Closed-form estimation of panels with attrition and refreshment

Grigory Franguridi
University of Southern California
Center for Economic and Social Research

Lidia Kosenkova
University of Virginia
Department of Economics

DSE 2024
August 7

Attrition and refreshment

Nonrandom attrition in panel data is well-documented:
Rubin (1976), Hausman & Wise (1979), Fitzgerald, Gottschalk & Moffitt (1998) and others

Refreshment samples to reduce attrition bias:
Kish & Hess (1959), Wissen & Meurs (1989), Ridder (1992), Lin & Shaeffer (1995), Bhattacharya (2008)

Additively nonignorable attrition:
Hirano, Imbens, Ridder & Rubin (2001): identification $T = 2$
Deng, Hillygus, Reiter, Si & Zheng (2013): review
Si, Reiter & Hillygus (2015): Bayesian approach
Chen, Felt & Hyunh (2017): payment innovations and cash usage
Sadinle & Reiter (2019): general missingness patterns
Hoonhout & Ridder (2019): identification $T > 2$

Alternative identification assumptions: Nevo (2003)

Results

- New identification assumption
- Nonparametric approach **without tuning parameters**
- Closed-form "plug-in" estimator of the parameter defined by moment conditions
- Consistency, inference
- Nonparametric bootstrap
- Monte Carlo simulations
- Empirical illustration

Outline

Framework

Examples

Identification

Estimation

Asymptotics

MC

Conclusion

Framework

Panel $Z_t = (Y_t, X_t) \in R^{d_t}$ over $T = 2$ periods

Attrition in period 2: stay if $W = 1$

Data:

- Period 1: Z_1
- Period 2: $Z_2|W = 1$ and *refreshment* Z_2^r (independent sample)

Put differently, we have access to

- balanced panel $(Z_1, Z_2)|W = 1$ (notation: CDF F^w)
- period marginals Z_1 and Z_2^r (notation: CDFs F_1 and F_2, resp.)

Target parameter θ_0 satisfying

$$Em(Z_1, Z_2; \theta_0) = \int m(z_1, z_2; \theta_0)\, dF(z_1, z_2) = 0,$$

where $F(z_1, z_2)$ is the full-panel (unselected) CDF

Example 1: linear regression with two-way fixed effects

$$y_{it} = \alpha_i + f_t + x_{it}'\beta + \varepsilon_{it}$$

$(\alpha_i, y_{i1}, x_{i1}, y_{i2}, x_{i2})_{i=1}^n \sim \text{IID}$,
allow **arbitrary correlation** between α_i and x_{it} ("fixed effects")

Drop index i:

$$y_t = \alpha + f_t + x_t'\beta + \varepsilon_t$$

Within transform in population:

$$\ddot{\zeta}_t := \psi_t(\zeta_1, \zeta_2, E\zeta_1, E\zeta_2) := \zeta_t - \frac{1}{2}(\zeta_1 + \zeta_2) - E\zeta_t + \frac{1}{2}E(\zeta_1 + \zeta_2)$$

Then

$$\ddot{y}_t = \ddot{x}_t'\beta + \ddot{\varepsilon}_t$$

Example 1: linear regression with two-way fixed effects

Under strict exogeneity and rank condition,

$$\beta = \left(E\left[\ddot{x}_t \ddot{x}_t'\right]\right)^{-1} E\left[\ddot{x}_t \ddot{y}_t\right]$$
$$= \left(E\left[\psi_t(x_1, x_2, Ex_1, Ex_2)\psi_t(x_1, x_2, Ex_1, Ex_2)'\right]\right)^{-1} \times$$
$$\times E\left[\psi_t(x_1, x_2, Ex_1, Ex_2)\psi_t(y_1, y_2, Ey_1, Ey_2)\right]$$

Therefore, β is a **functional of the joint distribution** of (y_1, x_1, y_2, x_2)

Our framework: $(z_1, z_2) = (y_1, x_1, y_2, x_2)$

Example 2: diff-in-diff

Outcomes y_{it} are tracked for individuals i over periods $t = 1, 2$, with some individuals treated in period 2

Classical diff-in-diff estimator:

$$DID := E\left[y_{i2} - y_{i1} \,|\, d_{i2} = 1\right] - E\left[y_{i2} - y_{i1} \,|\, d_{i2} = 0\right]$$

Under parallel trends, identifies
the **average treatment effect on the treated**

Our framework:
DID as a functional of the joint distribution F of (y_{i1}, y_{i2}, d_{i2}),

$$DID = \int \frac{(y_2 - y_1)}{P(d_2 = 1)}\, dF(y_1, y_2, 1) - \int \frac{(y_2 - y_1)}{P(d_2 = 0)}\, dF(y_1, y_2, 0)$$

