NONPARAMETRIC REGRESSION FOR MARGINAL MODELS IN LONGITUDINAL DATA

Raymond J. Carroll

Texas A&M University

Departments: Statistics, Epidemiology & Biostatistics, Nutrition, Toxicology

This work was done with Xihong Lin (University of Michigan) and Alan Welsh (Australian National University).

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

research papers,

paper on time to review and

training program announcement for Biostatistics, Bioinformatics and the Biology of Nutrition and Cancer
Basic Conclusions

• I will define a marginal longitudinal model of GEE type for nonparametric regression

• I will then describe the surprising (distressing?) asymptotic theory for kernel methods.
  
  • Basically, kernel methods in the literature suggest that you ignore the correlation structure

• I will then contrast kernel methods to spline methods
  
  • Basically, spline methods suggest that you make use of the correlation structure

• Either the wrong kernel methods have been suggested in the literature

• Or kernel methods are the wrong way to attack marginal longitudinal models
  
  • and I’ll show why it is the former
**Panel Data**

- Panel data: Observations are obtained at the same time points (waves), e.g., every 3 months.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Wave</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>( \times )</td>
</tr>
<tr>
<td>2</td>
<td>( \times )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>( n )</td>
<td>( \times )</td>
</tr>
</tbody>
</table>
MAIN QUESTIONS:

• In ordinary parametric longitudinal marginal models, it is well known that generalized least squares gives more efficient parameter estimates than does ignoring correlations

  - Correct Covariance vs working independence

• Question Do results such as this hold in nonparametric regression?

• This is a purely theoretical talk.

  - However, there are gobs of longitudinal data applications where nonparametric regression is of use.

• The kernel results easily extend to general (GLIM–type) models.

  - I will restrict to the Gaussian–type case for simplicity
**Some Basic Background**

- The data $Y_{ij}, X_{ij}$:
  
  - **subject** $i = 1, \ldots, n$
  
  - **observation** $j = 1, \ldots, M < \infty$
  
  - We assume that the number of observations per subject is bounded (not time series!)

- A marginal longitudinal GLIM specifies the marginal means and variances, e.g.,

  $$E(Y_{ij}|X_{ij}) = X_{ij}\beta;$$
  $$\text{var}(Y_{ij}|X_{ij}) = \sigma^2$$

- GEE’s fit generalized least squares with a working correlation matrix

  - The most efficient estimates are obtained by correctly specifying the correlation structure.
SOME BASIC BACKGROUND

- **Goal #1:** Define a marginal nonparametric model
  
  - Develop working covariance matrix methods
  
  - Answer the question: *is there any benefit to estimating or knowing the covariance matrix?*
  
  - We claim that for kernel regression, the answer is **NO!!** for common methods.

- **Goal #2:** Contrast the behavior of kernels with splines.
  
  - Univariate equivalence of splines and kernels
  
  - Develop working covariance matrix methods for splines
  
  - Answer the question: *is there any benefit to estimating or knowing the covariance matrix?*
  
  - We claim that for splines, the answer is a qualified **YES!!**
**NONPARAMETRIC MARGINAL MODEL**

- \( i = \) subject \( i = 1, \cdots, n \)

- \( j = \) observation \( j = 1, \cdots, M < \infty \)

\[
E(Y_{ij} \mid X_{ij}) = \theta(X_{ij}),
\]

\( \theta(\cdot) = \) smooth unknown function.

- Note that the basic function \( \theta(\cdot) \) does not change over time

- The values and possibly the distribution of the covariates changes over time
**Review of Parametric GEEs:**

- Parametric $p$th polynomial model:
  \[
  \mu(X_{ij}) = G_p(X_{ij})^T \beta,
  \]
  where $G_p(x) = (1, x, \cdots, x^p)^T$.

- Parametric GEEs:
  \[
  \sum_{i=1}^{n} G_{ip}^T V^{-1}(Y_i - \mu_i) = 0,
  \]
  where
  \[
  V = \text{working correlation matrix}
  \]

- Remember that I’m assuming constant variance over time

- Properties of Parametric GEEs:
  - $\hat{\beta}$ is consistent when $\mu(\cdot)$ is correctly specified even when $V$ is misspecified.
  - $\hat{\beta}$ is most efficient when $V$ = true correlation.
The idea of kernel smoothing is to fit the polynomial locally.

