Inference by Stochastic Optimization: A Free-Lunch Bootstrap

Jean-Jacques Forneron, Boston University
Serena Ng, Columbia University and NBER

November 16, 2019

Introduction: Challenging Inferences

- Extremum estimation: GMM, NLS, MLE, ...
  - where computing the asymptotic variance is not tractable
    e.g. rely on transformed/generated data, multi-step estimation, complicated moments/likelihood
  - or counterfactuals st. Δ-method is challenging to implement

Introduction: Challenging Inferences

- Extremum estimation: GMM, NLS, MLE, ...
  - where computing the asymptotic variance is not tractable
    e.g. rely on transformed/generated data, multi-step estimation, complicated moments/likelihood
  - or counterfactuals st. Δ-method is challenging to implement
- But estimation procedure needs to be repeated many times
  $\Rightarrow$ rely on Bootstrap inference
  - common in structural estimation (IO, labour, macro, etc.)
- Classical Bootstrap: run the optimizer <u>many times</u>

Introduction: Challenging Inferences

- Extremum estimation: GMM, NLS, MLE, ...
  - where computing the asymptotic variance is not tractable
    e.g. rely on transformed/generated data, multi-step estimation, complicated moments/likelihood
  - or counterfactuals st. Δ-method is challenging to implement
- But estimation procedure needs to be repeated many times
  $\Rightarrow$ rely on Bootstrap inference
  - common in structural estimation (IO, labour, macro, etc.)
- Classical Bootstrap: run the optimizer many times
- THIS PAPER: focused on a **Stochastic Newton-Raphson** algorithm, a single run produces
  - a consistent estimator by simple averaging
  - asymptotically valid Bootstrap draws (free-lunch)

The Setup

- Interested in the sample GMM/MLE/MD estimator:

$$\hat{\theta}_n = \text{argmin}_{\theta \in \Theta} Q_n(\theta)$$

$Q_n(\cdot)$ is the sample objective function

The Setup

- Interested in the sample GMM/MLE/MD estimator:

$$\hat{\theta}_n = \text{argmin}_{\theta \in \Theta} Q_n(\theta)$$

  $Q_n(\cdot)$ is the sample objective function

- Ideal Bootstrap sample:

$$\hat{\theta}^b_m = \text{argmin}_{\theta \in \Theta} Q^{(b)}_m(\theta)$$

  for $b = 1, \ldots, B$, where $Q^{(b)}_m(\cdot)$ is an m out of n re-sampled objective (or re-weighted/multiplier)

The Setup

- Interested in the sample GMM/MLE/MD estimator:

$$\hat{\theta}_n = \text{argmin}_{\theta \in \Theta} Q_n(\theta)$$

  $Q_n(\cdot)$ is the sample objective function

- Ideal Bootstrap sample:

$$\hat{\theta}_m^b = \text{argmin}_{\theta \in \Theta} Q_m^{(b)}(\theta)$$

  for $b = 1, \ldots, B$, where $Q_m^{(b)}(\cdot)$ is an m out of n re-sampled objective (or re-weighted/multiplier)

- **Goal: a 2-in-1 procedure for estimation and inference**

Related Literatures

This Paper

Algorithm: Stochastic Newton-Raphson

i. initialize: at θ^0, given

Bootstrap is '*free*' in the sense that we get both *an approximate minimizer* (iii) and *asym. valid standard errors* (iv) in one run

This Paper

Algorithm: Stochastic Newton-Raphson

i. initialize: at θ^0, given

ii. for $b = 1, \ldots, B$ compute:

$$\theta^b = \theta^{b-1} - \gamma \cdot [H_m^{(b)}(\theta^{b-1})]^{-1} G_m^{(b)}(\theta^{b-1})$$

$G_m^{(b)}, H_m^{(b)}$ re-sampled gradient, hessian; $m/n \to c \in (0, 1]$;
$\gamma \in (0, 1]$ fixed learning rate

Bootstrap is '*free*' in the sense that we get both *an approximate minimizer* (iii) and *asym. valid standard errors* (iv) in one run

This Paper

Algorithm: Stochastic Newton-Raphson

i. initialize: at θ^0, given

ii. for $b = 1, \ldots, B$ compute:

$$\theta^b = \theta^{b-1} - \gamma \cdot [H_m^{(b)}(\theta^{b-1})]^{-1} G_m^{(b)}(\theta^{b-1})$$

$G_m^{(b)}, H_m^{(b)}$ re-sampled gradient, hessian; $m/n \to c \in (0,1]$;

$\gamma \in (0,1]$ fixed learning rate

iii. $\frac{1}{B}\sum_{b=1}^{B} \theta^b \simeq \hat{\theta}_n$

Bootstrap is '*free*' in the sense that we get both *an approximate minimizer* (iii) and *asym. valid standard errors* (iv) in one run

This Paper

Algorithm: Stochastic Newton-Raphson

i. initialize: at θ^0, given

ii. for $b = 1, \ldots, B$ compute:

$$\theta^b = \theta^{b-1} - \gamma \cdot [H_m^{(b)}(\theta^{b-1})]^{-1} G_m^{(b)}(\theta^{b-1})$$

$G_m^{(b)}, H_m^{(b)}$ re-sampled gradient, hessian; $m/n \to c \in (0,1]$;
$\gamma \in (0,1]$ fixed learning rate

iii. $\frac{1}{B}\sum_{b=1}^{B} \theta^b \simeq \hat{\theta}_n$

iv. $var(\theta^b) \simeq \frac{\gamma^2}{1-[1-\gamma]^2} var(\hat{\theta}_n)$ (the asymptotic variance)

Bootstrap is '*free*' in the sense that we get both *an approximate minimizer* (iii) and *asym. valid standard errors* (iv) in one run

Outline

An Overview of Derivative-Based Methods

Gradient Descent and Newton-Raphson Methods

Algorithm: Newton-Raphson

i. initialize: at θ^0, given

ii. for $b = 1, \ldots, B$ compute:

$$\theta^b = \theta^{b-1} - \underbrace{\gamma_b}_{\text{learning rate}} \cdot [H_n(\theta^{b-1})]^{-1} G_n(\theta^{b-1})$$

- **Gradient-Descent**: $\theta^b = \theta^{b-1} - \gamma_b \cdot \cancel{[H_n(\theta^{b-1})]^{-1}} G_n(\theta^{b-1})$
  - less costly, slow convergence when $\lambda_{\max}(H_n)/\lambda_{\min}(H_n)$ large

Illustration OLS regression

- OLS regression: $y_i = x_i'\theta + u_i$; $\gamma_b = \gamma$ fixed
- **Newton-Raphson**:

$$\boxed{\theta^b - \hat{\theta}_n = (1-\gamma)^b[\theta^0 - \hat{\theta}_n]}$$

  - for $\gamma_b = 1$ convergence after one iteration

- **Gradient Descent**:

$$\boxed{\theta^b - \hat{\theta}_n = (I - 2\gamma[\sum_i x_i x_i'/n])^b[\theta^0 - \hat{\theta}_n]}$$

