NONPARAMETRICS, SEMIPARAMETRICS AND MEASUREMENT ERROR FOR SOME MARGINAL MODELS IN LONGITUDINAL DATA

Raymond J. Carroll
Texas A&M University
http://stat.tamu.edu/~carroll (most papers available)
http://stat.tamu.edu/B3NC (Training program in biology & bioinformatics)

Xihong Lin
University of Michigan
http://www.sph.umich.edu/~xlin
Basic Conclusions

- We study a particular class of marginal nonparametric and semiparametric models for correlated data.
- The central theme is that intuition from the iid case often fails, and in subtle ways.
- Each problem has to be thought through carefully, even if you assume working independence. Here are a few results:
  - Semiparametric profile methods need not be semiparametric efficient.
  - Blind application of error–correction methods is inefficient.
  - Classic kernel methods do not have the behavior one would expect.
  - Splines seem to be better than kernel methods.
FRAMEWORK: PANEL DATA

- For cluster $i$, wave $j$, we have studied models of the general form

$$E(Y_{ij} | X_{ij}, Z_{ij}) = Z_{ij} \beta_0 + \theta(X_{ij}); \quad (1)$$

$$E(Y_{ij} | X_{ij}, Z_{ij}) = E\{Y_{ij} | X_{ij}, Z_{ij}, (X_{ik}, Z_{ik})_{k \neq j}\} \quad (2)$$

- The function $\theta$ is unknown

- Pepe and Couper (1997, JASA) have noted that if (2) fails, then fitting (1) should largely be done using working independence.

- We have also allowed either $Z$ or $X$ to be measured with error.

- The results apply to general marginal models, but are illustrated here in the “linear” case for convenience.

- Most of the results apply to general marginal models, with varying time points, but panel data are used here to keep the notation simple
For cluster $i$, wave $j$, we have studied models of the general form

$$E(Y_{ij} | X_{ij}, Z_{ij}) = Z_{ij} \beta_0 + \theta(X_{ij})$$

We have also studied its fully nonparametric version

$$E(Y_{ij} | X_{ij}) = \theta(X_{ij})$$

We have studied specific problems but also made general conclusions (especially how intuition fails!)
**Measurement Error**

**Nonparametrics**

- \( E(Y_{ij} | X_{ij}) = \theta(X_{ij}) \)
- \( X_{ij} \) measured with error
- You observe in a cluster \( W = X + U \),
- \( U \sim \text{Normal}(0, \Sigma_u) \)
- If \( \theta \) were parametric, **except for simplest linear models**, likelihood analyses would require model for the **joint** distribution of the latent variables \( X \)
- In earlier work with Naisyin Wang, we showed that if you pretended the \( X \)'s were independent and they are not, even in the linear model biases could creep in.
The SIMEX method of Cook & Stefanski is completely general and can be applied to nonparametric regression.

- In the usual linear model it gives the “right” answer.

The SIMEX method is the “default” method in the literature.

- Models for latent $X_{ij}$ not necessary.

But in this problem (panel data), it is an inefficient method, even for working independence.

- The exception is if the marginal distributions $(X_{ij})_{i=1}^{n}$ are independent of $j$. 

$E(Y_{ij} | X_{ij}) = \theta(X_{ij})$

- $X_{ij}$ measured with error
MEASUREMENT ERROR NONPARAMETRICS

- $E(Y_{ij} | X_{ij}) = \theta(X_{ij})$

- For panel data, a more efficient method is to run a separate SIMEX regression on each wave
  - Then take a weighted average of the estimated functions.
  - The weights are, of course, the inverses of the variances of the fits.

- Of course?
  - The individual SIMEX function estimates are asymptotically independent

- Note: intuition fails
Semiparametric Cluster-Level Covariate

- $X$ is measured at the cluster level, e.g., baseline

\[ E(Y_{ij} | X_i, Z_{ij}) = Z_{ij} \beta_0 + \theta(X_i) \]

- Methods are simple

  - For any fixed $\beta$, let $\tilde{\theta}(x, \beta, \Sigma)$ be the nonparametric regression of $Y_{ij} - Z_{ij} \beta$ on $X_i$ with working covariance $\Sigma$.

  - Kernel methods exist for this, spline methods are well–known.