Example 3: quantile treatment effects

$T = 3$ periods, treatment d_3 in the last period

Data: random sample from (y_1, y_2, y_3, d_3)

Target: quantile treatment effect on the treated

$$QTT(\tau) = F^{-1}_{y_3(1)|d_3=1}(\tau) - F^{-1}_{y_3(0)|d_3=1}(\tau)$$

Callaway, Li (2019) show that,
under *distributional parallel trends* and *copula stability*,

$$F_{y_3(0)|d_3=1}(y) = P\Big[F^{-1}_{\Delta y_3|d_3=0}\left(F_{\Delta y_2|d_3=1}(\Delta y_2)\right) \leqslant y - F^{-1}_{y_2|d_3=1}\left(F_{y_1|d_3=1}(y_1)\right) \mid d_3 = 1\Big].$$

Our framework: $z_1 = (y_1, y_2)$, $z_2 = (y_3, d_3)$

Empirical illustration: income regression

Linear dynamic panel

$$\text{income}_{it} = \alpha_i + f_t + \theta \cdot \text{income}_{i,t-1} + \beta_1 \text{age}_{it} + \beta_2 \text{age}_{it}^2 + \varepsilon_{it}$$

Data: Understanding of America Survey by USC CESR

Waves 1-2: $N_1 = 4413$, $N_2 = 3738$ (**attrition** 18%),
refreshment sample: $N_r^1 = 4523$

Waves 2-3: $N_{2,total} = 8261$, $N_3 = 5686$ (**attrition** 31%),
refreshment sample: $N_r^2 = 1936, N_{3,total} = 7622$

Identification

Key identity:

$$\underbrace{F(z_1, z_2)}_{\text{target}} = \underbrace{\frac{P(W = 1)}{P(W = 1 | Z_1 \leqslant z_1, Z_2 \leqslant z_2)}}_{\text{weight}} \cdot \underbrace{F^w(z_1, z_2)}_{\text{balanced panel}}$$

No identification without further restrictions:

- need to identify $P(W = 1 | Z_1 \leqslant z_1, Z_2 \leqslant z_2)$
- extra information $F_1(z_1), F_2(z_2)$

Assumption (Identification)

$P(W = 1 | Z_1 \leqslant z_1, Z_2 \leqslant z_2) = G(k_1(z_1) + k_2(z_2))$ *for a know continuous strictly increasing function* $G : R \to R$ *and some unknown functions* $k_1 : R^{d_1} \to R$, $k_2 : R^{d_2} \to R$.

Identification assumption by Hirano et al 2001

$$P(W = 1|Z_1 \leqslant z_1, Z_2 \leqslant z_2) = G(k_1(z_1) + k_2(z_2))$$

Compare with AN (additive nonignorability)
Hirano, Imbens, Ridder, Rubin (2001):

$$P(W = 1|Z_1 = z_1, Z_2 = z_2) = G(k_1(z_1) + k_2(z_2)).$$

- Advantage: interpretation
- Disadvantage: computational complication

Comparing identification assumptions

Suppose $Z_1, Z_2 \in [0,1]^2$ and the conditional probability of staying is given by

$$P(W = 1 | Z_1 = z_1, Z_2 = z_2) = a z_1^2 + b z_1 z_2 + a z_2^2.$$

$$P(W = 1 | Z_1 \leqslant z_1, Z_2 \leqslant z_2) = \int_{-\infty}^{z_1} \int_{-\infty}^{z_2} \frac{P(W = 1 | Z_1 = t_1, Z_2 = t_2)}{F(z_1, z_2)} f(t_1, t_2)\, dt_1 dt_2.$$

- $a = 2/11$, $b = 7/11$: our assumption holds with, but HIRR does not;
- $a = 1/2$, $b = 0$: our does not hold, while HIRR holds;
- $a = 0$, $b = 1$: both assumptions hold.

Identification

Key identity:

$$\underbrace{F(z_1, z_2)}_{\text{target}} = \underbrace{\frac{P(W=1)}{P(W=1|Z_1 \leqslant z_1, Z_2 \leqslant z_2)}}_{\text{weight}} \cdot \underbrace{F^w(z_1, z_2)}_{\text{balanced panel}} .$$

Identifying restriction:

$$P(W=1|Z_1 \leqslant z_1, Z_2 \leqslant z_2) = G(k_1(z_1) + k_2(z_2)).$$

Then:

$$G(k_1(z_1) + k_2(z_2)) = \frac{P(W=1)}{P(W=1|Z_1 \leqslant z_1, Z_2 \leqslant z_2)} \cdot F^w(z_1, z_2).$$

Identification

Denote:

$$\Phi(p, F_1, F_2, F_1^w, F_2^w, F^w) = \frac{pF^w}{G\left(G^{-1}\left(\frac{pF_1^w}{F_1}\right) + G^{-1}\left(\frac{pF_2^w}{F_2}\right) - G^{-1}\left(p\right)\right)}.$$

Theorem (Identification)

$$F = \Phi(p, F_1, F_2, F_1^w, F_2^w, F^w)$$

Estimation

Step 1.