- \( K_h(u) = h^{-1}K(u/h), \) \( K(\cdot) \) = kernel function
- \( K_{ih}(x) \) is the diagonal matrix with weights.

- In parametric problems, you use generalized least squares (GLS) with a working covariance matrix \( V \)
- Nonparametric methods fit a polynomial locally at any \( x \)
  - They use GLS but with the inverse of the working covariance matrix being
  \[
  K_{ih}^{1/2}(x)V^{-1}K_{ih}^{1/2}(x)
  \]
- The function is easily computed on a grid.
ASYMPTOTIC RESULTS II: KERNEL GEEs

- For this standard method, the bias and variance expressions take on simple forms.

- In the local linear case, the bias has the usual form.

- In the local linear case, the variance is minimized at working independence.

- Hence, generally, the optimal estimator in the MSE sense is to ignore the correlations entirely.

- For specificity, if $v^{jj}$ is the $j$th diagonal element of the inverse of the working covariance matrix $V$, and if the marginal distributions of $X$ are the same, then the asymptotic variance is a constant independent of the working covariance matrix times

$$\frac{\sum_{j=1}^{M} v^{jj}}{\left(\sum_{j=1}^{M} v^{jj}\right)^{2}}$$
ASYMPTOTIC RESULTS: SUMMARY

• The most striking result is that working independence is generally optimal.

• There is no gain, and can be large cost, to spending effort trying to estimate the actual covariance matrix of observations within individuals.

• Remember, the variance is

\[ \frac{\sum_{j=1}^{M} v_{jj}^2}{\left( \sum_{j=1}^{M} v_{jj} \right)^2} \]

• Generally, the cost due to accurately estimating the correlations is greatest when one correlation is high and the others are not.

• \( M = 3, \rho_{12} = 0.9, \rho_{13} = 0.1, \rho_{23} = 0.1 \)

• Kernel-GLS has asymptotic variance 27% larger than working independence
**ASYMPTOTIC RESULTS: SPECULATION**

- Why is working independence reasonable in this? A partial explanation?

- Remember what kernel methods do. They have a bandwidth $h$, and weights at $x$ are of the form

  $$ K\{ (X_{ij} - x)/h \}. $$

- In asymptotics, $h \to 0$.

- The chance that $X_{i1}$ and $X_{i2}$ are both close to $x$ is small.

- Thus, the kernel method really forces independence.

- When you try to impose dependence, the mixed signal causes grief.
EXTENSIONS

- We have extended these basic results

- To marginal GLIM’s
  - The basic conclusions do not change

- To partially linear marginal GLIM’s
  \[ E(Y_{ij}|X_{ij}, Z_{ij}) = \theta(X_{ij}) + Z_{ij}\beta \]
  - The same results apply to estimating \( \theta(\cdot) \)
  - We have computed the semiparametric efficient estimate of \( \beta \)
  - It is not what has been claimed in the literature (very complex)
We have extended these basic results to measurement error models, where you do not observe $X_{ij}$, but instead observe

$$W_{ij} = X_{ij} + \text{random error}$$

- SIMEX method of Cook & Stefanski
- Strange results depending on how you apply their method: globally (bad) or by wave (better)
- Strange results on bandwidth asymptotics
- It is worth reemphasizing the basic result
  - Known kernel methods cannot take advantage of correlations in longitudinal data to improve efficiency
Splines and Equivalent Kernels

- In univariate iid problems, smoothing spline methods are asymptotically equivalent to kernel methods (Silverman)
  - Higher order kernel

- In working independence, you basically lump all the data together and then run your favorite nonparametric regression

- We have generalized Silverman’s result to show that in the marginal longitudinal model, under working independence, smoothing spline methods are asymptotically equivalent to kernel methods

- Working independence is optimal for kernels
- So, splines cannot be any worse than kernels
- Are splines generally equivalent to kernels?
LONGITUDINAL SMOOTHING AND P-SPLINES

