SUPPLEMENTARY INFORMATION

Accurate age estimation in small-scale societies

Yoan Diekmann, Daniel Smith, Pascale Gerbault, Mark Dyble, Abigail E. Page, Nikhil Chaudhary, Andrea Bamberg Migliano, Mark G. Thomas

SUP. MATERIALS AND METHODS

Estimating ages by Gibbs sampling

We consider a random variable $\boldsymbol{X} = (X_1, \dots, X_n)$ with ages of n individuals. Furthermore, we introduce an ordering R of these n individuals from youngest to oldest, which can always be re-labeled such as $R = (1, \dots, n)$. In a Bayesian framework, age estimation can thus be formalized as computing the posterior distribution

$$P(\boldsymbol{X}|R) = \frac{P(R|\boldsymbol{X})P(\boldsymbol{X})}{\int_{X \in \chi} P(R|\boldsymbol{X})P(\boldsymbol{X})d\boldsymbol{X}}$$

where $P(\boldsymbol{X})$ is an arbitrary prior distribution on the ages of the individuals satisfying $P(\boldsymbol{X}) = \prod_{i=1}^{n} P(X_i)$, and the likelihood function $P(R|\boldsymbol{X})$ is defined as

$$P\big(R = (1, \dots, n)\big|\boldsymbol{X} = (x_1, \dots, x_n)\big) = \begin{Bmatrix} 1 \; if \; x_i < x_j \; \forall \; i < j \\ 0 \; else \end{Bmatrix}.$$

In order to avoid explicit computation of the normalizing constant, we opted to approximate the posterior distribution by statistical sampling techniques. A naïve approach to sample from the posterior is to randomly draw an age for each of the n individuals independently, and then test if the resulting sample satisfies the ranking constraint. If not, the value is discarded. However, the more the individuals' prior age distributions overlap, the more samples generated by this approach would have to be discarded. To solve this more efficiently, we implement a Gibbs sampling approach, which samples from the posterior distribution directly without having to discard any age-vector. The key to achieve this lies in considering only univariate conditional distributions, i.e. the age distribution of one individual when all other individuals are assigned a fixed value from their respective range (3, p. 16), i.e. $P_X(x_i \mid x_1, \dots x_{i-1}, x_{i+1}, \dots, x_n)$. How an initial set of values $\boldsymbol{x}$ satisfying the age ranking can be found is described below (point 1). Iterating over all individuals in this manner generates a sample $\boldsymbol{x}$, and it can be shown that the sequence of samples $\boldsymbol{x}$ thereby generated converges to the desired target posterior distribution (3, p. 17) .

In our case, a Gibbs sampler can be constructed in the following manner. First, we observe an ordering R of all individuals and label them accordingly, i.e. individual labeled 1 is younger than individual 2 etc., the oldest being individual n. Next, iterative rounds of sampling are performed. Denote the k^{th} sample of ages $\boldsymbol{x}$ by $\boldsymbol{x}^{(k)} = (x_1^{(k)}, \dots, x_n^{(k)})$. Assume for example that $P_X(x) \sim Unif(l, u)$, i.e. the *a priori* age of any individual is distributed uniformly within an interval bounded by values l and u. We note that alternative distributions for $P_X(x)$ – such as a normally distributed *a priori* age – are easily accommodated in a way analogous to the one described below. Setting $x_0 \coloneqq -\infty$ and $x_{n+1} \coloneqq \infty$ for the sake of simplicity, our Gibbs sampler proceeds as follows:

1) Initialize the first sample $k = 0$:

$x_i^{(0)} = \max(l_i, x_{i-1}^{(0)})$, for $i \in \{1, \dots, n\}$

2) Iterate K times to generate $K + 1$ samples, i.e. $k \in \{1, \dots, K\}$:

$x_i^{(k)} \sim Unif(\max(l_i, x_{i-1}^{k}), \min\left(u_i, \dots, u_n, x_{i+1}^{(k-1)}\right))$, for $i \in \{1, \dots, n\}$

This procedure generates as many samples as desired. As always with empirical distributions, the general trade-off is that more samples occupy more memory space and require longer computation time, but reduce the stochastic sampling error and therefore better approximate the underlying distribution.

Figure 5 in the main text illustrates the type of input required and output generated by our method for five fictitious individuals.

Implementation details

We have implemented the Gibbs sampling algorithm in Python 2.7 (5). In order to find sensible parameter values for the total number of iterations, burn-in and thinning, we analysed 50,000 sampling iterations for the toy example with five individuals presented in Figure 5B of the main text.

Panel A of Supplementary Figure S3 shows perfect mixing, with low autocorrelation (see Panel D) also confirmed by a high effective sample size of 33521.62, meaning that for the estimation of the posterior mean 50,000 samples correspond to 33,522 independent samples. This suggests that no thinning is required. Panel B and C illustrate how the sample mean changes in the course of the sampling process. Based on visual inspection, we chose a burn-in of 50 iterations, largely exceeding Raftery-Lewis (9) method's recommendation of two to four. Panel B already suggests that convergence is achieved relatively quickly, as means remain stable after 10,000 iterations. Gelman and Rubin's shrink factor (8), a formal test for convergence presented in Panels E and F and computed on 4 independent runs of the Gibbs sampler with the first 10,000 iterations discarded, shows a shrink factor of 1 after 10,000 additional iterations. Therefore, we set our default to 20,050 iterations in total, resulting in 20,000 ages sampled per individual with no thinning and 50 iterations discarded a burn-in.

All diagnostic statistics were computes and plotted in R version 3.1.3 (6) using functions from the 'coda' and 'mcmcplots' libraries.

Palanan Agta: data collection method

In order to construct relative age rankings, we took and printed photographs of all individuals in every camp. Individuals were then assigned to approximate age cohorts (0-4, 4-8, 8-12, 13-19, 20-45, and 45+). Those not easily assigned to one cohort were included in the two nearest cohorts (e.g., an individual aged ~45 would be included in both the 20-45 and 45+ cohorts). Either individually or in small groups, we presented these photographs to individuals from a target cohort, one at a time. The target cohort was the cohort the individual ('ego') was included in, as well as all cohorts younger than ego. Cohorts, especially for children, were often presented together, so that some rankings included, for instance, all individuals aged 0 to 12. Children under the age of five were often unable to make the age rankings themselves, and in this instance either their mothers or older siblings would conduct the ranking. Individuals from a specific camp were shown pictures of others from their camp and neighbouring camps. More distant camps were not included due to a lack of familiarity, unless ego knew individuals from more distant camps particularly well (e.g. they grew up in the same camp and moved apart upon marriage). For cohorts including ego, ego's picture was displayed first. Participants were first asked if they knew the individual on the photograph (i.e. the target), and if so they were then asked if they knew the target well enough to give their approximate date of birth relative to other individuals. Each photograph was put into one of three categories; 'don't know', 'know but not the age', and 'age known'. If ego knew both the target and their age, they were asked to rank the age of the target relative to others. Although similar to the method by Hill and Hurtado (2), rather than having two piles of simply older and younger (with ego as reference), our method produced a relative age list from youngest to oldest. This process was repeated multiple times with different subjects producing a total of 266 partial ranks, including 587 individuals.

The second stage involved deriving age estimates for these 587 individuals. One invaluable source of information, especially for older individuals, was the Headlands' database from Casiguran (4), since some individuals from our study population were included in this database, with relatively accurate dates of birth assigned. Absolute ages of individuals were ascertained via various other methods, including; asking individuals if they knew their own or their children's age (which could be from various sources, such as, birth certificates, other documentation, school grades, own estimates, etc.), births near dated events (such as martial law in 1970 or various known typhoons), and age-mates of individuals with known birthdays. For children up to the age of 12 years, it was also possible to estimate age brackets by dental development.

There are, however, some issues with methods used to estimate absolute ages, especially estimates given by individual Agta, the dental aging and school grade. For example, many individuals gave various conflicting dates and/or ages, including; saying a child was four years old, yet born in 2004 (during the 2013 fieldwork season), or giving a birth date for one child as 2004 (~eight years old) yet saying a younger child was nine years old, and age conflicts between parents (for example, one child was given an age of seven months by one parent and two years by the other). For both teeth ages and school grades, the margins of error were often quite large (+/- half a year), which was especially problematic regarding school ages, as the grade reached was often variable for individuals of a similar age, and most children in the community either do not go to school, or start school at older ages than their agricultural neighbours. Therefore, strict criteria were used to select accurate ages/birth dates. First, if an individual was given two markedly different birth dates, that person was excluded from the absolute age list. Second, if ages for an entire sibling-set were provided, but at least one age was wrong (e.g., did not correspond to teeth ages, or did not allow at least nine months pre- or post-birth of the nearest sibling), then ages for the whole sibling-set were excluded. Furthermore, for all children, the birth date had to fall within the range of teeth ages to be accepted, and a similar protocol of matching with teeth ages was established for estimating the ages of individuals from school grade. For ages estimated based on comparisons to individuals with known birth dates, these individuals with estimated ages were given a year of birth with a +/- one year margin to account for error. Using these methods, 98 individuals (out of 587; 16.7%) were given an exact birthday, while many others were given age estimates within +/- one year (Supplementary Table S3).

For individuals which we could not attach a secure date or estimate, three of the field researchers (DS, AEP, & MD), as well as the principle investigator (ABM) estimated the ages based on cues such as dental development, school grade, birth order (if older or younger siblings have a known age), age of ego's children (if known), number of children, and visual inspection. Independently, each of the four researchers estimated an upper and lower age bound for each individual. In collating these estimates, the youngest lower bound and oldest upper bound of the four estimates were used in order to include as much uncertainty as possible. There was increased uncertainty for older individuals, as the average difference between upper and lower estimates increases with age (Supplementary Table S3).