  - very slow convergence when $\lambda_{\max}(X'X)/\lambda_{\min}(X'X)$ large

Stochastic Gradient Descent

- Full sample G_n, H_n costly to compute for n very large
- **Solution**: use a *minibatch* (small) of subsamples $m \ll n$
- In practice: $m = 1$ is popular

Algorithm: Stochastic Gradient-Descent

i. initialize: at θ^0, given

ii. for $b = 1, \ldots, B$ compute:

$$\theta^b = \theta^{b-1} - \gamma_b \cdot G_m^{(b)}(\theta^{b-1})$$

Simple Illustration: OLS estimation ($m = 1$)

- Mini-batch with $m = 1$
- **Stochatic Gradient Descent**:

$$\theta^b - \hat{\theta}_n = (I - 2\gamma_b \underbrace{x_i^{(b)} x_i^{(b)\prime}}_{\text{noisy}})(\theta^{b-1} - \hat{\theta}_n) - 2\gamma_b \underbrace{x_i^{(b)} \hat{u}_i^{(b)}}_{\text{noisy}}$$

Simple Illustration: OLS estimation ($m = 1$)

- Mini-batch with $m = 1$
- **Stochatic Gradient Descent**:

$$\theta^b - \hat{\theta}_n = (I - 2\gamma_b \underbrace{x_i^{(b)} x_i^{(b)\prime}}_{\text{noisy}})(\theta^{b-1} - \hat{\theta}_n) - 2\gamma_b \underbrace{x_i^{(b)} \hat{u}_i^{(b)}}_{\text{noisy}}$$

- For $\theta^b \xrightarrow{p^\star} \hat{\theta}_n$ we need $\gamma_b \searrow 0$
  - fast enough so that $\mathbb{E}^\star \| 2\gamma_b [x_i^{(b)} x_i^{(b)\prime}] \theta^{(b-1)} \|^2 \to 0$
  - not too fast so that $\mathbb{E}^\star \| (1 - 2\gamma_b x_i^{(b)} x_i^{(b)\prime})(\theta^{b-1} - \hat{\theta}_n) \|^2 \to 0$

  $\Rightarrow$ convergence can be very slow
  - in practice: adaptive methods (adagrad, RMSprop,...)

Simple Illustration: OLS estimation ($m = 1$)

- Mini-batch with $m = 1$
- **Stochatic Gradient Descent**:

$$\theta^b - \hat{\theta}_n = (I - 2\gamma_b \underbrace{x_i^{(b)} x_i^{(b)\prime}}_{\text{noisy}})(\theta^{b-1} - \hat{\theta}_n) - 2\gamma_b \underbrace{x_i^{(b)} \hat{u}_i^{(b)}}_{\text{noisy}}$$

- For $\theta^b \overset{p^\star}{\to} \hat{\theta}_n$ we need $\gamma_b \searrow 0$
  - fast enough so that $\mathbb{E}^\star \| 2\gamma_b [x_i^{(b)} x_i^{(b)\prime}] \theta^{(b-1)} \|^2 \to 0$
  - not too fast so that $\mathbb{E}^\star \| (1 - 2\gamma_b x_i^{(b)} x_i^{(b)\prime})(\theta^{b-1} - \hat{\theta}_n) \|^2 \to 0$

  $\Rightarrow$ convergence can be very slow
  - in practice: adaptive methods (adagrad, RMSprop,...)
- **Stochatic Newton-Raphson**: $H_1^{(b)}$ often noisy/near singular
  - e.g. $x_i = (1, x_{i,1})$, $x_{i,1} \sim \mathit{Bernoulli}(p)$

    $\Rightarrow x_i x_i'$ singular wp. 1 for any $p \in [0, 1]$

This Paper: S-NR with larger batches

- Three changes over S-GD:
  a. re-introduce the Hessian $H_m^{(b)}(\theta^{b-1})$ (NR)
  b. sample m out of n observations, $m/n \to c \in (0,1]$
  c. fixed learning rate $\gamma_b = \gamma \in (0,1]$

Algorithm: Stochastic Newton-Raphson

i. initialize: at θ^0, given

ii. for $b = 1, \ldots, B$ compute:

$$\theta^b = \theta^{b-1} - \gamma \cdot [H_m^{(b)}(\theta^{b-1})]^{-1} G_m^{(b)}(\theta^{b-1})$$

Simple Illustration: OLS estimation ($m/n \to c \in (0,1]$)

- **Stochatic Newton-Raphson**:

$$\theta^b - \hat{\theta}_n = \underbrace{(1-\gamma)(\theta^{b-1} - \hat{\theta}_n)}_{\text{deterministic cv.}} + \underbrace{\gamma(\hat{\theta}_m^{(b)} - \hat{\theta}_n)}_{\text{noise}}$$

Simple Illustration: OLS estimation ($m/n \to c \in (0,1]$)

- **Stochatic Newton-Raphson**:

$$\theta^b - \hat{\theta}_n = \underbrace{(1-\gamma)(\theta^{b-1} - \hat{\theta}_n)}_{\text{deterministic cv.}} + \underbrace{\gamma(\hat{\theta}_m^{(b)} - \hat{\theta}_n)}_{\text{noise}}$$

- For $\gamma = 1$, $\theta^b = \hat{\theta}_m^{(b)}$ the bootstrapped estimate
- $\theta^b \overset{p}{\not\to} \hat{\theta}_n$ with γ fixed but
  - $\mathbb{E}^\star(\theta^b) \simeq \hat{\theta}_n$ and $var^\star(\theta^b) \simeq \frac{\gamma^2}{1-[1-\gamma]^2} var^\star(\hat{\theta}_m^{(b)})$

Simple Illustration: OLS estimation ($m/n \to c \in (0,1]$)

- **Stochatic Newton-Raphson**:

$$\theta^b - \hat{\theta}_n = \underbrace{(1-\gamma)(\theta^{b-1} - \hat{\theta}_n)}_{\text{deterministic cv.}} + \underbrace{\gamma(\hat{\theta}_m^{(b)} - \hat{\theta}_n)}_{\text{noise}}$$

- For $\gamma = 1$, $\theta^b = \hat{\theta}_m^{(b)}$ the bootstrapped estimate
- $\theta^b \overset{p}{\not\to} \hat{\theta}_n$ with γ fixed but
  - $\mathbb{E}^\star(\theta^b) \simeq \hat{\theta}_n$ and $var^\star(\theta^b) \simeq \frac{\gamma^2}{1-[1-\gamma]^2} var^\star(\hat{\theta}_m^{(b)})$
- Now: extend this result to a class of non-linear models

Asymptotic Results

Assumptions: Q_n

Assumption (Sample Objective Function)

i. $\|H_n(\theta)^{-1}G_n(\theta)\|_2 \leq \overline{C}\|\theta - \hat{\theta}_n\|_2$,

ii. $\underline{C}\|\theta - \hat{\theta}_n\|_2^2 \leq \langle \theta - \hat{\theta}_n, H_n(\theta)^{-1}G_n(\theta)\rangle$,

iii. $\underline{c}_H \leq \lambda_{\min}(H_n(\theta)^{-1}) \leq \lambda_{\max}(H_n(\theta)^{-1}) \leq \overline{C}_H$,

iv. $\|H_n(\theta) - H_n(\hat{\theta}_n)\|_2 \leq C_{n,1} \times \|\theta - \hat{\theta}_n\|_2$,

v. $\|\sup_{\theta\in\Theta} G_n(\theta)\|_2 \leq \overline{C}_n$

Remark: conditions i-iii. imply strong convexity

Newton-Raphson

Lemma (Newton-Raphson)

