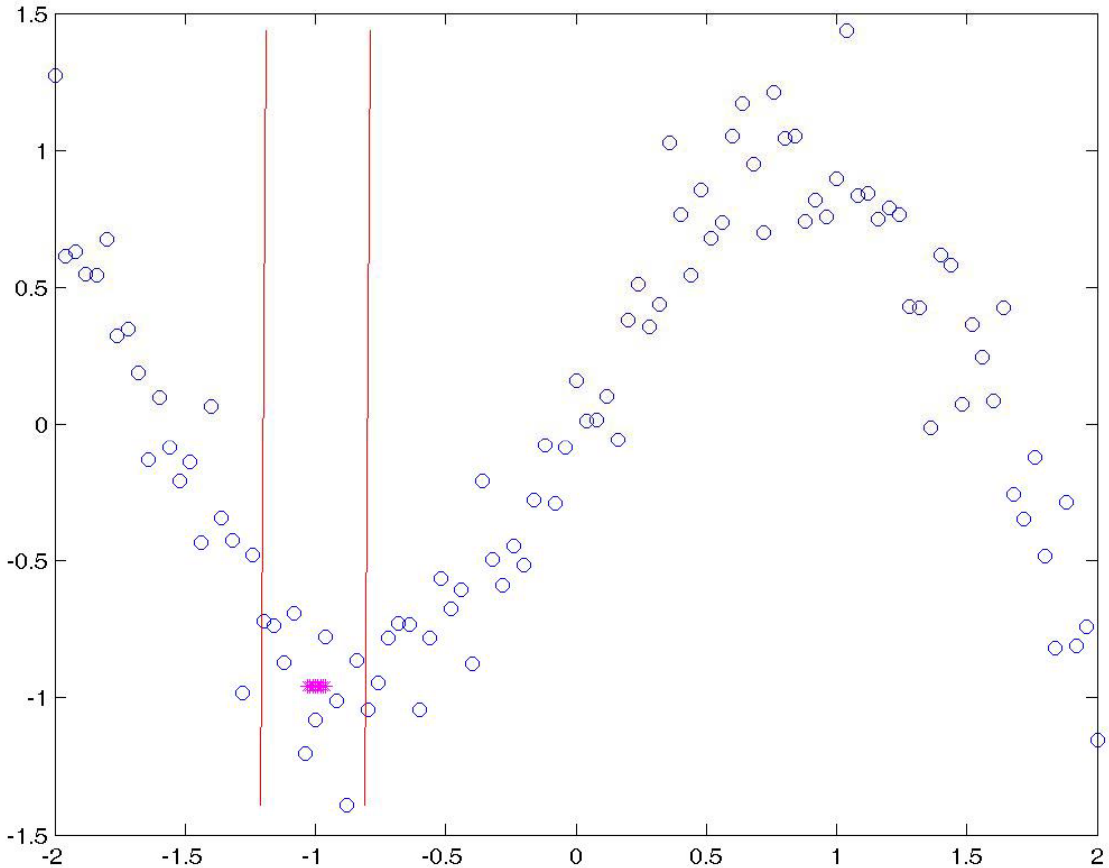
# Independent Data

- **Kernels** (local averages, local linear, etc.), with kernel density function **K** and bandwidth **h**

$$\hat{\Theta}(t) = \frac{n^{-1}\sum_{i=1}^{n} Y_i K\left(\frac{X_i - t}{h}\right)}{n^{-1}\sum_{i=1}^{n} K\left(\frac{X_i - t}{h}\right)}$$

- As the bandwidth **h** → 0, only observations with X near **t** get any weight in the fit

# Kernel Regression



# Independent Data

- **Major methods**
  - **Splines**
  - **Kernels**
- Smoothing parameters required for both
- **Fits**: similar in many (most?) datasets
- **Expectation**: some combination of bandwidths and kernel functions look like splines