SUP. RESULTS

Validation and benchmarking

Table 1 and Figure 1 in the main text show that the Gibbs sampler provides more accurate age estimates than the regression approach. However, the performances may be influenced by the specific cross-validation parameters chosen, i.e. k=5 partitions of n=13 individuals each for which ages are assumed to be known exactly. Therefore, we tested other parameter values from k=2 partitions, resulting in n=32 individuals, to k=13, with n=5 individuals per partition. We considered each partition in turn to estimate the regression equation and then deduced the ages of the remaining individuals. This procedure enabled us to assess how the number of individuals with known ages affects each method's accuracy.

Supplementary Figure S1 shows that the accuracy for the fifth-degree polynomial approach massively drops when more than five partitions are chosen (i.e. k>5). This is expected, as fewer known ages are available for the regression, resulting in a less constraint curve leading to overfitting. Note that although the LOESS approach also shows reduced accuracy in smaller partitions, the magnitude of the error is much smaller.

A flexible method for fieldwork data: dealing with multiple partial ranks

We relax the assumption of a single complete ordering R of all n individuals from youngest to oldest, and rather allow for multiple partial ranks. The approach we describe in the following is heuristic. Describing the problem of multiple partial ranks in a formal manner and finding optimal solutions is an important and interesting problem for future research.

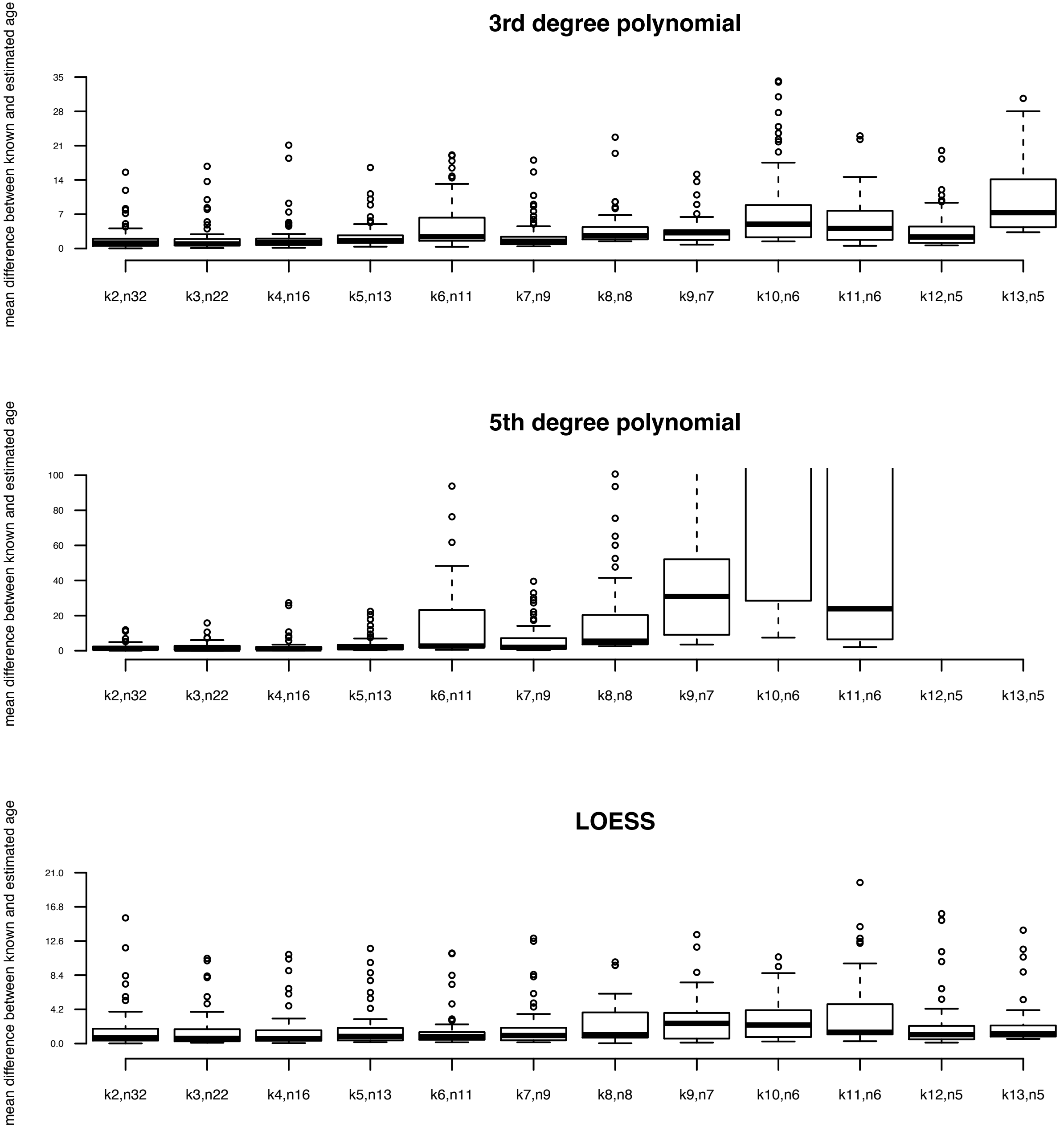
Let $\mathfrak{R} = \{R_1, \dots, R_m\}$ be a set of partial rankings of individuals. As described in the main text, we first merge partial ranks that are compatible, resulting in a modified set of partial ranks $\{\boldsymbol{R'}_1, \dots, \boldsymbol{R'}_l\}$, $l \leq m$, where each $\boldsymbol{R'}_j$ represents a subset of mutually compatible partial ranks from the initial full set, i.e. $\boldsymbol{R'}_j \subseteq \boldsymbol{\mathfrak{R}}.$ Merging is not always possible without ambiguity, as various different ways in which rankings could be merged may exist, e.g. if R_1 is compatible with R_2 and R_3, but R_2 and R_3 are not compatible with each other. In this case, we leave the corresponding ranks separate ($\{\boldsymbol{R'}_1, \dots, \boldsymbol{R'}_l\}$ is therefore a partition of the set $\mathfrak{R}$). It should be noted that alternative heuristics can easily be envisaged at this stage, for example a greedy strategy. The next step is to compute the posterior $P\left(\boldsymbol{X}\middle|\boldsymbol{R'}_j\right)$ separately for all merged partial ranks $\boldsymbol{R'}_j, j \in \{1, \dots, l\}$, by Gibbs sampling. Finally, we merge the resulting distributions per individual by forming a weighted finite mixture:

$$P(X_i = x|\mathfrak{R}) = \sum_{j=1}^{l} \frac{w_i(\boldsymbol{R'}_j)}{w_i(\mathfrak{R})} P\left(X_{\boldsymbol{i}} = x\middle|\boldsymbol{R'}_{\boldsymbol{j}}\right)$$

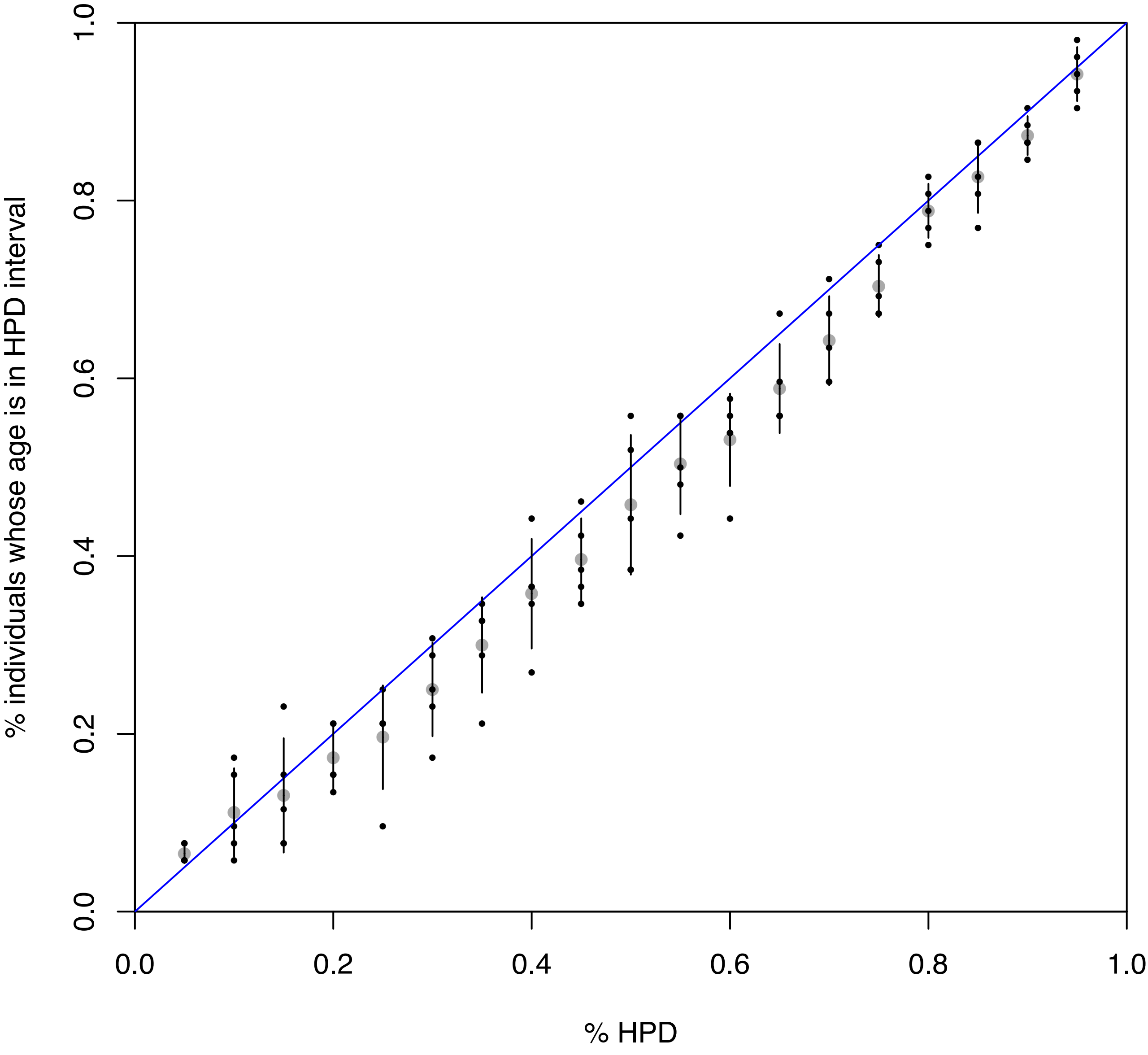
where $w_i()$ denotes the number of times individual i occurs in the corresponding set of rankings. The nominator term $w_i(\boldsymbol{R'})$ therefore preserves the information how many times an individual has been ranked consistently in a certain way in the initial set of unmerged partial rankings $\mathfrak{R}$.