Suppose Assumption 1 holds, then for $\gamma \in (0,1)$ *small enough,* $\exists \bar{\gamma} \in (0,1)$ *such that for any* θ^0*:*

$$\|\theta^b_{NR} - \hat{\theta}_n\|_2 \leq (1-\bar{\gamma})\|\theta^{b-1} - \hat{\theta}_n\|_2 \leq (1-\bar{\gamma})^b\|\theta^0 - \hat{\theta}_n\|_2$$

Assumptions: $Q_m^{(b)}$

Assumption (Re-Sampled Objective Function)

Suppose the following holds uniformly over $\theta \in \Theta$:

i. $\|[H_m^{(b)}(\theta)]^{-1}[G_m^{(b)}(\theta) - G_m^{(b)}(\hat{\theta}_n) - H_m^{(b)}(\theta)(\theta - \hat{\theta}_n)]\|_2 \leq C_{m,1} \times \|\theta - \hat{\theta}_n\|_2^2$,

ii. $\mathbb{E}^\star \left(\sup_{\theta\in\Theta} \|[H_n(\theta)]^{-1} - [H_m^{(b)}(\theta)]^{-1}\|_2^2\right)^{1/2} \leq C_{m,2} \times m^{-1/2}$,

iii. $\left[\mathbb{E}^\star \left(\sup_{\theta\in\Theta} \|H_n(\theta) - H_m^{(b)}(\theta)\|_2^2\right)\right]^{1/2} \leq C_{m,3} \times m^{-1/2}$,

iv. $\left[\mathbb{E}^\star \left(\sup_{\theta\in\Theta} \|\mathbb{G}_m^{(b)}(\theta)\|_2^2\right)\right]^{1/2} \leq \overline{C}$, *for* $\mathbb{G}_m^{(b)}(\theta) \stackrel{def}{=} \sqrt{m}[G_m^{(b)}(\theta) - G_n(\theta)]$.

where $C_{m,1/2/3}$ and $(C_n)_{n\geq 1}$ are bounded above, $\overline{C} < +\infty$.

Lemma (Linearization of the S-NR Markov-Chain)

Suppose Assumptions 1-3 hold, then for $\gamma \in (0,1)$ *small enough,* $\exists \bar{\gamma} \in (0,1)$ *such that* $\forall \theta^0$*, uniformly in* $b \geq 1$*:*

$$\mathbb{E}^\star\Big(\|\theta^b_{NR} - \hat{\theta}_n + \gamma \sum_{j=0}^{b-1} (1-\gamma)^j \mathbb{Z}_m^{b-j}\|_2^2\Big)^{1/2}$$

$$\lesssim m^{-1} + b\rho^b[d_{0,n} + d_{0,n}^2]$$

where $\rho = \max(1-\gamma, 1-\bar{\gamma}) \in [0,1)$*;* $d_{0,n} = \mathbb{E}^\star\left(\|\theta^0 - \hat{\theta}_n\|_2^2\right)^{1/2}$ *and* $\mathbb{Z}_m^{b-j} = [H_n(\hat{\theta}_n)]^{-1} G_m^{(b-j)}(\hat{\theta}_n)$

Bootstrap behaviour of the S-NR draws

Theorem (Convergence in Distribution)

Suppose Assumptions 1-3 hold, let $\mathbb{Z}_m^b = [H_n(\hat{\theta}_n)]^{-1} G_m^{(b)}(\hat{\theta}_n)$ *and* $\Sigma_n = var^\star(\mathbb{Z}_m^b)$*. Suppose* $0 < \underline{\lambda} \leq \lambda_{\min}(\Sigma_n) \leq \overline{\lambda} \leq \lambda_{\max}(\Sigma_n) < +\infty$*, and conditions on the characteristic function of* $\mathbb{Z}_m^b$ *hold then:*

$$\sqrt{m}\Sigma_n^{-1/2}(\theta^b - \hat{\theta}_n) \xrightarrow{d^\star} \mathcal{N}\left(0, \frac{\gamma^2}{1 - [1-\gamma]^2} I\right),$$

as $m, b \to \infty$*; if* $\log(m)/b \to 0$ *and* $d_{0,n} = O(1)$*,* $n/m = O(1)$*.*

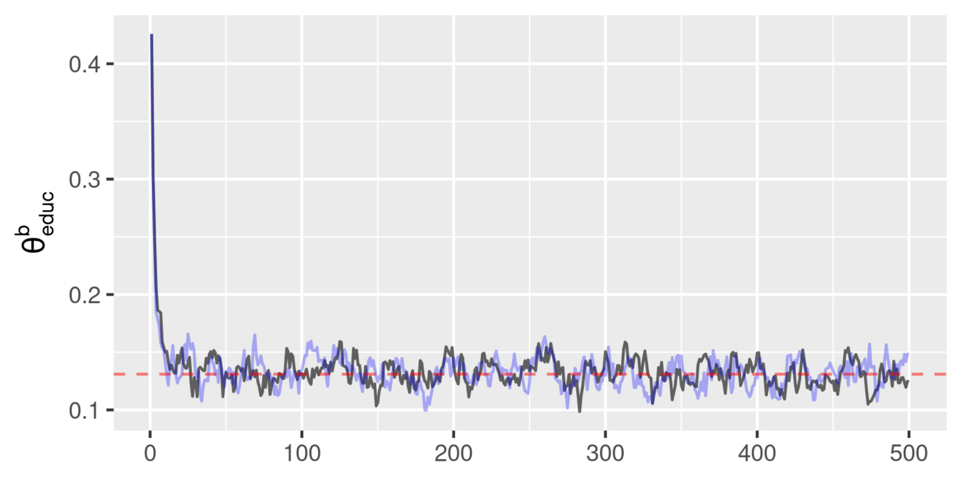
Empirical Illustration

Simple Example: Mroz (1987) Probit model

- Probit model: $\mathbb{P}(y_i = 1|x_i) = \Phi(x_i'\theta)$
- Sample of $n = 753$ observations, $m = n$, $\gamma = 0.3$
- $SNR_{np/m}$: iid re-sampling and multiplier Bootstrap
- Compare $\hat{\theta}_{n,MLE}$, asym. & boot. standard errors with SNR

	nwifeinc	educ	exper	exper2	age	kidslt6	kidsge6	constant
$\hat{\theta}_{n,MLE}$	-0.012	0.131	0.123	-0.002	-0.053	-0.868	0.036	0.270
Asym.	(0.005)	(0.025)	(0.019)	(0.001)	(0.008)	(0.119)	(0.043)	(0.509)
$\bar{\theta}_{n,boot}$	-0.012	0.134	0.124	-0.002	-0.054	-0.883	0.036	0.275
Boot.	(0.005)	(0.027)	(0.020)	(0.001)	(0.009)	(0.122)	(0.046)	(0.524)
SNR_{np}	-0.012	0.133	0.123	-0.002	-0.053	-0.873	0.038	0.263
	(0.005)	(0.026)	(0.019)	(0.001)	(0.010)	(0.119)	(0.046)	(0.510)
SNR_m	-0.012	0.132	0.123	-0.002	-0.053	-0.872	0.036	0.267
	(0.006)	(0.026)	(0.020)	(0.001)	(0.008)	(0.119)	(0.046)	(0.512)