# Independent Data

- Splines and kernels are linear in the responses

$$\hat{\Theta}(t) = n^{-1} \sum_{i=1}^{n} G_n(t, X_i) Y_i$$

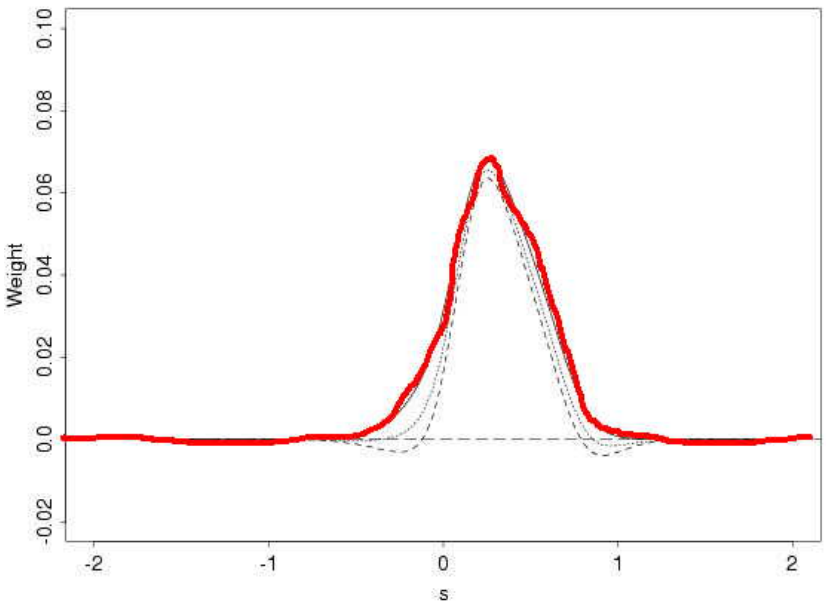
- Silverman showed that there is a kernel function and a bandwidth so that the weight functions

  $G_n(t, x)$ are asymptotically equivalent

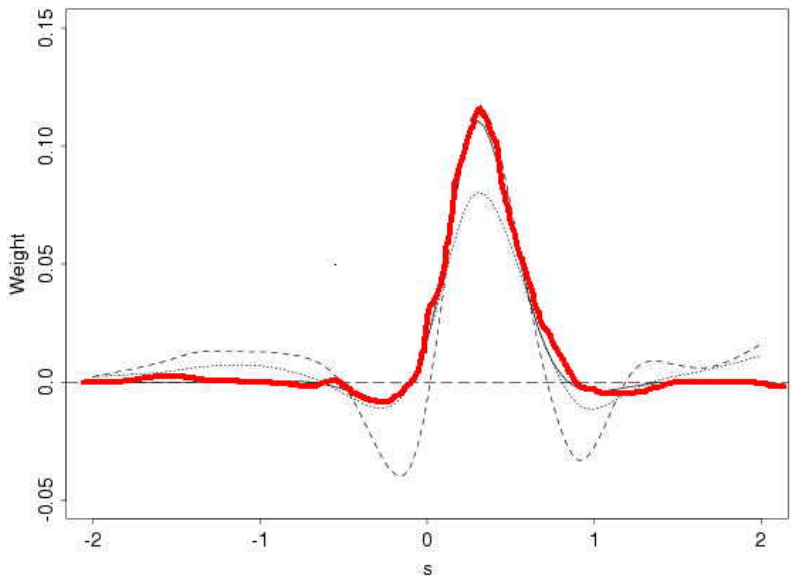
- In this sense, **splines = kernels**

# The weight functions $G_n(t=.25,x)$ in a specific case for independent data

**Kernel**




**Smoothing Spline**




**Note the similarity of shape and the locality: only X's near t=0.25 get any weight**

# Accounting for Correlation

- Splines have an obvious analogue for non-independent data
- Let $\Sigma_w$ be a working covariance matrix
  - Penalized **Generalized** least squares (GLS)

$$\sum_{i=1}^{n} \left\{ \underline{Y}_i - \underline{\Theta}(\underline{X}_i) \right\}^T \Sigma_w^{-1} \left\{ \underline{Y}_i - \underline{\Theta}(\underline{X}_i) \right\} + \lambda \int \left\{ \Theta''(t) \right\}^2 dt$$

- GLS ridge regression
- Because splines are based on likelihood ideas, they generalize quickly to new problems

