

---

**NONPARAMETRIC REGRESSION FOR  
MARGINAL MODELS IN LONGITUDINAL  
DATA WHEN THE PREDICTOR IS  
MEASURED WITHOUT/WITH ERROR**

---

Xihong Lin

Raymond J. Carroll

---

## OUTLINE

---

- An example motivating the work
- Marginal longitudinal models and nonparametric regression
  - **Does generalized least squares give the same efficiency gains as in the parametric case?**
- The SIMEX method for measurement error corrections
  - **In panel data**, should one fit all the data together or fit at each time point and then combine?
- Basic framework
- Main results described

---

## PANEL DATA

---

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

| Subject | Wave |   |     |     |
|---------|------|---|-----|-----|
|         | 1    | 2 | ... | $M$ |
| 1       | ×    | × | ... | ×   |
| 2       | ×    | × | ... | ×   |
| ⋮       |      |   | ⋮   |     |
| $n$     | ×    | × | ... | ×   |

---

## 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**
- When we add **measurement error** in longitudinal data, the SIMEX method can be applied to all the data.
- In panel data, SIMEX can be applied to all data simultaneously, or it can be applied to fit the model at each time point, and then the fits can be optimally combined.
  - In parametric problems, the former is efficient
- **Question** Do these results hold in nonparametric regression?

---

## PRACTICAL MOTIVATION: AIDS COSTS AND SERVICES UTILIZATION SURVEY

---

- A panel study with 273 HIV+ white males interviewed every 3 months for 18 months.
- Main question of interest:  
how do CD4 counts affect the hospitalization (y/n) risk?
- Complications:
  - The relationship does not even fit a **cubic** model
  - Measurement errors in CD4 counts (**CV=50%**)
- Analyses needed:
  - Nonparametric regression for longitudinal data.
  - Accounting for measurement error.

---

## SOME BASIC BACKGROUND

---

- The data  $Y_{ij}, X_{ij}$ :
  - **subject**  $i = 1, \dots, n$
  - **observation**  $j = 1, \dots, M_i \leq M < \infty$
  - We assume that the number of observations per subject is bounded (**not time series!**)
  - In panel data,  $M_i \equiv M$
- A marginal longitudinal GLIM specifies the marginal means and variances, e.g.,

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

- 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:** Define the SIMEX method for handling the measurement error in the covariates.
  - Study specifically the case of panel data.
  - Since the model is marginal, we can fit the function at each time point and combine.
  - Is this worth doing, over fitting simultaneously?
  - We claim that in kernel regression, the answer is **YES!!**

---

## NONPARAMETRIC MODEL: NO MEASUREMENT ERROR

---

- $i$ =subject  $i$  ( $i = 1, \dots, n$ )
- $j$ =observation  $j$  ( $j = 1, \dots, M_i \leq M < \infty$ )
- $Y_{ij}$ =outcome (Normal, Binomial, Poisson, etc)
- $X_{ij}$ =Covariate

$$E(Y_{ij}|X_{ij}) = \mu_{ij}, \text{ var}(Y_{ij}|X_{ij}) = \sigma_j^2 V(\mu_{ij}),$$

- **Model:**

$$g(\mu_{ij}) = \theta(X_{ij}),$$

where  $g(\cdot)$ =link function,  $\theta(\cdot)$ =smooth function.

- **Basic Technique** Estimate  $\theta(\cdot)$  using kernel methods.

---

## REVIEW OF PARAMETRIC GEEs:

---

- Parametric  $p$ th polynomial model:

$$g(\mu_{ij}) = \mathbf{G}_p(X_{ij})^T \boldsymbol{\beta},$$

where  $\mathbf{G}_p(x) = (1, x, \dots, x^p)^T$ .

- Parametric GEEs:

$$\sum_{i=1}^n \mathbf{G}_{ip}^T \boldsymbol{\Delta}_i \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0,$$

where

$$\boldsymbol{\Delta}_i = \text{diag}\{\mu^{(1)}(X_{ip})\}$$

$$\mathbf{V}_i = \mathbf{S}_i^{1/2} \mathbf{R}_i(\boldsymbol{\delta}) \mathbf{S}_i^{1/2}$$

$\mathbf{S}_i$  = diagonal matrix with marginal variances

$\mathbf{R}_i$  = working correlation

- Properties of Parametric GEEs:

- $\widehat{\boldsymbol{\beta}}$  is most efficient when  $\mathbf{R}_i$  = true correlation.

---

## KERNEL GEES

---

- The idea of kernel smoothing is to fit the polynomial locally.

