Imputing Top-Coded Income Data in Longitudinal Surveys

Li Tan†

October 2017

The incomes of top earners are typically top-coded in survey data to protect individuals' identities. Common imputation methods used to recover top-coded income values are limited in several ways when applied to longitudinal data. I show that the accuracy of imputed income values for top earners in longitudinal surveys can be improved significantly by incorporating information from multiple time periods into the imputation process in a simple way. Moreover, I introduce an innovative, nonparametric empirical Bayes imputation method that further improves imputation quality. With a sample of individuals for whom incomes are pseudo top-coded (i.e., in which the exact income figures are accessible but temporarily expunged), I show that the Bayesian imputation method reduces the root mean squared error of imputed income values by 19-46% relative to standard approaches in the literature. After documenting this improvement in performance, I illustrate the benefits of the Bayesian method for investigating multi-year income inequality.

† Department of Economics, University of Missouri. I thank Cory Koedel, David Kaplan and Peter Mueser for helpful comments and suggestions. Financial supports from the University of Missouri Census Regional Data Center Interdisciplinary Doctoral Fellowship Program are gratefully acknowledged. All errors are my own. The Bayesian imputation method introduced in this paper can be implemented with the R package BayesImp available in https://github.com/ltanecon/BayesImp.

1. Introduction

Longitudinal surveys are used in research in many fields. Among the most widely used longitudinal surveys in the United States are the Survey of Income and Program Participation (SIPP), the Panel Study of Income Dynamics (PSID) and the National Longitudinal Survey of Youth (NLSY). A common feature shared by the surveys is that income values of the highest earners are censored to protect individuals' identities. This censoring practice is referred to as income top-coding.

Three types of methods have typically been employed in previous research to handle income top-coding in research applications. The first type circumvents the top-coding problem by excluding top earners from the analysis – e.g., by redefining the income distribution of interest, such as dropping the top-coded observations (e.g., see Jensen and Shore, 2015; Shin and Solon, 2011). The second type replaces top-coded income values with basic, non-model-based imputed values. Survey-provided "default" values are a frequently used example (e.g., Christie-Mizell, 2006; Heywood and O'Halloran, 2005); another example is the use of a "common multiplier," in which imputed values are obtained by multiplying the top-coding threshold value by an *ad hoc* common multiplier (e.g., Autor, Katz and Kearney, 2008; Lemieux, 2006; Katz and Murphy, 1992). The third method type is model-based. It estimates a population income distribution under parametric assumptions with the truncated sample and then randomly draws imputed income values from the right tail for top-coded earners. The distributional assumptions used to recover the population income distribution vary some across studies, but the fundamental approach is the same. I refer to this latter approach as the standard imputation method hereafter (e.g., Armour, Burkhauser and Larrimore, 2016; Attanasio and Pistaferri, 2014; Burkhauser et al., 2012; Kopczuk, Saez and Song, 2010).

Despite the popularity of these methods, they possess important limitations, particularly in longitudinal-data applications. For example, the standard imputation method – the most rigorous of

the three – has been shown to be effective in recovering the population income distribution accurately in any given year (Burkhauser et al., 2012), but it originates from cross-sectional data applications and is not suited to be applied to longitudinal data, which is quite common (Bhuller, Mogstad and Salvanes, 2011; Nichols, 2008; Orthofer, 2016). The reason is that it does not leverage longitudinal information to improve imputation accuracy.

This paper develops and tests methods that incorporate information in longitudinal data into the imputation process to account for income dependency across time periods within individuals. I first show that the standard imputation method can be improved significantly by incorporating available longitudinal information in a simple way. Without substantially modifying the analytic framework of the standard method, the suggested approach, which I call "rank-based imputation," generates a ranking algorithm based on longitudinal information to assign imputed income values to top-coded individuals. An appeal of rank-based imputation is its simplicity and effectiveness in improving imputation accuracy, but it has some limitations: perhaps most importantly, the "rank assignment" for individuals is static and does not account for the variation in ranking caused by income fluctuations over time.

Next, I introduce an innovative, Bayesian-based imputation method that further improves imputation accuracy by combining the analytic insights of the rank-based method and the methodological advantages of the nonparametric empirical Bayes framework recently developed by Gu and Koenker (2016, 2017; also see Koenker and Mizera, 2014). The advantage of this approach is that it leverages additional information from the sample to increase imputation accuracy, and it does not subject to the static-rank limitation of the rank-based method.

I compare these new imputation methods to the standard method using data from the 1996 SIPP. In addition to the actual income top-coding in the SIPP, applied to the roughly 0.5% of highest

earners, I pseudo top-code income values starting from the 98th percentile of the income distribution. This follows the NLSY top-coding protocol to create a realistic top-coding scenario (i.e., the NLSY top-codes the highest 2% of income observations). Individuals affected by the pseudo top-coding are valuable for examining the efficacy of the methods because their true income values are available in the SIPP but temporarily expunged, which allows for straightforward comparisons between imputed and actual income values. The results from the comparison of methods are not qualitatively sensitive to alternative pseudo top-coding thresholds (see Appendix A).

Using the pseudo top-coded sample, I show that the income values imputed by the standard method exhibit excessive income volatility within individuals.[1] This is because the values are imputed independently across survey waves for the same individuals. The bias in income volatility is greatly reduced by the rank-based method because it generates positive correlation between income observations within top-coded individuals over time. The Bayesian method further improves the estimation of income volatility. Both methods also improve imputation accuracy measured at the individual level. The Bayesian method improves accuracy the most, reducing the root mean squared error (RMSE) of the imputed income values by 19% relative to the standard imputation method.

After documenting the improved performance of the new methods I introduce, I illustrate the advantage of the Bayesian method for investigating multi-year income inequality with longitudinal survey data. This application is research and policy relevant (Aaberge and Mogstad, 2015; Björklund, Jäntti and Roemer, 2012; Friesen and Miller, 1983). I show that the standard method significantly under-predicts the income gap among the top 1% of earners because the independent income draws compress the distribution of aggregated multi-period income. Consistent with my documentation of

[1] Given the short time horizon (4 years) of the 1996 SIPP panel, I define income volatility simply as the standard deviation of individuals' income values, where the income in each year is re-centered to be mean zero (to remove the income trends).

the improved performance of the new methods, they are both effective in recovering the actual multi-period gap among high earners. Moreover, whereas results from the standard imputation method have the undesirable property that they are sensitive to the censoring threshold (e.g., 2% or 0.5%), both the rank-based and Bayesian methods produce similar results over a reasonable range of censoring values.

2. Background

In this section I review the three most commonly used approaches to handle income top-coding in research briefly mentioned above. Again, the first two approaches are quite simple and are sufficient when top-coded earners are not of interest. The last approach, the standard method, is more sophisticated and model-based. It works well for cross-sectional data but not longitudinal data, as will become clear, although it is frequently employed in analyses of longitudinal data (e.g., Attanasio and Pistaferri, 2014; Bhuller, Mogstad and Salvanes, 2011; Nichols, 2008; Orthofer, 2016).[2]

2.1 Circumventing the Top-Coding Problem by Ignoring Top Earners

Studies often exclude top earners by redefining the income distribution of interest and/or employing percentile-based measures. For example, numerous studies that model income inequality or income volatility exclude top-coded earners by either focusing on an interior range of income values – e.g., the 10^{th} to 90^{th} percentiles – or dropping top earners completely (e.g., Gottschalk and Danziger, 2005; Shin and Solon, 2011; Jensen and Shore, 2015). However, estimates based on redefined samples need to be interpreted with caution at the population level, given the potentially large income variability across the highest earners. Studies show that the highest 1% of earners have extraordinarily volatile income dynamics (Guvenen et al., 2015; Splinter, Bryant and Diamond, 2010). In fact, the

[2] Also see Jenkins et al. (2011), "Although our proposed methods [which belong to the standard imputation method] are illustrated with reference to the CPS, they are applicable more widely since the CPS is not the only survey with top-coded data. For example, in the USA, the National Longitudinal Survey of Youth top-codes some of its sources of income as does the Panel Study of Income Dynamics".

recent trend of increased income inequality in the U.S. is largely driven by the highest 1% of earners. This has been shown by studies that use tax return data from the Internal Revenue Service (IRS), which are not subject to income top-coding (Piketty and Saez, 2003).

2.2 "Default" Values or Common Multipliers

Perhaps the most frequently used imputation values for top-coded income in research are the "default" values provided in major surveys. Different surveys create default values differently. As an example, the 1996 SIPP uses the cell means of income values for individuals subject to top-coding, with the cell means calculated by gender, race and work status (full time/not full time) across all survey waves.[3] It is common for studies using longitudinal survey data to not explicitly address the top-coding issue, in which case the implication is that researchers are relying on the default values provided by the survey (e.g., Heywood and O'Halloran, 2005 and Christie-Mizell, 2006). Alternatively, Katz and Murphy (1992) replace top-coded income values with the top-coding threshold multiplied by 1.45, which is their "common multiplier" (also see Autor, Katz and Kearney, 2008; Lemieux, 2006). This type of method works well for some research questions, but not all: a limitation is that it minimizes the influence of top-coded earners (Jenkins et al., 2011).

For perspective on the substantial variation in income among the top earners that is removed using these methods, note that the income gap between the 99th and 99.9th percentiles is many times larger than the income gap between the 10th and 99th percentiles (Guvenen et al., 2015). By using the default value and common multiplier approaches, researchers essentially collapse the entire right tail of the income distribution into one or several isolated points. Not only is income variability across

[3] In contrast, the PSID places all top-coded values at the top-coding threshold, and the NLSY uses average income values for individuals subject to top-coding within each survey year.

individuals substantially under-predicted, information about income volatility within individual is also lost in this process.

2.3 The Standard Imputation Method

The standard imputation method covers variants of imputation methods widely used in many economic studies (Armour, Burkhauser and Larrimore, 2016; Attanasio and Pistaferri, 2014; Burkhauser et al., 2012; Kopczuk, Saez and Song, 2010; Western, Bloome and Percheski, 2008). These methods rely on the fact that unlike typical unobserved values, top-coded income observations reveal specific information about individuals' income levels; i.e., we know that the top-coded income values are larger than the threshold income value(s), but we do not know the exact figures. What we know from a top-coded observation in cross-sectional data is that a top-coded value belongs to an entire section of the income distribution – the right tail. The standard imputation method leverages this information and replaces top-coded income values with random draws from the right tail of the income distribution after re-constructing the truncated portion.

An important conceptual distinction that separates the standard imputation method from the default value or common multiplier methods described above is that the simpler methods are not stochastic. They only generate one unique dataset with imputed income values and belong to the "single imputation" family of methods. By contrast, an advantage of the standard imputation method is that the imputed income values are constructed in a stochastic way via "multiple imputation." This means that the method is executed multiple times to produce an independently generated series of output datasets with different imputed income values for individuals. The research question of interest is independently analyzed on all output datasets and results from these analyses are consolidated, often by taking the average. This general process is referred to as multiple imputation.

The standard imputation method estimates the unobserved portion of the income distribution under parametric assumptions, mostly under either the Generalized Beta of the Second Kind (GB2) or Pareto distributional assumption. Throughout this paper I employ the GB2 assumption because of its popularity and the well-established evidence from the literature that under this assumption, cross-sectional income distributions can be fitted quite well in a variety of settings (Bordley, McDonald and Mantrala, 1997; Feng, Burkhauser and Butler, 2006; Jenkins, 2009; Burkhauser et al., 2012).[4]

The probability density function (pdf) of the GB2 distribution is

$$f_{GB2}(y;a,b,p,q) = \frac{ay^{ap-1}}{b^{ap}B(p,q)(1+(y/b)^{a})^{p+q}} \qquad for\ y>0.^{5} \tag{1}$$

In equation (1), $B(\cdot)$ is the beta function, b is the scale parameter, and a, p, q are shape parameters. The cumulative distribution function (cdf) is denoted by $F_{GB2}(\cdot;a,b,p,q)$.

The parameters a,b,p,q can be estimated by maximum likelihood estimation (MLE). The log likelihood function for estimating the GB2 distribution parameters is

$$\begin{aligned}\Pi(a,b,p,q) = \sum\nolimits_{i=1}^{n} \{ & I(y_i \le C_1)\cdot \log F_{GB2}(C_1;a,b,p,q) + I(C_1 < y_i < C_2)\cdot \log f_{GB2}(y_i;a,b,p,q) \\ & + I(y_i \ge C_2)\cdot (1-\log F_{GB2}(C_2;a,b,p,q))\}.\end{aligned} \tag{2}$$

In equation (2), y_i is income for person i, C_1 is the income bottom-coding threshold (or a small number if there is no bottom-coding), C_2 is the income top-coding threshold, and $I(\cdot)$ is the

[4] The qualitative conclusions in this paper are not affected by using the Pareto distributional assumption (see Appendix A).

[5] Note that the beta distribution is not a special case of the GB2 distribution.

indicator function that equals one if the condition in parenthesis is satisfied, and zero otherwise.[6] Again, the procedure is best suited for cross-sectional data. Standard practice in the literature when this method has been applied to longitudinal data is to estimate a, b, p, q separately for each survey year (Bhuller, Mogstad and Salvanes, 2011; Nichols, 2008; Orthofer, 2016).

Figure 1 illustrates output from the standard imputation method compared to the kernel density of income for the wave-1 1996 SIPP sample. The dotted vertical line is the 2% (pseudo) top-coding threshold, and the top-coded observations are stacked at the threshold for the kernel density. Consistent with evidence from the literature, the kernel and GB2-MLE estimated densities before the threshold are very close. The imputed income values of the standard method are randomly populated from the estimated density beyond the threshold.

As noted previously, the standard method has been shown to be effective in capturing the unobserved right tail of the income distribution in cross-sectional applications (Burkhauser et al., 2012; Jenkins et al., 2011). However, despite its general effectiveness with cross-sectional data, it has critical limitations when applied to longitudinal data. Naturally, in longitudinal data, it is common for individuals to have multiple top-coded income values, and the standard imputation method greatly over-predicts the income volatilities of these individuals because it draws imputed values for the same individual across multiple periods independently. Clearly, the assumption of independence does not hold since income values for the same individual are correlated over time. The objective of this paper is to build available longitudinal information into the imputation process to account for this dependency, thereby improving accuracy.