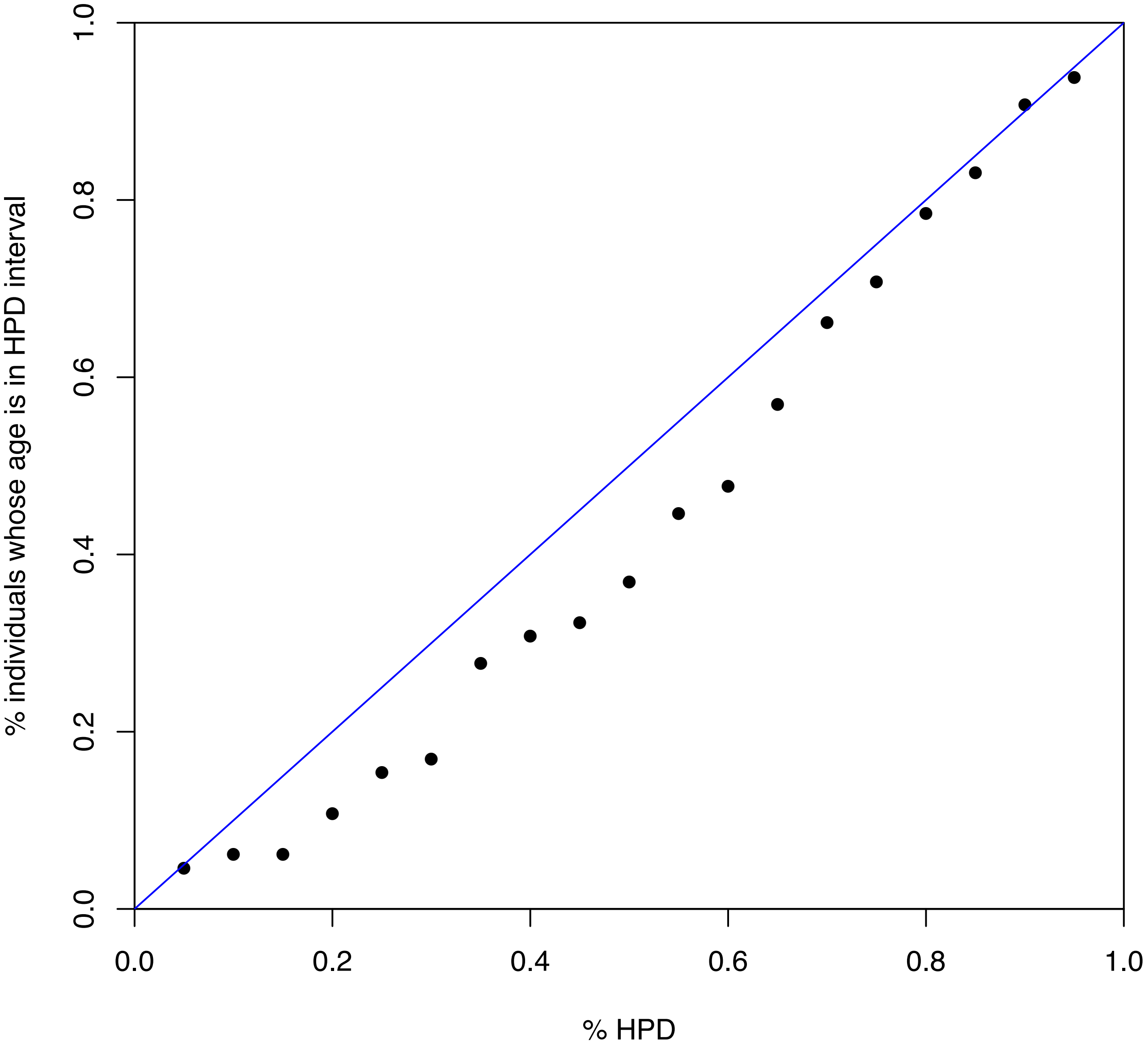
SUP. FIGURES

Supplementary Figure S1. *Differences in estimation accuracy under varying cross-validation parameters.* Boxplots of the mean of the differences between known ages and those estimated using regression analyses; top: third-order (3rd degree) polynomial, middle: fifth-order (5th degree) polynomial, bottom: local regression (LOESS; 7). The x-axis shows the number of partitions used (‘k’) and the number of individuals (‘n’) in these corresponding partitions; ‘k2,n32’ for example means 2 partitions of 32 individuals whose ages are known and used to estimate the regression coefficients. The y-axis shows the mean of the differences between known and estimated ages per individuals over the k partitions. Note that the scale of the y-axis of these three panels is not the same.

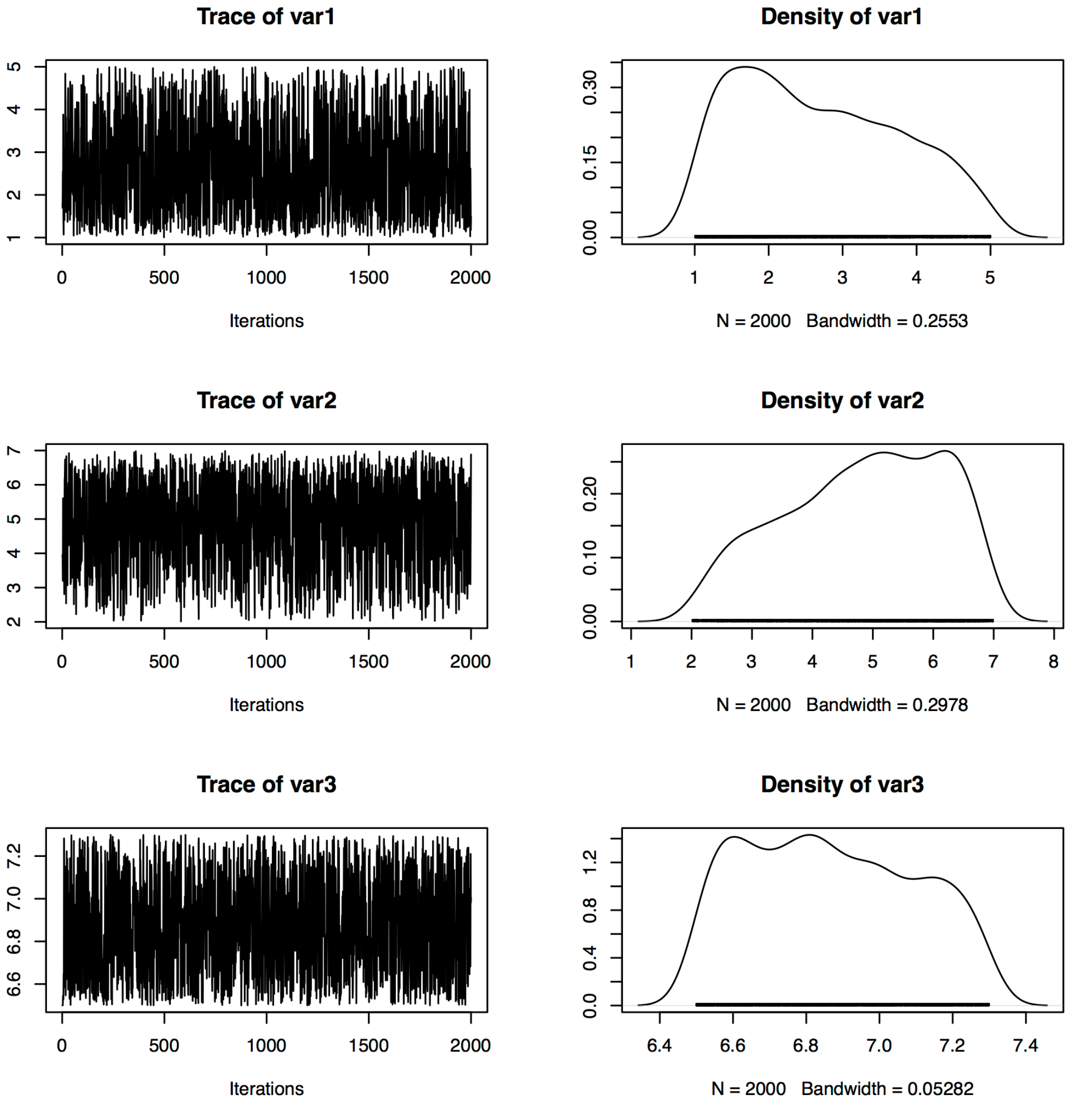

3rd degree polynomial
mean difference between known and estimated age
0
7
14
21
28
35
k2,n32
k3,n22
k4,n16
k5,n13
k6,n11
k7,n9
k8,n8
k9,n7
k10,n6
k11,n6
k12,n5
k13,n5
5th degree polynomial
mean difference between known and estimated age
0
20
40
60
80
100
k2,n32
k3,n22
k4,n16
k5,n13
k6,n11
k7,n9
k8,n8
k9,n7
k10,n6
k11,n6
k12,n5
k13,n5
LOESS
mean difference between known and estimated age
0.0
4.2
8.4
12.6
16.8
21.0
k2,n32
k3,n22
k4,n16
k5,n13
k6,n11
k7,n9
k8,n8
k9,n7
k10,n6
k11,n6
k12,n5
k13,n5