- Remember that the working covariance matrix is $V$
  - In the presence of correlated data, you **pretend** that the true covariance matrix is $V$

- Formulae for smoothing splines with correlated data are standard but too complex for here, see e.g., Green & Silverman

- Let $Y$ be the vector of responses for an individual

  - **Effectively**, $Y^* = V^{-1/2}Y$ has a (working) diagonal covariance matrix
    - **Effectively**, do what comes naturally to $Y^*$ and fix up
**P–Splines**

- Splines have basis functions $B_j(x)$ and a penalty matrix $\Omega$

- The regression P–spline of order $p$ using the plus–function basis as defined here has the following properties

  - If uses $L$ knots $\xi_1, \ldots, \xi_L$
  - $B_j(x) = x^j$ for $j = 0, \ldots, p$
  - $B_j(x) + (x - \xi_{j-p})_+^p$ for $j > p$
  - $x_+ = x I(x > 0)$

- Fitted function is $\sum_j B_j(x) \beta_j$

- $\Omega = \text{diag}(0_{p+1}, 1_L)$
\textbf{P–SPLINES}

- If number of knots $L$ is fixed but large, generally competitive with smoothing splines

- In the univariate case, minimize

\[\sum_i \{Y_i - \sum_j B_j(X_i)\}^2 + \lambda \beta^T \Omega \beta\]

- Used here for some asymptotic convenience and also as a robustness check on the smoothing spline results.
**Some Exact Calculations**

- We did exact calculations first, to compare with optimal MSE’s for kernel methods.
  - P–splines, plus function basis, 35 knots
  - \( n = 50 \) individuals, \( M = 3 \) observations per
  - \( X \) uniform\([-2,2]\)
  - \( \theta(x) = \sin(2x) \)
  - Computed MSE exactly, averaged over the distribution of \( X \), then minimized over smoothing parameter \( \lambda \)
  - \( \text{corr}(Y_1, Y_2) = \text{corr}(Y_2, Y_3) = 0.8, \text{corr}(Y_1, Y_3) = 0.5 \), all known

- The P–spline optimal efficiency compared to the best kernel method: 1.96!!

- Splines are better than kernels?
A THEORETICAL SIMULATION

- $\theta(x) = \sin(2x)$

- $n = 50$ individuals

- $M = 3$ observations per individual

- $X$’s are independent uniform $[-2, 2]$

- $\sigma^2 = 1$

- $\text{corr}(Y_1, Y_2) = \text{corr}(Y_2, Y_3) = 0.8$

- $\text{corr}(Y_1, Y_3) = 0.5$

- This is theoretical in the sense that
  
  - Use true correlations instead of estimating them
  
  - Smoothing parameters for splines and kernels chosen to minimize average MSE
  
  - Done for working independence and true correlations
Case #3: \( \sin(2x) \)

Figure 1: The function \( \sin(2x) \) with \( X \)'s uniform on \([-2,2]\) and the best fitting quadratic function.
THEORETICAL SIMULATION

- Optimized efficiency when using GLS versus working independence
  - Kernel method: 0.96
  - P–spline, 35 knots: 2.03
  - Smoothing spline: 1.97

- Optimized efficiency of spline GLS compared to the kernel method with working independence
  - P–spline, 35 knots: 2.40
  - Smoothing spline: 2.46

- Conclusion: splines are more efficient than kernels
  - Many simulations of various functions confirm this result.
REAL SIMULATION

- True covariance estimated by residuals derived from working independence
- GCV used to estimate smoothing parameters for smoothing and P–splines
- EBBS used to estimate bandwidths for kernels using working independence
- Efficiency of spline GLS compared to the kernel method with working independence
  - P–spline, 35 knots: 2.01
  - Smoothing spline: 2.08
- Conclusion: **splines are more efficient than kernels**
What’s Going On?