[6] In Burkhauser et al. (2012), C_1 is set to be the 30th income percentile to "ensure that model fit is maximized at the top of the distribution." I follow their approach in the main setting, but using alternative C_1 values (i.e., the 1st or 70th income percentile) has no noticeable effect on method performance.

3. Methodology

In this section I develop the rank-based and Bayesian imputation methods. The standard imputation method is nested in the rank-based method, and analytic insights of the rank-based method is incorporated in the Bayesian method. Similar to the standard method, the rank-based and Bayesian methods also produce noise in the imputed income values and thus need to be executed in a multiple imputation framework. The rest of this section describes a single iteration of the rank-based and Bayesian imputation methods.[7]

3.1 The Rank-based Method

Income values observed in other periods are useful predictors of top-coded income(s) for the same individual, as is information about top-coding status in other periods. The rank-based method leverages available longitudinal information to improve inference. It shares common analytic insights with rank-based methods employed in other economic applications (e.g., Richiardi and Poggi, 2014).[8] The additional model complexity introduced by the rank-based method is very modest. It does not generate "new" imputed values, it only reassigns the imputed values generated by the standard method to top-coded earners based on available longitudinal information.

The first step of the rank-based method is identical to the version of the standard imputation method described in Burkhauser et al. (2012). For each survey year, let $\hat{a}_t, \hat{b}_t, \hat{p}_t, \hat{q}_t$ denote the estimated parameters for the log likelihood function in equation (2). The pdf of the estimated GB2 distribution for survey year *t* can therefore be written as $f_{GB2}(y; \hat{a}_t, \hat{b}_t, \hat{p}_t, \hat{q}_t)$. Let $\tilde{f}^t_{GB2}(\cdot)$ be the pdf of

[7] As an additional note, the imputation methods are illustrated without survey weights for presentational convenience. Survey weights are straightforward to incorporate conceptually and empirically.
[8] In Richiardi and Poggi (2014), they use a rank-based method to assign individual effects in binary response models.

the estimated GB2 distribution left-truncated at the top-coding threshold, and let l_t denote the number of top-coded income values in survey year *t*. The rank-based imputation method independently draws l_t imputed income values from $\tilde{f}^t_{GB2}(\cdot)$. These imputed values are assigned to top-coded earners at random, mirroring the standard method thus far. Let y_{it} denote the income for person i at time t inclusive of the imputed income values, in real dollars.

After standard imputation, the next step in the rank-based method is to remove the sample time trend in income with the following basic regression:

$$y_{it} = \lambda_t + \varepsilon_{it} \tag{3}$$

In equation (3), λ_t is a fixed effect for the survey period and ε_{it} is the residual after removing the flexible time trend in income.[9]

Let α_i denote the average time invariant income for individual i, and $\hat{\alpha}_i$ the estimator of α_i. The corresponding estimator $\hat{\alpha}_i$ is calculated as the sample average of the regression residuals, $\hat{\varepsilon}_{it}$. There are two layers of error when estimating α_i. The first layer is measurement error in the observed income data caused by top-coding: the exact income figures of the top-coded income values are unknown and populated based on the standard imputation method. Therefore, for individuals with at least one top-coded income value, $\hat{\alpha}_i$ is estimated partly on imputed income values from the standard method, which will bring random error into the rank-based procedure. This random error reflects the

[9] Demographic and socioeconomic variables are not included because sample-average coefficients are not particularly informative for top-coded earners. As an example, the gender income gap among the top-coded earners is much larger than the average gender income gap among the entire population (Guvenen, Kaplan and Song, 2014). That said, including demographic and socioeconomic variables in equation (3) has no qualitative bearing on method performance.

uncertainty in estimating α_i due to top-coding protocol. Similarly to the standard method, the impact of the random error can be reduced by running the rank-based method in a multiple-imputation framework.

The second layer of error is estimation error. Similar to most panel data applications, given the small number of observations within each individual and the incidental parameter problem, even ignoring the random error above, $\hat{\alpha}_i$ is still both imprecisely and inconsistently estimated. However, it contains useful longitudinal information to be exploited: namely, $\hat{\alpha}_i$ is positively correlated with individuals' actual income levels. For example, for individuals with a larger fraction of top-coded income values, $\hat{\alpha}_i$ will be higher on average. This simple fact provides the foundation for the ranking algorithm, in which the values of $\hat{\alpha}_i$ are used to re-assign new imputed values, which are the final values.

The process is as follows. First, as described thus far, for each survey year among individuals with any top-coded incomes, I obtain the $\hat{\varepsilon}_{it}$ values using the standard imputation method in each year. Next I produce $\hat{\alpha}_i$ and sort individuals on $\hat{\alpha}_i$. Finally, I detach individuals' initially imputed values, then re-assign the imputed values by rank such that individuals with the highest values of $\hat{\alpha}_i$ receive the highest imputed values. The residual series after this sorting is denoted as $\tilde{\varepsilon}_{it}$. Because the rank-based method uses the same distribution of imputed values as the standard method in any given year, but simply changes the values assigned to individuals based on rank, within any year the two approaches produce identical income distributions.

A special case of the ranking algorithm is individuals with top-coded income values in all sample periods, which account for 6.5% of all individuals with at least one top-coded value in my analytic sample from the 1996 SIPP (see below for sample construction). The income information for individuals in this special case is the same: all observations are top-coded. Consequently, rankings among these individuals are purely random. However, the rank-based method still offers an improvement over the standard method because these individuals are ranked systematically higher than partially censored individuals, which properly reflects their higher position in the income distribution generally.

A limitation of the ranking algorithm is that the ranking of individuals by $\hat{\alpha}_i$ is static, but in reality income rankings fluctuate quite often. Because of this limitation, the rank-based method under-predicts individual income volatility on average. In practice, however, the under-prediction is modest and the magnitude of bias in income volatility is much smaller in absolute value compared with the standard method. I further improve on this limitation with the Bayesian imputation method.

3.2 The Bayesian Imputation Method

The Bayesian method developed in this section incorporates additional information about other individuals from the same population to further improve imputation accuracy. It also addresses the problem of static rankings in the rank-based method. The approach builds on the non-parametric empirical Bayes framework used in Gu and Koenker (2017). This framework has been shown to be quite effective in various applications to project income over time when unobserved income heterogeneity in levels and volatility is important (Gu and Koenker, 2016, 2017; Tan and Koedel, 2017).

The Bayesian imputation method starts with the sorted residuals $\tilde{\varepsilon}_{it}$ produced by the rank-based method. Similar to Gu and Koenker (2017), I decompose $\tilde{\varepsilon}_{it}$ into an individual effect and transitory random error. The magnitudes of the transitory errors are individual-specific, which accounts for the income volatility of individuals

$$\tilde{\varepsilon}_{it} = \alpha_i + \sqrt{\theta_i}\xi_{it}, \qquad \xi_{it} \sim N(0,1). \tag{4}$$

Equation (4) can be viewed as a classic Gaussian location-scale mixture model, with α_i and θ_i serving as the location and variance parameters.[10] The transitory error term with individual income volatility ensures income rankings are not static. Per Kiefer and Wolfowitz (1956), under the assumption that α_i and θ_i are drawn i.i.d. from an unknown (to the statistician) joint distribution F, then F can be consistently estimated via nonparametric MLE. Koenker and Mizera (2014) provide an estimation framework based on convex optimization that can be applied efficiently to large datasets. I obtain the distribution F by the same estimation scheme (Gu and Koenker, 2016, 2017; Koenker and Gu, 2016; Koenker and Mizera, 2014).[11] The estimated F is subsequently treated as the prior distribution of α_i and θ_i since it represents the distribution of income heterogeneity for a generic individual in the sample without any individualized information.

The residual values of top earners are used as inputs to estimate F and taken from the imputed income values generated by the rank-based method. This raises the concern that estimated F may be

[10] In Appendix A, I discuss the sensitivity check if ξ_{it} is allowed to be serially correlated.

[11] Similar to these studies, the estimated pdf of the prior distribution is approximated by a 50*50 discrete function. More refined grids (e.g., 60*60) in the prior estimation increase the computing burden significantly but have little effect on the imputation performance.

unrepresentative of the sample of true incomes inclusive of top-coded values. However, this concern is eased by the fact that the rank-based method can recover the population distribution of income levels and volatility closely (see below). As an additional validity test, I calculate two prior distributions with the 1996 SIPP sample subject to different top-coding thresholds. The first prior distribution, F_1, is generated with the sample subject to pseudo top-coding, the second one, F_2, is generated with the sample subject to true top-coding. The two prior distributions are plotted in panels A and B of Figure 2; they are very similar. On average, imputation accuracy results are not meaningfully different using F_1 and F_2, which is not what one would expect if the method were biased by the level of noise in imputed values produced by the rank-based method.

After the prior distribution is established, I calculate the likelihood function for each individual to obtain the individualized posterior distribution. Let N_i denote the number of income observations for person i. The likelihood function is written as

$$\frac{1}{N_i}\sum_{t=1}^{N_i}\tilde{\varepsilon}_{it} \mid \alpha_i,\theta_i \sim N(\alpha_i,\theta_i / N_i), \tag{5}$$

$$\sum_{t=1}^{N_i}(\tilde{\varepsilon}_{it} - \frac{1}{N_i}\sum_{j=1}^{N_i}\tilde{\varepsilon}_{ij})^2 / \theta_i \mid \theta_i \sim \chi^2(N_i - 1). \tag{6}$$

Given Bayes' formula, the individualized posterior distribution, φ_i, is calculated based on the prior distribution F and the likelihood function.

After constructing φ_i, I introduce two new concepts, $\eta_i(\cdot)$ and $\boldsymbol{y}_i^I$, which are important intermediate products to populate final imputed income values using the Bayesian method. Let

$\eta_i(y_{i1}, y_{i2}, \cdots, y_{iN_i})$ denote the joint distribution of income for individual i across all time periods, and let $\boldsymbol{y_i^I} = (y_{i1}^I, y_{i2}^I, \cdots, y_{iN_i}^I)$ denote one draw of income values from $\eta_i(\cdot)$. The values within vector $\boldsymbol{y_i^I}$ can be obtained by executing the following stepwise procedure: (a) draw one set of (α_i^I, θ_i^I) from the individualized posterior distribution, φ_i; (b) simulate residual values $\tilde{\varepsilon}_{it}^I$ by the formula $\tilde{\varepsilon}_{it}^I = \alpha_i^I + \sqrt{\theta_i^I}\xi_{it}$, where ξ_{it} is randomly populated from the standard normal distribution; and (c) recover the imputed income values y_{it}^I by $y_{it}^I = \hat{\lambda}_t + \tilde{\varepsilon}_{it}^I$, in which $\hat{\lambda}_t$ is the estimated time fixed effect in period *t* from equation (3).

There are two important conceptual issues involving $\boldsymbol{y_i^I}$ and $\eta_i(\cdot)$. First, although the empirical Bayes framework is capable of producing imputed income values for individuals across all time periods, the imputation procedure need focus only on top-coded values. Let $\tilde{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{I}}$ denote the subset of $\boldsymbol{y_i^I}$ with top-coding for individual i.[12] Second, because we know the income value in a top-coded period is certainly larger than the top-coding threshold, the imputed income values can be generated from the truncated $\eta_i(\cdot)$. Thus, I keep simulating $\tilde{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{I}}$ via step (a) – (c) above until every element in the simulated $\tilde{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{I}}$ vector is at or above its top-coding threshold.[13] The resultant $\tilde{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{I}}$ vector is the vector of imputed income values provided by the Bayesian imputation method in a single

[12] In principle imputation could be performed for all income values for an individual, but it introduces unnecessary noise to impute the income not subject to top-coding.

[13] This repetition of steps is necessary and does not lead to systematic bias, because $\eta_i(\cdot)$ is the unconditional income distribution, but the imputed income values need to be populated from the income distribution conditional on top-coding status.

iteration of the Bayesian method. Again, the $\tilde{\boldsymbol{y}}_i^I$ vector contains random errors, which are addressed by running the Bayesian method in a multiple-imputation framework.

4. Data

I evaluate the performance of the imputation methods using data from the 1996 SIPP. The SIPP is useful because of its large sample size and popularity in research. As a nationally representative household survey, the SIPP consists of a continuous series of panels. Individuals were interviewed three times each year over a four-year period for the 1996 SIPP panel. Income figures over $50,000 during any survey interval (tri-annual wave) are top-coded. I limit the analytic sample to individuals born between 1936 to 1971 (or equivalently, aged 25 to 60 in 1996). This is a typical age restriction that many studies impose (Shin and Solon, 2011; Guvenen, et al., 2015). Zero-income records are dropped from the analytic sample on a wave-by-wave basis (i.e., individuals with at least one non-zero income value, including censored values, will remain in the analysis), since they are typically not used in imputation procedures (i.e., the GB2 distribution is defined on positive income only).

Around 0.5% of the income observations are truly top-coded in the SIPP. As noted previously, for the purpose of evaluating the imputation methods, I impose an additional pseudo top-coding on the analytic sample: for each survey wave, I also top-code the highest 2% of income observations inclusive of the truly top-coded values, mirroring the NLSY top-coding protocol. Consequently, in the analytic sample there are two types of top-coded observations, pseudo and truly top-coded. Imputed income values are generated across the two types of top-coded observations for all three imputation methods. The methods can be evaluated by comparing the imputed and actual income values based on the pseudo top-coded portion of the sample. Some suggestive comparisons can also be made across methods with the actual top-coded data, but they are less definitive because the true values are unknown.

Table 1 provides basic descriptive statistics for the analytic sample (with survey weights). Note that data panel is unbalanced, either due to individuals sometimes skipping survey interviews or having zero income observations. This is not a problem for the standard and rank-based methods, but a balanced panel is required to estimate the prior distribution F used in the Bayesian method (Koenker and Gu, 2016). Thus, in equation (4), I restrict $\tilde{\varepsilon}_{it}$ values to individuals with records in all 12 waves, and use this balanced panel of residuals to estimate F. My findings are not sensitive to alternative sample restrictions used to estimate F (see Appendix A). After obtaining F with the restricted sample, I proceed with the Bayesian imputation method using the full analytic sample. That is, I calculate the individual likelihood function via equations (5) and (6) for every individual with at least one top-coded income observation.[14]

5. Imputation Performance

I report results from tests designed to evaluate the relative performance of the three methods. Method performance is measured along two dimensions: distributional accuracy and individual imputation accuracy. Distributional accuracy refers to the ability of the method to produce imputed income values that are scattered along the lengthy right tail of the population income distribution, with the correct amount of income volatility on average. Individual accuracy is measured by the RMSE of the imputed values, calculated as the square root of the average squared distance between the imputed and actual income values for individuals. Smaller RMSE values suggest more accurate imputed values at the individual level. For ease of presentation, I exclude all individuals without any top-coded income data in any wave (inclusive of both pseudo and truly top-coded incomes).