# Accounting for Correlation

- **Splines** have an obvious analogue for non-independent data
- **Kernels** are not so obvious
  - One can do theory with kernels
- **Local likelihood** kernel ideas are standard in independent data problems
- Most attempts at kernels for correlated data have tried to use local likelihood kernel methods

# Kernels and Correlation

- **Problem**: how to define locality for kernels?
- **Goal**: estimate the function at $\mathbf{t}$
- Let $\underline{\mathbf{K}}(\mathbf{t}, \underline{\mathbf{X}}_i)$ be a diagonal matrix of standard kernel weights
- **Standard Kernel** method: GLS pretending inverse covariance matrix is

$$\underline{\mathbf{K}}^{1/2}(\mathbf{t}, \underline{\mathbf{X}}_i)\boldsymbol{\Sigma}_w^{-1}\underline{\mathbf{K}}^{1/2}(\mathbf{t}, \underline{\mathbf{X}}_i)$$

- The estimate is inherently local

# Kernels and Correlation

**Specific case: m=3, n=35**

**Exchangeable correlation structure**

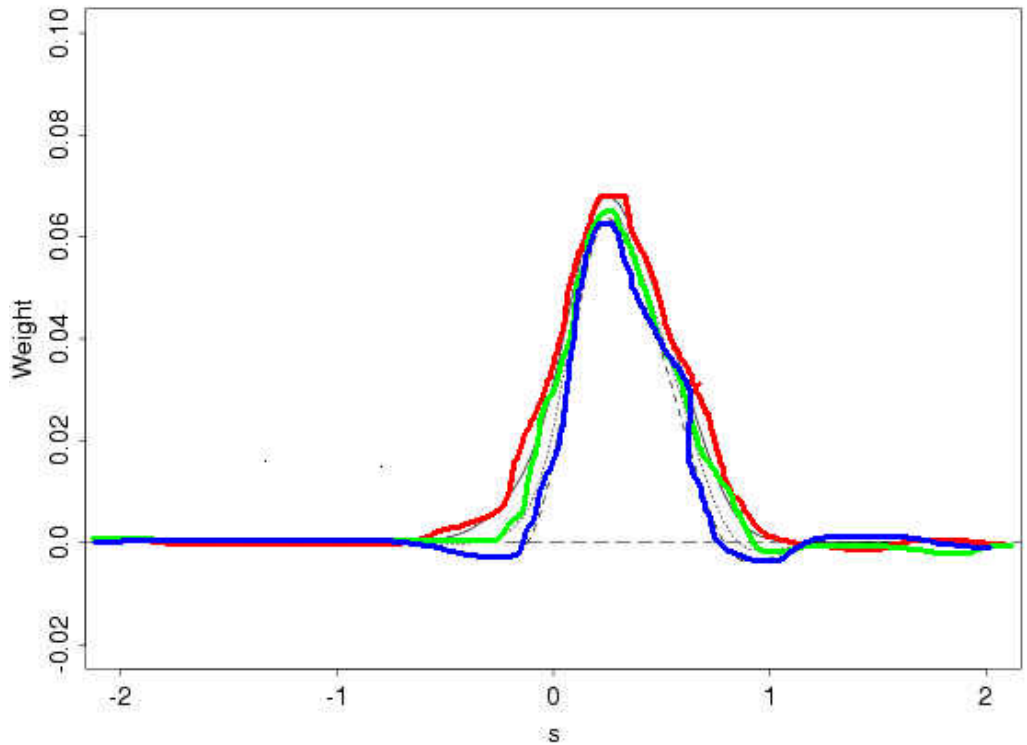
**Red: $\rho$ = 0.0**

**Green: $\rho$ = 0.4**

**Blue: $\rho$ = 0.8**

**Note the locality of the kernel method**

**The weight functions $G_n(t=.25,x)$ in a specific case**



# Splines and Correlation

**Specific case: m=3, n=35**

**Exchangeable correlation structure**

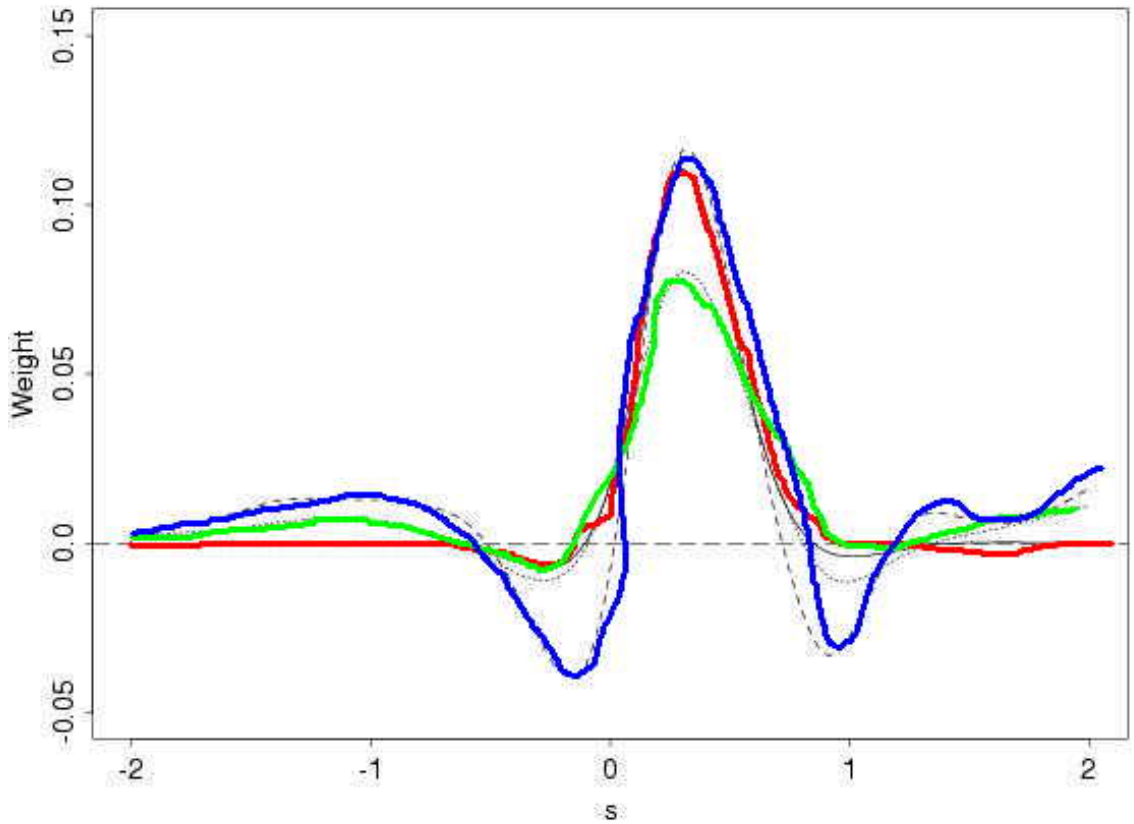