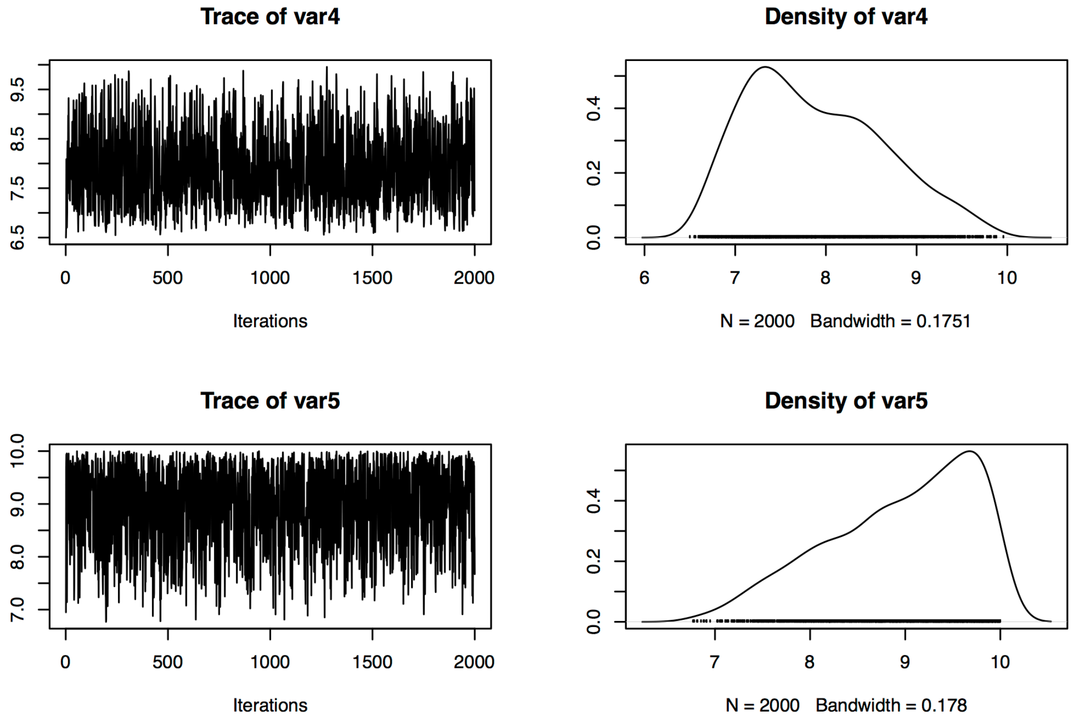
Supplementary Figure S2. *Error calibration of posterior distributions.* For the cross-validation experiment corresponding to Figure 1, we show that the highest posterior densities (HPD) contain the true age as often as the size of the interval suggests, and the posterior therefore correctly quantifies estimation uncertainty. For example, the 95% HPD covers the true age in 95% of the individuals. Panel A shows the results for each of the 5 cross-validation partitions (black points), their average (grey points) and standard deviation (black bars). Panel B shows the same analysis for the case where no age has been fixed, i.e. all priors were proper intervals.

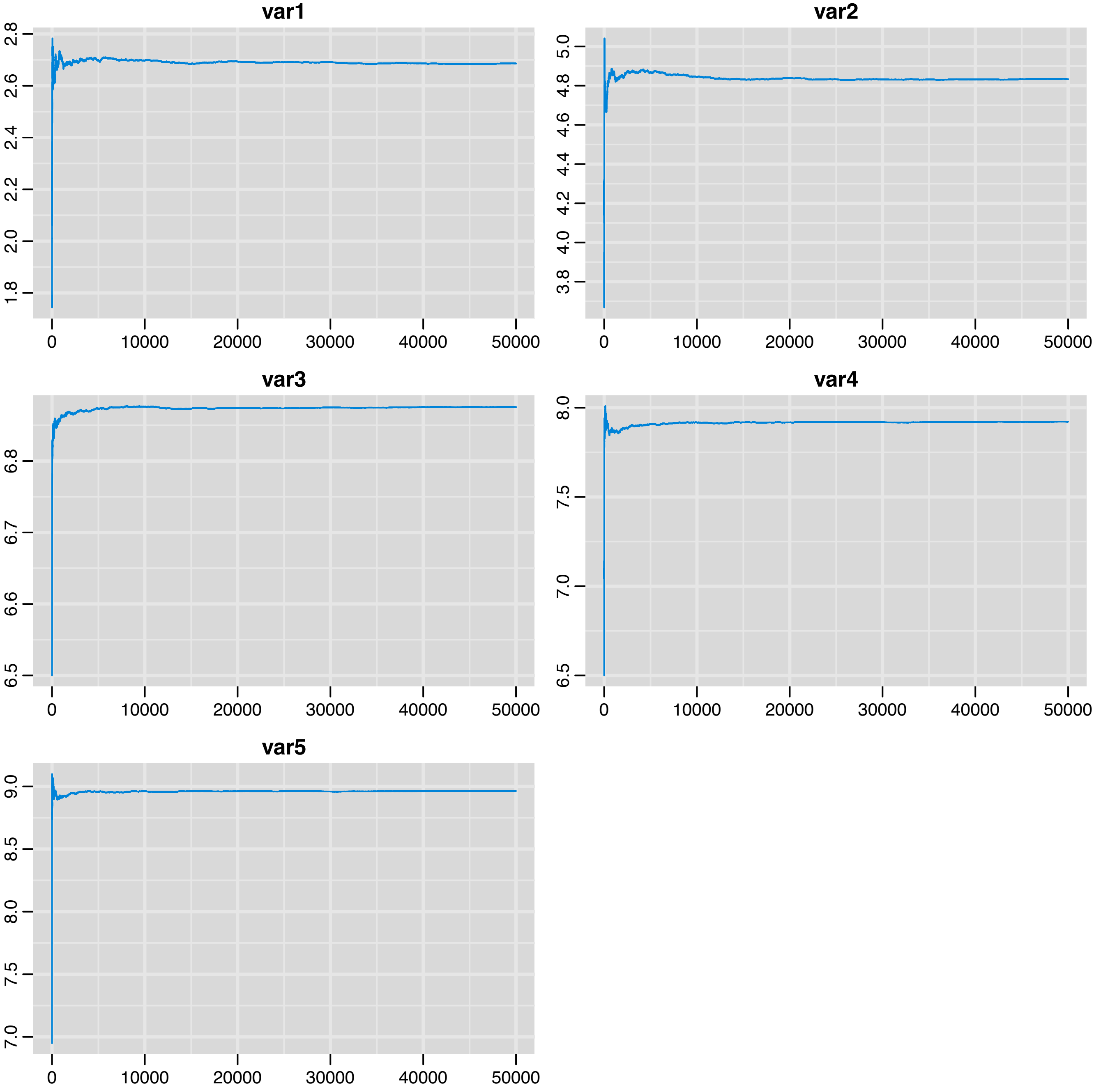

% individuals whose age is in HPD interval
0.0
0.2
0.4
0.6
0.8
1.0
% HPD
0.0
0.2
0.4
0.6
0.8
1.0