[14] Three percent of top-coded observations are from individuals with only one or two total survey records, for whom the likelihood functions cannot be calculated reliably. For these observations, the imputed income values are drawn following the standard method.

In order to minimize the impact of random noise on imputation performance, the imputation procedures are executed independently 250 times. The imputation results are consolidated as is typical in the multiple imputation framework by taking the average.[15] In each single iteration, I create three duplicates of the analytic sample, denoted as the "standard," "rank-based" and "Bayesian" samples, by replacing the pseudo and truly top-coded observations with imputed values populated by the standard, rank-based and Bayesian methods, respectively. I refer to the sample where the pseudo top-coded observations are revealed, but the truly top-coded income values remain missing, as the "actual" sample, which I also use in the comparisons.

5.1 Distributional Accuracy

I measure distributional accuracy along two dimensions: income levels and income volatility.

5.1.1 Distributional Accuracy of Income Levels

To evaluate the distributional accuracy of income levels, I rely on extensive evidence from the literature showing that the standard imputation method can recover the population income distribution very well (Bordley, McDonald and Mantrala, 1997; Feng, Burkhauser and Butler, 2006; Jenkins, 2009; Burkhauser et al., 2012). Therefore, I set the distribution produced by the standard method as the "baseline" to evaluate distributional accuracy of the imputation methods wave-by-wave in the SIPP. The distributions of income levels generated by the rank-based and Bayesian method can be compared to that generated by the standard method. The advantage of this approach is that the evaluation of distributional accuracy is inclusive of truly top-coded individuals with income values in the top 0.5%.

[15] The distributional accuracy results are extremely similar across iterations. The individual imputation accuracy results vary modestly across iterations, but 250 simulation loops are enough to ensure stable results. All of the results are similar if calculated with or without survey weights; results with survey weights are shown (results without survey weights are available upon request).

Figure 3 presents the pooled income distributions across SIPP waves generated by the three imputation methods. The income distributions generated by the standard and rank-based methods overlap exactly because the rank-based method simply reassigns the imputed values generated by the standard method. The income distribution generated by the Bayesian method is also very close to the other two. Overall, Figure 3 suggests all three imputation methods perform similarly in terms of distributional accuracy of income levels.

5.1.2 Distributional Accuracy of Income Volatility

Evaluating the distributional accuracy of income volatility is more difficult because there is no similar evidence in the literature about how to create a "baseline" distribution of income volatility for truly top-coded individuals. Consequently, I use a different strategy to examine the distributional accuracy of imputed income volatility. Specifically, I limit the actual, standard, rank-based, and Bayesian samples to individuals for whom income volatility can be calculated accurately – that is, I use individuals with at least one pseudo top-coded income value but no truly top-coded values in any wave. This comparison is informative, although limited because individuals with truly top-coded income values are excluded.[16]

The distributional comparisons of income volatility are presented in Figure 4. The standard imputation method generates an excessively long tail of individuals with extraordinarily high income volatility, which does not appear in the actual data for the individuals in the sample. As mentioned above, this is because the standard method generates imputed income for same individual independently in each period and thus creates excess variance within individuals. This limitation does not apply to either the rank-based or Bayesian imputation methods. As alluded to above, the rank-

[16] I am in the process of obtaining access to fully uncensored SIPP data via the Regional Data Center at the University of Missouri. These data will allow for a complete evaluation of income volatility inclusive of individuals with truly top-coded incomes. See the conclusion for a brief discussion of extensions of this work.

based method modestly under-predicts income volatility on average due to its use of a static ranking. In contrast, the Bayesian method is not static and uses the full longitudinal sample in a more comprehensive and effective way. The income volatility distribution generated by the Bayesian imputation method is a very close match to the actual distribution.

5.2 Individual Imputation Accuracy

Individual imputation accuracy is defined by the average RMSE between imputed income levels and volatilities and individuals' true values. All calculations are based on the portion of the sample where the true values can be accurately obtained (i.e., pseudo top-coded values). Table 2 documents the substantial improvements in imputation accuracy accomplished by the rank-based and Bayesian imputation methods. Compared to the standard method, the rank-based method reduces imputation errors by 9% and 40% for income levels and volatility, respectively. The Bayesian imputation method reduces imputation errors even more, by 19% and 46% relative to the standard method.

In order to show the different types of biases caused by the imputation methods, in the last panel of Table 2 I present the over-prediction ratio for each method. The over-prediction ratio is defined by the share of imputed values (either income levels or volatility) larger than actual values. A ratio larger than 0.5 indicates the imputed values are systematically more likely to be higher than the actual values, while a ratio smaller than 0.5 indicates the opposite. Note that all three methods over-predict income levels, on average, for this sample. The reason is that the methods cannot fully account for the artificial censoring at the top 0.5 percent of income (i.e., they impute some values into that range even though only individuals with pseudo top-coded values from the 98^{th} to 99.5^{th} percentiles are included in the sample), which I must impose in order to compute the RMSE of the imputed values. The systematic over-prediction would be expected to disappear if the methods were applied

to impute values in the full upper tail, inclusive of the 0.5 percent of income records.[17] Nonetheless, the primary takeaway from the over-prediction ratios for income levels is that all of the methods perform similarly along this dimension, on average.

In contrast, the methods vary considerably in terms of how they predict income volatility within individuals, per the preceding discussion. The standard method consistently overstates income volatility by independently imputing censored values across waves, whereas the rank-based method understates income volatility by assigning imputed values based on the static ranking. The Bayesian method produces essentially balanced income volatility values for individuals.

Finally, to visually illustrate the improvements of the Bayesian method, in Figure 5 I plot the kernel densities of the imputed income values from 250 iterations for three example individuals with low, medium and high income values in the same survey year (in the pseudo top-coded range). The actual income values are labeled by the dotted vertical line. By construction, the kernel densities of the imputed values populated from the standard imputation method are the same for every individual. In contrast, the kernel densities from the rank-based and Bayesian methods exhibit highly idiosyncratic curvatures, with means much closer to the true income values.

6. Investigating Multi-year Income Inequality

In this section I illustrate the benefit of improved imputation for use in measuring multi-year income inequality with longitudinal survey data.[18] The multi-year income variable is constructed by averaging annual income across all four years of the 1996 SIPP panel.[19] Although measuring inequality

[17] Again, in future work with access to the uncensored SIPP data via the University of Missouri Regional Data Center, I will be able to confirm this expectation and provide full results.

[18] Alternatively, one can investigate this research question using administrative earnings records. However, public-use survey data has its own advantages: it is more accessible with a richer set of variables (e.g. health or educational information).

[19] More specifically, I impute values for all 12 tri-annual survey waves per the above-described processes, aggregate the values into annual measures, and then take the average across the four years.

by annual income is common, annual income is influenced by transitory income shocks and thus overstates persistent inequality (Friesen and Miller, 1983; Haider and Solon, 2006). Inequality measured by multi-year income is a more informative measure (e.g., Aaberge and Mogstad, 2015; Björklund, Jäntti and Roemer, 2012; Friesen and Miller, 1983).

The standard imputation method will under-predict longer-term income variability across top earners. To see this, let $\bar{Y}_i$ denote the average annual income value for individual i over the 4-year period. For individuals without top-coded observations, $\bar{Y}_i$ is a constant, but for a top earner with at least one top-coded observation, $\bar{Y}_i$ is a random variable even conditional on the sample values, due to random sampling from the imputed distribution. The income variability across top earners largely depends on the variance of $\bar{Y}_i$. Because imputed income values are independently populated across waves using the standard method, the variance of $\bar{Y}_i$ is too small.[20] The magnitude of the bias increases with the time horizon assuming a constant proportion of top-coded observations. By contrast, both the rank-based and Bayesian methods allow for income dependency within individuals and thus should not cause similar compression of estimated longer-term income across individuals. In fact, the rank-based method should overstate the variance in $\bar{Y}_i$ because of the static rankings per above, although as a practical matter this seems to have only a small impact on the results in this application (relative to the Bayesian method – see below).

To evaluate imputation performance, I define income inequality among high earners as the gap between the 99th and 99.9th income percentiles. Income percentiles below the 99th percentile are not included because they are much less affected by changes in the imputation procedure. In addition,

[20] This under-prediction in multi-year income variability across top earners is conceptually different from the over-prediction in income volatility within individuals, although both phenomena are caused by the same underlying reason.

the highest 1% of earners are particularly important for investigating total income inequality (Burkhauser et al., 2012; Piketty and Saez, 2003).

A challenge in testing method performance is that I cannot directly observe the top 0.5% of income records. Despite this limitation, some insight can be gained by comparing how the methods perform using two separately imputed samples: the 2% (pseudo top-coded) and 0.5% (truly top-coded) samples.[21] The reason that this comparison is useful for assessing method performance is that the actual population income distribution is the same under any top-coding threshold. Therefore, conditional on a valid imputation method, the percentile estimates from samples with different top-coding thresholds should be identical. Evidence of divergence in the imputed values due to different top-coding thresholds is an indicator of poor imputation performance.

Figure 6 presents the comparisons. Panel A-1 compares percentile income estimates from the 2% sample (the solid line) against estimates from the 0.5% sample (the dashed line) using the standard imputation method. The solid line diverges from the dashed line substantially, especially for higher income percentiles. This shows that the standard method produces meaningfully different results depending on the censoring threshold even though the underlying income distribution is the same. Panels A-2 and A-3 show the same comparison for the rank-based and Bayesian methods, respectively. Although the two lines in panels A-2 or A-3 do not overlap perfectly, they are very close. Overall, the Bayesian method is slightly better than the rank-based method, and they both perform much better than the standard method.

To illustrate the results further, I reorganize the information presented in panels A-1, A-2 and A-3 in panels B-1 and B-2. Whereas each panel A-1, A-2 and A-3 holds the imputation method fixed

[21] For the 2% sample, the 99th to 99.9th income percentile estimates primarily rely on the imputed income values, while for the 0.5% sample, both observed and imputed income values are important for the calculation.

and varies the censoring threshold, each panel B-1 and B-2 holds the censoring threshold fixed and varies the imputation method. Panel B-1 shows that the estimates from the standard method using the 0.5% sample are compressed and have less variability, albeit mildly, relative to the estimates from the rank-based and Bayesian methods. The difference is much more pronounced for the 2% sample in panel B-2, where there are more top-coded observations. The sensitivity of the standard imputation method to adjustments to the top-coding threshold is a clear indication of poor performance; conversely, the general robustness of the other methods is consistent with the above evidence showing that they generate more accurate imputed values.

7. Conclusion

I build on the standard imputation method to address its limitations when applied to longitudinal data. Tests of performance show that the rank-based and Bayesian imputation methods developed here greatly improve both distributional and individual imputation accuracy. These methods also allow for a more accurate investigation of multi-year income inequality, inclusive of top earners. The Bayesian method consistently performs better than the rank-based method, although both methods are a considerable improvement on the standard method, and the rank-based method has the virtue of being very simple to implement.

While my analysis makes the imputation improvements of the new methods quite clear, an obvious extension is to quantify the differences across methods more thoroughly by gaining access to the truly top-coded income values (i.e., the top 0.5%). To this end, I am in the process of obtaining access to the restricted-use 1996 SIPP panel, which includes fully uncensored income data, via the University of Missouri Census Regional Data Center (MU RDC). It will be straightforward to apply the comparative methods here to these new data, which will allow for more thorough documentation

of the improved accuracy of these new imputation methods. I was awarded a doctoral fellowship for AY2017-18 from the MU RDC to support this work.

Finally, the Bayesian method used here has other applications. As a related example, the income variable in the Current Population Survey (CPS) has a high missing rate, which has drawn much research attention (Bollinger and Hirsch, 2006; Hirsch and Schumacher, 2004). The Bayesian imputation method can be expanded on to impute the missing income values in the CPS. A strength of the Bayesian method is that it is transparent and accessible, and the framework can be modified in a straightforward way for related imputation work. During this age of rapid growth in the availability of data, there is great potential to improve research quality through improving the methods by which missing information in large datasets is recovered.

References

Aaberge, R., & Mogstad, M. (2015). Inequality in current and lifetime income. *Social Choice and Welfare*, 44(2), 217-230.

Armour, P., Burkhauser, R. V., & Larrimore, J. (2016). Using the Pareto distribution to improve estimates of topcoded earnings. *Economic Inquiry*, *54*(2), 1263-1273.

Attanasio, O., & Pistaferri, L. (2014). Consumption inequality over the last half century: some evidence using the new PSID consumption measure. *American Economic Review*, *104*(5), 122-126.

Autor, D. H., Katz, L. F., & Kearney, M. S. (2008). Trends in US wage inequality: Revising the revisionists. *Review of Economics and Statistics*, 90(2), 300-323.

Bhuller, M., Mogstad, M., & Salvanes, K. G. (2011). Life-cycle bias and the returns to schooling in current and lifetime earnings. Statistics Norway, Research Department Discussion Papers No. 666.

Björklund, A., Jäntti, M., & Roemer, J. E. (2012). Equality of opportunity and the distribution of long-run income in Sweden. *Social Choice and Welfare*, 39(2), 675-696.

Bollinger, C. R., & Hirsch, B. T. (2006). Match bias from earnings imputation in the Current Population Survey: The case of imperfect matching. *Journal of Labor Economics*, 24(3), 483-519.

Bordley, R. F., McDonald, J. B., & Mantrala, A. (1997). Something new, something old: parametric models for the size of distribution of income. *Journal of Income Distribution*, *6*(1), 5-5.

Burkhauser, R. V., Feng, S., Jenkins, S. P., & Larrimore, J. (2012). Recent trends in top income shares in the United States: reconciling estimates from March CPS and IRS tax return data. *Review of Economics and Statistics*, *94*(2), 371-388.

Christie-Mizell, C. A. (2006). The effects of traditional family and gender ideology on earnings: Race and gender differences. *Journal of Family and Economic Issues*, *27*(1), 48-71.

