A Bayesian Approach to Imputing a Consumption-Income Panel Using the PSID and CEX

Matthew E. Smith
Hutchin Hill Capital

Christopher Tonetti
Stanford GSB

13 June 2014

Imputing a Panel of Consumption and Income

- Many important questions can only be answered with a panel of income and consumption data
- The PSID is a panel that includes individual characteristics, income, and food consumption
- The CEX is a repeated cross-section that contains individual characteristics, food and total consumption
- Our paper imputes total consumption in the PSID using information from the PSID and CEX

Blundell, Pistaferri, and Preston (2006)

- Consider the food demand equation

$$\ln F_{itj} = D'_{itj}\beta + \gamma \ln C_{itj} + e_{itj}$$

- F_{itj} is food consumption of agent i at time t in data set j
- D_{itj} is vector of agent i's characteristics at time t in data set j and some aggregate variables
- C_{itj} is total consumption of agent i at time t in data set j

BPP Imputation

1. Using data in the CEX, estimate the food demand equation

$$\ln F_{itj} = D'_{itj}\beta + \gamma \ln C_{itj} + e_{itj}$$

2. Imputed measure of consumption in the PSID is,

$$\widehat{C}_{itp} = exp\left(\frac{\ln F_{itp} - D'_{itp}\widehat{\beta}}{\widehat{\gamma}}\right)$$

Our Basic Idea

- Keep similar statistical model and innovate on imputation algorithm
- 'Stack' both data sets, Y_t
- Parameters shared across data sets θ
- Treat total consumption for each individual at each time as a hidden variable, grouped into X_t
  - Dimension of $X \approx 25,000 \Rightarrow$ very difficult sampling problem

The goal is to draw from $p(\theta, X_{1:T} | Y_{1:T})$

Contributions

- Efficiently use information from both data sets to impute total consumption in the PSID panel
  - Efficiency gains from one step procedure
  - Efficiency gains from likelihood based estimation
- Quantify uncertainty about imputation error
  - Posterior distribution vs. point estimate
  - Uncertainty matters: consumption inequality example
- Flexibility of methodology allows for gains from improved statistical model
  - Richer demand systems
  - More data: aggregate and disaggregate
  - Rich modelling of measurement error

Motivation: Back of the Envelope Calculation I

1. BPP estimate $\widehat{\gamma} = 0.8503$ with standard error $= 0.1511$
2. So one std around 0.8503 is $\approx (0.7, 1)$
3. Invert to get $1/\widehat{\gamma} \in (1, 1.42)$. Note $1/0.8503 = 1.17$
4. Hence the imputed measure of consumption in the PSID is,

$$\widehat{C}_{itp} = exp\left(1.17(\ln F_{itp} - D'_{itp}\widehat{\beta})\right)$$

5. 1 standard deviation (in γ) is

$$\widehat{C}_{itp} \in \left[exp\left(1(\ln F_{itp} - D'_{itp}\widehat{\beta})\right), exp\left(1.42(\ln F_{itp} - D'_{itp}\widehat{\beta})\right)\right]$$

6. $\rightarrow$ 1 std band spans roughly $\widehat{C}_{itp} \pm 20\%(\approx 1.42/1.17)$
7. 2 std is $\gamma \in (0.5478, 1.1522)$ or $1/\widehat{\gamma} \in (0.85, 1.83)$

Motivation:Back of the Envelope Calculation II

1. BPP actually include interaction terms in the food demand equation

$$\ln F_{itj} = D'_{itj}\beta + \ln C_{itj}(\gamma + H'_{itj}\alpha) + e_{itj}$$

2. Imputed measure of consumption in the PSID is,

$$\widehat{C}_{itp} = exp\left(\frac{\ln F_{itp} - D'_{itp}\widehat{\beta}}{\widehat{\gamma} + H'_{itp}\widehat{\alpha}}\right)$$