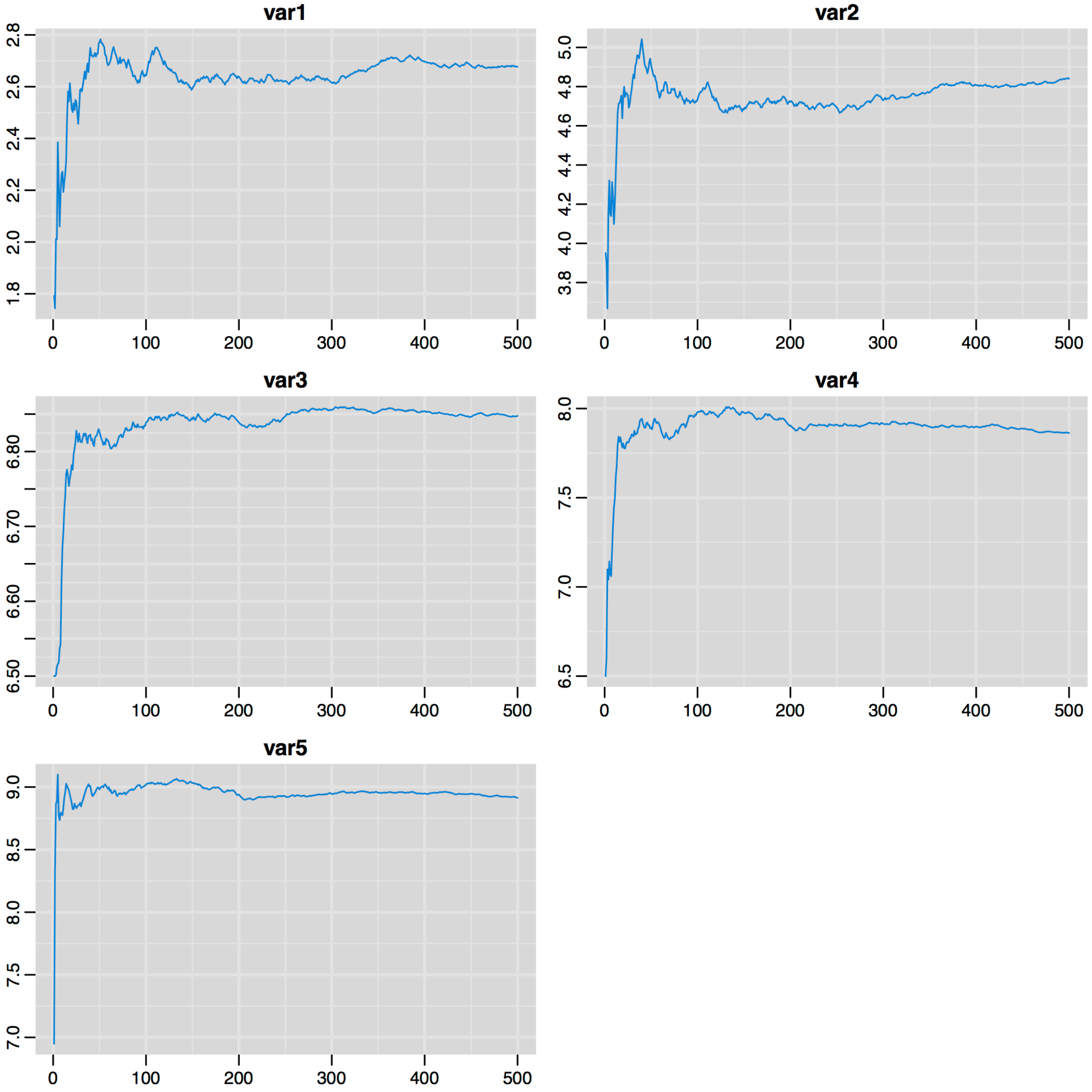

% individuals whose age is in HPD interval
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
% HPD

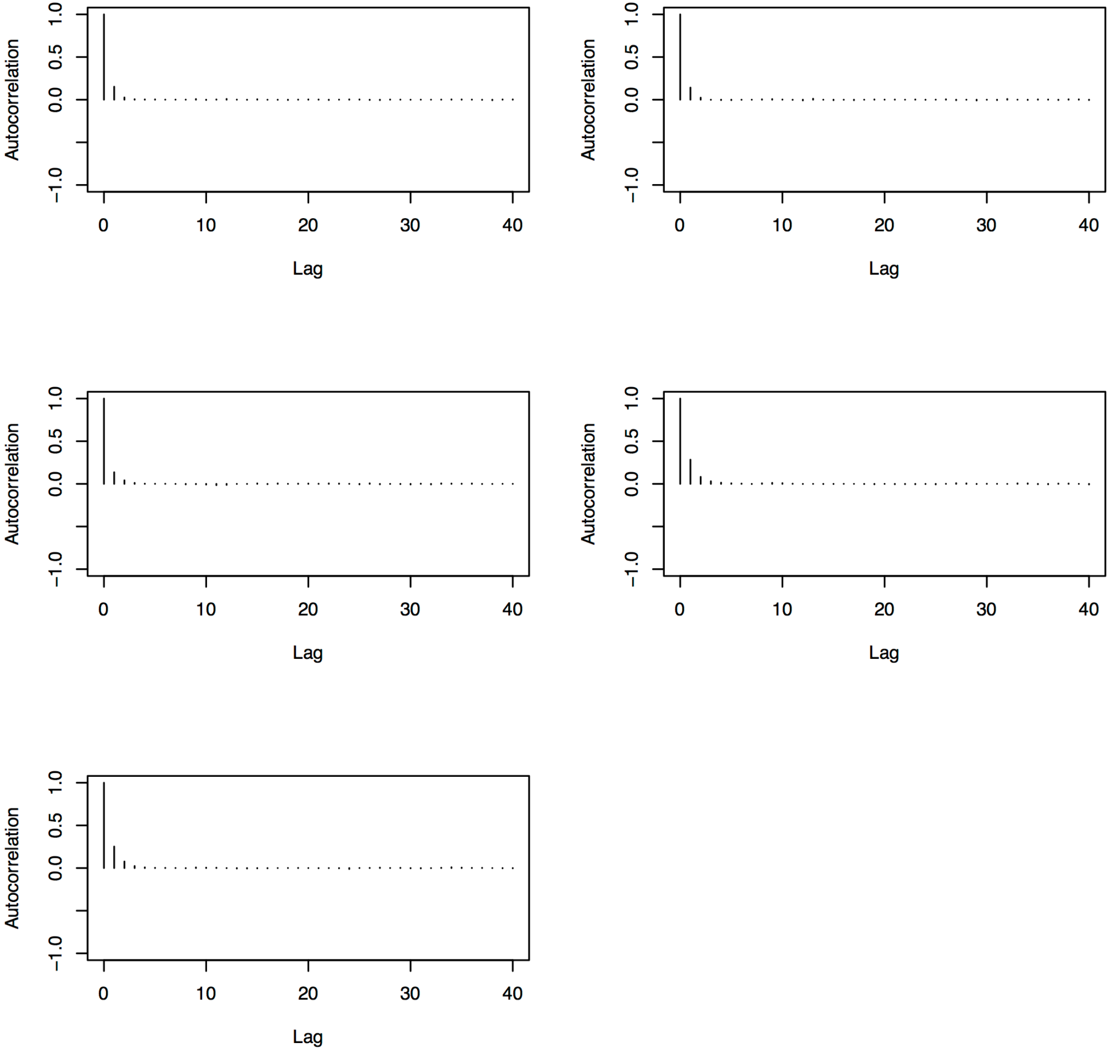
Supplementary Figure S3. *Gibbs sampler diagnostic statistics.* 50,000 sampling iterations were performed for the toy example with five individuals presented in Figure 5B of the main text. In Panels A to D, all sampling iterations are included, i.e. no burn-in is discarded. Panel A shows the trace and resulting density estimates (less smoothed versions of densities shown in Figure 5B) for the first 2000 iterations. Panel B and C show the running mean age for all 50,000 respectively for the first 500 samples. Panel D visualises the autocorrelation between consecutive samples. Panel E and F show Gelman and Rubin's shrink factor (8) on all respectively the first 2000 samples after discarding the first 10,000 in 4 independent runs of the Gibbs sampler.