**Red: $\rho$ = 0.0**

**Green: $\rho$ = 0.4**

**Blue: $\rho$ = 0.8**

**Note the lack of locality of the spline method**

**The weight functions $G_n(t=.25,x)$ in a specific case**



# Splines and Correlation

**Specific case: m=3, n=35**

**Complex correlation structure**

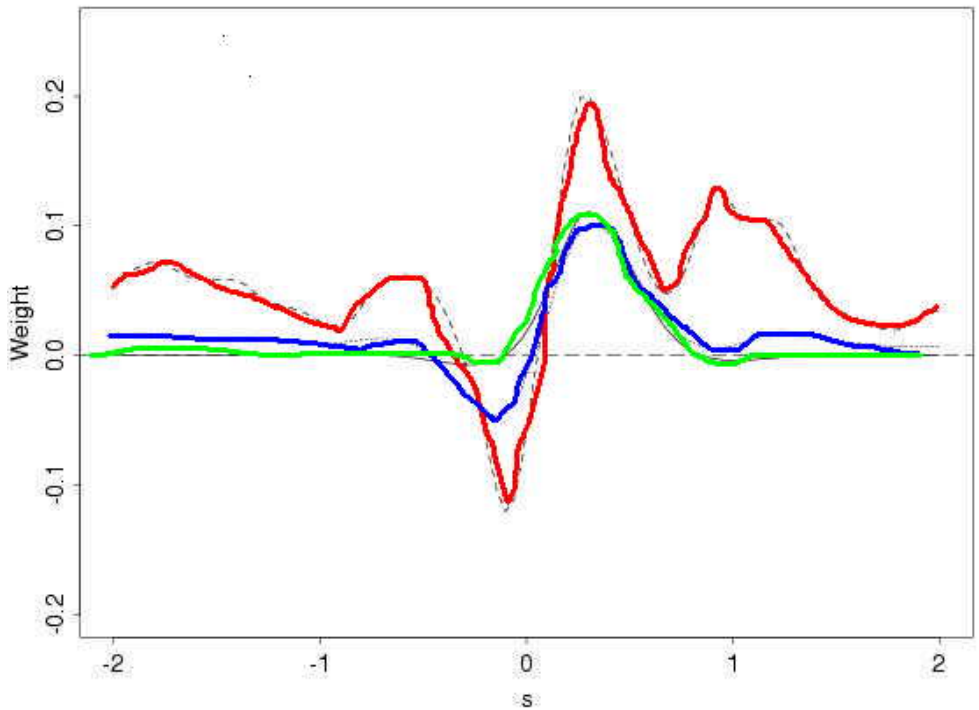
**Red: Nearly singular**

**Green: $\rho$ = 0.0**

**Blue: $\rho$ = AR(0.8)**

**Note the lack of locality of the spline method**

**The weight functions $G_n(t=.25,x)$ in a specific case**



# Splines and Standard Kernels

- **Accounting** for correlation:
  - **Standard kernels** remain local
  - **Splines** are not local
- Numerical results have been confirmed theoretically

# Results on Kernels and Correlation

- GLS with weights

$$\underline{K}^{1/2}(t, \underline{X}_i)\Sigma_w^{-1}\underline{K}^{1/2}(t, \underline{X}_i)$$

- **Optimal** working covariance matrix is **working independence**!
- Using the correct covariance matrix
  - Increases variance
  - Increases MSE
  - **Splines ≠ Kernels** (or at least these kernels)

# Pseudo-Observation Kernel Methods

- **Better** kernel methods are possible
- **Pseudo-observation**: original method