Simple Example: Mroz (1987) Probit model



red: $\hat{\theta}_n$; black/blue: SNR with iid/multiplier Bootstrap

Application:
Sensitivity Analyses

Influential Observations

- **Influential observations:**
  a subset $\mathcal{I} \subset \{1, \ldots, n\}$ of the data which impacts the conclusions significantly
- e.g. leads to very different point estimates or standard errors
  - outliers
  - leverage points
  - ...
- Idea: under H_0 (no influential observations) removing $\mathcal{I}$ during the iterations should not significantly affect the Markov-Chain
- Under H_1 (influential observations) should lead to a structural break in the levels/variance

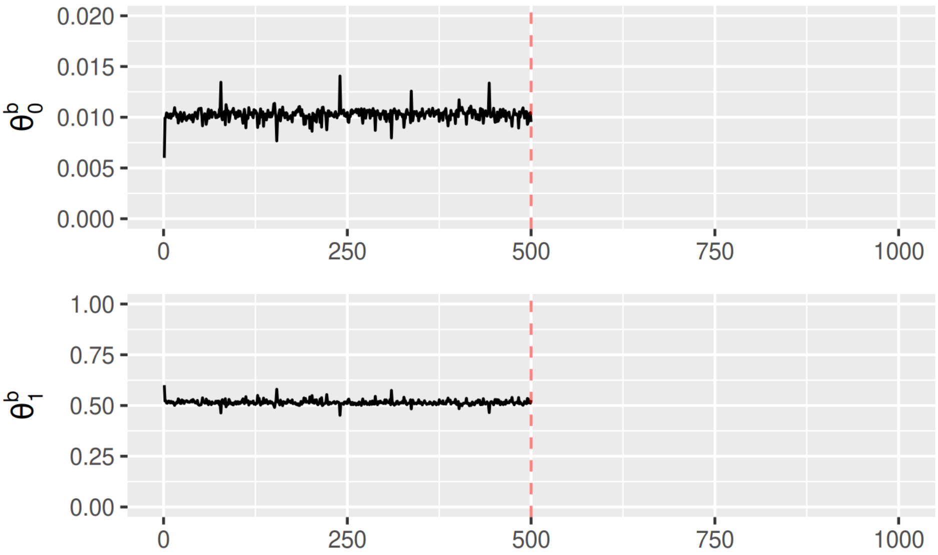
Simple Example: IBES

- Institutional Brokers' Estimate System (IBES)
- Large database of analyst earnings estimates vs. realized
- Predictive regression:

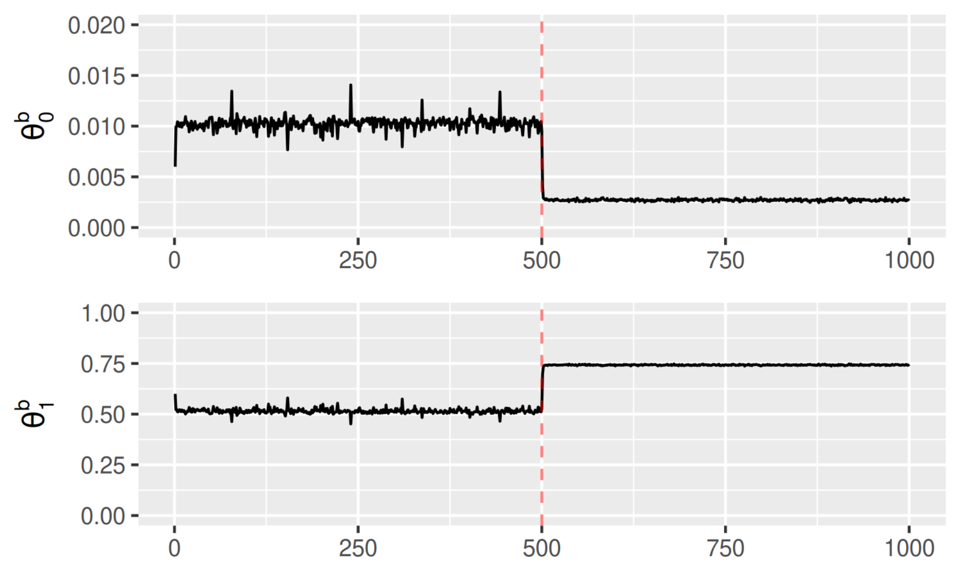
$$y_i^{\text{actual}} = \theta_0 + \theta_1 y_i^{\text{medest}} + e_i$$

- medest = median estimate
- Rational expectations: $\theta_0 \simeq 0, \theta_1 \simeq 1$
- Experiment: remove the 2% most influential obs. & compare
- $n = 9,278$ firms, $t = 01/1985 - 12/2017$
- $\gamma = 0.8$; $m = 6,000$ (re-sample firms)

Simple Example: IBES - Full Sample



Simple Example: IBES - Without Outliers



Main Example: PSID Income Dynamics

- Panel Study of Income Dynamics (PSID)
- Moffitt and Zhang (2018) earnings volatility
- $3,508$ males ($36,403$ person-year obs.)
- Model permanent and transitory components:

$$
\begin{aligned}
y_{iat} &= \alpha_t \mu_{ia} + \beta_t \nu_{ia} \\
\mu_{ia} &= \mu_{i0} + \sum_{s=1}^{a} \omega_{is} \\
\nu_{ia} &= \varepsilon_{ia} + \sum_{s=1}^{a-1} \psi_{a,a-s} \varepsilon_{is}, \quad a \geq 2
\end{aligned}
$$

- $a = \text{age} \in [24, 54]$

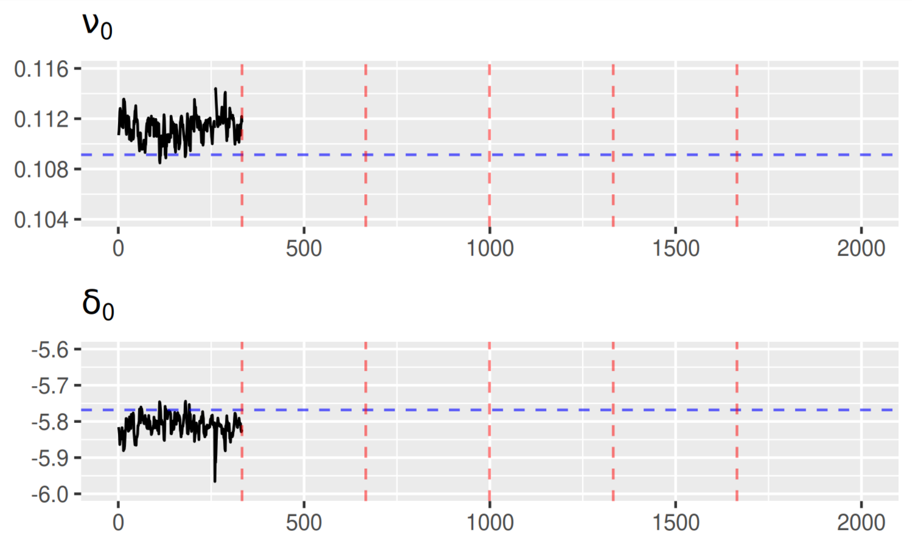
Empirics

- De-trend the data using OLS with polynomial regressors
- Aggregate residual autocovariances by age-group
- Match sample with model-based autocovariance matrix
- Warning:
  - original paper estimates 11 variance parameters
  - we only estimate 4 because of identification issues