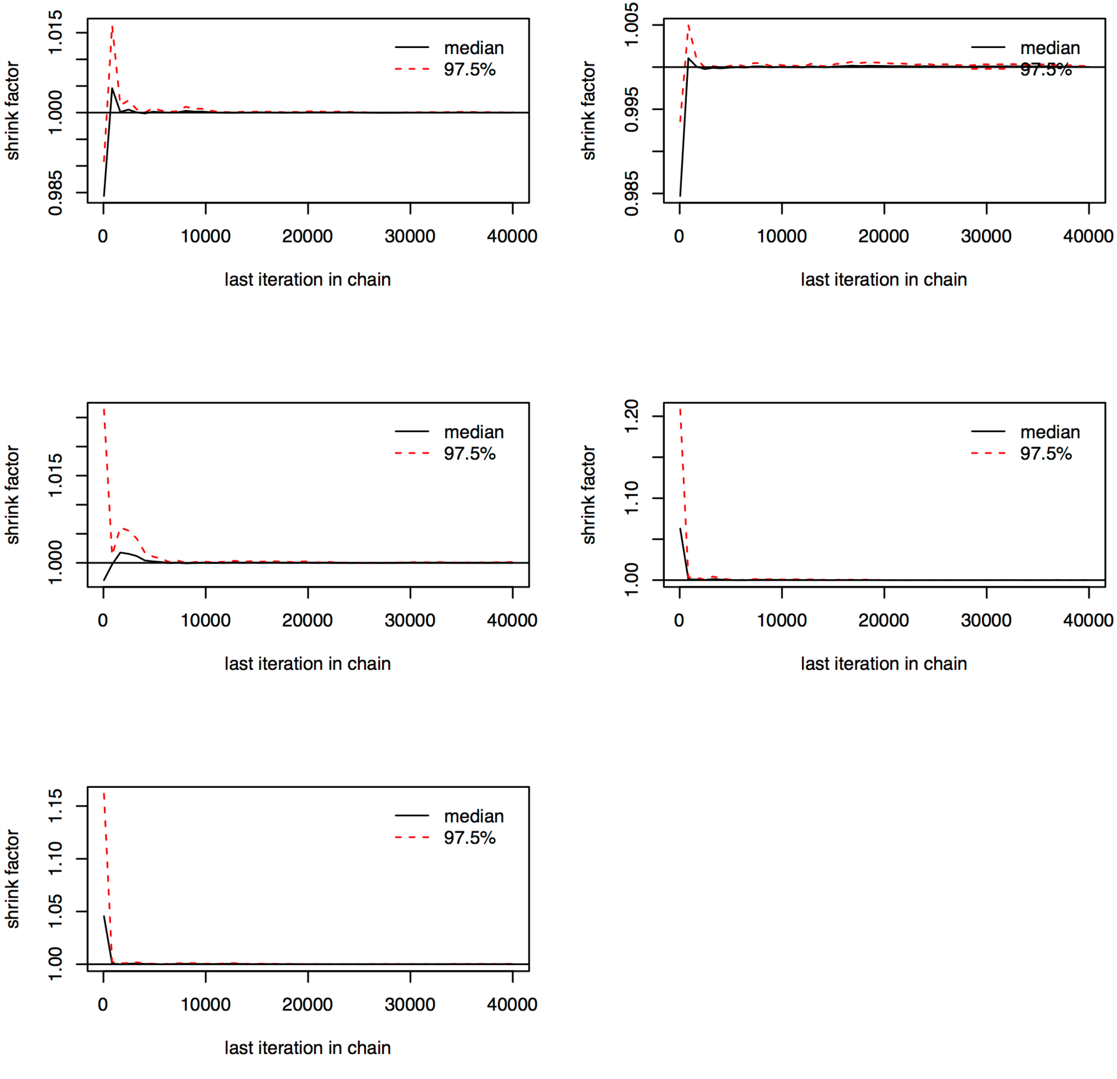

Trace of var1
5
4
3
2
1
0
500
1000
1500
2000
Iterations
Density of var1
0.30
0.15
0.00
1
2
3
4
5
N = 2000   Bandwidth = 0.2553
Trace of var2
7
6
5
4
3
2
0
500
1000
1500
2000
Iterations
Density of var2
0.20
0.10
0.00
1
2
3
4
5
6
7
8
N = 2000   Bandwidth = 0.2978
Trace of var3
7.2
7.0
6.8
6.6
0
500
1000
1500
2000
Iterations
Density of var3
1.2
0.8
0.4
0.0
6.4
6.6
6.8
7.0
7.2
7.4
N = 2000   Bandwidth = 0.05282

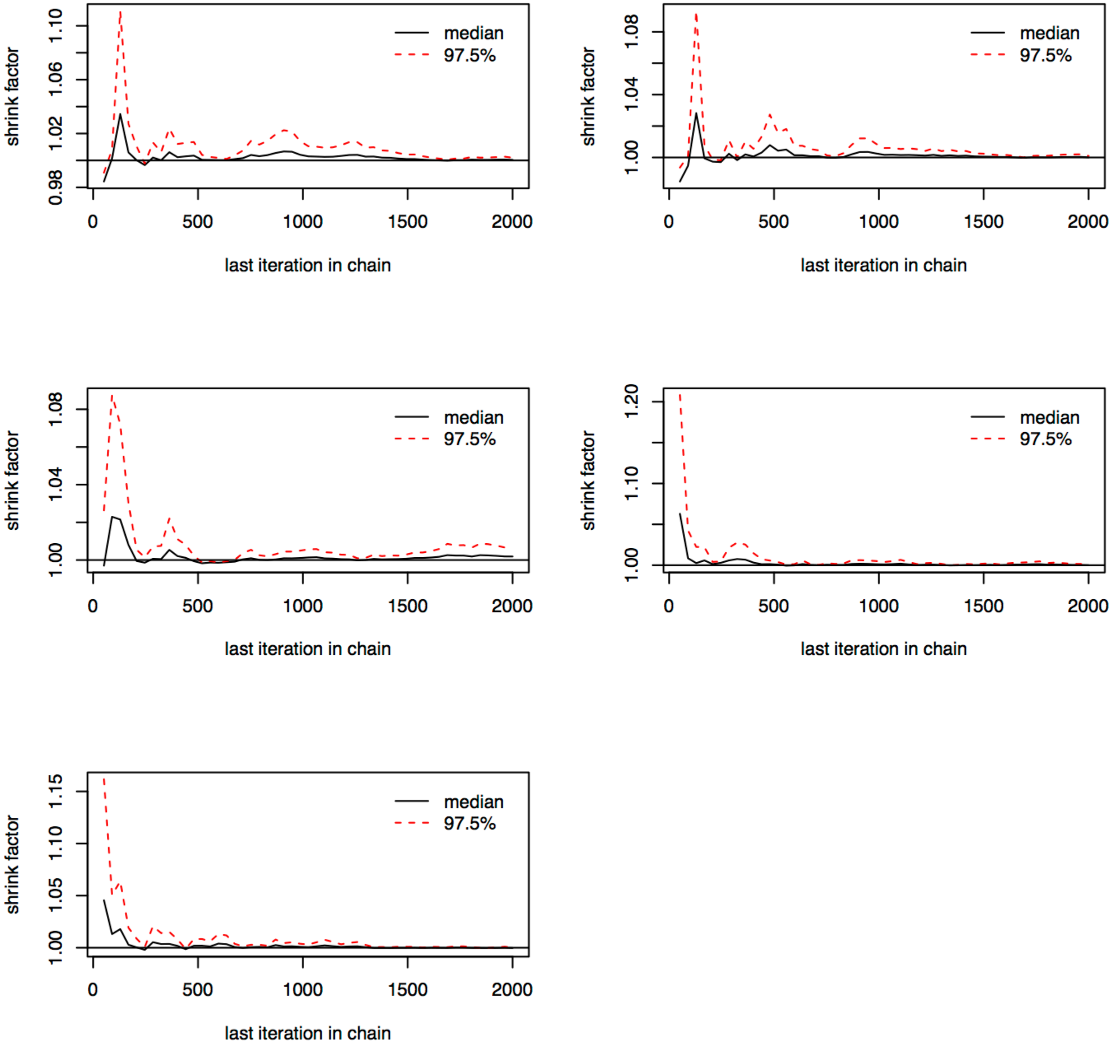

Trace of var4
6.5
7.5
8.5
9.5
0
500
1000
1500
2000
Iterations
Density of var4
0.0
0.2
0.4
6
7
8
9
10
N = 2000 Bandwidth = 0.1751
Trace of var5
7.0
8.0
9.0
10.0
0
500
1000
1500
2000
Iterations
Density of var5
0.0
0.2
0.4
7
8
9
10
N = 2000 Bandwidth = 0.178


var1
2.8
2.6
2.4
2.2
2.0
1.8
0
10000
20000
30000
40000
50000
var2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
0
10000
20000
30000
40000
50000
var3
6.8
6.7
6.6
6.5
0
10000
20000
30000
40000
50000
var4
8.0
7.5
7.0
6.5
0
10000
20000
30000
40000
50000
var5
9.0
8.5
8.0
7.5
7.0
0
10000
20000
30000
40000
50000

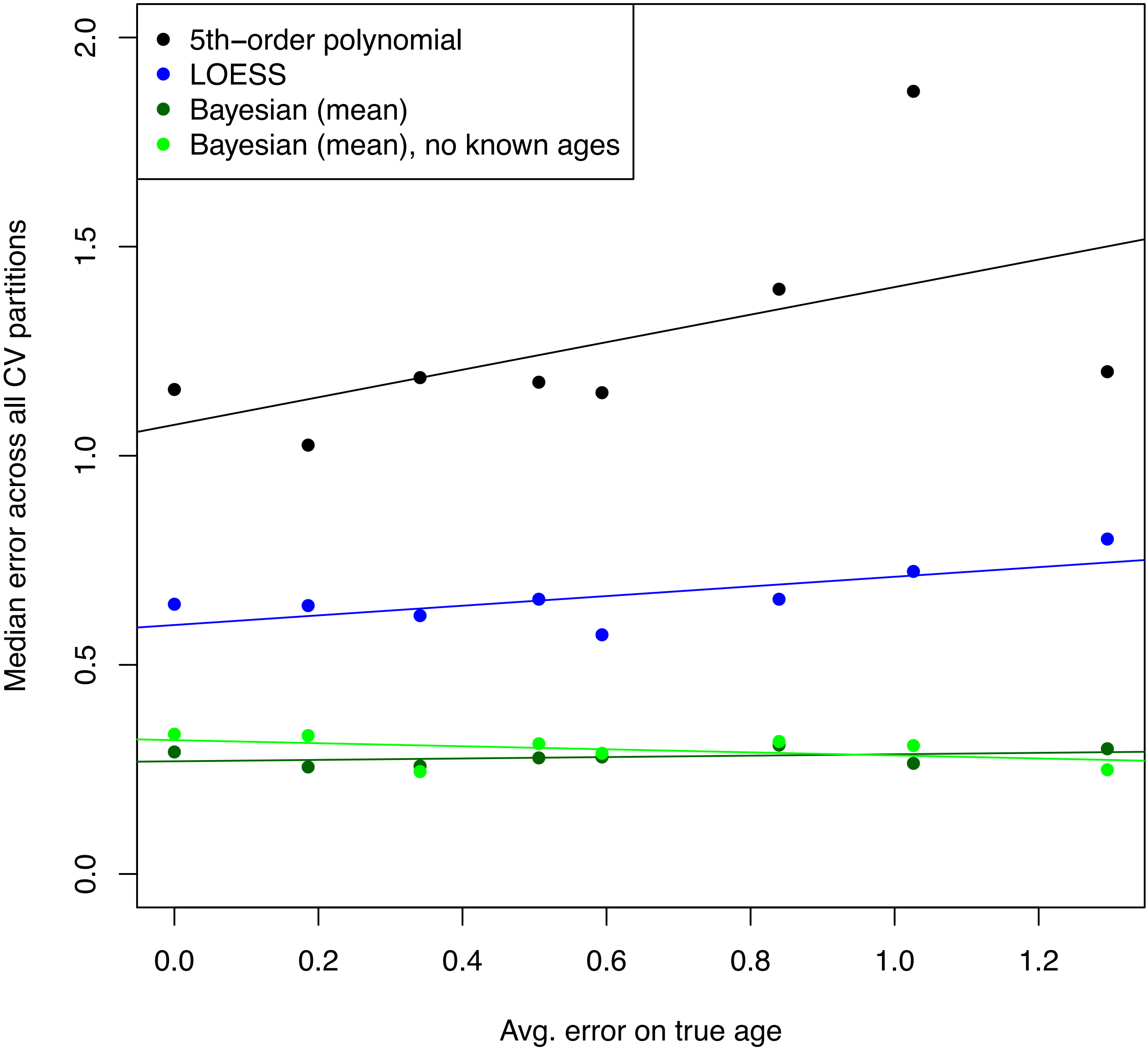

var1
1.8
2.0
2.2
2.4
2.6
2.8
0
100
200
300
400
500
var2
3.8
4.0
4.2
4.4
4.6
4.8
5.0
0
100
200
300
400
500
var3
6.50
6.60
6.70
6.80
0
100
200
300
400
500
var4
6.5
7.0
7.5
8.0
0
100
200
300
400
500
var5
7.0
7.5
8.0
8.5
9.0
0
100
200
300
400
500