- Kernels are inherently local

- Smoothing splines are local (like kernels) in univariate or working independence situations
  - Silverman basically shows that the “hat” matrix is banded asymptotically for small values of the smoothing parameter

- Splines are not local in GLS situation
  - Theoretically, the hat matrix is not banded asymptotically for small values of the smoothing parameter
  - The very fact that splines are not local means that they can take advantage of the correlation structure
The methods have equivalent kernel algebraic formula

If you want to predict the function at $X = x$, the methods satisfy

$$\tilde{\theta}(x) = \sum_{i=1}^{n} \sum_{j=1}^{M} G_{ij}(x, X_{ij}) Y_{ij}$$

$G_{ij}(x, X_{ij})$ is called the equivalent kernel

We have computed the formulae for the equivalent kernels for all the estimators ($X_{ij} = 0.25$, $X$’s uniform[-2,2])

From our asymptotics, we expect GLS kernel methods to behave locally

From our simulations and asymptotics, we expect GLS spline methods to not behave locally
Figure 2: The equivalent kernel for $n = 35$ individuals, $M = 3$ observations per individual, using a GLS kernel method, when the true correlation matrix has constant correlation with $\rho = 0.0$ (solid line), $\rho = 0.4$ (dotted line), $\rho = 0.8$ (dashed line).
Figure 3: The equivalent kernel for $n = 35$ individuals, $M = 3$ observations per individual, using a GLS smoothing spline method, when the true correlation matrix has constant correlation with $\rho = 0.0$ (solid line), $\rho = 0.4$ (dotted line), $\rho = 0.8$ (dashed line).
It’s possible to do better with kernels, and use the correlations

Remember splines:

- \( Y \): vector of responses for a panel member.
- \( B \): matrix of basis functions of the spline
- Splines think \( E(Y) = B\beta_0 \)
- This means a linear transformation is a spline with identity covariance matrix

\[
E(\Sigma^{-1/2}Y) = E(Y_*) = B_*\beta_0 = \Sigma^{-1/2}B\beta_0
\]

- Estimate \( \beta_0 \) by penalized least squares

Virtually any simulation you do shows that the GLS spline is more efficient than the working independence spline

- Can we use the idea to improve kernels?
Figure 4: The equivalent kernel for the GLS smoothing spline when $n = 35$, $m = 3$. The solid line is working independence, the dotted autocorrelation $\rho = 0.8$, and the dashed line is an unstructured and nearly singular correlation matrix.
KERNELS AND SPLINES

- Assume common variance over waves
- Let $\tilde{\theta}(x)$ be a working independence, undersmoothesd kernel estimator
- $Z(\theta) = Y + (\Sigma^{-1/2} - I)\{Y - \theta(X)\}$ has mean $\theta(X)$ and identity covariance matrix
- Regress $Z_{ij}(\tilde{\theta})$ on $X_{ij}$ with working independence
- Can be shown to have smaller variance than $\tilde{\theta}(x)$, and same theoretical bias behavior
KERNELS AND SPLINES

- $X$’s uniform on [-2,2]
- 3 waves, 50 (100) clusters/panels
- Autoregression, $\rho = 0.6$
- A typical simulation has $\theta(x) = \sin(2x)$ (not fit well by a low-order polynomial)
- Efficiency results $n = 50$ (100)
  - GLS spline / independence spline = 1.49 (1.50)
  - GLS kernel / independence kernel = 2.29 (1.39)
  - Efficiency GLS spline / GLS kernel = 1.47 (1.61)
- These are fairly typical: splines always seem better
I defined a marginal longitudinal model of GEE type for nonparametric regression

I then described the surprising (distressing?) asymptotic theory for kernel methods.

- Basically, kernel methods suggest that you ignore the correlation structure

I then contrasted kernel methods to spline methods

- GLS splines are not local
- Basically, spline methods suggest that you make use of the correlation structure

My conclusion is that the wrong kernel methods in the literature

Splines are better: kernel methods are the wrong way to attack marginal longitudinal models
More precisely, with either working independence or not, the fitted values are of the form $GY$ 

- $G$ is the “hat” matrix

- With working independence, can show that for sufficiently small $\epsilon$, there exists a constant $H$ such that

$$|G_{jk}| < \epsilon \text{ if } |j - k| > H$$

- With the equicorrelated case, there exists $c$ and $H_*$ such that

$$|G_{jk}| > c \text{ if } |j - k| > H_*$$