Daly, M. C., & Valletta, R. G. (2006). Inequality and poverty in United States: the effects of rising dispersion of men's earnings and changing family behaviour. *Economica*, *73*(289), 75-98.

Feng, S., Burkhauser, R. V., & Butler, J. S. (2006). Levels and long-term trends in earnings inequality: overcoming current population survey censoring problems using the GB2 distribution. *Journal of Business & Economic Statistics*, *24*(1), 57-62.

Friesen, P. H., & Miller, D. (1983). Annual inequality and lifetime inequality. *Quarterly Journal of Economics*, 98(1), 139-155.

Gottschalk, P., & Danziger, S. (2005). Inequality of wage rates, earnings and family income in the United States, 1975–2002. *Review of Income and Wealth*, *51*(2), 231-254.

Gu, J., & Koenker, R. (2016). Empirical bayesball remixed: empirical bayes methods for longitudinal data. *Journal of Applied Econometrics*, *32*(3), 575–599.

Gu, J., & Koenker, R. (2017). Unobserved heterogeneity in income dynamics: an empirical Bayes perspective. *Journal of Business & Economic Statistics*, *35*(1), 1-16.

Guvenen, F., Kaplan, G., & Song, J. (2014). The glass ceiling and the paper floor: Gender differences among top earners, 1981-2012. NBER Working Paper No. 20560.

Guvenen, F., Karahan, F., Ozkan, S., & Song, J. (2015). What do data on millions of US workers reveal about life-cycle earnings risk? NBER Working Paper No. 20913.

Haider, S., & Solon, G. (2006). Life-cycle variation in the association between current and lifetime Earnings. *American Economic Review*, 96(4), 1308-1320.

Heywood, J. S., & O'Halloran, P. L. (2005). Racial earnings differentials and performance pay. *Journal of Human Resources*, *40*(2), 435-452.

Hirsch, B. T., & Schumacher, E. J. (2004). Match bias in wage gap estimates due to earnings imputation. *Journal of Labor Economics*, 22(3), 689-722.

Jenkins, S. P. (2009). Distributionally-sensitive inequality indices and The GB2 income distribution. *Review of Income and Wealth*, *55*(2), 392-398.

Jenkins, S. P., Burkhauser, R. V., Feng, S., & Larrimore, J. (2011). Measuring inequality using censored data: a multiple-imputation approach to estimation and inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(1), 63-81.

Jensen, S. T., & Shore, S. H. (2015). Changes in the distribution of earnings volatility. *Journal of Human Resources*, *50*(3), 811-836.

Katz, L. F., & Murphy, K. M. (1992). Changes in relative wages, 1963–1987: supply and demand factors. *Quarterly Journal of Economics*, 107(1), 35-78.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, *27*(4), 887-906.

Koenker R. & Gu, J. (2016). REBayes: An R Package for Empirical Bayes Mixture Methods. Working Paper.

Koenker, R., & Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association*, *109*(506), 674-685.

Kopczuk, W., Saez, E., & Song, J. (2010). Earnings inequality and mobility in the United States: evidence from social security data since 1937. *Quarterly Journal of Economics*, *125*(1), 91-128.

Lemieux, T. (2006). Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill?. *American Economic Review*, 96(3), 461-498.

Nichols, A. (2008). Trends in Income Inequality, Volatility, and Mobility Risk: Via Intertemporal Variability Decompositions, Urban Institute, Washington DC.

Orthofer, A. (2016). Wealth inequality in South Africa: Evidence from survey and tax data. REDI3x3 Working Papers No. 15.

Piketty, T., & Saez, E. (2003). Income Inequality in the United States, 1913-1998. *Quarterly Journal of Economics*, *118*(1), 1-39.

Richiardi, M., & Poggi, A. (2014). Imputing Individual Effects in Dynamic Microsimulation Models An application to household formation and labour market participation in Italy. *International Journal of Microsimulation*, 7(2), 3-39.

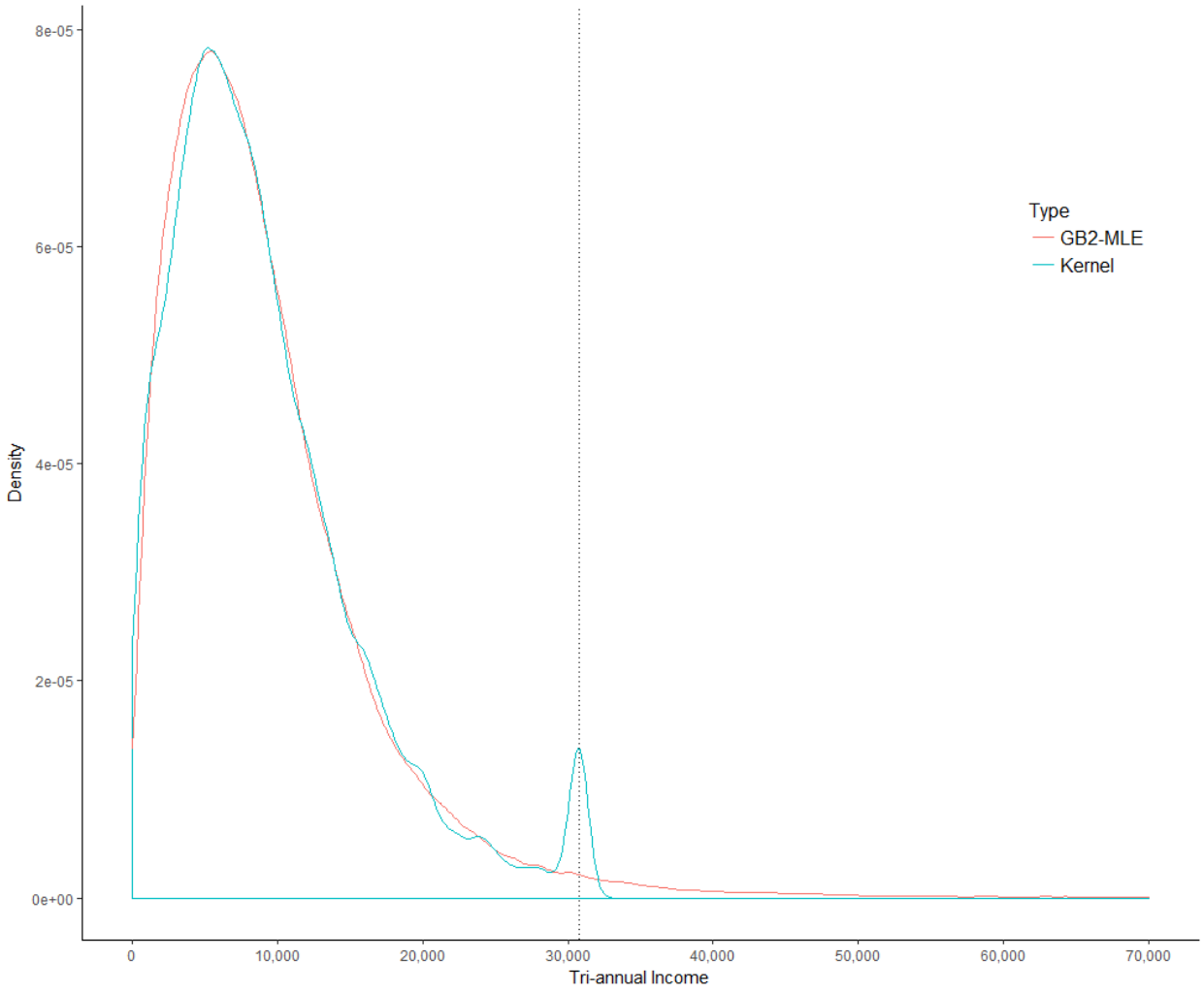
Shin, D., & Solon, G. (2011). Trends in men's earnings volatility: What does the Panel Study of Income Dynamics show?. *Journal of Public Economics*, *95*(7), 973-982.

Splinter, D., Bryant, V., & Diamond, J. W. (2010). Income Volatility and Mobility: US Income Tax Data, 1999-2007. *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association*. *102*(1), 1-10.

Tan, L., & Koedel, C. (2017). New Estimates of the Redistributive Effects of Social Security. University of Missouri Working Paper No. WP 17-01.

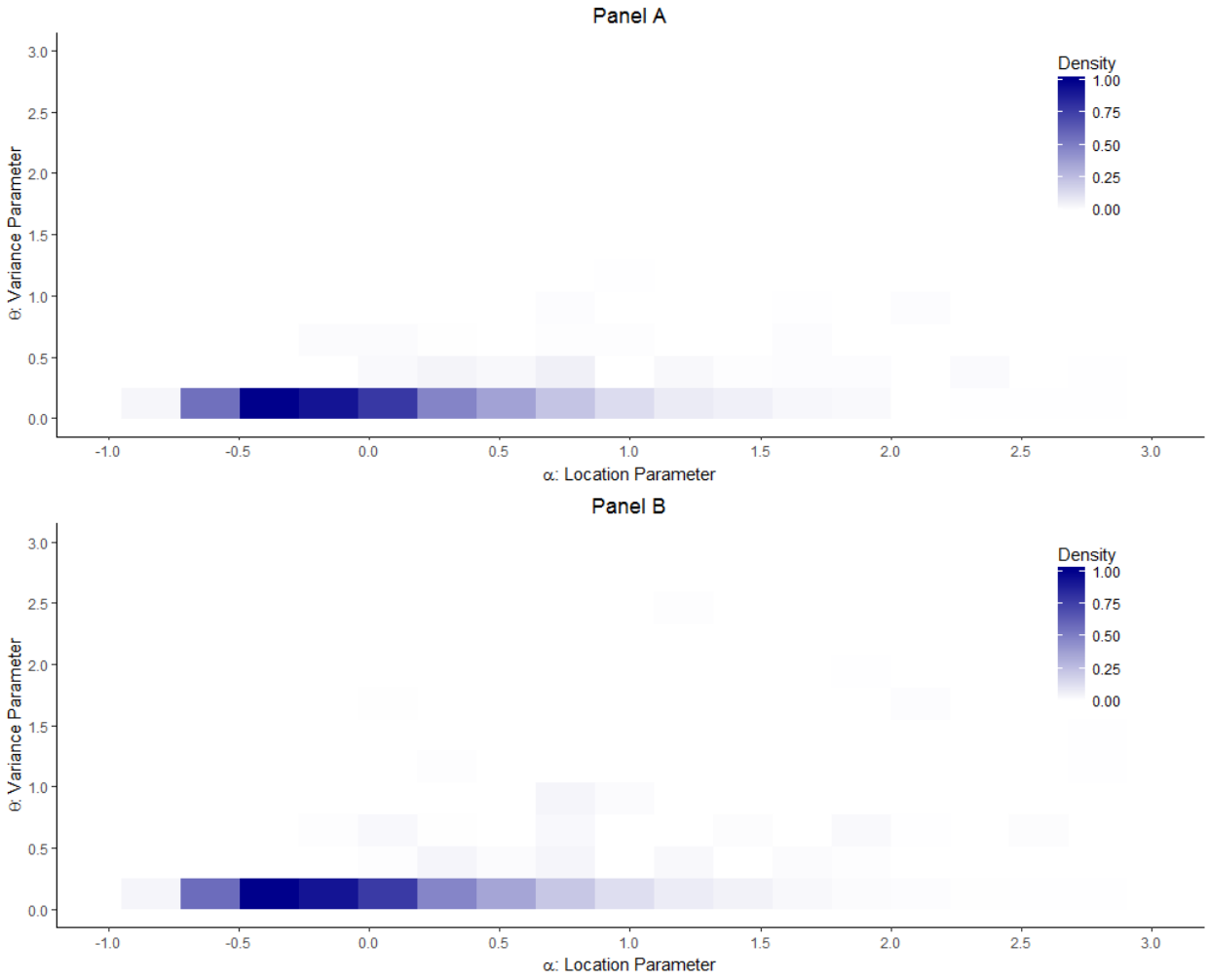
Western, B., Bloome, D., & Percheski, C. (2008). Inequality among American families with children, 1975 to 2005. *American Sociological Review*, *73*(6), 903-920.

Figure 1. Kernel and GB2-MLE Estimated Density of Income for the Wave-1 1996 SIPP Sample



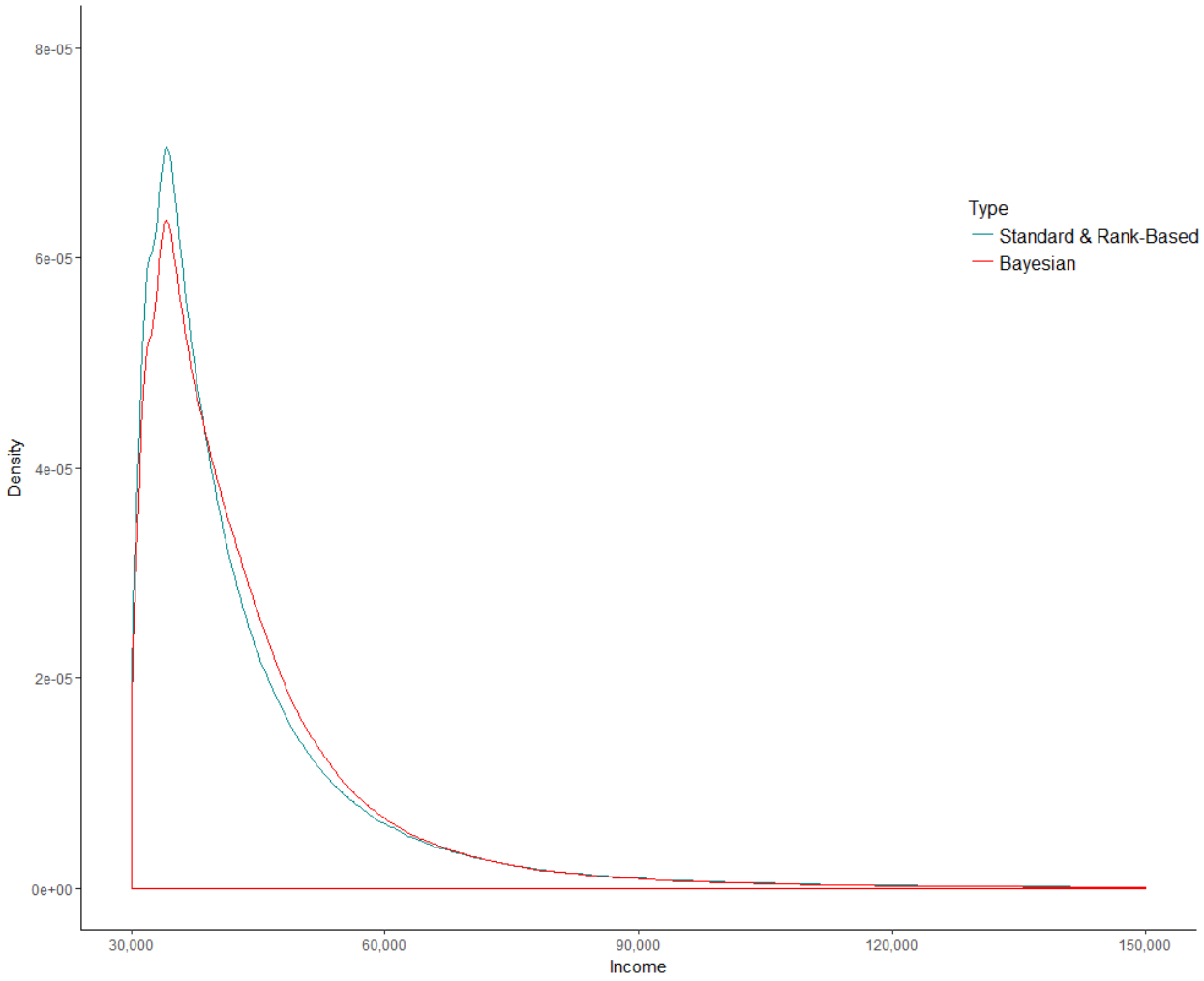
Notes: The dotted vertical line indicates the 2% (pseudo) top-coding threshold. Top-coded observations are stacked at the threshold. Income values are based on four-month intervals per the tri-annual surveys in the 1996 SIPP.