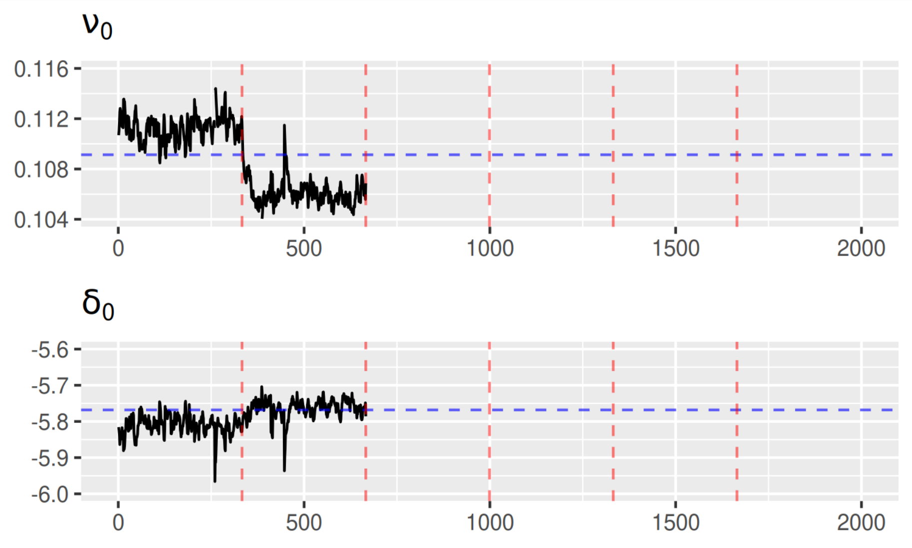
$$var(\mu_{i,0}) : \nu_0$$
$$var(\omega_{ir}) : \delta_0, \cancel{\delta_1}$$
$$var(\varepsilon_{ir}) : \gamma_0, \gamma_1, \cancel{k}$$
$$\psi_{a,a-r} : \cancel{\pi}, \cancel{\lambda_1}, \cancel{\eta_1}, \cancel{\eta_2}, \cancel{\eta_3}$$

- **Goal: Are the results sensitive to particular age groups?**

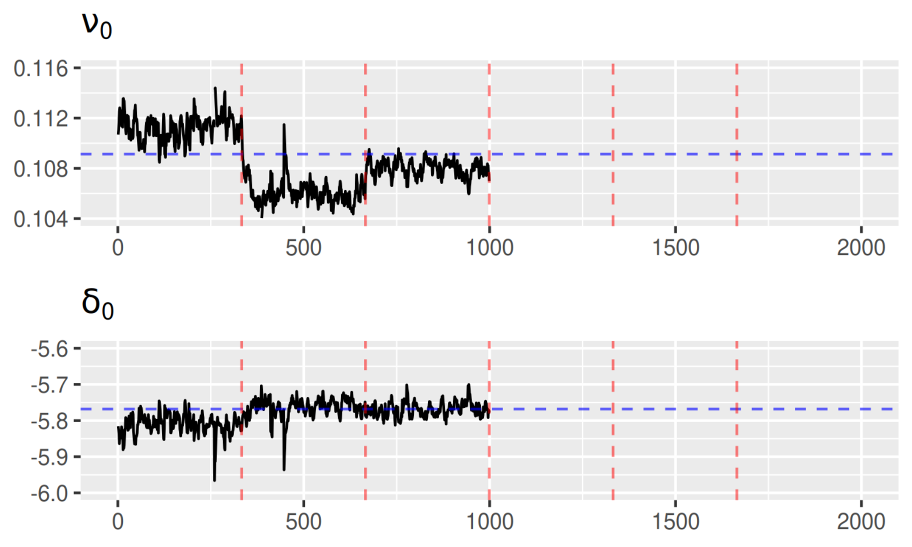
Influence: estimates without 24-28 age group



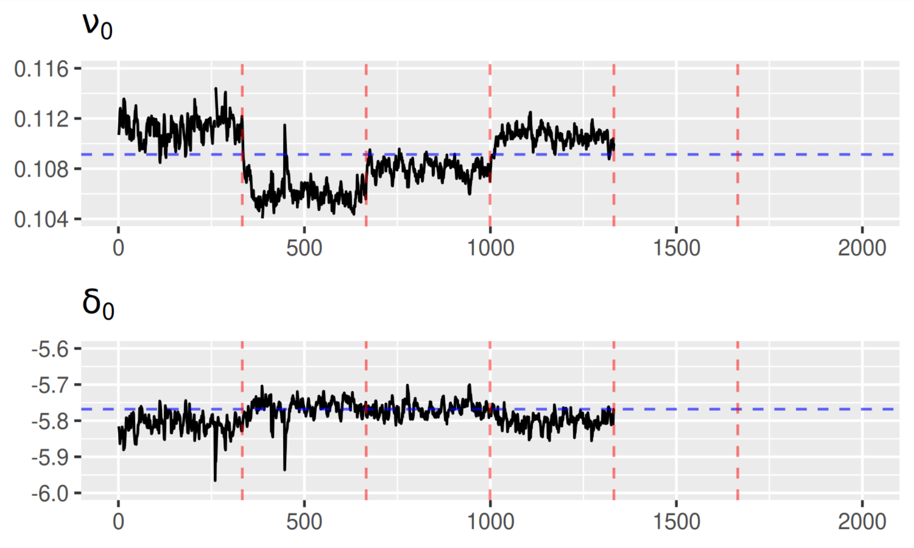
Influence: estimates without 29-33 age group



