Semiparametric Estimation of Covariance

for Analysis of Longitudinal Data

Jianqing Fan

Princeton University

http://www.princeton.edu/~jqfan

Joint work: Tao Huang and Runze Li

March 12, 2007

Outline

— Introduction

— Estimation of Regression Coefficients

■Profile least-squares; ■Sampling properties.

— Estimating Covariance Function

■kernel method; ■Sampling properties.

correlation: ■pseudo-likelihood; ■variance minimization.

— Numerical results and trajectory projection

Introduction

Covariance matrix: are important for longitudinal data analysis:

- Improve efficiency for regression coefficients

  —**Parametric models**: ♠GMM (Hansen, 1982); ♠GEE (Liang and Zeger, 86) ♠QIF (Qu, Lindsay and Li, 2000).

Introduction

<u>Covariance matrix</u>: are important for longitudinal data analysis:

- Improve efficiency for regression coefficients

—**Parametric models**: ♠GMM (Hansen, 1982); ♠GEE (Liang and Zeger, 86) ♠QIF (Qu, Lindsay and Li, 2000).

—**Nonparametric models**: WI can be improved by

•spline to incorporate the inter-subject correlation (Lin and Carroll, 01);

•innovative two-step kernel method (Wang, 2003; Wang, *et al.*2005);

•a minimax view is offered by Chen, Fan and Jin (2007).

Introduction

Covariance matrix: are important for longitudinal data analysis:

- Improve efficiency for regression coefficients

  —**Parametric models**: ♠GMM (Hansen, 1982); ♠GEE (Liang and Zeger, 86) ♠QIF (Qu, Lindsay and Li, 2000).

  —**Nonparametric models**: WI can be improved by

  •spline to incorporate the inter-subject correlation (Lin and Carroll, 01);

  •innovative two-step kernel method (Wang, 2003; Wang, *et al*.2005);

  •a minimax view is offered by Chen, Fan and Jin (2007).

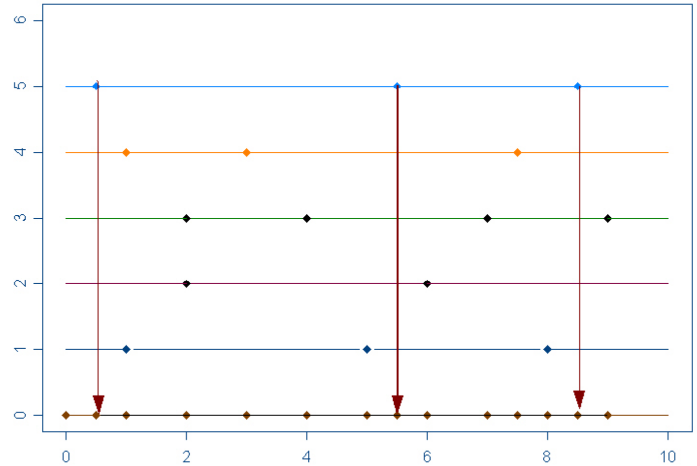
- Predict trajectories of individuals

Challenges

Sampling: Data collected at **irregular and subject-specific** time points.

- Limited information for nonparametric;
- nonnegative definite constraints.

Challenges



Sampling: Data collected at **irregular and subject-specific** time points.

■Limited information for nonparametric;

■nonnegative definite constraints.

Recent advances: Nearly balanced design / parametric model

♠Nonparametric: Wu and Pourahmadi (03).

♠Penalized MLE and Cholesky decomposion: Huang, *et al.*(06).

♠Regularized large covariance and banding (Bickel and Levina, 07).

♠FDA (Yao, Müller and Wang 05a, b)

Semiparametric models

■ Noise processes have covariance structure:

$$\text{var}\{\varepsilon(\mathbf{t})\} = \sigma^{\mathbf{2}}(\mathbf{t}) \quad \text{and} \quad \text{corr}\{\varepsilon(\mathbf{t}), \varepsilon(\mathbf{s})\} = \rho(\mathbf{t}, \mathbf{s}; \boldsymbol{\theta})$$

★always positive definite; ★irregular designs; ★well-approximated.

Semiparametric models

■Noise processes have covariance structure:

$$\mathrm{var}\{\varepsilon(\mathbf{t})\} = \sigma^2(\mathbf{t}) \quad \text{and} \quad \mathrm{corr}\{\varepsilon(\mathbf{t}), \varepsilon(\mathbf{s})\} = \rho(\mathbf{t}, \mathbf{s}; \boldsymbol{\theta})$$

★always positive definite; ★irregular designs; ★well-approximated.

Examples: ♠ARMA models; ♠Factor (random effects) models.

General strategy:

—Embed working correlation $\rho_0(s,t)$ into $\rho(s,t,\boldsymbol{\theta})$.

—Improve efficiency even when $\rho(s,t,\boldsymbol{\theta})$ is **wrong**.

Outline

— Introduction

— **Estimation of Regression Coefficients**

■Profile least-squares; ■Sampling properties.

— Estimating Covariance Function

■kernel method; ■Sampling properties.

correlation: ■pseudo-likelihood; ■variance minimization.

— Numerical results and trajectory projection

Varying-coefficient partially linear model

$$\mathbf{y(t)} = \mathbf{x(t)}^{\mathbf{T}}\boldsymbol{\alpha}\mathbf{(t)} + \mathbf{z(t)}^{\mathbf{T}}\boldsymbol{\beta} + \varepsilon\mathbf{(t)},$$

■ $\boldsymbol{\alpha}(t)$: p smooth functions ■ $\boldsymbol{\beta}$: q parameters.

cross-sectional: Zhang, Lee and Song (2002) and Fan and Huang (2005)

longitudinal: Scheike and Martinussen (2002), Sun and Wu (2005).

