Nonparametric Regression
and Clustered/Longitudinal Data

Raymond J. Carroll
Texas A&M University

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

carroll@stat.tamu.edu
Outline

- Longitudinal nonparametric model
- Kernel Methods
  - Working independence
  - Pseudo observation methods
- Comparison With Smoothing Splines
  - Locality
**Panel Data (for simplicity)**

- $i = 1, \ldots, n$ clusters/individuals
- $j = 1, \ldots, m$ observations per cluster

<table>
<thead>
<tr>
<th>Subject</th>
<th>Wave 1</th>
<th>Wave 2</th>
<th>...</th>
<th>Wave m</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>n</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
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?
- **We can do so for parametric problems (GLS)**
**Independent Data**

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

$$
\text{minimize } \sum_{i=1}^{n} \left\{ \mathbf{y}_i - \Theta(\mathbf{x}_i) \right\}^T \left\{ \mathbf{y}_i - \Theta(\mathbf{x}_i) \right\} + \lambda \int \left\{ \Theta''(t) \right\}^2 dt
$$

- 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 $h \to 0$, only observations with $X$ near $t$ get any weight in the fit
Kernel Methods

- Largely based on *working independence*

- **Ignores correlation structure** entirely in the fitting
  - Fixes up standard errors afterwards

- Large literature

- Significant loss of efficiency possible, as with any problem
Kernel Methods

• First kernel methods trying to account for correlation **failed**

• **Bizarre result**: Knowing the correct covariance matrix was worse than working independence

• **Justification** for working independence?

• **Difficulty**: defining “locality” for multivariate observations with the same mean function
**Pseudo-observation Kernel Methods**

- **Pseudo-observations** transform the responses

- **Construction**: linear transformation of $Y$

- **Mean** = $\Theta(X)$ remains unchanged

- **Obvious**(?): make the covariance matrix diagonal

- **Apply** standard kernel smoothers to independent pseudo-observations
Pseudo-observation Kernel Methods

- **Choices**: infinite, but one works

\[ \Sigma_w = \text{working covariance matrix} \]

\[ \Omega = \Sigma_w^{-1/2} \quad \Lambda = \text{diag}(\Omega) \]

\[ Y_i^* = Y_i + \Lambda^{-1}(\Omega - \Lambda)\{Y_i - \Theta(X_i)\} \]

- **Note**: The mean is unchanged

- **Iterate**: Start with W.I., transform, apply working independence smoother, etc.

- **Efficiency**: Always better than working independence
**Pseudo-observation Kernel Methods**

- **Construction**: Mean = \( \Theta(X) \) unchanged
  - Covariance = diagonal, back to independence
- **Generalizes** to Time Series, say AR(1)

\[
Y_t^0 = Y_t - \rho \{ Y_{t-1} - \Theta(X_{t-1}) \}
\]

- Efficiency with respect to working independence

\[
\frac{1}{1 - \rho^2} \to \infty \quad \text{as} \quad \rho \to 1
\]
Pseudo-observation Kernel Methods

- **Time Series**: Generalizations to finite order ARMA process possible
  - Multiple transformations chosen so that resulting estimates are asymptotically independent, then average
  - In AR(1), reverse the roles of current and lagged variables
- It is not clear, however, that insisting on a transformation to independence is efficient
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)

\[ \sum_{i=1}^{n} \left\{ Y_i - \Theta(X_i) \right\}^T \Sigma_w^{-1} \left\{ Y_i - \Theta(X_i) \right\} + \lambda \int \left\{ \Theta''(t) \right\}^2 dt \]

- GLS ridge regression
- Because splines are based on likelihood ideas, they generalize quickly to new problems
Efficiency (with respect to WI) of Splines and Pseudo-Observation Kernels: Splines Superior

**Exchng:**
Exchangeable with correlation 0.6

**AR:**
Autoregressive with correlation 0.6

**Near Sing:** A nearly singular matrix
New Construction

• Due to Naisiyin Wang (*Biometrika*, 2003)
• Multiple steps
• Get initial estimate \( \hat{\Theta} \)
• \( m \) observations per cluster/individual
• Consider observation \( j=1 \). Assume that \( \Theta \) is known and equal to \( \hat{\Theta} \) for \( k=2,\ldots,m \)
• Form local likelihood score with only the 1st component mean unknown
New Construction

- Continue. Consider observation $j$. Assume that $\Theta$ is known and equal to $\Theta$ for $k \neq j$
- Form local likelihood score with only the $j^{th}$ component mean unknown
- Repeat for all $j$
- Sum local likelihood scores over $j$ and solve
- Gives new $\hat{\Theta}$
- Now iterate.
Efficiency (with respect to WI) of Splines and Wang-type Kernels: Nearly identical

**Exchng:** Exchangeable with correlation 0.6

**AR:** autoregressive with correlation 0.6

**Near Sing:** A nearly singular matrix
GLS Splines and New Kernels

- **Relationship** between GLS Splines and the new kernel methods
- Both are **pseudo-observation methods**
- **Identical** pseudo-observations
- **Working independence** is applied to both pseudo-observations
- **Fitting methods** at each stage differ (splines versus kernels!)
- **Independence?** the pseudo-observations are not
GLS Splines and New Kernels