3. Error in $\widehat{\beta}$ and $\widehat{\alpha} \rightarrow$ *individual specific* confidence interval size

4. What about correlation in $(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma})$?

Model

$$
\begin{aligned}
f_{itx} &= D_{itx}^{'}\beta + \gamma c_{itx}^{*} + \sigma_f e_{itx} \\
c_{itx} &= c_{itx}^{*} + \sigma_{cx} v_{itx} \\
f_{itp} &= D_{itp}^{'}\beta + \gamma c_{itp}^{*} + \sigma_f e_{itp}
\end{aligned}
$$

$$
e_{itx}, e_{itp}, v_{itx} \sim N(0,1)
$$

$$
\begin{aligned}
\theta &:= \{\beta, \gamma, \sigma_f, \sigma_{cx}\} \\
X &:= \{c_{itx}^{*}, c_{itp}^{*}\} \\
Y &:= \{f_{itx}, f_{itp}, D_{itx}, D_{itp}, c_{itx}\}
\end{aligned}
$$

We can obtain analytical expression for likelihood.

Metropolis-Hastings

- Target: $f(s)$
- Want to be able to draw from f to characterize f when high dimensions
- Proposal: $s' \sim q(s, \cdot)$, with density $q(s, s')$
- Given chain is at some point $s^m = s$, set $s^{m+1} = s'$ with probability

$$\alpha(s, s') = \min \left\{ 1, \frac{f(s')q(s', s)}{f(s)q(s, s')} \right\}$$

- M-H generates a Markov chain whose invariant distribution is the target distribution

What proposal to use?

- Most popular: Random-Walk
- Proposal: $s' \sim N(s, \Sigma)$. Has density $q(s, s') = q(s', s)$
- Acceptance probability

$$\alpha(s, s') = \min\left\{1, \frac{f(s')}{f(s)}\right\}$$

- Does not work too well in high dimensions. Why?
  - Random walk proposal does not take into account shape of target density
  - Highly likely you "fall off a cliff" with the proposal
  - Σ is of dimension n^2
- Sampling challenges potentially why this approach is not popular.

Rubin 1988 on Multiple Imputation

"Multiple imputations ideally should be drawn according to the following general scheme. For each model being considered, the M imputations of the missing values, Ymis, are M repetitions from the posterior predictive distribution of Ymis, each repetition being an independent drawing of the parameters and missing values under appropriate Bayesian models for the data and the posited response mechanism."

Metropolis Adjusted Langevin Algorithm (MALA)

- Langevin SDE:

$$dS = \frac{1}{2}\nabla log(f(S))dt + dW$$

- This SDE has $f(s)$ as its invariant distribution
- MALA uses a (Euler) discretization as a proposal distribution in the Metropolis-Hastings Algorithm, i.e.,

$$S_{t+1} = S_t + \frac{h^2}{2}\nabla log(f(S_t)) + h\epsilon_{t+1}$$

- Thus, when $S = \{\theta, X\}$

$$q((\theta, X), (\cdot, \cdot)) = N((\theta, X) + \frac{h^2}{2}\nabla log(p(\theta, X|Y)), h^2 I)$$

More on MALA

- Other enhancements
  - Can have a constant or state-dependent preconditioning matrix, $\Lambda(S)$

$$S_{t+1} = S_t + \frac{h^2}{2}\Lambda(S_t)\nabla log(f(S_t)) + h\sqrt{\Lambda(S_t)}\epsilon_{t+1}$$

  - Truncate or further augment the drift term (formulas in paper)

$$S_{t+1} = S_t + \frac{h^2}{2}D(S_t) + h\sqrt{\Lambda(S_t)}\epsilon_{t+1}$$

- Theoretical out-performance of MALA relative to RWMH in high dimensions.
  - Performance depends on properties of target density
  - MALA has a higher optimal acceptance rate than RWMH (0.574 vs 0.234)
  - MALA explores posterior faster, "mixing time" is $O(n^{1/3})$ vs $O(n)$
    - For us 29 vs. 25,000

I now want to show you that this works and that it matters.