Varying-coefficient partially linear model

$$\mathbf{y}(\mathbf{t}) = \mathbf{x}(\mathbf{t})^{\mathbf{T}}\boldsymbol{\alpha}(\mathbf{t}) + \mathbf{z}(\mathbf{t})^{\mathbf{T}}\boldsymbol{\beta} + \varepsilon(\mathbf{t}),$$

■$\boldsymbol{\alpha}(t)$: p smooth functions ■$\boldsymbol{\beta}$: q parameters.

cross-sectional: Zhang, Lee and Song (2002) and Fan and Huang (2005)

longitudinal: Scheike and Martinussen (2002), Sun and Wu (2005).

—**partially linear models** (Wahba, 1984; Engle, et al. 1984, Heckman, 1986; Speckman, 1988; ...; Härdle, Liang and Gao, 2000),

—**Varying-coefficient models** and **functional linear models** (Hastie and Tibshirani, 1993, Works by Wu, Rice, Fan, Huang, Müller,...).

—**semiparametric models** studied by Lin and Carroll (2001) (with identify link), Wang, Carroll and Lin (2005), and Huang and Zhang (2004).

Estimation of Regression Coefficients

Profile LS: Let $y^*(t) = y(t) - \mathbf{z}(t)^T \boldsymbol{\beta}$. Then,

$$y^*(t) = \mathbf{x}(t)^T \boldsymbol{\alpha}(t) + \varepsilon,$$

■Use a (local) linear smoother to estimate $\boldsymbol{\alpha}(t)$.

■Plug-in $\hat{\boldsymbol{\alpha}}(\cdot)$ and obtain **synthetic** linear model:

$$(I - \mathbf{S})\mathbf{y} = (I - \mathbf{S})\mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Z}$: the design matrix for $\mathbf{z}_i(t_{ij})$, and

$\mathbf{S}$: smoothing matrix depends only on t_{ij} and $\mathbf{x}_i(t_{ij})$

■Apply **profile** weighted LS with a weight matrix $\mathbf{W}$:

$$\hat{\boldsymbol{\beta}} = \{\mathbf{Z}^T(I-\mathbf{S})^T\mathbf{W}(I-\mathbf{S})\mathbf{Z}\}^{-1}\mathbf{Z}^T(I-\mathbf{S})^T\mathbf{W}(I-\mathbf{S})\mathbf{y}.$$

Covariance matrix: With $\mathbf{D} = \mathbf{Z}^T(I-\mathbf{S})^T\mathbf{W}(I-\mathbf{S})\mathbf{Z}$ and $\mathbf{V} = \text{cov}\{\mathbf{Z}^T(I-\mathbf{S})^T\mathbf{W}\boldsymbol{\varepsilon}\}$,

$$\text{cov}\{\hat{\boldsymbol{\beta}}|\mathbf{t_{ij}}, \mathbf{x_i}(\mathbf{t_{ij}}), \mathbf{z_i}(\mathbf{t_{ij}})\} = \mathbf{D}^{-1}\mathbf{V}\mathbf{D}^{-1} \hat{=} \boldsymbol{\Gamma}(\sigma^2, \boldsymbol{\theta}),$$

Efficiency of $\hat{\beta}$ depends on $\mathbf{W}$ $= \text{diag}\{\mathbf{W}_1, \cdots, \mathbf{W}_n\}$,

Sampling assumption

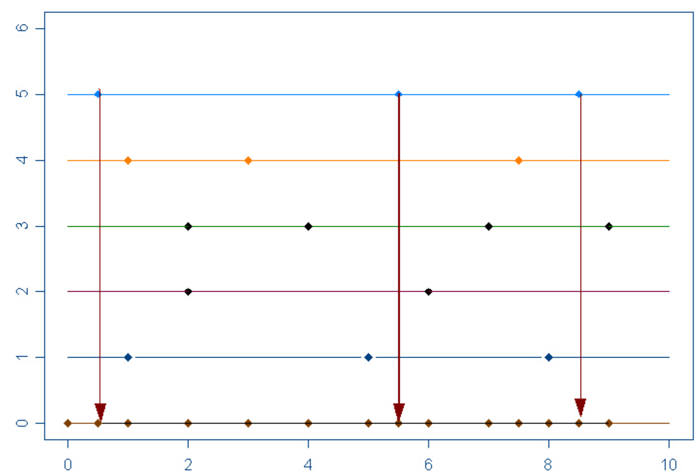
Data: a sample from a process $\{y(t), \mathbf{x}(t), \mathbf{z}(t)\}, t \in [0, T]$.

Sampling points: Assume that J_i, $i = 1, \cdots, n$ are iid with $0 < E(J_i) < \infty$, and for given J_i, t_{ij}, $j = 1, \cdots, J_i$ are iid according to a density $f(t)$.

Counting process:

Lin and Ying (01);

see Fan and Li (04)



Sampling properties

Theorem 1: We have asymptotic representation,

$$\sqrt{\mathbf{n}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{0}}) = \sqrt{\mathbf{n}}\boldsymbol{\Sigma}_{\mathbf{n}}^{-\mathbf{1}}\xi_{\mathbf{n}} + \mathbf{o}_{\mathbf{P}}(\mathbf{1}),$$

—$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}\{\mathbf{Z}_i - \tilde{\mathbf{X}}_i\}^T \mathbf{W}_i \{\mathbf{Z}_i - \tilde{\mathbf{X}}_i\}$,

—$\xi_n = \frac{1}{n}\sum_{i=1}^{n}\{\mathbf{Z}_i - \tilde{\mathbf{X}}_i\}^T \mathbf{W}_i \varepsilon_i$,