- **Construction**: specific linear transformation of Y
  - Mean = $\Theta(X)$
  - Covariance = diagonal matrix

$$\Omega = \Sigma_w^{-1/2} \qquad \Lambda = \mathbf{diag}(\Omega)$$

$$\underline{\mathbf{Y}}_i^* = \underline{\mathbf{Y}}_i + \Lambda^{-1}(\Omega - \Lambda)\left\{\underline{\mathbf{Y}}_i - \Theta(\underline{\mathbf{X}}_i)\right\}$$

  - This adjusts the original responses without affecting the mean

# Pseudo-Observation Kernel Methods

$$\underline{\mathbf{Y}}_i^* = \underline{\mathbf{Y}}_i + \Lambda^{-1}(\Omega - \Lambda)\left\{\underline{\mathbf{Y}}_i - \Theta(\underline{\mathbf{X}}_i)\right\}$$

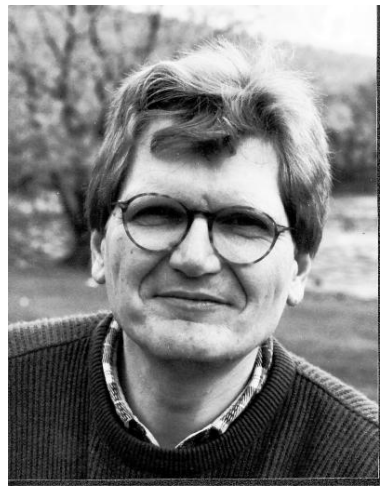
- **Construction**: specific linear transformation of Y
  - Mean = $\Theta(X)$
  - Covariance = diagonal
- **Iterative**:
- **Efficiency**: More efficient than working independence
- **Proof of Principle**: kernel methods can be constructed to take advantage of correlation

# Time Series Problems

- **Time series** problems: many of the same issues arise
- **Pseudo-observation** method
  - Two stages
  - Linear transformation
  - Mean $\Theta(X)$
  - Independent errors
  - Standard kernel applied
- **Potential** for great gains in efficiency (even infinite for AR problems with large correlation)