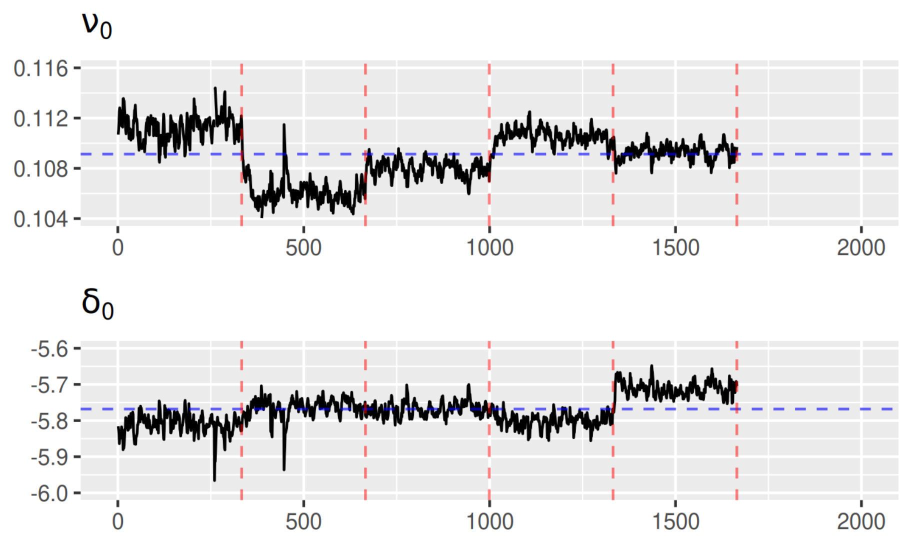
Influence: estimates without 34-38 age group



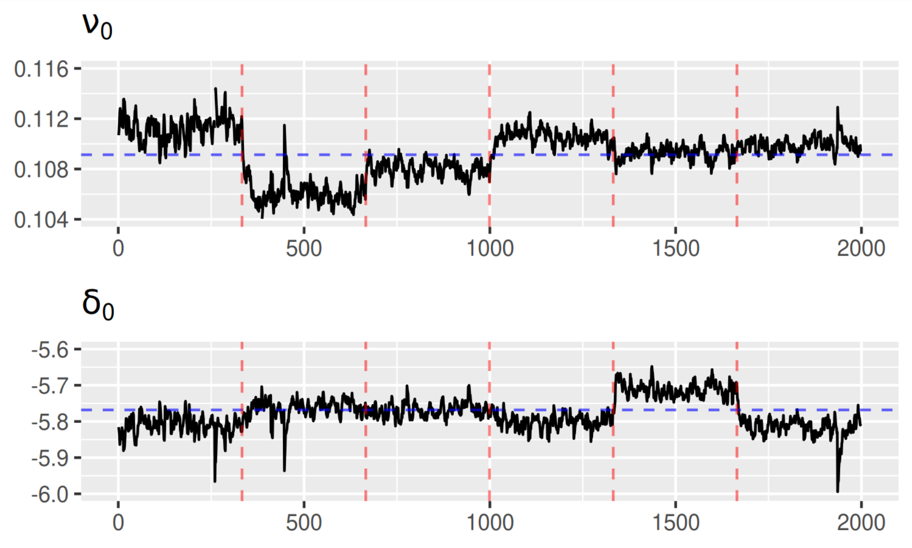
Influence: estimates without 39-43 age group



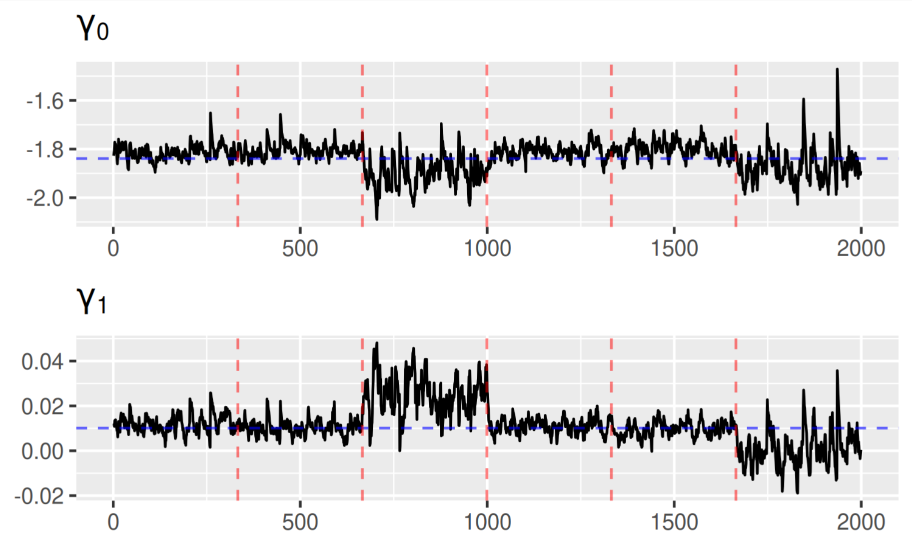
Influence: estimates without 44-48 age group



Influence: estimates without 49-54 age group



Influence of individual age-groups



Trimming

Conclusion

Conclusion

- SNR: simultaneous estimation and Bootstrapping
- Appealing for two-step estimators with complicated variance
- Potential avenues of research:
  - Stochastic quasi-Newton Methods (S-BFGS)
    computationally very attractive
  - Alternative sampling schemes
    look for theoretical guarantees in non-convex settings

References

Andrews, D. W. K. (2002). Higher-Order Improvements of a Computationally Attractive k-Step Bootstrap for Extremum Estimators. Econometrica, 70(1):119–162.

Armstrong, T. B., Bertanha, M., and Hong, H. (2014). A fast resample method for parametric and semiparametric models. Journal of Econometrics, 179(2):128–133.

Bach, F. and Moulines, E. (2011). Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. Nips.

Byrd, R. H., Hansen, S. L., Nocedal, J., and Singer, Y. (2016). A Stochastic Quasi-Newton Method for Large-Scale Optimization. SIAM Journal on Optimization, 26(2):1008–1031.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. Journal of Econometrics, 115(2):293–346.

Davidson, R. and MacKinnon, J. G. (1999). Bootstrap Testing in Nonlinear Models. International Economic Review, 40(2):487–508.

Dvoretzky, A. (1956). On stochastic approximation. Technical report, Columbia University New York City United States.

Honoré, B. E. and Hu, L. (2017). Poor (Wo)man's Bootstrap. Econometrica, 85(4):1277–1301.

Kline, P. and Santos, A. (2012). A Score Based Approach to Wild Bootstrap Inference. Journal of Econometric Methods, 1(1).

Li, T., Kyrillidis, A., Liu, L., and Caramanis, C. (2018). Approximate Newton-based statistical inference using only stochastic gradients.

Mandt, S., Hoffman, M. D., and Blei, D. M. (2017). Stochastic gradient descent as approximate bayesian inference. The Journal of Machine Learning Research, 18(1):4873–4907.

Moffitt, R. and Zhang, S. (2018). Income Volatility and the PSID: Past Research and New Results. AEA Papers and Proceedings.

Moritz, P., Nishihara, R., and Jordan, M. (2016). A Linearly-Convergent Stochastic L-BFGS Algorithm. In Gretton, A. and Robert, C. C., editors, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, volume 51 of Proceedings of Machine Learning Research, pages 249–258, Cadiz, Spain. PMLR.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization, 30(4):838–855.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. The Annals of Mathematical Statistics, 22(3):400–407.

Ruppert, D. (1988). Efficient estimators from a slowly convergent Robbins-Monro procedure. School of Oper. Res. and Ind. Eng., Cornell Univ., Ithaca, NY, Tech. Rep, 781.