Sampling properties

Theorem 1: We have asymptotic representation,

$$\sqrt{\mathbf{n}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{0}}) = \sqrt{\mathbf{n}}\boldsymbol{\Sigma}_{\mathbf{n}}^{-\mathbf{1}}\xi_{\mathbf{n}} + \mathbf{o}_{\mathbf{P}}(\mathbf{1}),$$

—$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}\{\mathbf{Z}_i - \tilde{\mathbf{X}}_i\}^T \mathbf{W}_i \{\mathbf{Z}_i - \tilde{\mathbf{X}}_i\}$,

—$\xi_n = \frac{1}{n}\sum_{i=1}^{n}\{\mathbf{Z}_i - \tilde{\mathbf{X}}_i\}^T \mathbf{W}_i \boldsymbol{\varepsilon}_i$,

—$\boldsymbol{\varepsilon}_i = (\varepsilon_i(t_{i1}), \cdots, \varepsilon_i(t_{iJ_i}))^T$,

—$\tilde{\mathbf{X}}_i = (\Psi(t_{i1})\Gamma^{-1}(t_{i1})\mathbf{x}_i(t_{i1}), \cdots, \Psi(t_{iJ_i})\Gamma^{-1}(t_{iJ_i})\mathbf{x}_i(t_{iJ_i}))^T$.

—$\Gamma(t) = E\mathbf{x}(t)\mathbf{x}^T(t)$, $\Psi(t) = E\mathbf{x}(t)\mathbf{z}^T(t)$.

Asymptotic Normality — Parametric Part

$$\sqrt{\mathbf{n}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{0}}) \xrightarrow{\mathcal{D}} \mathbf{N}(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}),$$

— $\mathbf{A} = E\{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}^T \mathbf{W}_1 \{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\},$

— $\mathbf{B} = E\{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}^T \mathbf{W}_1 \boldsymbol{\varepsilon}_1 \boldsymbol{\varepsilon}_1^T \mathbf{W}_1 \{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}.$

Asymptotic Normality — Parametric Part

$$\sqrt{\mathbf{n}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{0}}) \xrightarrow{\mathcal{D}} \mathbf{N}(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}),$$

—$\mathbf{A} = E\{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}^T \mathbf{W}_1 \{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}$,

—$\mathbf{B} = E\{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}^T \mathbf{W}_1 \boldsymbol{\varepsilon}_1 \boldsymbol{\varepsilon}_1^T \mathbf{W}_1 \{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}$.

■If $\mathbf{W}_i = \text{cov}^{-1}\{\boldsymbol{\varepsilon}_i | \mathbf{x}_i(t_{ij}), \mathbf{z}_i(t_{ij})\}$, then $\mathbf{A} = \mathbf{B}$. Asymp var is $\mathbf{B}_0^{-1}$

$$\mathbf{B}_0 = E\{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}^T \text{cov}^{-1}(\boldsymbol{\varepsilon}_1 | \mathbf{X}_1, \mathbf{Z}_1)\{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}.$$

—**Most efficient** estimate among profile WLSE.

Asymptotic Normality — Parametric Part

$$\sqrt{\mathbf{n}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{0}}) \xrightarrow{\mathcal{D}} \mathbf{N}(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}),$$

—$\mathbf{A} = E\{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}^T \mathbf{W}_1 \{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}$,

—$\mathbf{B} = E\{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}^T \mathbf{W}_1 \boldsymbol{\varepsilon}_1 \boldsymbol{\varepsilon}_1^T \mathbf{W}_1 \{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}$.

■If $\mathbf{W}_i = \text{cov}^{-1}\{\boldsymbol{\varepsilon}_i | \mathbf{x}_i(t_{ij}), \mathbf{z}_i(t_{ij})\}$, then $\mathbf{A} = \mathbf{B}$. Asymp var is $\mathbf{B}_0^{-1}$

$$\mathbf{B}_0 = E\{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}^T \text{cov}^{-1}(\boldsymbol{\varepsilon}_1 | \mathbf{X}_1, \mathbf{Z}_1)\{\mathbf{Z}_1 - \tilde{\mathbf{X}}_1\}.$$

—**Most efficient** estimate among profile WLSE.

■Working independent: $\mathbf{W}$ is diagonal.

—$\hat{\boldsymbol{\beta}}$ is still root n consistent.

Asymptotic Normality — Nonparametric Part

Substituting $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ $\Longrightarrow$ an estimate for $\boldsymbol{\alpha}(t)$

Theorem 2. If $nh^5 = O(1)$, then

$$\sqrt{nh}(\hat{\boldsymbol{\alpha}}(t) - \boldsymbol{\alpha}(t) - \frac{1}{2}\mu_2 h^2 \ddot{\boldsymbol{\alpha}}(t)) \xrightarrow{\mathcal{D}} N(0, \frac{\nu_0}{f(t)E(J_1)}\sigma^2(t)\Gamma^{-1}(t)).$$

where $\mu_i = \int u^i K(u)du$, and $\nu_i = \int u^i K^2(u)du$.

Asymptotic Normality — Nonparametric Part

Substituting $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}} \Longrightarrow$ an estimate for $\boldsymbol{\alpha}(t)$

Theorem 2. If $nh^5 = O(1)$, then

$$\sqrt{nh}(\hat{\boldsymbol{\alpha}}(t) - \boldsymbol{\alpha}(t) - \frac{1}{2}\mu_2 h^2 \ddot{\boldsymbol{\alpha}}(t)) \xrightarrow{\mathcal{D}} N(0, \frac{\nu_0}{f(t)E(J_1)}\sigma^2(t)\Gamma^{-1}(t)).$$

where $\mu_i = \int u^i K(u) du$, and $\nu_i = \int u^i K^2(u) du$.