Estimation using Simulated Data

- Specify priors
- Simulate data using known $(C_{itx}^{*sim}, C_{itp}^{*sim}, \beta, \sigma_f, \sigma_{cx}, \gamma)$
- Use $(C_{itx}^{sim}, f_{itx}^{sim}, f_{itp}^{sim}, D_{itx}, D_{itp})$ as a dataset and try to estimate $(C_{itx}^{*sim}, C_{itp}^{*sim}, \beta, \sigma_f, \sigma_{cx}, \gamma)$
- Chain length 20 million, keeping every 5,000th draw, 4,000 for inference

Results on θ from Simulation Exercise

Table: Estimation Results for θ on Simulated Data

Parameter	True	Mean	StDev	2.5%	97.5%
β_6	0.0176	0.0153	0.0048	0.0062	0.0246
β_{16}	-0.3930	-0.4282	0.1022	-0.6263	-0.2326
β_{23}	0.0128	0.0070	0.0116	-0.0160	0.0294
β_{24}	0.0507	0.0387	0.0075	0.0241	0.0530
σ_f	0.2340	0.2348	0.0019	0.2310	0.2386
σ_{cx}	0.0206	0.0374	0.0065	0.0259	0.0515
γ	0.8343	0.8200	0.0039	0.8122	0.8273

Summary Statistics from Simulation Exercise: c_x^*

Table: Estimation Results for c_x^* on Simulated Data

Object	Mean	StDev	2.5%	97.5%
Truth	9.3980	0.4144	8.5646	10.2035
$\widehat{c}_{itx}^*$	9.3983	0.4136	8.5666	10.1967
Point($\widehat{c}_{itx}^*$)	9.3983	0.4119	8.5748	10.1972
StDev($\widehat{c}_{itx}^*$)	0.0375	0.0005	0.0366	0.0384

Summary Statistics from Simulation Exercise: c_p^*

Table: Estimation Results for c_p^* on Simulated Data

Object	Mean	StDev	2.5%	97.5%
Truth	10.1199	0.4566	9.2141	11.0123
$\widehat{c}_{itp}^*$	10.1424	0.4642	9.2202	11.0450
Point($\widehat{c}_{itp}^*$)	10.1424	0.3950	9.3514	10.9057
StDev($\widehat{c}_{itp}^*$)	0.2439	0.0028	0.2384	0.2493

Summary Statistics from Simulation Exercise: c_x^*

Table: Estimation Results for c_x^* on Simulated Data

Mean(ε_{itx})	StDev(ε_{itx})	Max(Abs(ε_{itx}))	mean(ε_{itx}^2)	StDev(ε_{itx}^2)
0.0003	0.0208	0.0752	0.0004	0.0006