Figure 2. Estimated Prior Distribution of the Analytic Sample



Notes: The highest-density cell is rescaled to a density value of 1.0. The numbers on the horizontal axis are scaled in $10,000 units (in 1996 real dollars). The figure concentrates on the noticeable part of the density only; density values in the $\alpha > 3$ range are too small to be visible. Panel A is calculated based on the analytic sample under pseudo (2%) top-coding, and Panel B is calculated based on the analytic sample under true (0.5%) top-coding.

Figure 3. Distributions of Income Levels Generated by the Imputation Methods



Notes: The density plot of the standard imputation method overlaps perfectly with that of the rank-based method as described in the text.

Figure 4. Distributions of Income Volatility Generated by the Imputation Methods

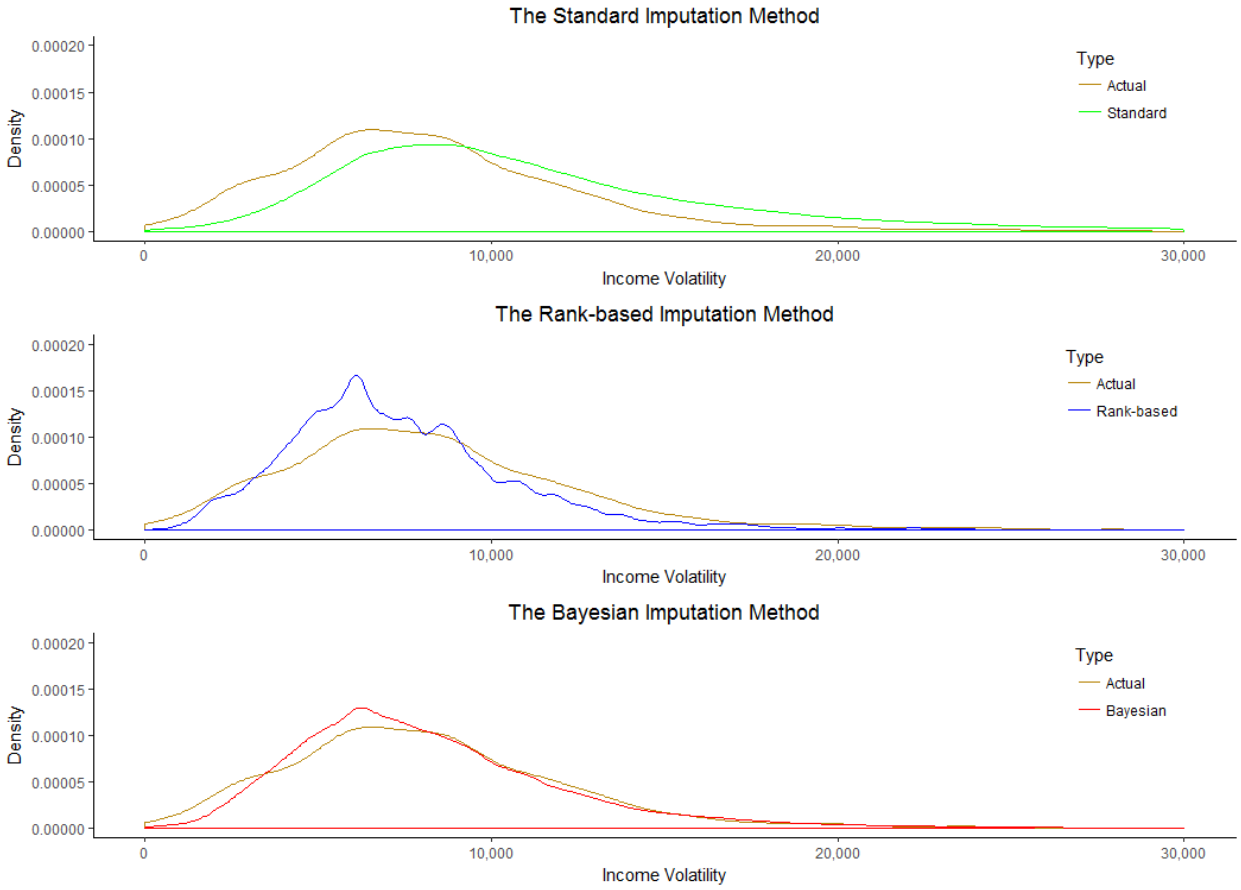
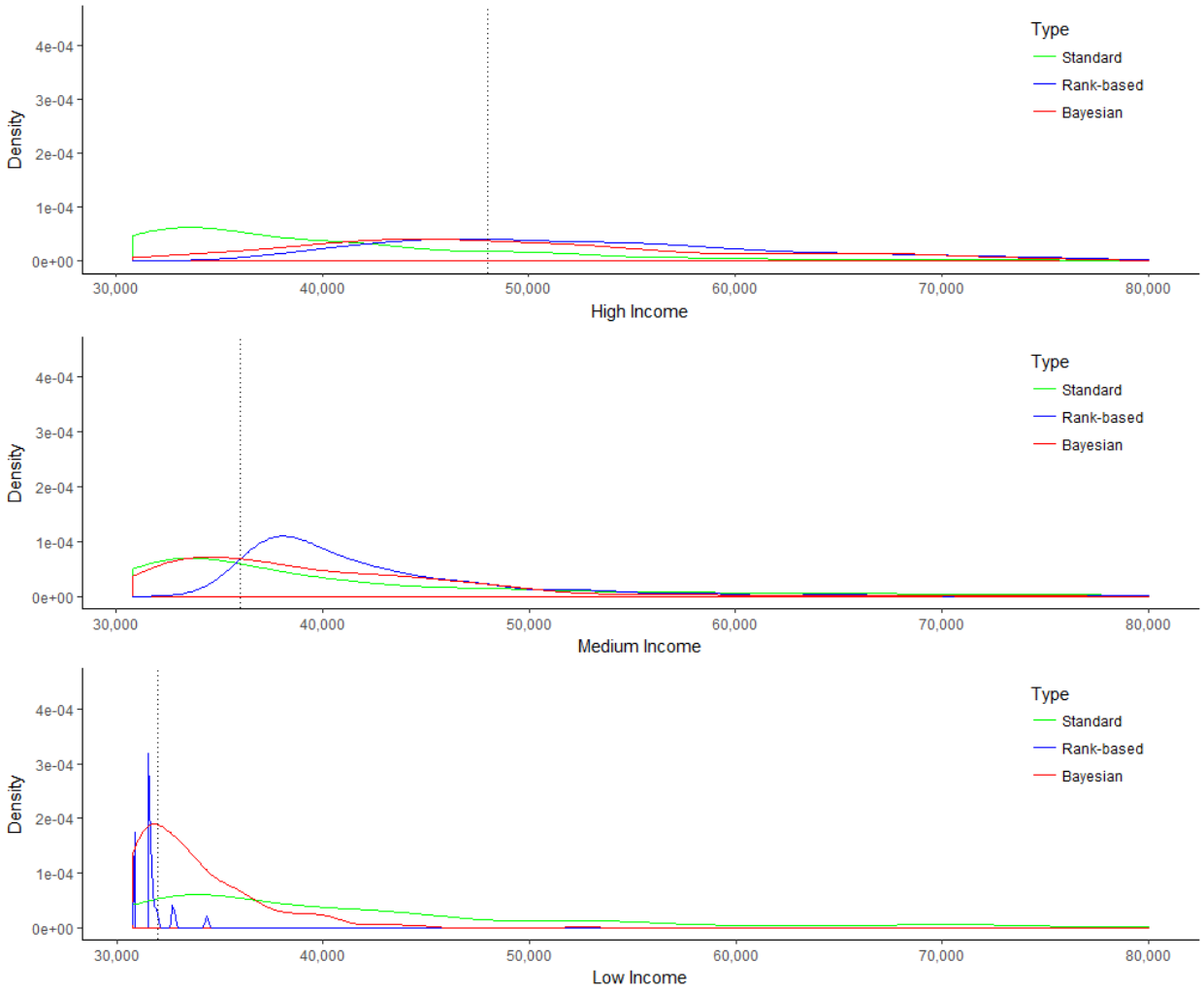
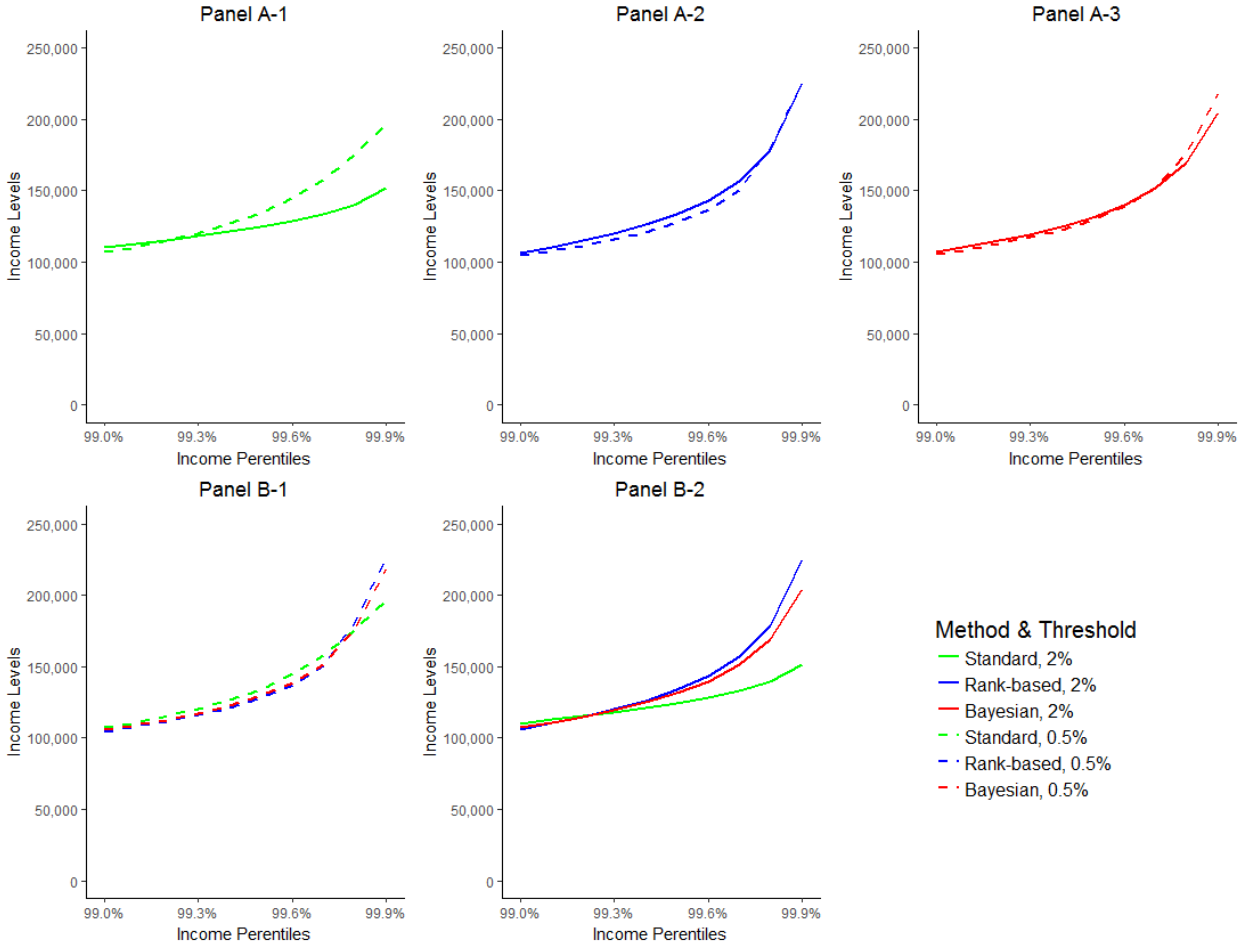


Figure 5. Kernel Densities of Imputed Income Values for Three Individuals



Notes: The dotted vertical line marks the actual income value. The low, medium and high income categories are among pseudo top-coded income values.

Figure 6. Comparison of Income Differences from the 99th to 99.9th Percentiles of the Longer-term Income Distribution Using Different Imputation Methods and Different Top-Coding Thresholds



Note: Panel A-1 compares the percentiles of average annual income over the 4-year period calculated using the 2% and 0.5% samples and standard imputation. Panel A-2 and A-3 show similar comparisons for the rank-based and Bayesian methods, respectively. Panel B-1 and B-2 reorganize the same information to compare method performance holding the top-coding threshold fixed.