# Acknowledgments

Oliver Linton,
London School
of Economics



Enno Mammen,
University of Mainz

# Time Series: AR(1) Illustration, Pseudo Observation Method

- AR(1), correlation $\rho$:

$$\varepsilon_t - \rho\varepsilon_{t-1} = u_t \text{ (white noise)}$$

$$Y_t^0 = Y_t - \rho\left\{Y_{t-1} - \Theta(X_{t-1})\right\}$$

- Regress $Y_t^0$ on $X_t$. Efficiency of original pseudo-observation method to working independence:

$$\frac{1}{1-\rho^2} :\to \infty \text{ as } \rho \to 1$$

# Time Series: AR(1) Illustration

- The obvious question: can we do better?
- There are many possible pseudo-observations

$$Y_t^0 = Y_t - \rho\{Y_{t-1} - \Theta(X_{t-1})\}, \text{mean} = \Theta(X_t), \text{var} = \sigma_u^2$$

$$Y_t^0 = Y_{t-1} - \rho^{-1}\{Y_t - \Theta(X_t)\}, \text{mean} = \Theta(X_{t-1}), \text{var} = \sigma_u^2 / \rho^2$$

- Obvious alternative: two separate regressions, plus the obvious weighted average

# Time Series Problems

- AR(1) errors with correlation $\rho$
- Efficiency of original pseudo-observation method to working independence:

$$\frac{1}{1-\rho^2} :\to \infty \text{ as } \rho \to 1$$

- Efficiency of new pseudo-observation method to original pseudo-observation method:

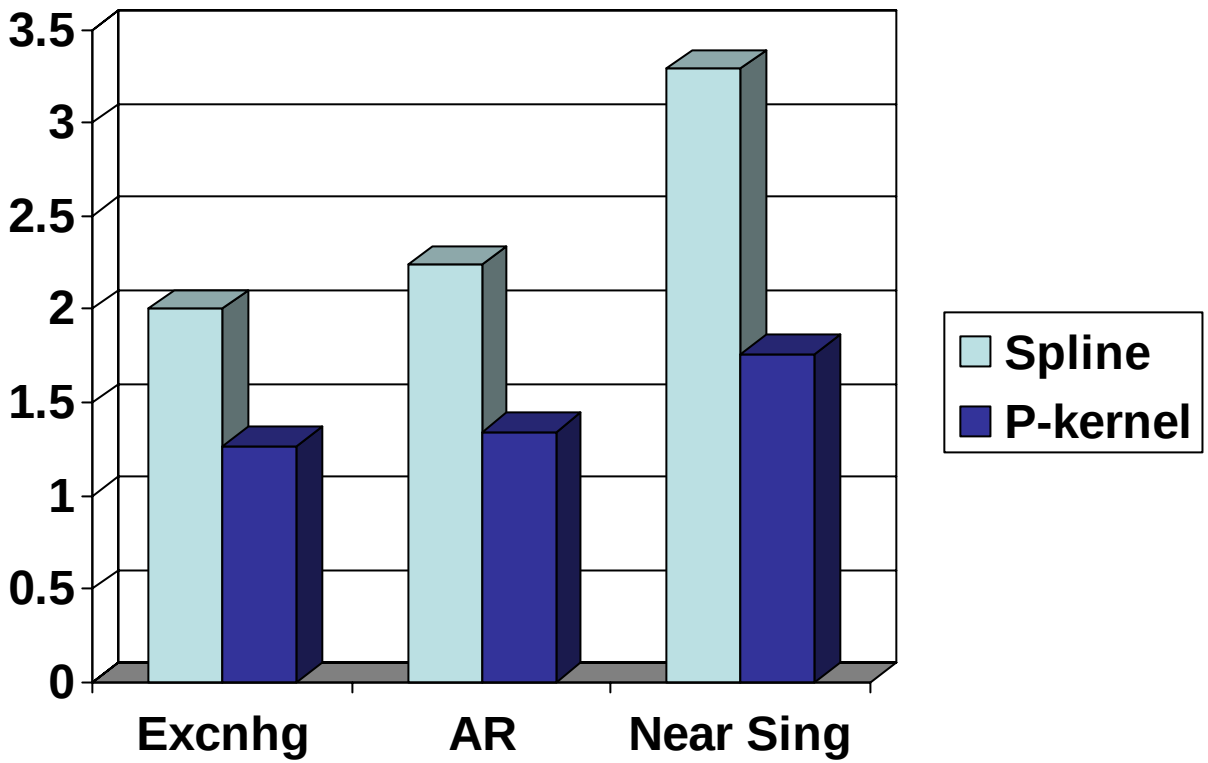
$$1+\rho^2 :\to 2 \text{ as } \rho \to 1$$