Plug-in estimator of the joint CDF:

$$\hat{F}(z_1, z_2) = \Phi\left(\hat{p}, \hat{F}_1(z_1), \hat{F}_2(z_2), \hat{F}_1^w(z_1), \hat{F}_2^w(z_2), \hat{F}^w(z_1, z_2)\right),$$

where $\hat{F}_1, \hat{F}_2, \hat{F}_1^w, \hat{F}_2^w, \hat{F}^w$ are empirical CDF's and $\hat{p} = \hat{P}(W = 1)$

Step 2.

Let $\hat{\theta}$ s.t.

$$\int m(z_1, z_2; \hat{\theta})\, d\hat{F}(z_1, z_2) = 0.$$

Estimation Algorithm

1. Calculate the plug-in estimator $\hat{F} = \Phi(\hat{p}, \hat{F}_1, \hat{F}_2, \hat{F}_1^w, \hat{F}_2^w, \hat{F}^w)$
2. Calculate its jump sizes $\hat{f}(z_1, z_2)$ at points $(z_1, z_2) \in \hat{\mathcal{Z}}_1 \times \hat{\mathcal{Z}}_2$:

$$\hat{f}(x) = \sum_{(i_1,\ldots,i_d)\in\{0,1\}^d} (-1)^{i_1+\cdots+i_d} \hat{F}\left(x_1 + (-1)^{i_1} h_1, \ldots, x_d + (-1)^{i_d} h_d\right).$$

3. Set $\hat{\theta}$ such that

$$\sum_{(z_1,z_2)\in\hat{\mathcal{Z}}_1\times\hat{\mathcal{Z}}_2} m(z_1, z_2; \hat{\theta})\hat{f}(z_1, z_2) = 0.$$

Consistency

Lemma (Uniform Convergence)

Let the identification assumption hold and

(i) $P(W = 1 | Z_1 \leqslant z_1, Z_2 \leqslant z_2)$ *is bounded away from zero*

(ii) $\theta_0 \in \Theta$ *is compact*

(iii) $m(z; \theta)$ *is of bounded variation for each* $\theta \in \Theta$*;*

(iv) $m(z; \theta)$ *is continuous at each* $\theta \in \Theta$ *with probability one in* F*;*

(v) *there exists a function* $d(z)$ *such that* $\|m(z; \theta)\| \leqslant d(z)$ *for all* $\theta \in \Theta$ *and* $\int d(z)\, dF(z) < \infty$.

Then

$$\sup_{\theta \in \Theta} \left| \int m(z; \theta) d\hat{F} - \int m(z; \theta) dF \right| \to 0 \quad a.s.$$

Consistency

Theorem (Consistency)

Let all assumptions of the uniform convergence lemma hold and

(i) θ_0 *is identified from the moment conditions;*

Then $\hat{\theta} \xrightarrow{p} \theta_0$.

Theorem (Inference)

Suppose $\hat{\theta}$ is a consistent estimator of θ_0 and

(i) $\theta_0 \in$ *interior* (Θ);

(ii) $m(z;\theta)$ *is differentiable in a neighborhood* $\mathcal{N}$ *of* θ_0;

(iii) $J := EDm(Z;\theta_0)$ *is nonsingular. Then*

$G(\cdot)$ *is differentiable*

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \rightsquigarrow J^{-1}\cdot\int m(z;\theta_0)\, d\Phi'_{F_\eta}(\mathbb{G}_{F_\eta})(z),$$

where $\eta=(W,Z_1,WZ_2,Z_2^r)$ *is data,* $\mathbb{G}_{F_\eta}$ *is the* F_η*-Brownian bridge and* Φ'_{F_η} *is Hadamard derivative.*

Bootstrap Validity

Fang & Santos (2019):

F_0 is a possibly infinite dimensional parameter and there exists an estimator $\hat{F}_n$ s.t.

$$r_n(\hat{F}_n - F_0) \rightsquigarrow G_0$$

The parameter is interest is $\theta_0 = \phi(F_0)$:

$$r_n(\phi(\hat{F}_n) - \phi(F_0)) \rightsquigarrow \phi'_{F_0}(G_0).$$

Theorem (Fang & Santos 3.1)

*Suppose the G_0 is **Gaussian** and technical assumptions hold. Then ϕ is **Hadamard differentiable** at $F_0 \in D_\phi$ tangentially to the support of G_0 **if and only if the bootstrap is valid** for $\phi(\hat{F}_n)$.*

Monte Carlo simulation: discrete data

DGP: discrete Markov process

$Z_1 \sim$ uniform over $\{1, \dots, m\}$

$Z_2 \in \{1, \dots, m\}$, positive transition matrix

Attrition rate $P(W = 0) = 0.3$

Target parameter $\theta(m) = P_m(Z_2 = 1 | Z_1 = 1)$
true value $\theta(5) = 0.23$, $\theta(10) = 0.12$, $\theta(20) = 0.05$

Monte Carlo: number of repetitions 1000, warp speed bootstrap.