The bias and variance of $\hat{\boldsymbol{\alpha}}(t)$ do not depend on **W**, since

- the root n consistency of $\hat{\boldsymbol{\beta}}$ does not depend on **W**
- the estimator is intrinsically local (Lin and Carroll, 2000).

Outline

— Introduction

— Estimation of Regression Coefficients

■Profile least-squares; ■Sampling properties.

— **Estimating Covariance Function**

■kernel method; ■Sampling properties.

correlation: ■pseudo-likelihood; ■variance minimization.

— Numerical results and trajectory projection

Estimation of Covariance Function

Residuals: $\mathbf{r}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (r_{i1}, \cdots, r_{iJ_i})^T$ with

$$r_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = y_i(t_{ij}) - \mathbf{x}_i(t_{ij})^T \boldsymbol{\alpha}(t_{ij}) - \mathbf{z}_i(t_{ij})^T \boldsymbol{\beta},$$

Pseudo-likelihood: Pretending $\boldsymbol{\varepsilon}_i \sim N(0, \Sigma_i)$, then,

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = -\frac{1}{2} \sum_{i=1}^{n} \log |\boldsymbol{\Sigma}_i| - \frac{1}{2} \sum_{i=1}^{n} \mathbf{r}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{r}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}).$$

Estimation of Covariance Function

Residuals: $\mathbf{r}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (r_{i1}, \cdots, r_{iJ_i})^T$ with

$$r_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = y_i(t_{ij}) - \mathbf{x}_i(t_{ij})^T \boldsymbol{\alpha}(t_{ij}) - \mathbf{z}_i(t_{ij})^T \boldsymbol{\beta},$$

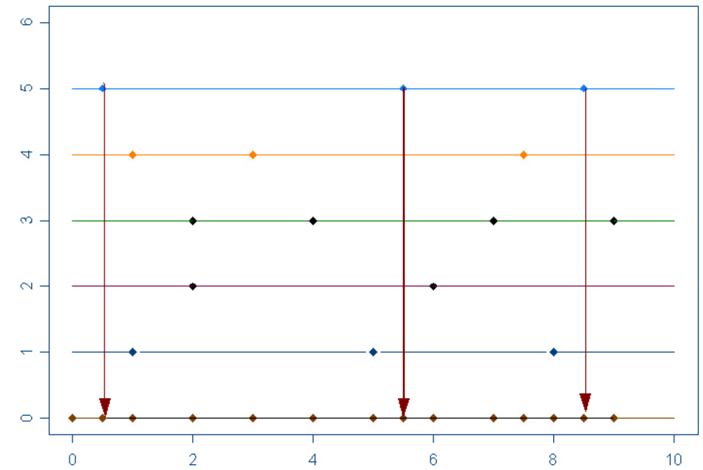
Pseudo-likelihood: Pretending $\boldsymbol{\varepsilon}_i \sim N(0, \Sigma_i)$, then,

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{n} \log |\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{n} \mathbf{r}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{r}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}).$$

■ Iterate between estimation of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $(\sigma^2, \boldsymbol{\theta})$.

Estimation of $\sigma^2(t)$

Residuals: Let $\hat{r}_{ij} = r_{ij}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$.



Kernel estimator: Since $\sigma^2(t_{ij}) = E\{\varepsilon^2(t)|t = t_{ij}\}$,

$$\hat{\sigma}^2(\mathbf{t}) = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \sum_{\mathbf{j=1}}^{\mathbf{J_i}} \hat{\mathbf{r}}_{\mathbf{ij}}^{\mathbf{2}} \mathbf{K}_{\mathbf{h_1}}(\mathbf{t} - \mathbf{t_{ij}})}{\sum_{\mathbf{i=1}}^{\mathbf{n}} \sum_{\mathbf{j=1}}^{\mathbf{J_i}} \mathbf{K}_{\mathbf{h_1}}(\mathbf{t} - \mathbf{t_{ij}})}.$$

where $K_{h_1}(x) = {h_1}^{-1} K(x/h_1)$ with a kernel K and a bandwidth h_1. (Ruppert, et al. 1997, Fan & Yao, 1998).

Asymptotic Properties

Theorem 3. If $c < nh_1^5 < C$, and $c < h/h_1 < C$, then

$$\sqrt{nh_1}(\hat{\sigma}^2(t) - \sigma^2(t) - b(t)) \xrightarrow{\mathcal{D}} N(0, v(t)).$$

$$b(t) = \frac{h_1^2}{2}\left\{\ddot{\sigma}^2(t) + \frac{2\dot{\sigma}^2(t)\dot{f}(t)}{f(t)}\right\}\mu_2 \text{ and } \quad v(t) = \frac{\mathrm{var}\{\varepsilon^2(t)\}\nu_0}{f(t)E(J_1)}.$$

Asymptotic Properties

Theorem 3. If $c < nh_1^5 < C$, and $c < h/h_1 < C$, then

$$\sqrt{nh_1}(\hat{\sigma}^2(t) - \sigma^2(t) - b(t)) \xrightarrow{\mathcal{D}} N(0, v(t)).$$

$$b(t) = \frac{h_1^2}{2}\left\{\ddot{\sigma}^2(t) + \frac{2\dot{\sigma}^2(t)\dot{f}(t)}{f(t)}\right\}\mu_2 \text{ and } \quad v(t) = \frac{\text{var}\{\varepsilon^2(t)\}\nu_0}{f(t)E(J_1)}.$$

■ The asymptotic bias and variance do not depend on **W**.

$\Longrightarrow$ Use the residuals w/ working indep to estimate $\sigma^2(t)$.

■ Consistent with our empirical experience

Estimation of correlation function

Challenges: •bivariate functions; •positive definite.