# Efficiency of Splines and Pseudo-Observation Kernels

**Exchng**: Exchangeable with correlation 0.6

**AR**: autoregressive with correlation 0.6

**Near Sing**: A nearly singular matrix



# What Do GLS Splines Do?

- GLS Splines are really working independence splines using pseudo-observations
- Let

$$\Sigma^{-1} = \left(\sigma^{jk}\right)_{jk}$$

- GLS Splines are working independence splines

$$\text{weights} = \sigma^{jj}$$

$$\text{pseudo-obs } Y_{ij}^{*} = Y_{ij} + \sum_{k \neq j} \frac{\sigma^{jk}}{\sigma^{jj}} \left\{ Y_{ik} - \Theta(X_{ik}) \right\}$$

# GLS Splines and SUR Kernels

- GLS Splines are working independence splines

$$\Sigma^{-1} = \left(\sigma^{jk}\right)_{jk} \qquad \text{weights} = \sigma^{jj}$$

$$\text{pseudo-obs } Y_{ij}^{*} = Y_{ij} + \sum_{k \neq j} \frac{\sigma^{jk}}{\sigma^{jj}} \left\{ Y_{ik} - \Theta(X_{ik}) \right\}$$

- Algorithm: iterate until convergence
- Idea: for kernels, do same thing
- This is Naisyin Wang's SUR method (Biometrika, 2003)

# SUR Kernel Methods: Motivation

- Basic idea: we have m observations per cluster.
- Suppose we know the means for observations j=2,…,m, i.e., $E(Y_{ij})=\Theta_j(X_{ij})$
- How would we estimate the mean for the 1st observation, i.e., $E(Y_{i1})=\Theta_1(X_{i1})$
- Method: local likelihood but using all the responses $Y_{i1}, Y_{i2}, \ldots, Y_{im}$
- Wang simply sums the resulting estimating equations

# SUR Kernel Methods

- It is well known that the GLS spline has an exact, analytic expression
- We have shown that the SUR kernel method has an exact, analytic expression
- Both methods are linear in the responses
- Nontrivial (for me!) calculations show that Silverman's result still holds
- **Splines = SUR Kernels**

# Nonlocality

- The lack of locality of GLS splines and SUR kernels is surprising
- Suppose we want to estimate the function at **t**
- **Result**: All observations in a cluster contribute to the fit, not just those with covariates near **t**
- **Locality**: Defined at the cluster level

$$\textbf{pseudo-obs } Y_{ij}^{*} = Y_{ij} + \sum_{k \neq j} \frac{\sigma^{jk}}{\sigma^{jj}} \left\{ Y_{ik} - \Theta(X_{ik}) \right\}$$

# The Semiparametric Model

- Y = Response
- X,Z = time-varying covariates

$$Y_{ij} = Z_{ij}\beta + \Theta(X_{ij}) + \varepsilon_{ij}$$
$$\mathrm{cov}(\varepsilon_{ij}) = \Sigma$$

- **Question**: can we improve efficiency for $\beta$ by accounting for correlation, i.e., what method is semiparametric efficient?