Summary Statistics from Simulation Exercise: c_p^*

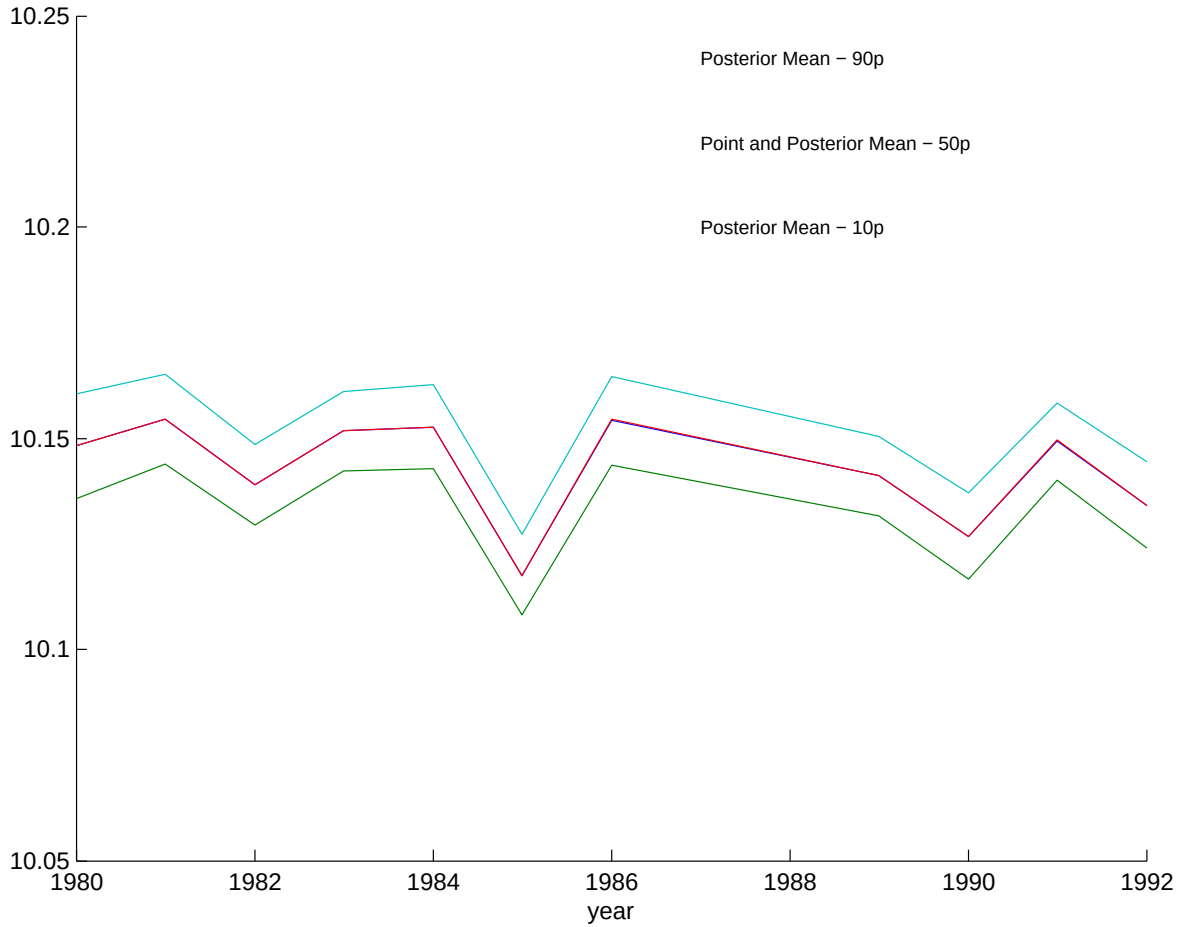
Table: Estimation Results for c_p^* on Simulated Data

Mean(ε_{itp})	StDev(ε_{itp})	Max(Abs(ε_{itp}))	mean(ε_{itp}^2)	StDev(ε_{itp}^2)
0.0225	0.2394	1.0312	0.0578	0.0847