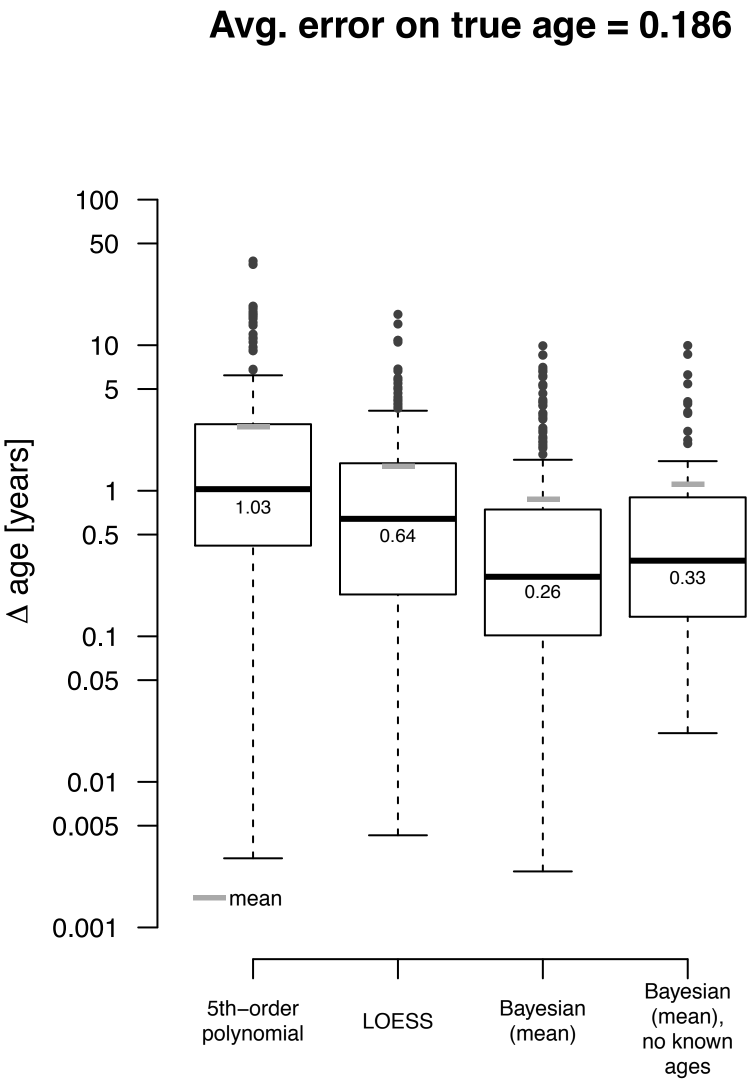

Autocorrelation
1.0
0.5
0.0
−1.0
0
10
20
30
40
Lag
Autocorrelation
1.0
0.5
0.0
−1.0
0
10
20
30
40
Lag
Autocorrelation
1.0
0.5
0.0
−1.0
0
10
20
30
40
Lag
Autocorrelation
1.0
0.5
0.0
−1.0
0
10
20
30
40
Lag
Autocorrelation
1.0
0.5
0.0
−1.0
0
10
20
30
40
Lag

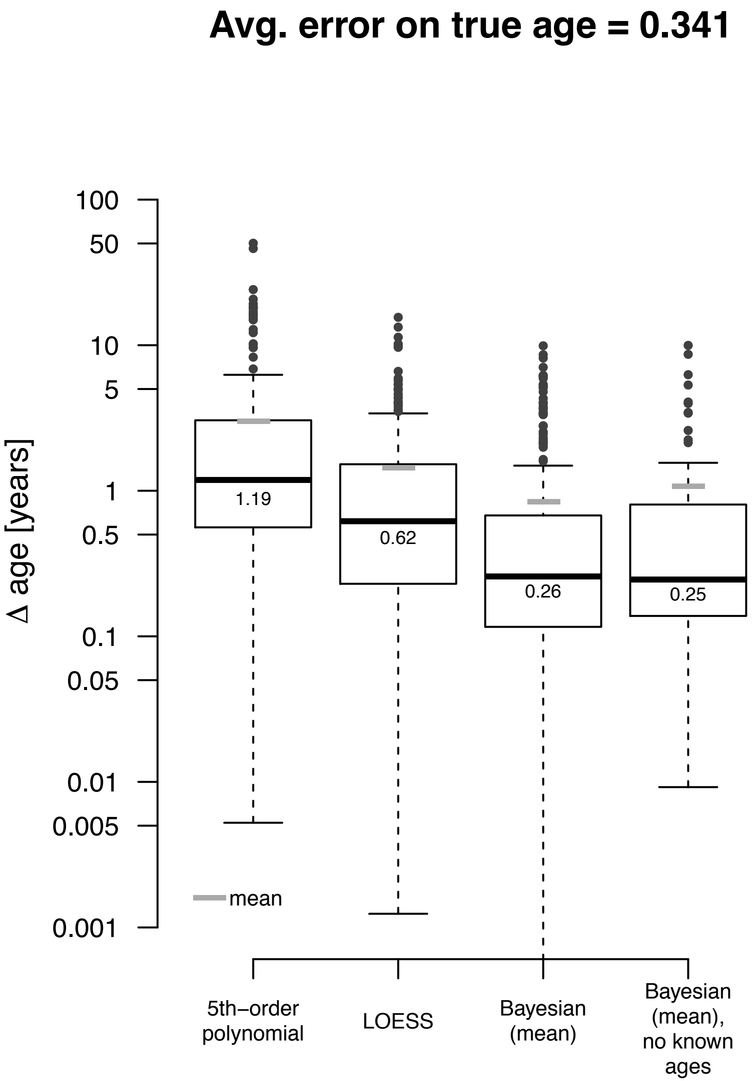

shrink factor
0.985
1.000
1.015
0
10000
20000
30000
40000
last iteration in chain
median
97.5%
shrink factor
0.985
0.995
1.005
0
10000
20000
30000
40000
last iteration in chain
median
97.5%
shrink factor
1.000
1.015
0
10000
20000
30000
40000
last iteration in chain
median
97.5%
shrink factor
1.00
1.10
1.20
0
10000
20000
30000
40000
last iteration in chain
median
97.5%
shrink factor
1.00
1.05
1.10
1.15
0
10000
20000
30000
40000
last iteration in chain
median
97.5%

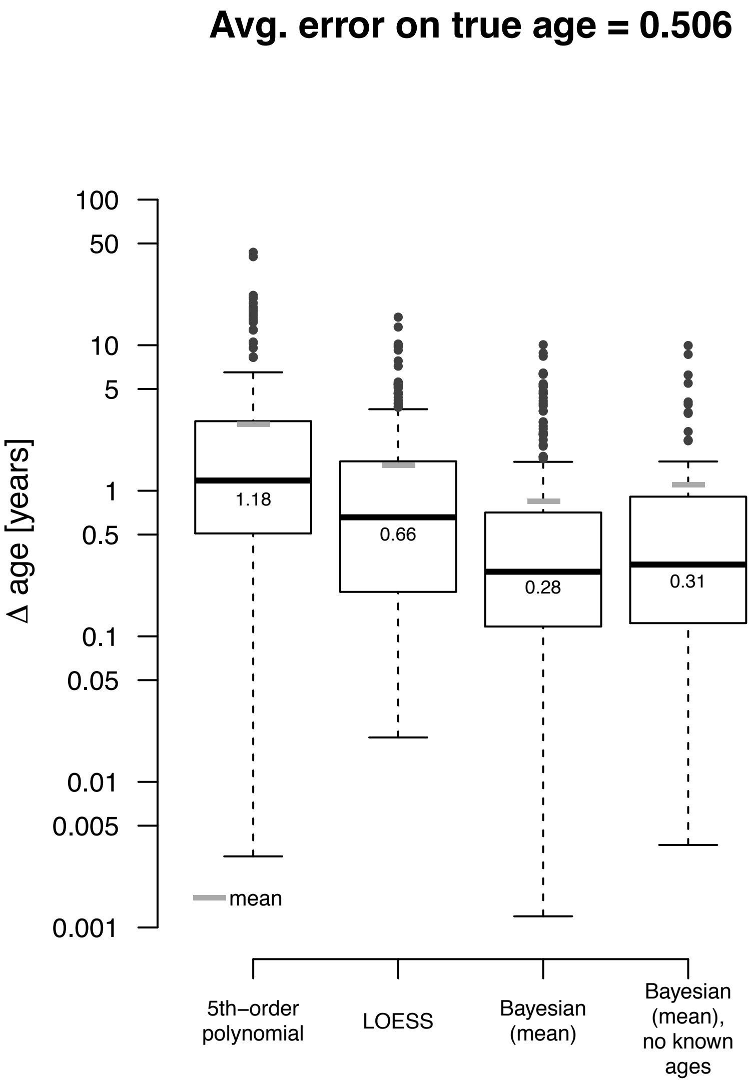

shrink factor
0.98
1.02
1.06
1.10
median
97.5%
0
500
1000
1500
2000
last iteration in chain
shrink factor
1.00
1.04
1.08
median
97.5%
0
500
1000
1500
2000
last iteration in chain
shrink factor
1.00
1.04
1.08
median
97.5%
0
500
1000
1500
2000
last iteration in chain
shrink factor
1.00
1.10
1.20
median
97.5%
0
500
1000
1500
2000
last iteration in chain
shrink factor
1.00
1.05
1.10
1.15
median
97.5%
0
500
1000
1500
2000
last iteration in chain