Table 1. Basic Descriptive Statistics for the 1996 SIPP Sample

	SIPP 1996
Female	0.48
Non-Hispanic Black	0.13
Hispanic	0.11
Non-Hispanic Non-Black	0.76
Less than High School	0.11
High School Graduate	0.63
4-Year College Graduate	0.17
Master Degree or Higher	0.09
Number of Individuals	43,238
Average Income Per Record	9,859
Ratio of Records Top-coded	0.005
Number of Records	334,071

Notes: Column 1 includes all individuals in 1996 SIPP born between 1936 to 1971 with at least one record. Only records with non-zero income are kept. Each person can have up to 12 records. Each record is based on a 4-month interval.

Table 2. Individual Imputation Accuracy of the Three Imputation Methods

Method	RMSE		RMSE Reduction		Over-prediction Ratio	
	Income Level	Income Volatility	Income Level	Income Volatility	Income Level	Income Volatility
Standard	19,708	9,037			0.58	0.72
Rank-based	17,931	5,432	9.02%	39.88%	0.59	0.31
Bayesian	15,998	4,893	18.82%	45.85%	0.61	0.48

Notes: The RMSE panel shows the average RMSE of imputed income values produced by each imputation method. All calculations are based on the portion of sample where the true values can be accurately obtained (i.e., pseudo top-coded values). The RMSE Reduction panel shows the average RMSE reductions of the rank-based or Bayesian imputation methods relative to the standard method. The over-prediction panel shows the average proportion of observations for which the imputed values are larger than the actual values.

Appendix A

Sensitivity Analyses

This appendix reports on sensitivity analyses under four different scenarios mentioned in the main text. Individual imputation accuracy results are shown in Tables A.1 through A.4. For completeness, distributional accuracy results (Figure A.1 to A.8) are shown as well, although they are less sensitive to the adjustments considered.

First, in the main setting the standard imputation method is implemented under the GB2 distribution assumption. I replace this with the Pareto distribution assumption, where I estimate the parameter of the Pareto distribution by equation (4) in Armour, Burkhauser and Larrimore (2016). The Pareto distribution has just one parameter and is therefore less flexible than the GB2 distribution. Burkhauser et al. (2012) show that the Pareto distribution assumption is not ideal in the sense that the estimated parameter is sensitive to the top-coding threshold used. This is consistent with my findings using the SIPP, as the imputation methods perform generally worse under the Pareto assumption. As shown in Figure A.1, the Pareto distribution puts slightly more mass on the highest income values relative to the GB2 distribution. As a result, the standard method over-predicts income volatility even more. Income volatility as estimated using imputed values from the Bayesian method is also more biased, as shown in Figure A.2. From Table A.1, the individual accuracy results are also much worse for all three methods. Still, the Bayesian method performs the best of the three, and the advantage of the Bayesian method relative to the standard method in terms of individual imputation accuracy is even higher.

Second, I change the pseudo top-coding threshold from 2% to either 1% or 3% (always inclusive of the top 0.5%, as in the main text). Table A.2 shows that the relative advantage of the

Bayesian method remains the same. Thus, my findings are not qualitatively sensitive to reasonable changes to the top-coding threshold.

Third, recall that in order to estimate the prior distribution F employed in the Bayesian method, a balanced sample of residuals is needed (Koenker and Gu, 2016). In the main text I restrict the estimation sample for this portion of the analysis to individuals with income observations in all 12 waves of the SIPP. Here I use an alternative restricted sample created by randomly selecting 6 residuals from all individuals with income records in 6 or more waves. This alternative sample is shorter but with more individuals. The imputation accuracy results in Table A.3 are unaffected by this research design change.

Fourth, in equation (4) in the main text the transitory random shocks, ξ_{it}, may be serially correlated within individuals, but this serial correlation is set to zero for the analysis because no straightforward measurement is available.[22] In this sensitivity check I examine the potential impact of parameterizing the serial correlation in ξ_{it} above zero. Following Gu and Koenker (2017), I decompose ξ_{it} further:

$$\xi_{it} = \rho \xi_{it-1} + \kappa_{it} \tag{A.1}$$

The parameter ρ controls the magnitude of the serial correlation. In Gu and Koenker (2017), ρ is estimated by assuming it is the same across all individuals. However, this homogeneity assumption is troublesome for my application because ρ is likely much smaller for the top earners

[22] Note that there is an important conceptual distinction between serial correlation in income across waves and the serial correlation in ξ_{it}. The correlation in income is primarily accounted for by the time invariant income level estimate ($\hat{\alpha}_i$); the serial correlation in income shocks will only have at most a small contribution to the overall serial correlation in income.

due to high levels of mean reversion (Guvenen, et al., 2015). To test the sensitivity of my results to variation in the serial correlation of ξ_{it}, I parameterize ρ to 0.49 (taken from Gu and Koenker, 2017) and 0.32 (estimated based on the analytic sample here excluding top-coded observations). Note that these two values are likely to be higher than an appropriate value for top earners. Overall, the results in Table A.4 show that the performance of the Bayesian method is modestly worse, especially when ρ equals 0.49. Even so, the Bayesian method still performs better than the other methods. In results not shown but available upon request, I test multiple ρ values from 0.00 to 0.49, the performance of the Bayesian method is the best when ρ equals 0. Overall, the incorporation of this type of serial correlation does not increase the imputation performance of the Bayesian method.

Table A.1. Sensitivity Check for the Pareto Distribution Assumption: Individual Imputation Accuracy

Main Setting: The GB2 Distribution Assumption (from Table 2)

Method	RMSE		RMSE Reduction		Over-prediction Ratio	
	Income Level	Income Volatility	Income Level	Income Volatility	Income Level	Income Volatility
Standard	19,708	9,037			0.58	0.72
Rank-based	17,931	5,432	9.02%	39.88%	0.59	0.31
Bayesian	15,998	4,893	18.82%	45.85%	0.61	0.48

Sensitivity Check: The Pareto Distribution Assumption

Method	RMSE		RMSE Reduction		Over-prediction Ratio	
	Income Level	Income Volatility	Income Level	Income Volatility	Income Level	Income Volatility
Standard	28,338	13,908			0.62	0.75
Rank-based	26,338	9,187	7.06%	33.95%	0.62	0.34
Bayesian	21,385	6,887	24.54%	50.48%	0.64	0.51

Notes: The first panel is taken from Table 2. The RMSE panel shows the average RMSE of the imputed income values produced by each imputation method. All calculations are based on the portion of sample where the true values can be accurately obtained (i.e., pseudo top-coded values). The RMSE Reduction panel shows the average RMSE reductions of the rank-based or Bayesian imputation methods relative to the standard method. The over-prediction panel shows the average proportion of observations for which the imputed values are larger than the actual values.

Table A.2. Sensitivity Check for Alternative Pseudo Top-coding Thresholds: Individual Imputation Accuracy

Main Setting: 2% Threshold (from Table 2)

Method	RMSE		RMSE Reduction		Over-prediction Ratio	
	Income Level	Income Volatility	Income Level	Income Volatility	Income Level	Income Volatility
Standard	19,708	9,037			0.58	0.72
Rank-based	17,931	5,432	9.02%	39.88%	0.59	0.31
Bayesian	15,998	4,893	18.82%	45.85%	0.61	0.48

Sensitivity Check: 1% Threshold

Method	RMSE		RMSE Reduction		Over-prediction Ratio	
	Income Level	Income Volatility	Income Level	Income Volatility	Income Level	Income Volatility
Standard	24,277	10,090			0.66	0.75
Rank-based	20,515	5,722	15.50%	43.29%	0.66	0.38
Bayesian	18,015	5,030	25.79%	50.15%	0.68	0.54

Sensitivity Check: 3% Threshold

Method	RMSE		RMSE Reduction		Over-prediction Ratio	
	Income Level	Income Volatility	Income Level	Income Volatility	Income Level	Income Volatility
Standard	17,710	8,482			0.55	0.72
Rank-based	16,109	5,165	9.04%	39.10%	0.56	0.29
Bayesian	14,606	4,614	17.53%	45.61%	0.58	0.47

Notes: The first panel is taken from Table 2. The RMSE panel shows the average RMSE of the imputed income values produced by each imputation method. All calculations are based on the portion of sample where the true values can be accurately obtained (i.e., pseudo top-coded values). The RMSE Reduction panel shows the average RMSE reductions of the rank-based or Bayesian imputation methods relative to the standard method. The over-prediction panel shows the average proportion of observations for which the imputed values are larger than the actual values.

Table A.3. Sensitivity Check for Using an Alternative Sample to Estimate the Prior Distribution: Individual Imputation Accuracy

Main Setting (from Table 2)

Method	RMSE		RMSE Reduction		Over-prediction Ratio	
	Income Level	Income Volatility	Income Level	Income Volatility	Income Level	Income Volatility
Standard	19,708	9,037			0.58	0.72
Rank-based	17,931	5,432	9.02%	39.88%	0.59	0.31
Bayesian	**15,998**	**4,893**	**18.82%**	**45.85%**	**0.61**	**0.48**

Sensitivity Check: Alternative Sample

Method	RMSE		RMSE Reduction		Over-prediction Ratio	
	Income Level	Income Volatility	Income Level	Income Volatility	Income Level	Income Volatility
Standard	19,708	9,037			0.58	0.72
Rank-based	17,931	5,432	9.02%	39.88%	0.59	0.31
Bayesian	**16,329**	**4,962**	**17.14%**	**45.09%**	**0.61**	**0.49**

Notes: The first panel is taken from Table 2. The standard and rank-based methods do not change due to this sensitivity check but are included in the table for completeness. The RMSE panel shows the average RMSE of the imputed income values produced by each imputation method. All calculations are based on the portion of sample where the true values can be accurately obtained (i.e., pseudo top-coded values). The RMSE Reduction panel shows the average RMSE reductions of the rank-based or Bayesian imputation methods relative to the standard method. The over-prediction panel shows the average proportion of observations for which the imputed values are larger than the actual values.

Table A.4. Sensitivity Check for Incorporating Serial Correlations in Transitory Shocks: Individual Imputation Accuracy

Main Setting: 0 Correlation (from Table 2)

Method	RMSE		RMSE Reduction		Over-prediction Ratio	
	Income Level	Income Volatility	Income Level	Income Volatility	Income Level	Income Volatility
Standard	19,708	9,037			0.58	0.72
Rank-based	17,931	5,432	9.02%	39.88%	0.59	0.31
Bayesian	**15,998**	**4,893**	**18.82%**	**45.85%**	**0.61**	**0.48**

Sensitivity Check: 0.49 Correlation

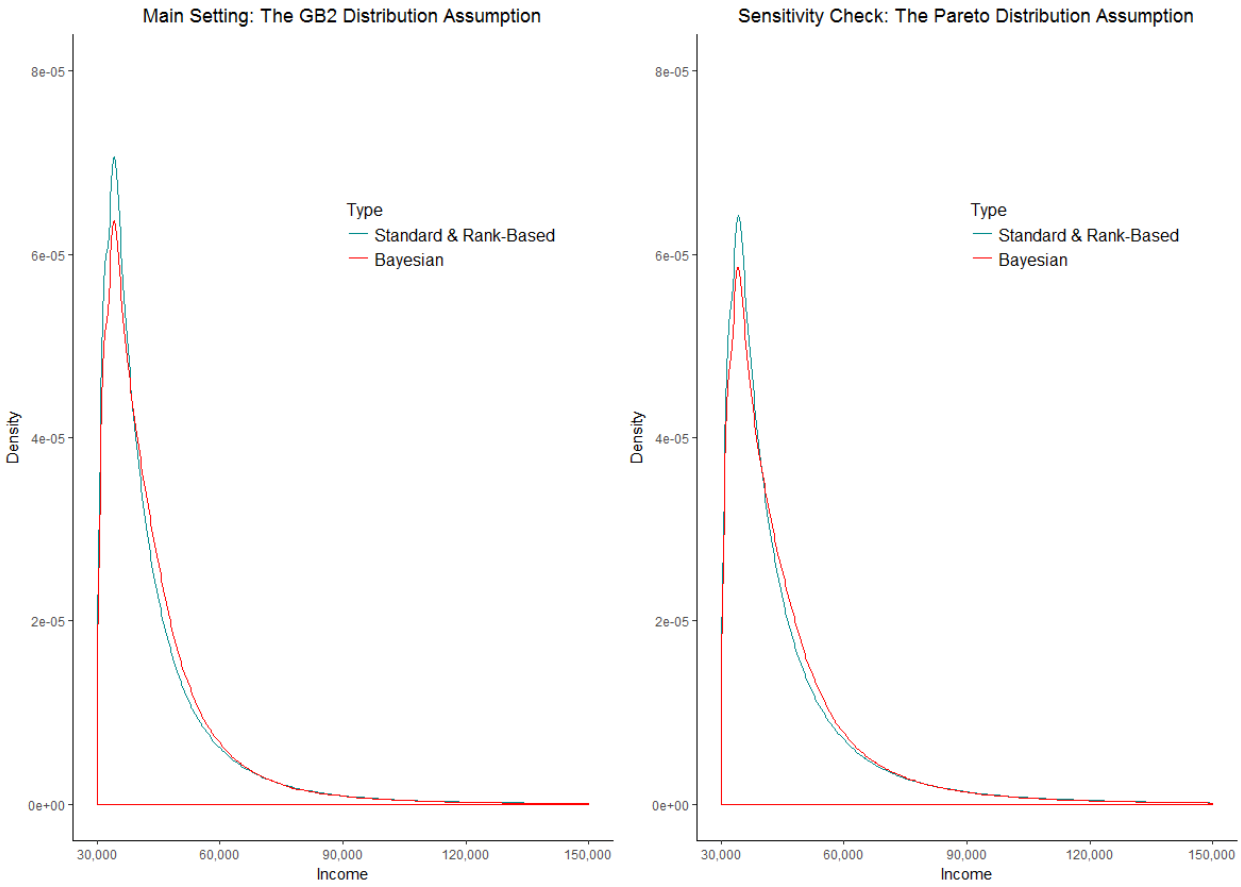
Method	RMSE		RMSE Reduction		Over-prediction Ratio	
	Income Level	Income Volatility	Income Level	Income Volatility	Income Level	Income Volatility
Standard	19,708	9,037			0.58	0.72
Rank-based	17,931	5,432	9.02%	39.88%	0.59	0.31
Bayesian	**16,394**	**5,412**	**16.81%**	**40.11%**	**0.60**	**0.48**