		$n_1 = n_r = 1000$		$n_1 = n_r = 10,000$	
		$\hat{\theta}$	$\hat{\theta}_{naive}$	$\hat{\theta}$	$\hat{\theta}_{naive}$
$m = 5$	bias	0.000	-0.018	-0.001	-0.019
	rmse	0.017	0.024	0.024	0.030
	mae	0.014	0.020	0.019	0.025
	coverage 99%	0.993		0.979	
	coverage 95%	0.954		0.946	
	coverage 90%	0.887		0.897	
$m = 10$	bias	0.000	-0.013	0.000	-0.014
	rmse	0.019	0.022	0.027	0.029
	mae	0.015	0.018	0.022	0.023
	coverage 99%	0.993		0.992	
	coverage 95%	0.945		0.944	
	coverage 90%	0.909		0.912	
$m = 20$	bias	0.000	-0.005	0.001	-0.005
	rmse	0.019	0.018	0.028	0.025
	mae	0.015	0.015	0.022	0.020
	coverage 99%	0.992		0.993	
	coverage 95%	0.949		0.953	
	coverage 90%	0.885		0.922	

Monte Carlo simulation: continuous data

DGP: $(Z_1, Z_2) = (Z_{11}, Z_{12}, Z_{21}, Z_{22}) \in [0,1]^4$, where

- Z_{11}, Z_{21} are independent of Z_{12}, Z_{22}
- $Z_{11}, Z_{21} \sim$ iid uniform[0,1]
- Z_{12}, Z_{22} have CDF

$$\text{Gumbel}(z_{12}, z_{22}; \nu) = \exp\left[-\left((-\log z_{11})^{\nu} + (-\log z_{22})^{\nu}\right)^{1/\nu}\right]$$

(Gumbel copula with dependence parameter $\nu > 1$)

Attrition rate $P(W = 0) = 0.70$

Target parameter $\theta(\nu) = E_{\nu}[Z_{12}Z_{22}]$,
true values $\theta(2) \approx \theta(10) \approx \theta(20) = 0.3$

		$n_1 = n_r = 1000$		$n_1 = n_r = 5000$	
		$\hat{\theta}$	$\hat{\theta}_{naive}$	$\hat{\theta}$	$\hat{\theta}_{naive}$
$\nu = 2$	bias	0.009	0.009	0.004	0.009
	rmse	0.024	0.018	0.012	0.011
	mae	0.020	0.015	0.009	0.010
	coverage 99%	0.998		0.997	
	coverage 95%	0.985		0.984	
	coverage 90%	0.958		0.948	
$\nu = 10$	bias	0.003	0.012	0.000	0.011
	rmse	0.028	0.021	0.014	0.013
	mae	0.022	0.017	0.011	0.012
	coverage 99%	0.997		0.992	
	coverage 95%	0.976		0.964	
	coverage 90%	0.942		0.921	
$\nu = 20$	bias	0.004	0.014	0.001	0.013
	rmse	0.030	0.022	0.014	0.015
	mae	0.024	0.018	0.011	0.013
	coverage 99%	0.997		0.994	
	coverage 95%	0.985		0.968	
	coverage 90%	0.949		0.951	

Empirical illustration

Static linear model

$$\sinh^{-1}(\text{income}_{it}) = \alpha_i + f_t + \theta_1 \cdot \text{age}_{it} + \theta_2 \cdot \text{age}_{it}^2 + \varepsilon_{it}$$

Data: Understanding of America Survey (USC CESR)
Period 1: $N_1 = 7909$, **period 2**: $N_2 = 5424$ (**attrition** 31%),
refreshment sample: $N_r = 1894$

	naive		with refreshment	
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
coeff.	0.128**	-0.0004	0.116***	-0.000
s.e.	0.047	0.0003	0.034	0.112

Conclusion

Panels with **attrition** and **refreshment**
This project:

- New identification assumption
- Nonparametric approach **without tuning parameters**
- Closed-form "plug-in" estimator of the parameter defined by moment conditions
- Consistency, inference
- Nonparametric bootstrap
- Monte Carlo simulations
- Empirical illustration