Correlation: Parametric form $\rho(s, t, \boldsymbol{\theta})$.

Covariance: Semiparametric: $\Sigma_i = V_i C_i(\boldsymbol{\theta}) V_i$.

—$V_i = \text{diag}\{\sigma(t_{i1}), \cdots, \sigma(t_{iJ_i})\}$, $C_i(\boldsymbol{\theta}) = (\rho(t_{ik}, t_{il}, \boldsymbol{\theta}))_{J_i \times J_i}$.

Estimation of correlation function

Challenges: •bivariate functions; •positive definite.

Correlation: Parametric form $\rho(s, t, \boldsymbol{\theta})$.

Covariance: Semiparametric: $\Sigma_i = V_i C_i(\boldsymbol{\theta}) V_i$.

—$V_i = \text{diag}\{\sigma(t_{i1}), \cdots, \sigma(t_{iJ_i})\}$, $C_i(\boldsymbol{\theta}) = (\rho(t_{ik}, t_{il}, \boldsymbol{\theta}))_{J_i \times J_i}$.

Specifications: Embed the working correlation $\rho_0(s, t; \boldsymbol{\theta}_0)$ into the family of convex combinations:

$$\rho(s, t; \boldsymbol{\theta}) = \tau_{\mathbf{0}} \rho_{\mathbf{0}}(\mathbf{s}, \mathbf{t}; \boldsymbol{\theta}_{\mathbf{0}}) + \tau_{\mathbf{1}} \rho_{\mathbf{1}}(\mathbf{s}, \mathbf{t}) + \cdots + \tau_{\mathbf{m}} \rho_{\mathbf{m}}(\mathbf{s}, \mathbf{t}).$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_0, \cdots, \boldsymbol{\theta}_m, \tau_0, \cdots, \tau_m\}$, and $\tau_0 + \cdots + \tau_m = 1$.

Optimizing $\boldsymbol{\theta}$ always improves the efficiency of $\boldsymbol{\beta}$.

Variance minimization

Aim: Choose $\boldsymbol{\theta}$ to minimize $\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = \Gamma(\hat{\sigma}^2, \boldsymbol{\theta})$.

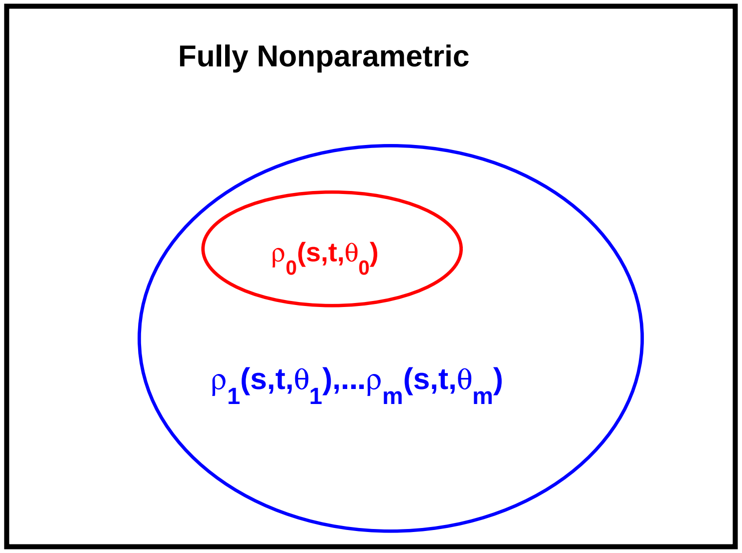
$\Longrightarrow$ **Improve the efficiency** of $\text{var}(\boldsymbol{\beta})$ even when **misspecified**.

Minimal generalized variance:

Choose $\boldsymbol{\theta}$ to minimize

$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} |\Gamma(\hat{\sigma}^2, \boldsymbol{\theta})|$,

the volume of the confidence set:



$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \Gamma^{-1}(\hat{\sigma}^2, \boldsymbol{\theta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) < c.$$

Quasi Maximum Likelihood

Maximize: $\ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, namely

$$-\frac{1}{2}\sum_{i=1}^{n}\left\{\log|C_i(\boldsymbol{\theta})| + \hat{\mathbf{r}}_i^T \hat{V}_i^{-1} C_i^{-1}(\boldsymbol{\theta}) \hat{V}_i^{-1} \hat{\mathbf{r}}_i\right\}$$

where $\hat{V}_i = \text{diag}\{\hat{\sigma}(t_{i1}), \cdots, \hat{\sigma}(t_{iJ_i})\}$, and $\hat{\mathbf{r}}_i = (\hat{r}_{i1}, \cdots, \hat{r}_{iJ_i})^T$.

Quasi Maximum Likelihood

Maximize: $\ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, namely

$$-\frac{1}{2}\sum_{i=1}^{n}\left\{\log|C_i(\boldsymbol{\theta})| + \hat{\mathbf{r}}_i^T \hat{V}_i^{-1} C_i^{-1}(\boldsymbol{\theta}) \hat{V}_i^{-1} \hat{\mathbf{r}}_i\right\}$$

where $\hat{V}_i = \text{diag}\{\hat{\sigma}(t_{i1}), \cdots, \hat{\sigma}(t_{iJ_i})\}$, and $\hat{\mathbf{r}}_i = (\hat{r}_{i1}, \cdots, \hat{r}_{iJ_i})^T$.

■When $\rho(s, t, \boldsymbol{\theta})$ is correctly specified, the QL may provide a good estimate for $\boldsymbol{\theta}$.

■When incorrectly specified, it improves efficiency for $\boldsymbol{\beta}$ using a different criterion.

Outline

— Introduction

— Estimation of Regression Coefficients

■Profile least-squares; ■Sampling properties.