Sensitivity Check: 0.32 Correlation

Method	RMSE		RMSE Reduction		Over-prediction Ratio	
	Income Level	Income Volatility	Income Level	Income Volatility	Income Level	Income Volatility
Standard	19,708	9,037			0.58	0.72
Rank-based	17,931	5,432	9.02%	39.88%	0.59	0.31
Bayesian	**16,151**	**5,192**	**18.05%**	**42.54%**	**0.60**	**0.47**

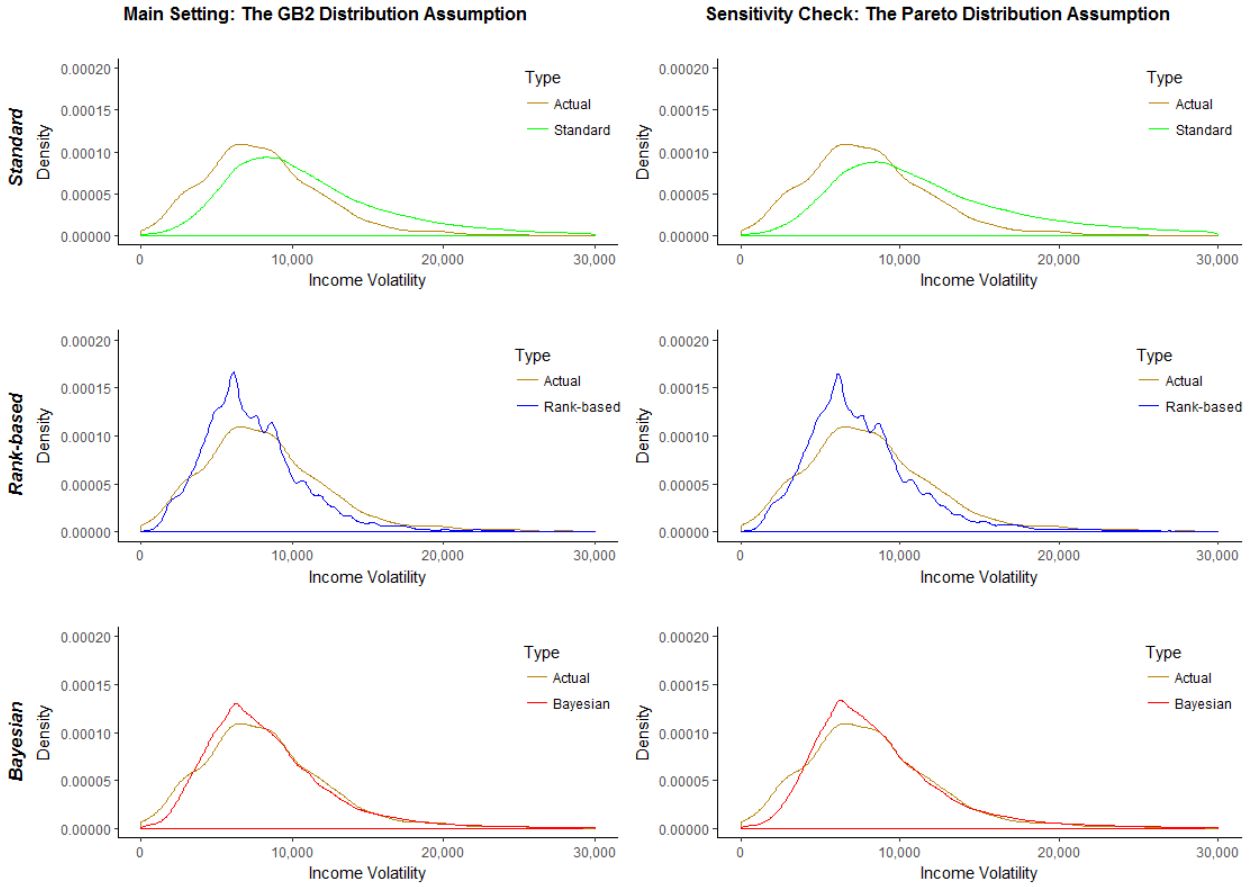
Notes: The first panel is taken from Table 2. The standard and rank-based methods do not change due to this sensitivity check but are included in the table for completeness. The RMSE panel shows the average RMSE of the imputed income values produced by each imputation method. All calculations are based on the portion of sample where the true values can be accurately obtained (i.e., pseudo top-coded values). The RMSE Reduction panel shows the average RMSE reductions of the rank-based or Bayesian imputation methods relative to the standard method. The over-prediction panel shows the average proportion of observations for which the imputed values are larger than the actual values.

Figure A.1. Sensitivity Check for the Pareto Distribution Assumption: Distributions of Income Levels



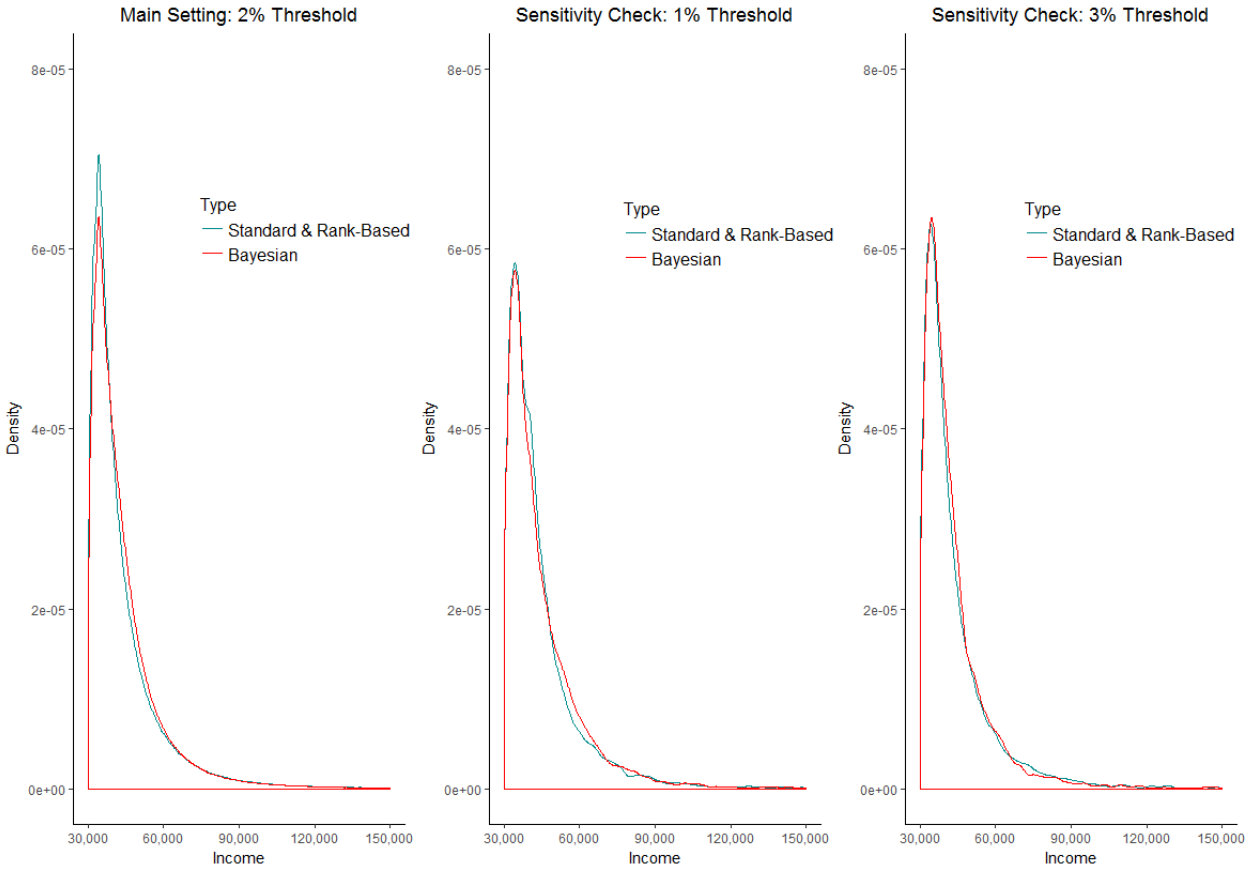
Notes: The first panel is taken from Figure 3. The density plot of the standard imputation method overlaps perfectly with that of the rank-based method as described in the text.

Figure A.2. Sensitivity Check for the Pareto Distribution Assumption: Distributions of Income Volatility



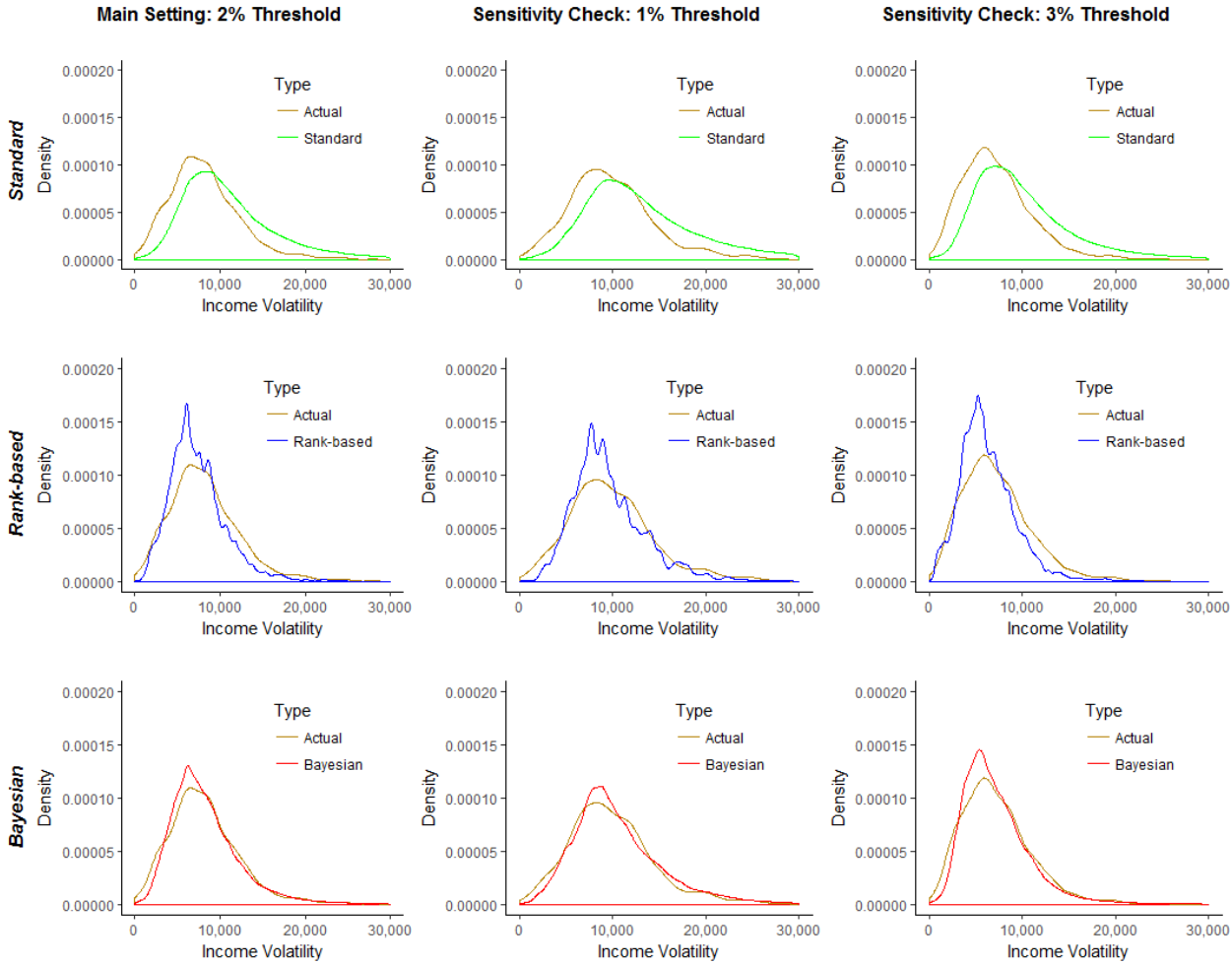
Notes: The first vertical panel is taken from Figure 4.