  - Then $\tilde{\beta}$ is the GLS of $Y_{ij}$ on $Z_{ij} \beta + \tilde{\theta}(X_i, \beta, \Sigma)$ with working covariance $\Sigma$. **No iteration involved**

  - Semiparametric efficient in the Gaussian case, $\Sigma = \text{consistent estimate of the true covariance matrix}$
**Semiparametric Cluster-Level Covariate**

- $X$ is measured at the cluster level, e.g., baseline
  
  $$E(Y_{ij}|X_i, Z_{ij}) = Z_{ij}\beta_0 + \theta(X_i)$$

- The method described is really a **semiparametric profile** method

- $\hat{\theta}(x, \beta, \Sigma)$ is the nonparametric regression of $Y_{ij} - Z_{ij}\beta$ on $X_i$ with working covariance $\Sigma$.

- Then do a likelihood analysis (this is how to generalize)
  
  - mean function $Z_{ij}\beta + \hat{\theta}(X_i, \beta, \Sigma)$
  - Covariance matrix $\Sigma$

- Semiparametric efficient, essentially independent of what nonparametric regression you do.
**INDIVIDUAL–LEVEL COVARIATE**

- $X$ is measured at the individual level

$$E(Y_{ij}|X_{ij}, Z_{ij}) = Z_{ij}\beta_0 + \theta(X_{ij})$$

- Semiparametric profile method??

  - For any fixed $\beta$, let $\tilde{\theta}(x, \beta, \Sigma)$ be the nonparametric regression of $Y_{ij} - Z_{ij}\beta$ on $X_{ij}$ with working covariance $\Sigma$.

  - Kernel methods exist for this, spline methods are well–known.

  - Then $\tilde{\beta}$ is the MLE of $Y_{ij}$ on $Z_{ij}\beta + \tilde{\theta}(X_{ij}, \beta, \Sigma)$, working covariance $\Sigma$. No iteration involved

  - Semiparametric efficient in the Gaussian case, $\Sigma = $ consistent estimate of the true covariance matrix?

  - Ah – No!!!! And it gets worse.
**INDIVIDUAL–LEVEL COVARIATE**

- $E(Y_{ij}|X_{ij}, Z_{ij}) = Z_{ij}\beta_0 + \theta(X_{ij})$

- Semiparametric profile method is semiparametric efficient only if the following occurs:
  - $E(Z_{ij}|X_{ij}) = E(Z_{ij}|X_{ij}, X_{ik})$ for $k \neq j$

- In general problems, the semiparametric efficient score is the solution to a Fredholm integral equation of the second kind, involving the regressions of $Z_{ij}$ on $X_{ik}$

- We have no idea how to implement this in real life.
**INDIVIDUAL–LEVEL COVARIATE**

- $E(Y_{ij} | X_{ij}, Z_{ij}) = Z_{ij} \beta_0 + \theta(X_{ij})$

- Theoretically, the semiparametric profile method does not in general result in an estimate $\hat{\beta}$ which is asymptotically normal with mean $\beta_0$ and variance $\sigma^2/n$

- The problem is that there is a bias term from the non-parametric regression that does not disappear

- This is a standard feature in semiparametric problems
  - If you have a semiparametric inefficient estimator, standard nonparametric regressions will not work (theoretically)

- Curiously, none of this is a problem for working independence, although the method is inefficient.
INDIVIDUAL–LEVEL COVARIATE

- \( E(Y_{ij} | X_{ij}, Z_{ij}) = Z_{ij}\beta_0 + \theta(X_{ij}) \)

- If you have a semiparametric inefficient estimator, standard nonparametric regressions will not work (theoretically), except for working independence

- This does work to give asymptotically correct inference if you undersmooth the nonparametric regression, causing less bias but more variance

- Also a general result

- Easy to see how to do this with kernel methods: we give an explicit formula

- Multiply your bandwidth by \( n^{-2/15} \)!

- Not known how to do this with splines, or even whether it is necessary
Kernels and Splines

- $E(Y_{ij}|X_{ij}) = \theta(X_{ij})$

- The standard kernel methods are due to Severini & Staniswalis

- Let $\Sigma$ be the working covariance matrix.

- Let $K_{ij}(x)$ be the diagonal matrix of kernel weights when fitting the function at $x$

- Function estimate is from the GLS (local) linear regression of $Y_{ij}$ on $X_{ij}$ with weights $K_{ij}^{1/2}(x) \Sigma^{-1} K_{ij}^{1/2}(x)$

- Whoops!
  - The best working method is working independence
Kernels and Splines

- It’s possible to do better, 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?
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
Basic Conclusions

- We studied a particular class of marginal nonparametric and semiparametric models for correlated data.
- A central theme is that intuition from the iid case often fails, and in subtle ways.
- Each problem has to be thought through carefully, even if you assume working independence. Here are a few results:
  - Semiparametric profile methods need not be semiparametric efficient.
  - Blind application of error-correction methods is inefficient.
  - Classic kernel methods do not have the behavior one would expect.
  - Splines seem to be better than kernel methods.