Related Literatures

- **Computationally Attractive Bootstrap:**

  Davidson and MacKinnon (1999); Andrews (2002); Kline and Santos (2012); Armstrong et al. (2014); Honoré and Hu (2017),. . .

  k-step re-sampling at a converged estimate of $\hat{\theta}_n$

- **Stochastic Derivative-Based Optimization:**

  Robbins and Monro (1951); Dvoretzky (1956); Ruppert (1988); Polyak and Juditsky (1992),[. . .], Bach and Moulines (2011); Moritz et al. (2016); Mandt et al. (2017),. . .

  interested in optimization on very large or online data sets

- **Stochastic Optimization and Inference:**

  Chernozhukov and Hong (2003),. . .

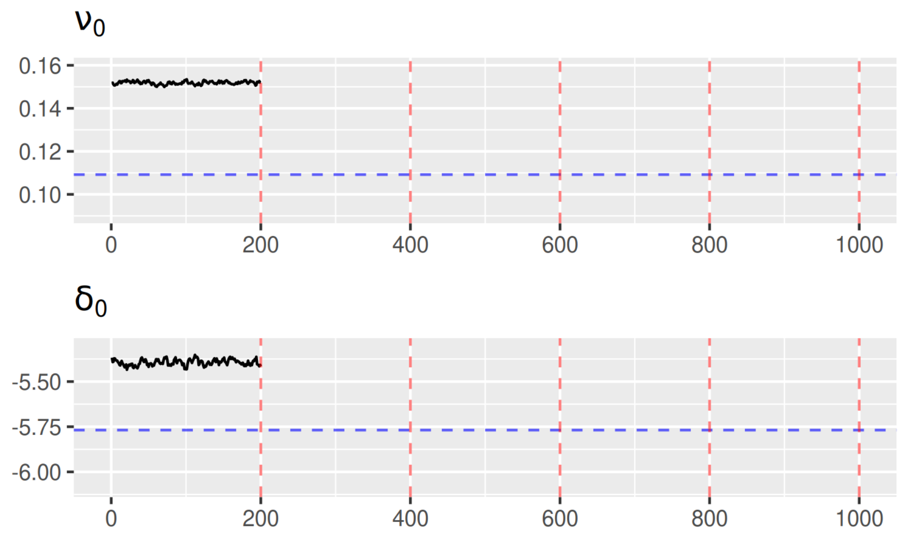
  MCMC similar to Simulated Annealing with a fixed temperature; (quasi)-posterior distribution asymptotically valid for inference

Main

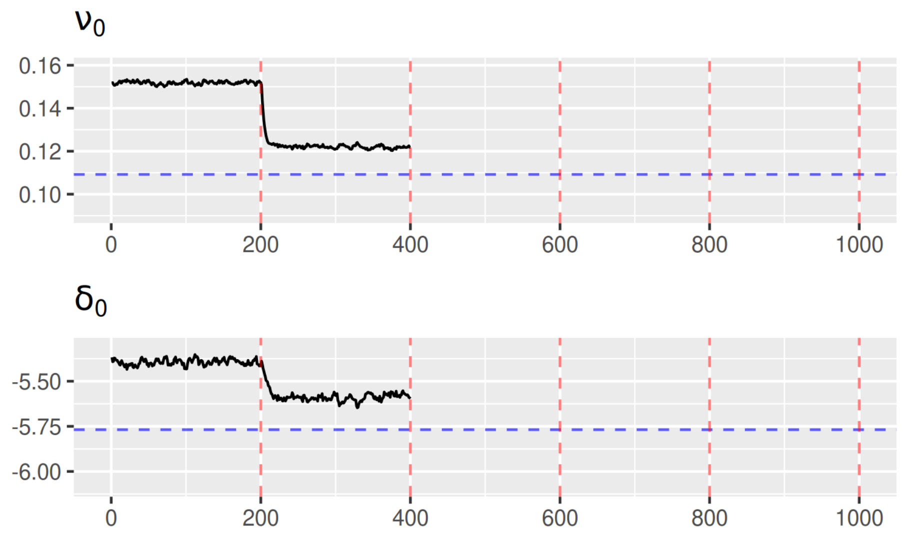
PSID Income Dynamics and Outlier Trimming

- Common empirical practice: remove extreme observations
- Here: the authors of the original paper trimmed the top and bottom 1% observations in each age-time group
  - are the results sensitive to the level of trimming?
- We look at a range of trimming levels: 0, 0.5, 1, 1.5 and 2%