# Semiparametric Efficiency

- The semiparametric efficient score is readily worked out.
- Involves a Fredholm equation of the 2nd kind
- Effectively impossible to solve directly:
  - Involves densities of each X conditional on the others
- The usual device of solving integral equations does not work here

# The Efficient Score

$\tilde{X}=(X_1,\ldots,X_m)$

$\tilde{Z}=(Z_1,\ldots,Z_m)$

**Efficient Score**

$$\{\tilde{X}-\phi_{eff}(\tilde{Z})\}\Sigma^{-1}\{\tilde{Y}-\tilde{X}\beta-\theta(\tilde{Z})\}$$

**Fredholm equation:**

$$0\equiv\sum_{j=1}^{m}\sum_{k=1}^{m}\sigma^{jk}E[\{X_k-\phi_{eff}(Z_k)\}\mid Z_j=z]f_j(t)$$

# Profile Methods

- It is worth remembering what one would do here as a likelihood person.

$$Y_{ij}=Z_{ij}\beta+X_{ij}\Theta+\varepsilon_{ij}$$

- The profile likelihood says:

**for every $\beta$, estimate $\theta$ by GLS**

**$Y_{ij}-Z_{ij}\beta$ on $X_{ij}$**

**Call this $\hat{\theta}(\beta)$.**

**Maximize likelihood in $\{\beta,\hat{\theta}(\beta)\}$.**

# Profile Method: General Case

- Given $\beta$, solve for $\Theta$, say $\hat{\Theta}(X_{ij}, \beta)$
- Then fit GLS or W.I. to the model with mean

$$Z_{ij}\beta + \hat{\Theta}(X_{ij}, \beta)$$

- In this general case, how you estimate $\Theta$ matters
  - Working independence
  - Standard kernel
  - Pseudo-observation kernel
  - SUR kernel: **no surprise that this is the best!**

# Semiparametric Efficiency

- The semiparametric efficient method is profiled GLS with a SUR-kernel or a smoothing spline

- If you use working independence kernels/splines in this context, less efficiency

# Longitudinal CD4 Count Data (Zeger and Diggle): X = time of exam

| | Working Independence | | | Semiparametric GLS Z-D | |
|---|---|---|---|---|---|
| | Est. | s.e. | | Est. | s.e. |
| Age | .014 | .035 | | .010 | .033 |
| # of Smokes | .984 | .192 | | .549 | .144 |
| Drug Use? | 1.05 | .53 | | .58 | .33 |
| # of Partners | -.054 | .059 | | .080 | .038 |
| Depression? | -.033 | .021 | | -.045 | .013 |

# Conclusions (1/3): Nonparametric Regression

- In nonparametric regression
  - Kernels = splines for working independence (W.I.)
  - Working independence is inefficient
  - Personally, I think we ignore the inefficiency of working independence methods at some peril
  - Most of the literature uses working independence

# Conclusions (2/3): Nonparametric Regression

- In nonparametric regression
  - Pseudo-observation methods improve upon working independence
  - SUR kernels = splines for correlated data
  - Splines and SUR kernels are not local
  - Splines and SUR kernels are local in pseudo-observations

# Conclusions (3/3): Semiparametric Regression

- In semiparametric regression
  - When X is time-varying, method of estimating affects properties of parameter estimates
    - Using SUR kernels or GLS splines as the nonparametric method leads to efficient results
  - Conclusions can change between working independence and semiparametric GLS

# Conclusions: Splines versus Kernels

- One has to be struck by the fact that all the grief in this problem has come from trying to define kernel methods
- At the end of the day, they are no more efficient than splines, and harder and more subtle to define
- Showing equivalence as we have done suggests the good properties of splines

# The Numbers in the Table

The decrease in s.e.'s is in accordance with our theory. The other phenomena are more difficult to explain. Nonetheless, they are not unique to semiparametric GEE methods.

Similar discrepant outcomes occurred in parametric GEE estimation in which $\theta(t)$ was replaced by a cubic regression function in time. Furthermore, we simulated data using the observed covariates but having responses generated from the multivariate normal with mean equal to the fitted mean in the parametric correlated GEE estimation, and with correlation given by Zeger and Diggle.

The level of divergence between two sets of results in the simulated data was fairly consistent with what appeared in the Table. For example, among the first 25 generated data sets, 3 had different signs in sex partners and 7 had the scale of drug use coefficient obtained by WI 1.8 times or larger than what was obtained by the proposed method.

# The Marginal Nonparametric Model

- **Important assumption**
- Covariates at other waves are not conditionally predictive, i.e., they are surrogates

$$E(Y_{ij} | X_{ij}, X_{ik} \text{ for } k \neq j) = \Theta(X_{ij})$$

- This assumption is required for any GLS fit, including parametric GLS