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

- Parametric estimating equation

$$\sum_{i=1}^n \mathbf{G}_{ip}^T \Delta_i \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0,$$

- One (other due to Severini and Staniswalis, 1994, JASA) nonparametric method for any  $x$ :

- **Symmetric Kernel GEE**

$$\sum_{i=1}^n \mathbf{G}_{ip}(x)^T \Delta_i(x) \mathbf{K}_{ih}^{1/2}(x) \mathbf{V}_i^{-1}(x) \mathbf{K}_{ih}^{1/2}(x) \{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\} = 0$$

---

## ASYMPTOTIC RESULTS I: KERNEL GEES

---

- The **theoretical calculations** are neither clever nor pretty.
- Basically, **brute-force** methods are used to derive an **asymptotic expansion** for the estimates.
  - Then special cases are calculated

The **asymptotic expansion** is however extremely useful for the consideration of measurement error.

---

## ASYMPTOTIC RESULTS II: KERNEL GEEs

---

- For **Symmetric Kernel GEE**, 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.**

---

## 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.

- If  $v^{jj}$  is the  $j$ th diagonal element of the inverse of the working covariance matrix  $V$ , and if  $\sigma_j^2$  is the actual variance, if  $M_i \equiv M$  and if the marginal distributions of  $X$  are the same, the variance is

$$\frac{\sum_{j=1}^M \{v^{jj}\}^2 \sigma_j^2}{\{\sum_{j=1}^M v^{jj}\}^2}$$

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

---

## NONPARAMETRIC MODEL: MEASUREMENT ERROR

---

- $i$ =subject  $i$  ( $i = 1, \dots, n$ )
- $j$ =observation  $j$  ( $j = 1, \dots, M$ ), panel data
- $Y_{ij}$ =outcome
- $X_{ij}$ =unobserved covariate (e.g., true log-CD4)
- $W_{ij}$ =Observed error-prone covariate (e.g., observed log-CD4)
- **Model:**

$$g(\mu_{ij}) = \theta(X_{ij}),$$

$$W_{ij} = X_{ij} + U_{ij},$$

- $U_{ij}$  = measurement error
- $\mathbf{U}_i = (U_{i1}, \dots, U_{im_i})^T \sim \text{Normal}(0, \Sigma_{i,uu})$ .
- We want to estimate  $\theta(\cdot)$  **without assuming a distribution for the unobserved  $X_{ij}$ .**

---

## REVIEW OF SIMEX

---

- SIMEX=Simulation Extrapolation
  - Cook and Stefanski, 1994, JASA
- Simple and Robust
  - **Simple to implement:** Does not need any additional measurement error software.
  - **Robust:** Does not require specification of the distribution of  $X$ .
  - **Computationally intensive**
  - Theory based on showing that the SIMEX method is asymptotically equivalent to solving an estimating equation: our work in JASA (1996) and Biometrika (1999).