Figure A.3. Sensitivity Check for Alternative Pseudo Top-coding Thresholds: Distributions of Income Levels



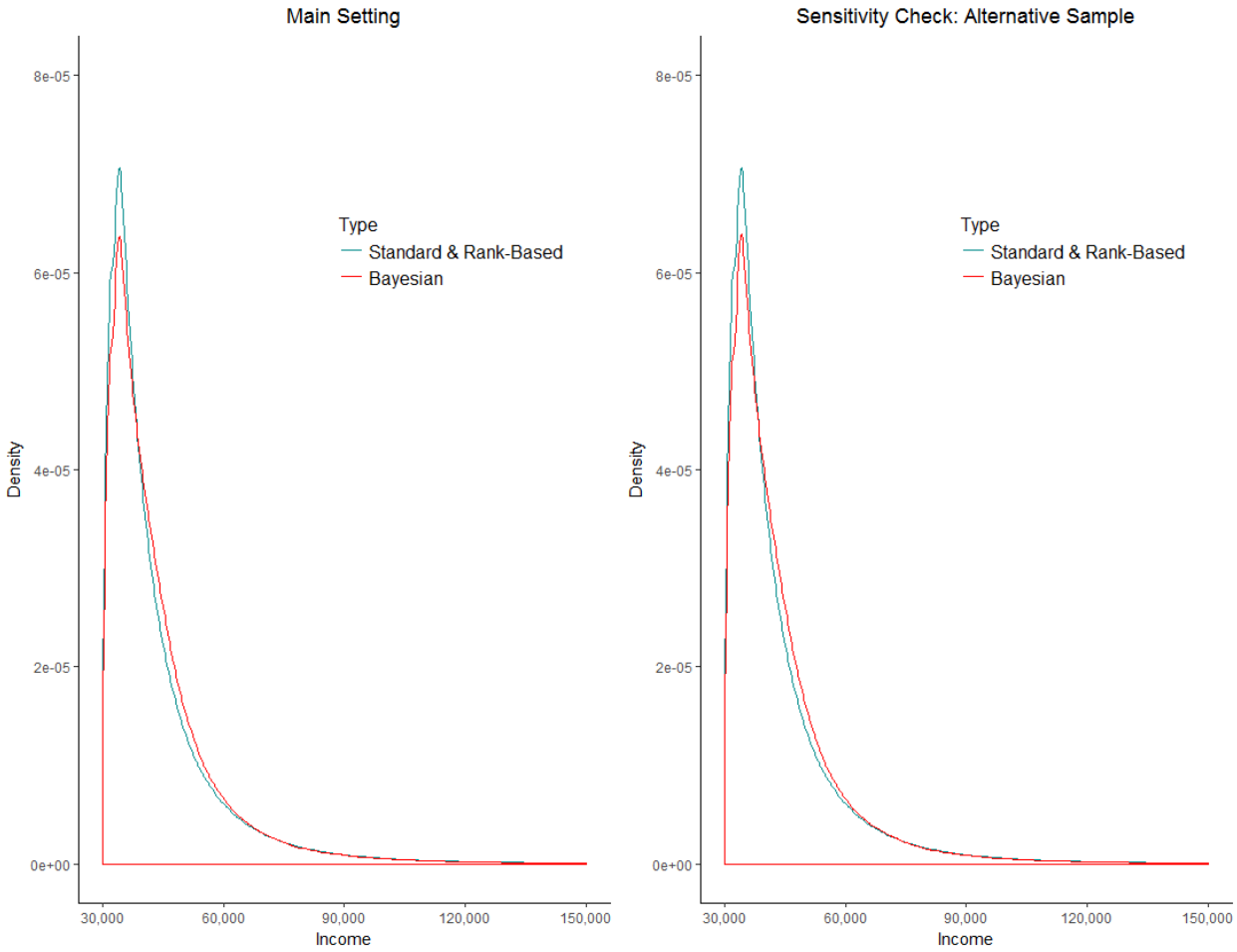
Notes: The first panel is taken from Figure 3. The density plot of the standard imputation method overlaps perfectly with that of the rank-based method as described in the text. All panels plot the income density of the top 2% of earners.

Figure A.4. Sensitivity Check for Alternative Pseudo Top-coding Thresholds: Distributions of Income Volatility



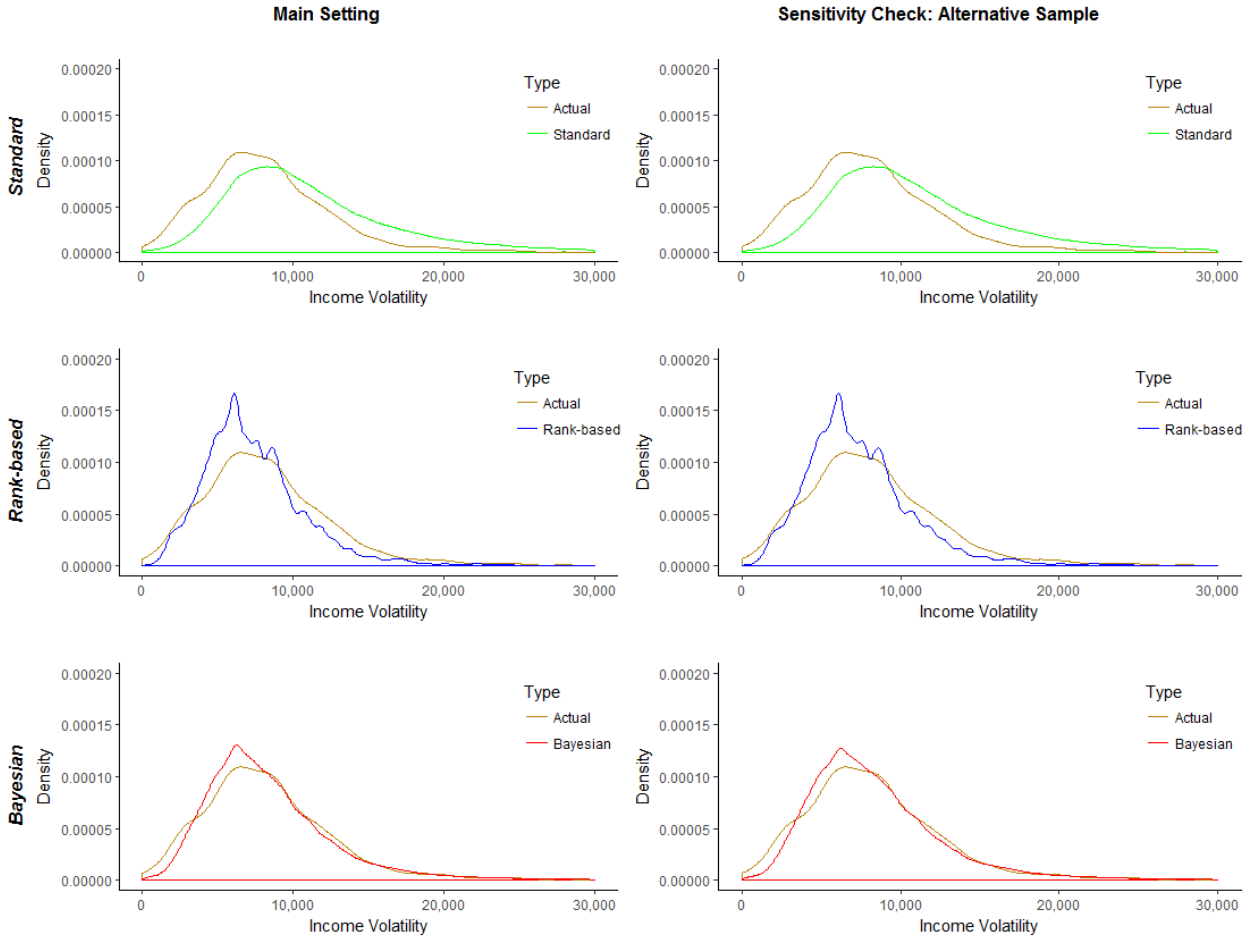
Notes: The first vertical panel is taken from Figure 4. The underlying sample used to plot the density in sensitivity checks are different due to changes in the top-coding thresholds as specified.

Figure A.5. Sensitivity Check for Using an Alternative Sample to Estimate the Prior Distribution Using the Bayesian Method: Distributions of Income Levels



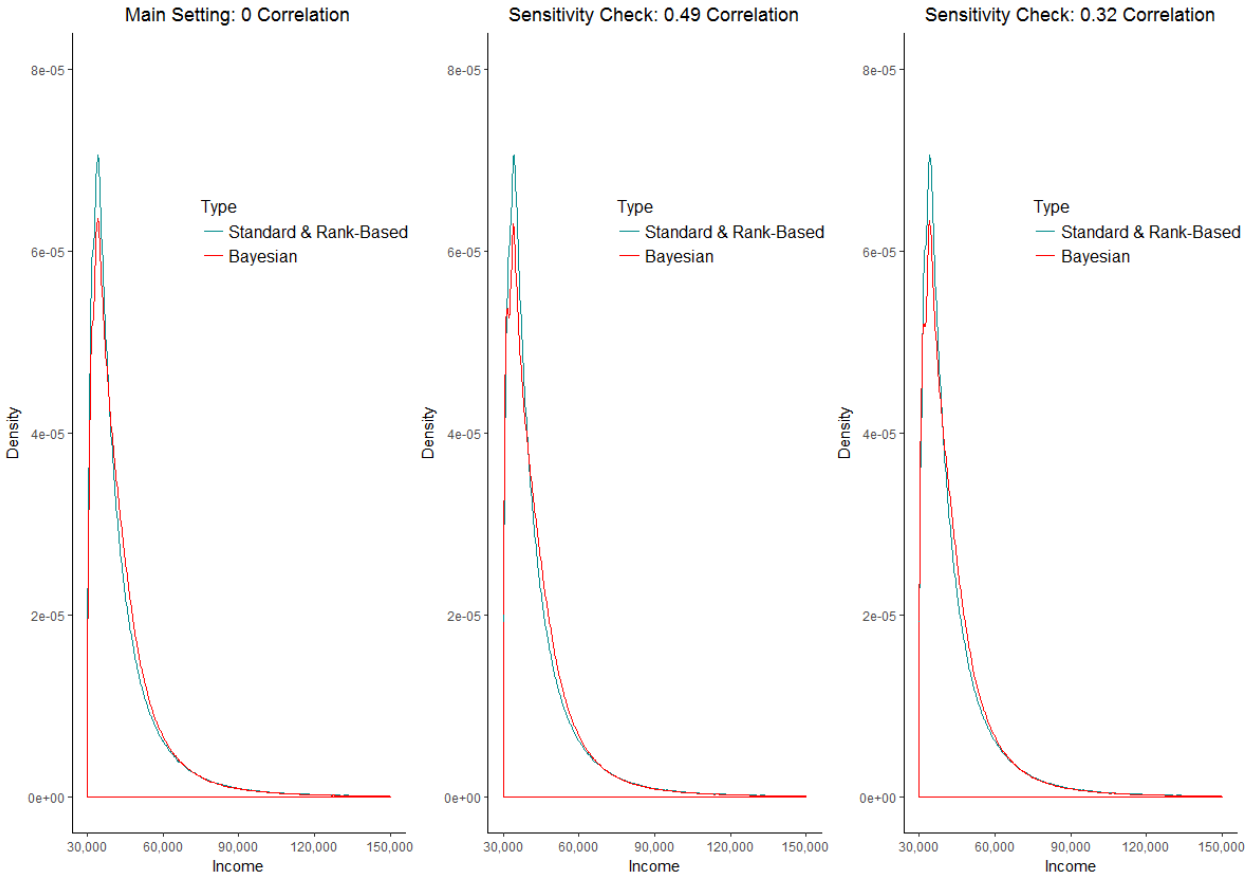
Notes: The first panel is taken from Figure 3. The density plot of the standard imputation method overlaps perfectly with that of the rank-based method as described in the text.

Figure A.6. Sensitivity Check for Using an Alternative Sample to Estimate the Prior Distribution Using the Bayesian Method: Distributions of Income Volatility



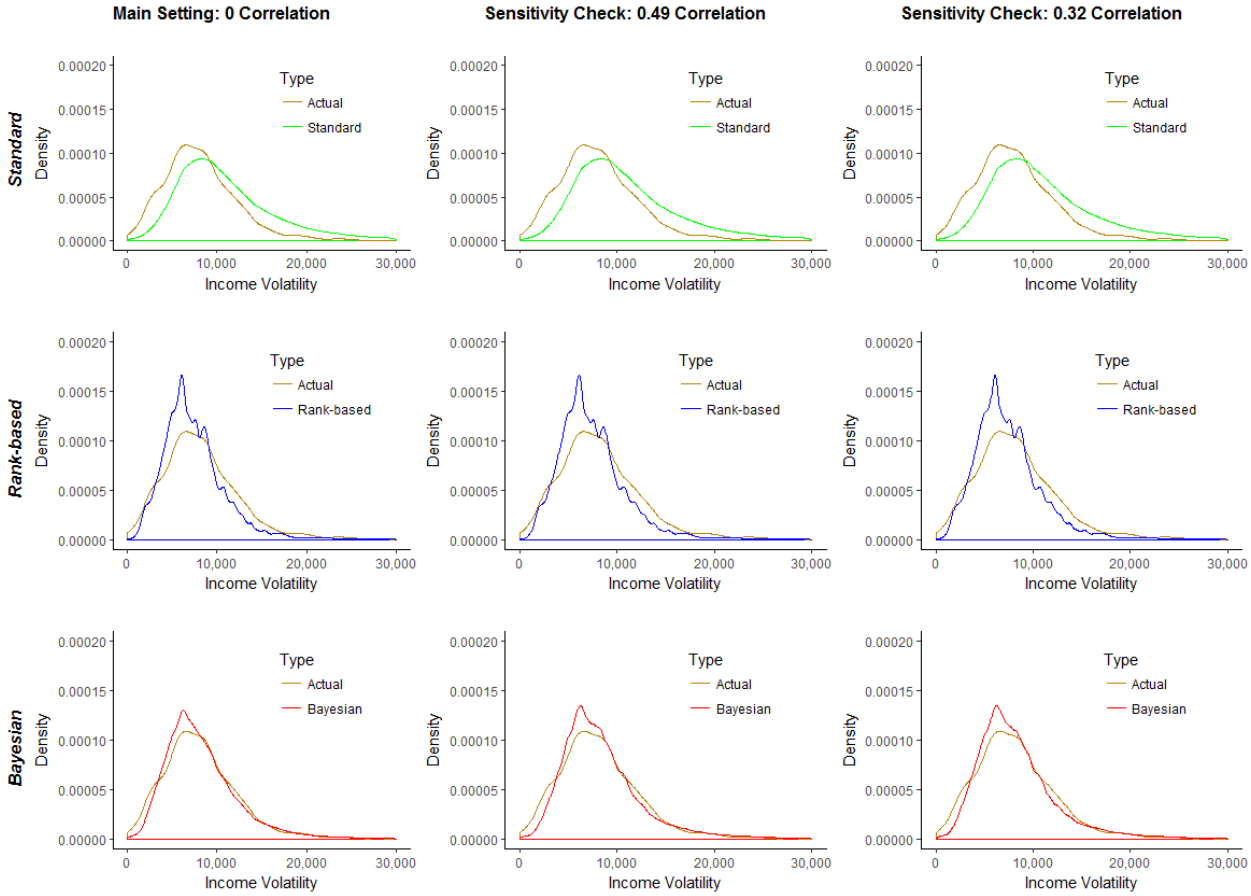
Notes: The first vertical panel is taken from Figure 4. The standard and rank-based methods do not change for this sensitivity check but are included for completeness.

Figure A.7. Sensitivity Check for Incorporating Serial Correlation in Transitory Shocks Using the Bayesian Method: Distributions of Income Levels



Notes: The first panel is taken from Figure 3. The density plot of the standard imputation method overlaps perfectly with that of the rank-based method as described in the text. The standard and rank-based methods do not change for this sensitivity check but are included for completeness.

Figure A.8. Sensitivity Check for Incorporating Serial Correlations in Transitory Shocks Using the Bayesian Method: Distributions of Income Volatility



Notes: The first vertical panel is taken from Figure 4. The standard and rank-based methods do not change for this sensitivity check but are included for completeness.