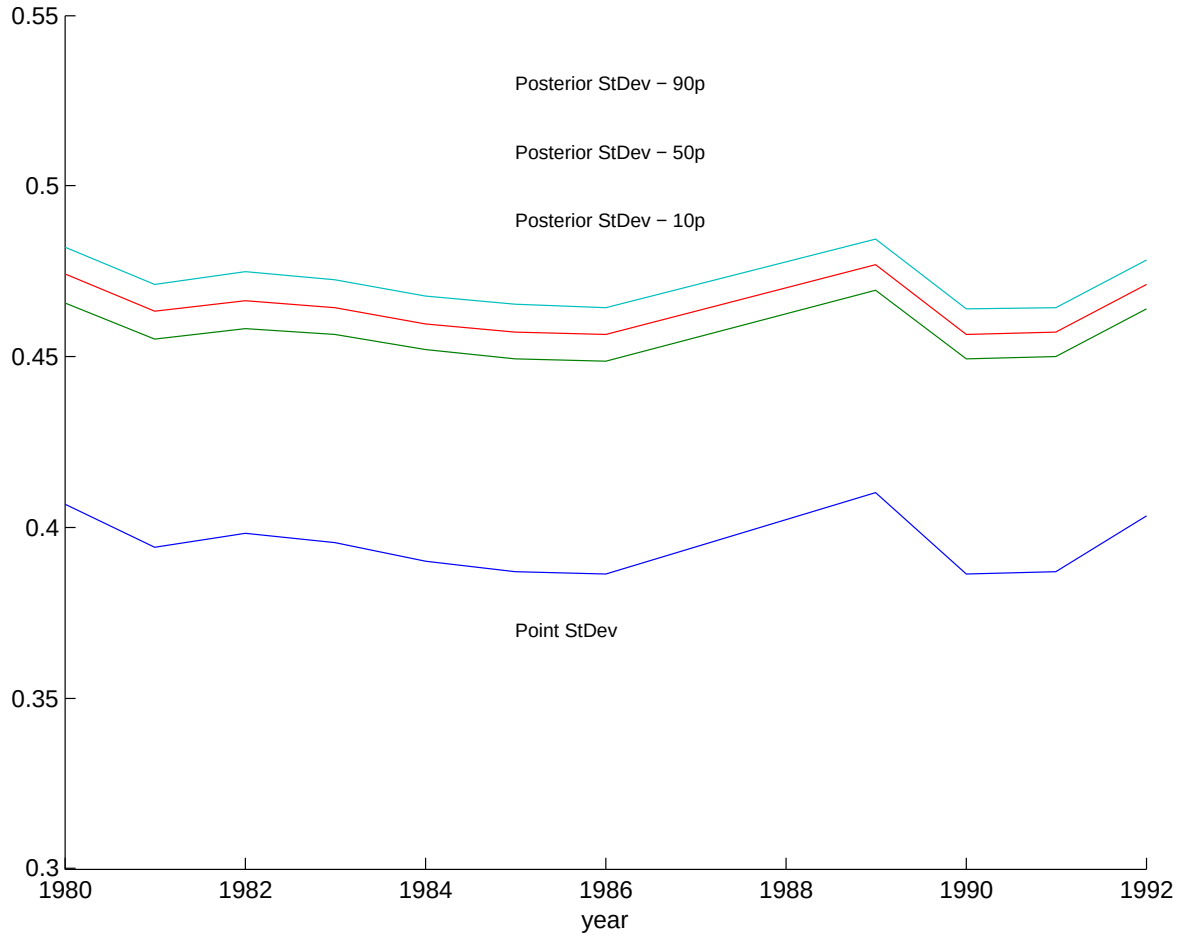
Estimate using PSID and CEX Data

- For now, use 1980-1992 data to match BPP sample
- Two methods of using the imputed “data”:
  1. Use the full posterior distribution (full information about the uncertainty surrounding imputed values)
  2. Use the mean of the posterior distribution as a point estimate