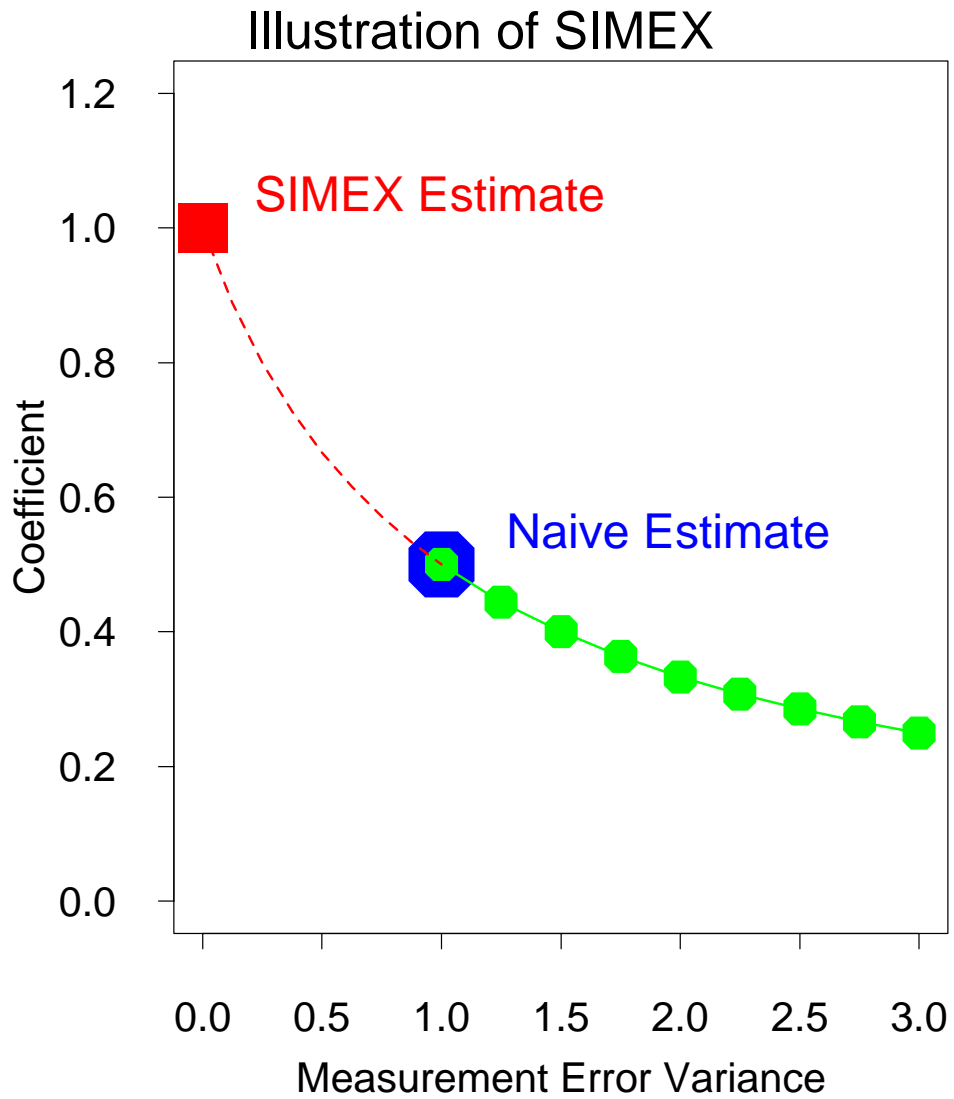


Figure 1: **Now extrapolate to the case of no measurement error.**

---

## THEORETICAL PROPERTIES

---

- Following our previous results, we used working independence to calculate the estimators.
- Panel data: Observations are obtained at the same time points (waves), e.g., every 3 months.
- **Working independence estimator:** (As before)
- **Weighted average estimator:** (new)
  - Calculate standard kernel estimators at each wave and average them using optimal weights (e.g., inverses of the variances).

| Subject | Wave |   |     |     |
|---------|------|---|-----|-----|
|         | 1    | 2 | ... | $M$ |
| 1       | ×    | × | ... | ×   |
| 2       | ×    | × | ... | ×   |
| ⋮       |      |   | ⋮   |     |
| n       | ×    | × | ... | ×   |

---

## ASYMPTOTIC RESULTS I: SIMEX

---

- The **theoretical calculations** are now very pretty.
- We have an expansion for the case of no measurement error.
- SIMEX asymptotic theory (1996, JASA) simply requires such an expansion.
  - Some nice little tricks handle the nonparametric part.

---

## THEORETICAL RESULTS FOR PANEL DATA

---

- **Without measurement error:** Both estimators are asymptotically equivalent.
- **With Measurement error:**
  - It is generally asymptotically superior to fit the function separately at each wave and then combine by a weighted average.
  - Same bias but smaller variance than using all the data simultaneously
  - The same variance occurs only if the marginal distributions of  $(Y, X, W)$  are independent of wave.
  - These results are counterintuitive.
  - **Very different from parametric regression with measurement error.**

---

## CONCLUSIONS

---

- We have discussed nonparametric kernel regression for longitudinal data when the covariate is accurately measured and measured with error.
- When the covariate is accurately measured, it is the best strategy asymptotically to entirely ignore the within-cluster correlation.
- Correctly specifying the correlation structure has adverse effects and results in a less efficient estimator.
- This is dramatically different from parametric GEEs.
- When the covariate is measured with error, the SIMEX method is an easy way to do nonparametric regression accounting for measurement error.
- For panel data with covariate measured with error, pooling data for kernel regression yields a less efficient estimator compared to averaging individual estimators using some weights.

---

## CONCLUSIONS

---

- It is worth re-emphasizing that the results assume that the number of observations per individual is bounded.
- Thus, the results do not apply to time series data, where it is known that accounting for correlation is a good thing.
- The results apply to the nonparametric component of semiparametric models, e.g.

$$E(Y|X, T) = \beta T + \theta(X)$$

- It is not known whether the same results would apply for marginal models fit by splines.