— Estimating Covariance Function

■kernel method; ■Sampling properties.

correlation: ■pseudo-likelihood; ■variance minimization.

— **Numerical results and trajectory projection**

Simulation Studies

Model: $y(t) = \mathbf{x}(t)^T \boldsymbol{\alpha}(t) + \mathbf{z}(t)^T \boldsymbol{\beta} + \varepsilon(t)$.

Sample size: $n = 50$

Observation times t_{ij}: Each individual has a set of 'scheduled' time points, **{0,1,2,...,12}**, each having **20%** chance being skipped. The observation times are random perturbations of the scheduled times.

Simulation Studies

Model: $y(t) = \mathbf{x}(t)^T \boldsymbol{\alpha}(t) + \mathbf{z}(t)^T \boldsymbol{\beta} + \varepsilon(t)$.

Sample size: $n = 50$

Observation times t_{ij}: Each individual has a set of 'scheduled' time points, **{0,1,2,...,12}**, each having **20%** chance being skipped. The observation times are random perturbations of the scheduled times.

Covariates: ★$x_1(t) \equiv 1$ — intercept.

★$(x_2(t), z_1(t))^T$: bivariate normal with $\rho = 0.5$.

★$z_2(t)$: Bernouli with success prob 0.5, indep. of $(x_2(t), z_1(t))$.

Model specifications

Coefficients: Parametric component: $\boldsymbol{\beta} = (1, 2)^T$.

Nonparametric: $\alpha_1(t) = \sqrt{t/12}$, and $\alpha_2(t) = \sin(2\pi t/12)$.

Model specifications

Coefficients: Parametric component: $\boldsymbol{\beta} = (1, 2)^T$.

Nonparametric: $\alpha_1(t) = \sqrt{t/12}$, and $\alpha_2(t) = \sin(2\pi t/12)$.

Error process $\varepsilon(t)$: a Gaussian process with zero mean,

$$\sigma^2(t) = 0.5 \exp(t/12), \quad \text{and} \quad \text{corr}(\varepsilon(s), \varepsilon(t)) = \gamma \rho^{|t-s|}$$

for $s \neq t$ with $(\gamma, \rho) = (0.85, 0.9), (0.85, 0.6)$ and (0.85, 0.3).

Number of Simulation: 1000 for each case.

Performance of parametric components β

Strong correlation:

Correct specification of $\rho(s,t,\boldsymbol{\theta})$ $((\gamma,\rho) = (0.85, \mathbf{0.9}))$

Method	SD	$\overline{\text{Bias}}$	MAD	Median(Bias)
Indep.	47.780	-1.9730	30.066	-1.2802
True	25.061	-1.2565	17.473	-0.7676
QL	25.156	-1.2545	17.224	-0.7709
MGV	25.205	-1.2040	17.250	-0.9126

★for $\hat{\beta}_1$; ★Values multiplied by 1000.

Efficiency gain: $(30.066/17.224)^2 \approx 3.$

Performance of Parametric Component β

Weak correlation:

Correct specification of $\rho(s, t, \boldsymbol{\theta})$ $((\gamma, \rho) = (0.85, 0.3))$

Method	SD	$\overline{\text{Bias}}$	MAD	Median(Bias)
Indep.	46.991	-2.8990	32.010	-1.6817
True	40.123	-1.9687	27.104	-2.1143
QL	**95.506**	-6.7632	28.222	-1.9187
MGV	40.389	-1.6740	27.442	-1.4153

★for $\hat{\beta}_1$; ★Values multiplied by 1000.

Efficiency gain: 36%. MGV is more robust QL.

Effect of misspecification of correlection structure

True: AMRA(1,1) with $(\gamma, \rho) = (0.85, 0.9)$. **Working**: AR(1)

	Method	SD	$\overline{\text{Bias}}$	MAD	Med(Bias)
Optimization	Indep.	47.780	-1.9730	30.066	-1.2802
Algorithm	QL	31.857	-0.4859	19.975	-0.0837
	MGV	33.121	-0.5275	21.449	0.0535
Grid	QL	31.939	-0.4489	19.792	-0.0714
Search	MGV	33.293	-0.5232	21.218	-0.2297

efficiency gain: $(30.066/19.975)^2 \approx 2.3$

Grid search: ρ over $\{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$.

Performance of nonparametric part: $\hat{\alpha}(t)$

RASE: Given grid points $\{t_g, : g = 1, \cdots, 200\}$, define

$$\text{RASE}\{\hat{\alpha}_j(\cdot)\} = \left[\frac{1}{G}\sum_{g=1}^{G}\{\hat{\alpha}_j(t_g) - \alpha_j(t_g)\}^2\right]^{1/2}.$$

Performance of $\hat{\boldsymbol{\alpha}}(\cdot)$

Correlation structure	$\hat{\alpha}_1(\cdot)$	$\hat{\alpha}_2(\cdot)$
Independence	0.1340(0.0545)	0.1168(0.0324)
QL with ARMA(1,1)	0.1328(0.0517)	0.1153(0.0319)
QL with AR(1)	0.1618(0.0598)	0.1270(0.0360)

Performance of variance: $\hat{\sigma}^2(t)$

RASEs for $\hat{\sigma}^2(t)$

	Scenario I: Independence		Scenario II: Oracle	
Bandwidth	Mean	Standard Error	Mean	Standard Error
1	**0.0886**	0.0555	**0.0899**	0.0606
1.5	**0.0809**	0.0561	**0.0834**	0.0620
2.25	**0.0777**	0.0577	**0.0815**	0.0631

Independence: Use working indep correlation to estimate $(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Oracle: Use the true value of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Real data example

Data: ★Multi-Center AIDS Cohort study; ★283 homosexual men infected with HIV during the period: 1984-1991.