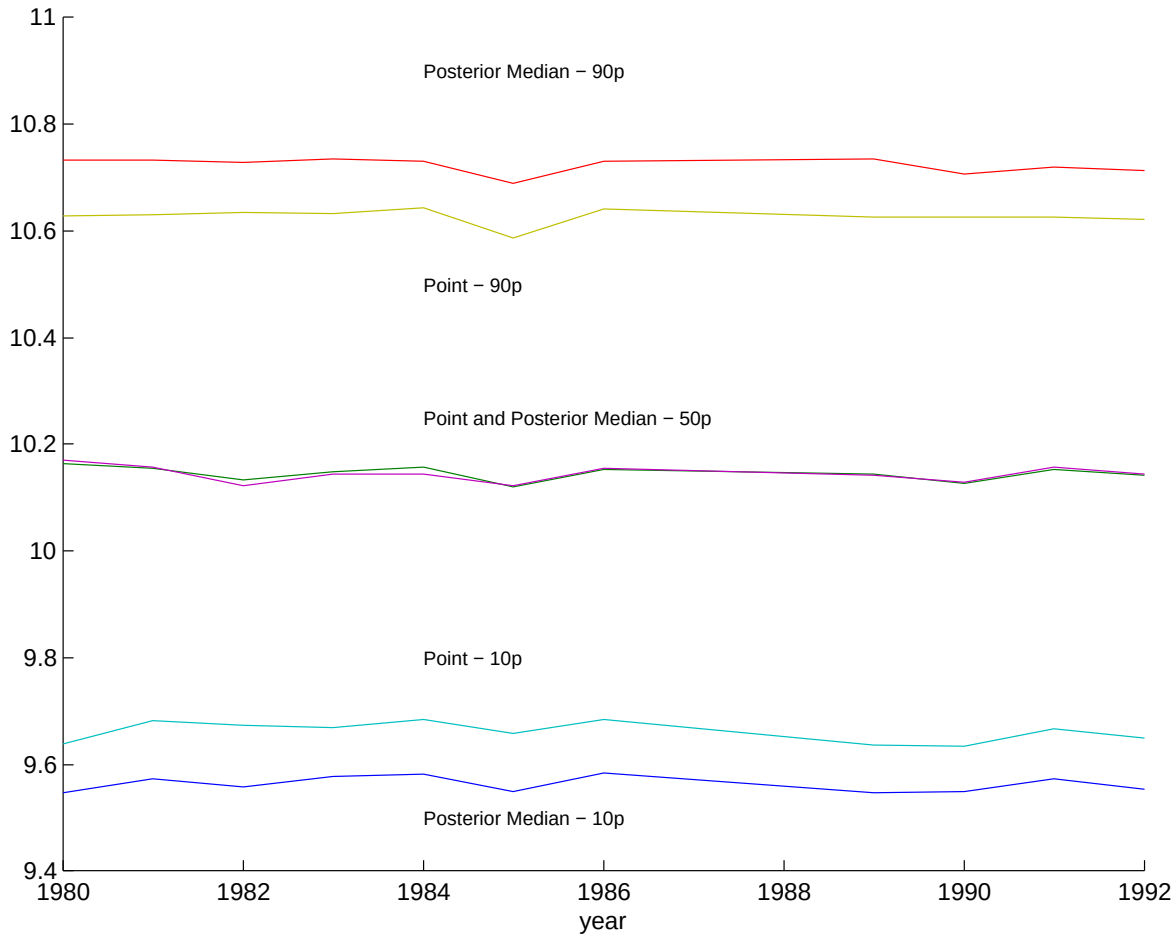
Evolution of Consumption Distribution: Mean



Evolution of Consumption Distribution: Std. Deviation



Evolution of Consumption Distribution: Quantiles



Summary

- We develop a Bayesian estimator to impute consumption into the PSID income panel
  - Overcome computational sampling challenges by implementing a Metropolis Adjusted Langevin Algorithm
- We establish the performance of the new estimator on simulated data
- We quantify relatively large uncertainty around imputed values
- We demonstrate that using our imputation methodology that incorporates imputation uncertainty could alter measurement of economically important objects
- Many ways to improve imputation and use data for economic applications

Flexible Approach Allows Many Potential Extensions

- Improving the imputation with more data
  - Subcategories of consumption, Diary and Survey CEX Data, Aggregate NIPA data, etc.
- Improving the imputation with a richer statistical model
  - Interaction terms
  - Richer demand systems
  - Richer measurement error modeling (Aguiar & Bils 2013)
  - Instrumental variables (error-in-variables - Lopes & Polson 2014)
  - Time series structure