• Let $\Sigma^{-1} = \left( \sigma_{jk} \right)_{jk}$ be the inverse covariance matrix
• Form the pseudo-observations:

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

• Weight the $j^{th}$ component: weights $= \sigma_{jj}$
• Algorithm: iterate until convergence
• Use your favorite method (splines, kernels, etc.)
• This is what GLS splines and new Kernels do
• Not a priori obvious!
GLS Splines and New Kernels

• It is easy to see that GLS splines have an exact formula (GLS ridge regression)
• Less obvious but true that the new kernel methods also have an exact formula
• Both are linear in the responses

$$\hat{\Theta}_s(t) = \sum_{i=1}^{n} \sum_{j=1}^{m} W_{s,ij}(t, \text{all } X's)Y_{ij}$$

$$\hat{\Theta}_k(t) = \sum_{i=1}^{n} \sum_{j=1}^{m} W_{k,ij}(t, \text{all } X's)Y_{ij}$$
GLS Splines and New Kernels: Locality

• Write the linear expressions

\[ \hat{\Theta}_s(t) = \sum_{i=1}^{n} \sum_{j=1}^{m} W_{s,ij}(t, \text{all X's})Y_{ij} \]

\[ \hat{\Theta}_k(t) = \sum_{i=1}^{n} \sum_{j=1}^{m} W_{k,ij}(t, \text{all X's})Y_{ij} \]

• We generated data, fixed the first X for the first person at \( X_{11} = 0.25 \)

• Then we investigated the weight functions as a function of \( t \), where you want to estimate the regression function
The weight functions $W_{S,ij}(t, X_{11} = 0.25)$ and $W_{K,ij}(t, X_{11} = 0.25)$ for a specific case for correlated data, working independence.

Red = Kernel
Blue = Spline

Note the similarity of shape and the locality: only if $t$ is near $=0.25$ does $X_{11} = 0.25$ get any weight.
The weight functions $W_{S,ij}(t, X_{11}=0.25)$ and $W_{K,ij}(t, X_{11}=0.25)$ for a specific case for correlated data, GLS

Red = Kernel
Blue = Spline

Note the similarity of shape and the lack of locality:
The weight functions $W_{S,ij}(t,X_{11}=0.25)$ and $W_{S,ij}(t,X_{11}=0.25)$ for a specific case for correlated data, GLS versus Working Independence

Red = GLS

Blue = Working Independence
Three Questions

- Why are neither GLS splines nor Kernels local in the usual sense?

- The weight functions look similar in data. Does this mean that splines and kernels are in some sense asymptotically equivalent?

- Theory for Kernels is possible. Can we use these results/ideas to derive bias/variance theory for GLS splines?
Locality

- GLS Splines and Kernels are iterative versions of working independence applied to

\[ Y_{ij}^* = Y_{ij} + \sum_{k \neq j} \frac{\sigma_{jk}}{\sigma_{jj}} \left\{ Y_{ik} - \Theta(X_{ik}) \right\} \]

- Nonlocality is thus clear: if any \( X \) in a cluster or individual, say \( X_{i1} \), is near \( t \), then all \( X \)'s in that cluster, such as \( X_{i2} \), get weight for \( \Theta(t) \)

- Locality is thus at the cluster/individual level
Spline and Kernel Equivalence

- We have shown that a result similar to Silverman’s for independent data hold.

- Asymptotically, the spline weight function is equivalent to the same kernel weight function described by Silverman

\[ \frac{1}{2} \exp(-|t|/\sqrt{2}) \sin(|t|/\sqrt{2} + \pi/4) \]
Spline and Kernel Equivalence

- The bandwidth though changes: for cubic smoothing splines with smoothing parameter $\lambda$, let
  \[ \Sigma^{-1} = (\sigma^{jk})_{jk} \]

- Let the density of $X_{ij}$ be $f_j$

- Then the effective bandwidth at $t$ is
  \[ \left\{ \frac{\lambda}{\sum_{j=1}^{m} \sigma^{jj} f_j(t)} \right\}^{1/4} \]

- Note how this depends on the correlation structure
Asymptotic Theory for GLS Splines

- GLS splines have smaller asymptotic variance than WI splines for the same bandwidth
- We have derived the bias and variance formulae for cubic smoothing splines with fixed penalty parameter $\lambda \to 0$
- Without going into technical details, these formulae are the same as those for kernels with the equivalent bandwidth
- Generalizes work of Nychka to non-iid settings
Conclusions

- Accounting for correlation to improve efficiency in nonparametric regression is possible
- Pseudo-observation methods can be defined, and form an essential link
- GLS splines and the “right” GLS kernels have the same asymptotics
- Locality of estimation is at the cluster level, and not the individual $X_{ij}$ level.
Coauthors

Raymond Carroll  Oliver Linton  Alan Welsh

Naisyin Wang  Enno Mammen  Xihong Lin

Series of papers summarizing these results and their history are on my website.
Semiparametric Regression
Regression via penalized regression splines
Cambridge University Press, 2003