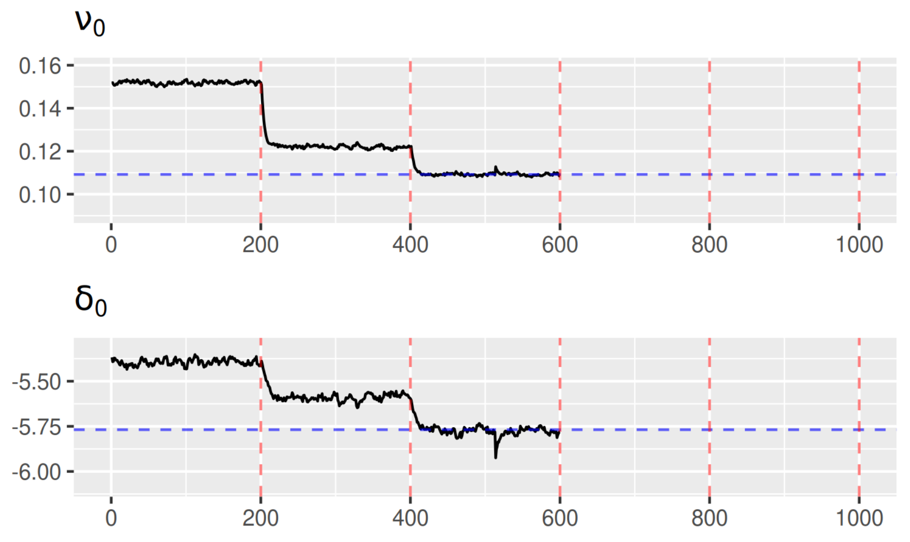
Influence of trimming: no trimming



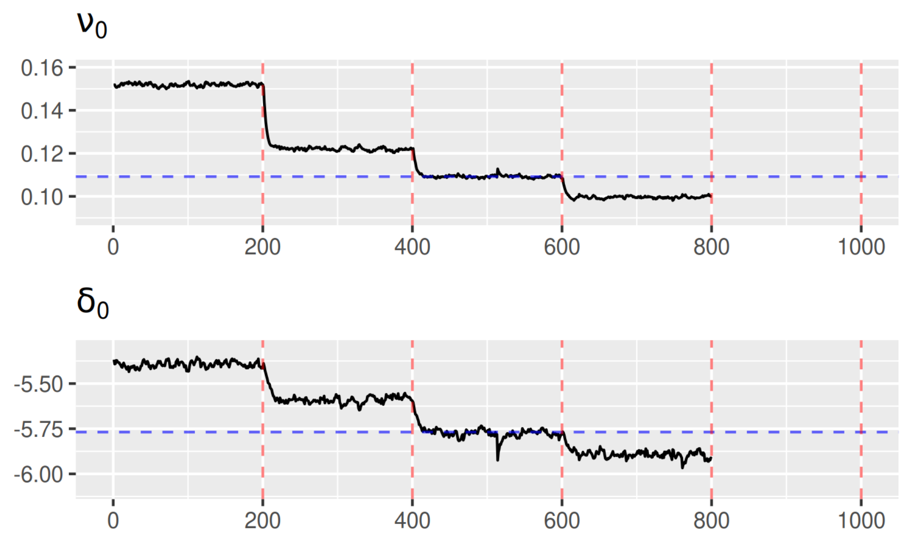
Influence of trimming: trim 0.5%



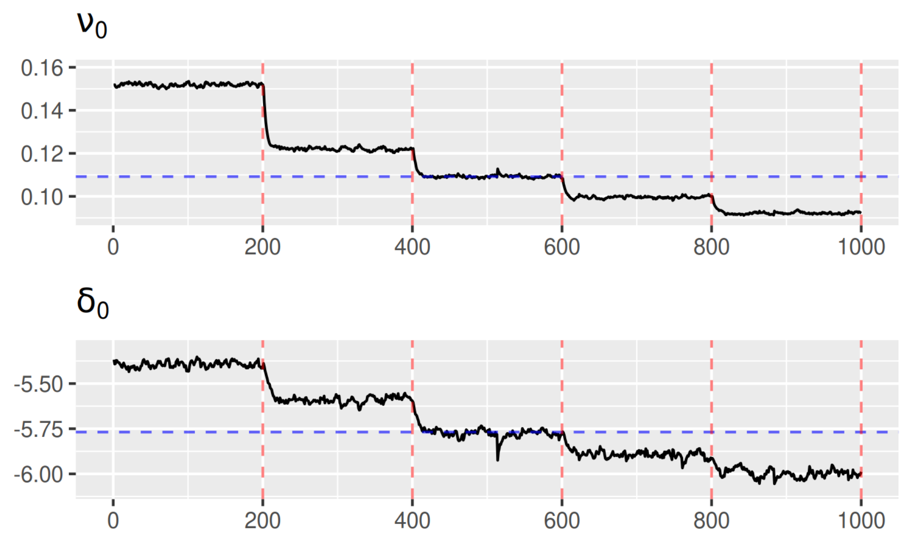
Influence of trimming: trim 1% (baseline)



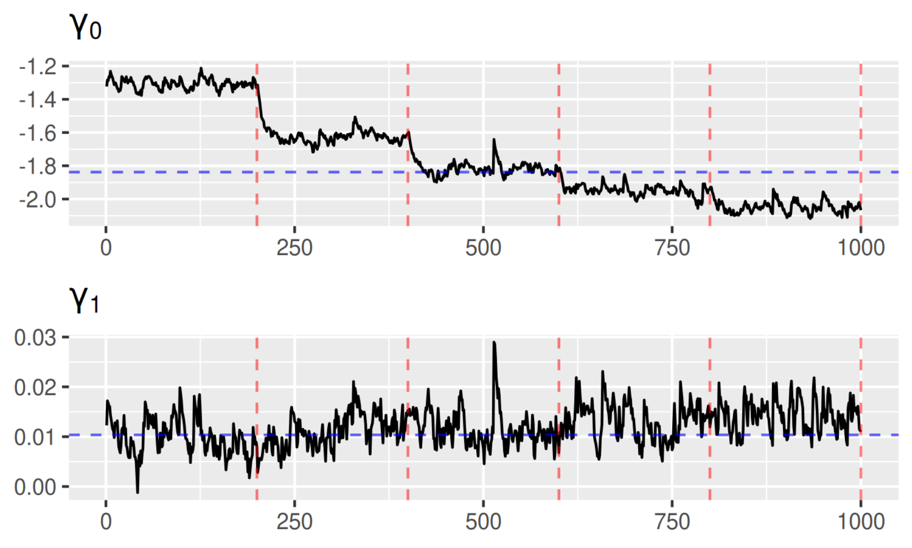
Influence of trimming: trim 1.5%



Influence of trimming: trim 2%



Influence of trimming



Age Groups

Simple Illustration: OLS estimation by Newton-Raphson

- OLS regression: $y_i = x_i'\theta + u_i$
- Sample objective: $Q_n(\theta) = \sum_{i=1}^n (y_i - x_i'\theta)^2/n$,
  $G_n(\theta) = -2\sum_i x_i(y_i - x_i'\theta)/n$, $H_n(\theta) = -2\sum_i x_i x_i'/n$
- **Newton-Raphson Iterations**:

$$\begin{aligned} \theta^b &= \theta^{b-1} + \gamma_b(\sum_i x_i x_i')^{-1}[\sum_i (x_i y_i - \theta^{b-1} x_i x_i')] \\ &= (1-\gamma_b)\theta^{b-1} + \gamma_b \hat{\theta}_n \end{aligned}$$

- For $\gamma_b = 1$ convergence after one iteration
- For $\gamma_b = \gamma \in (0,1]$ fixed, the error $\theta^b - \hat{\theta}_n$ is:

$$\theta^b - \hat{\theta}_n = (1-\gamma)^b[\theta^0 - \hat{\theta}_n]$$

Simple Illustration: OLS estimation by Gradient-Descent

- **Gradient Descent Iterations**:

$$\begin{aligned}\theta^b &= \theta^{b-1} + 2\gamma_b \sum_i x_i[y_i - x_i'\theta^{b-1}]/n \\ &= (I - 2\gamma_b \sum_i x_i x_i'/n)\theta^{b-1} + 2\gamma_b[\sum_i x_i x_i'/n]\hat{\theta}_n\end{aligned}$$

- Re-write the error $\theta^b - \hat{\theta}_n$ as:

$$\theta^b - \hat{\theta}_n = (I - 2\gamma_b[\sum_i x_i x_i'/n])(\theta^{b-1} - \hat{\theta}_n)$$

- For $\gamma_b = \gamma \leq \lambda_{\max}(\sum_i x_i x_i'/n)/2$ fixed, the error $\theta^b - \hat{\theta}_n$ is:

$$\theta^b - \hat{\theta}_n = (I - 2\gamma[\sum_i x_i x_i'/n])^b[\theta^0 - \hat{\theta}_n]$$

- Convergence after one iteration in one direction if $\gamma = \lambda_{\max}(\sum_i x_i x_i'/n)/2$

Issues with Mini-Batch Stochastic Newton-Raphson

- Deterministic case: $\theta^b_{NR} \to \hat{\theta}_n$ faster than $\theta^b_{GD} \to \hat{\theta}_n$
- Why is S-GD more popular than S-NR?
  - need to compute $[H_1^{(b)}(\theta)]^{-1}$ often (near)-singular
  - e.g. $x_i = (1, x_{i,1})$, $x_{i,1} \sim Bernoulli(p)$
    $\Rightarrow x_i x_i'$ singular wp. 1 for any $p \in [0, 1]$

$\Rightarrow$ mini-batch S-NR can be infeasible/unstable

- some solutions:
  - use more observations for H (Byrd et al., 2016; Li et al., 2018)
  - use accumulated gradient for scaling: adagrad, RMSprop,...