Variable: —$y(t)$: CD4 cell percentage;

—X_1: PreCD4 cell percentage (standardized);

—Z_1: smoking status. —Z_2: age at infection (standardized).

Model: $y(t) = \alpha_1(t) + \alpha_2(t)X_1 + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon(t)$.

Bandwidth selection: Q-fold cross-validation

$$CV(h) = \sum_{k=1}^{Q} \sum_{i \in d_k} \sum_{j=1}^{J_i} \{y_i(t_{ij}) - \hat{y}_{-d_k}(t_{ij})\}^2,$$

where $\hat{y}_{-d_k}(t_{ij})$ is a fitted value for the i-subject at observed time t_{ij} with the data in d_k deleted, and $Q = 15$.

Result: $h = 18.1710$, $\approx$ 30% of range of t_{ij}'s.

Bandwidth selection: Q-fold cross-validation

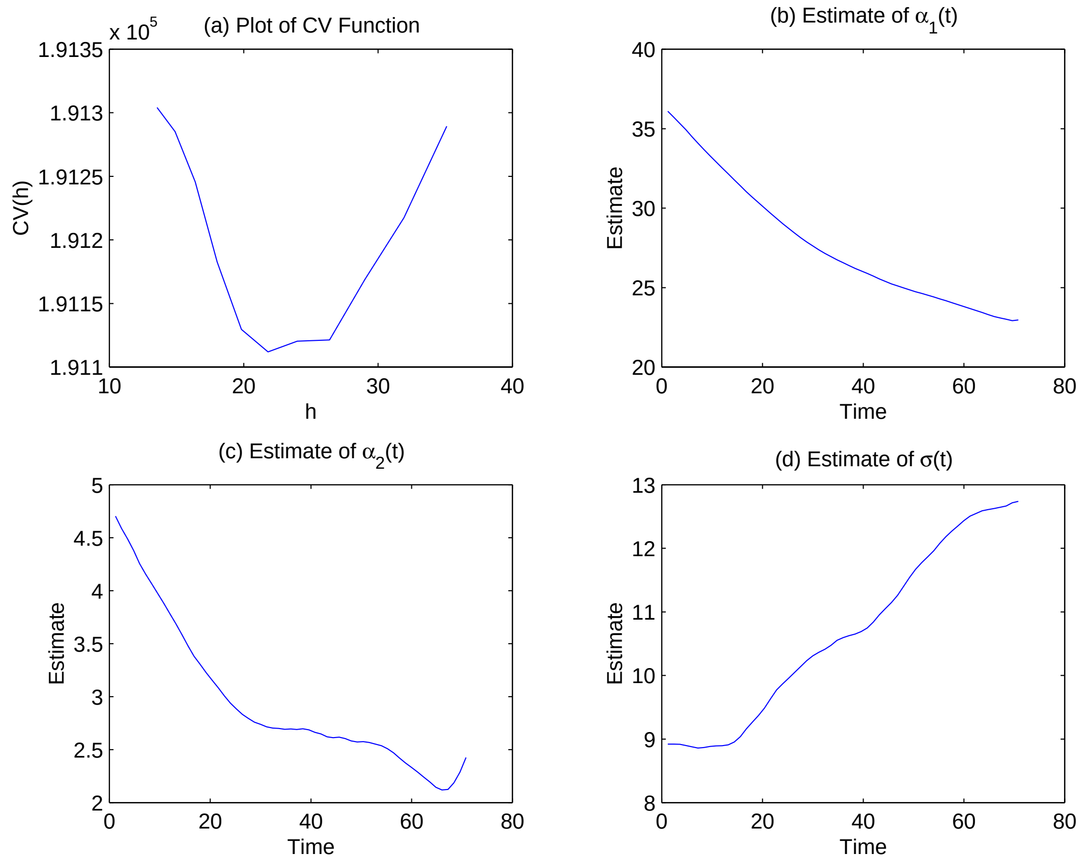
$$CV(h) = \sum_{k=1}^{Q} \sum_{i \in d_k} \sum_{j=1}^{J_i} \{y_i(t_{ij}) - \hat{y}_{-d_k}(t_{ij})\}^2,$$

where $\hat{y}_{-d_k}(t_{ij})$ is a fitted value for the i-subject at observed time t_{ij} with the data in d_k deleted, and $Q = 15$.

Result: $h = 18.1710$, $\approx$ 30% of range of t_{ij}'s.

Estimation of $\sigma^2(t)$: The bandwidth can be easily chosen with $h_1 = 10.2587$ by using plug-in method (Ruppert, Sheather and Wand, 1995).

Estimates:



Estimates of (γ, ρ) and $\boldsymbol{\beta}$

	Independence	QL	MGV
$\hat{\beta}_1$	0.8726(1.1545)	0.6772(0.9970)	0.6302(1.0860)
$\hat{\beta}_2$	-0.5143(0.6110)	0.0567(0.4716)	-0.3647(0.5484)

Prediction of individual trajectory

Individual data: collected at $t = t_1, \cdots, t_J$.

Aim: To predict $y(t)$ at $t = t^*$ with covariates $\mathbf{x}(t^*)$ and $\mathbf{z}(t^*)$.

Notation: —$\mathbf{y}_o = (y(t_1), \cdots, y(t_J))^T$

—$\boldsymbol{\mu} = (\mu(t_1), \cdots, \mu(t_J))^T$ with $\mu(t) = \mathbf{x}(t)^T\boldsymbol{\alpha}(t) + \mathbf{z}(t)^T\boldsymbol{\beta}$

—$\Sigma = \text{cov}\{(\varepsilon(t_1), \cdots, \varepsilon(t_J))^T\}$

—$\mathbf{c}^* = (c(t_1, t^*), \cdots, c(t_J, t^*))^T$.

Assumption: Joint normal.

Prediction formula

$$E\{y(t^*)|\mathbf{y}_o\} = \mu(\mathbf{t}^*) + \mathbf{c}^{*\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y_o} - \boldsymbol{\mu}) \equiv \hat{y}(t^*)$$

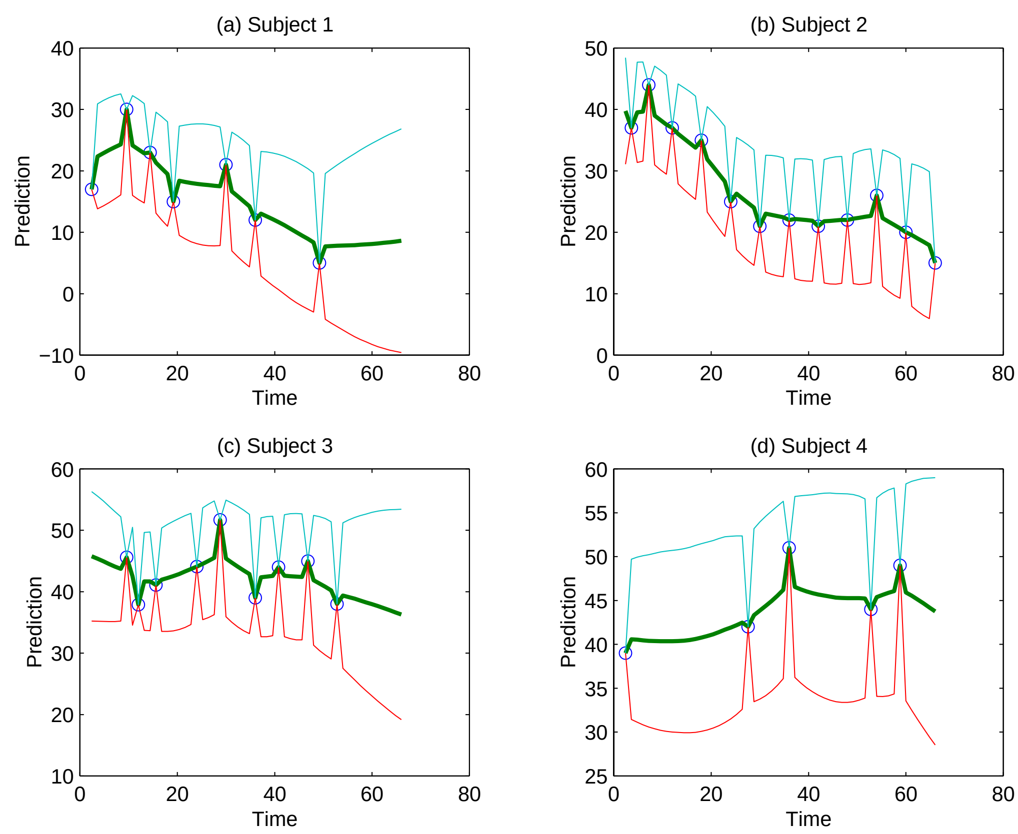
$$\text{var}\{y(t^*)|\mathbf{y}_o\} = \sigma^2(t^*) - \mathbf{c}^{*T}\Sigma^{-1}\mathbf{c}^* \equiv \hat{\sigma}(t^*).$$

Predictive interval:

$\hat{y}(t^*) \pm z_{1-\alpha/2}\hat{\sigma}^2(t^*)$.

If t^* is an observed time point,

the prediction error is zero.



Summary

Model: $y(t) = \mathbf{x}(t)^T \boldsymbol{\alpha}(t) + \mathbf{z}(t)^T \boldsymbol{\beta} + \varepsilon(t)$.

■Semiparametric covariance $\sigma(s)\sigma(t)\rho(s, t; \theta)$

to facilitate longitudinal data structure.

Summary

Model: $y(t) = \mathbf{x}(t)^T \boldsymbol{\alpha}(t) + \mathbf{z}(t)^T \boldsymbol{\beta} + \varepsilon(t)$.

■Semiparametric covariance $\sigma(s)\sigma(t)\rho(s,t;\theta)$

to facilitate longitudinal data structure.

★Profile WLS for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}(\cdot)$

★Kernel estimator for $\sigma^2(t)$

★QL and MGV approaches for $\boldsymbol{\theta}$

Extensions

Fan and Wu (2007) established

— difference-based estimator for $\boldsymbol{\beta}$.

- smoothness of $\boldsymbol{\alpha}(\cdot)$ with degree κ;
- no bandwidth selection involved;
- rates $O(n^{-\kappa} + n^{-1/2})$.

Extensions

Fan and Wu (2007) established

— difference-based estimator for $\boldsymbol{\beta}$.

- smoothness of $\boldsymbol{\alpha}(\cdot)$ with degree κ;
- no bandwidth selection involved;
- rates $O(n^{-\kappa} + n^{-1/2})$.

—
- rates of convergence for $\boldsymbol{\alpha}(\cdot)$;
- asymptotic normality of $\sigma(\cdot)$

Extensions

Fan and Wu (2007) established

— difference-based estimator for $\boldsymbol{\beta}$.

■ smoothness of $\boldsymbol{\alpha}(\cdot)$ with degree κ;

■ no bandwidth selection involved;

■ rates $O(n^{-\kappa} + n^{-1/2})$.

— ■ rates of convergence for $\boldsymbol{\alpha}(\cdot)$;

■ asymptotic normality of $\sigma(\cdot)$

— asymptotic normality of $\boldsymbol{\theta}$ in the correlation matrix.

Thank you!