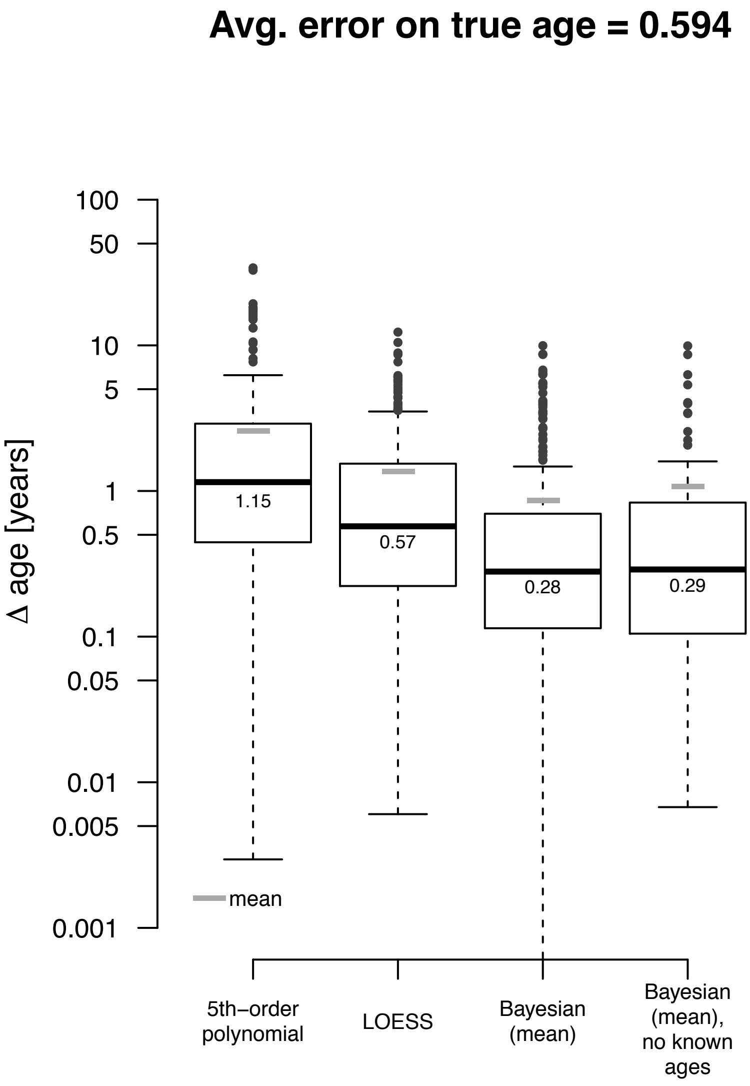
Supplementary Figure S4. *Estimation robustness to error in known ages.* We repeated the validation from Figures 1 and 2, however, added different amounts of error to the individuals' ages, where errors are constrained not to change the ranking order. Panel A summarizes how this affects the different methods: linear regression shows that estimation accuracy measured as the median of the differences between estimated and actual ages of the individuals across the 5 cross-validation partitions is reduced most for the polynomial regression approach, slightly for LOESS and not at all for our Bayesian method. Panel B gives the corresponding distributions in form of boxplots.

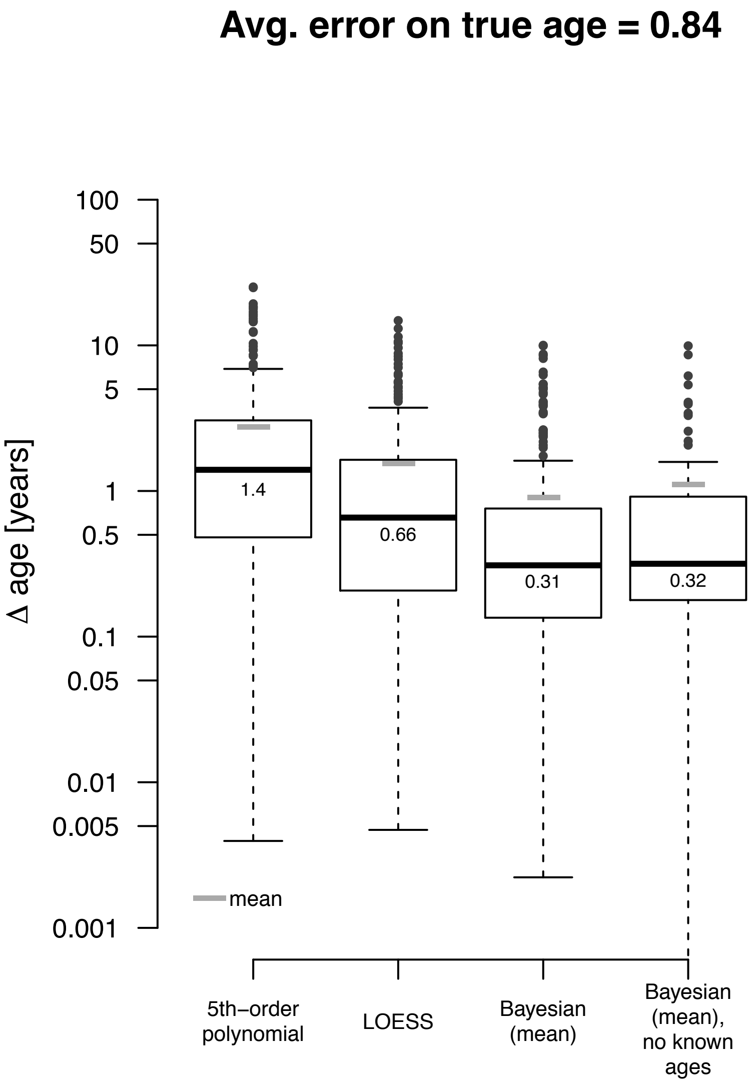

5th−order polynomial
LOESS
Bayesian (mean)
Bayesian (mean), no known ages
Median error across all CV partitions
Avg. error on true age
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2

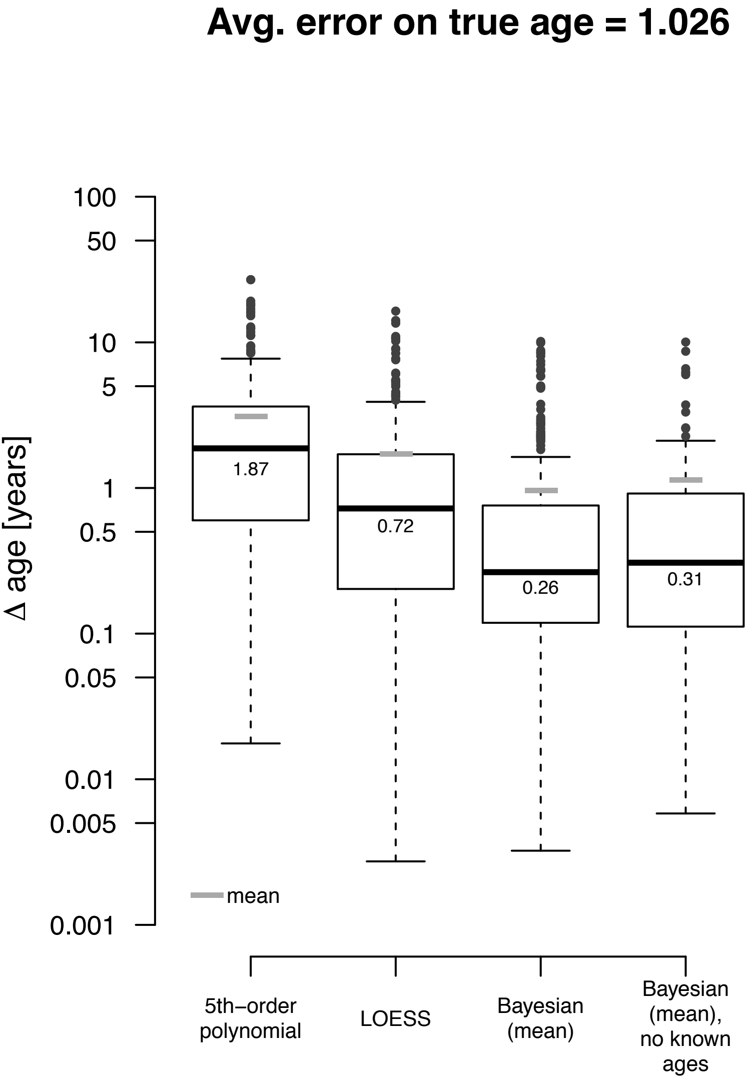

Avg. error on true age = 0.186
Δ age [years]
100
50
10
5
1
0.5
0.1
0.05
0.01
0.005
0.001
1.03
0.64
0.26
0.33
mean
5th−order polynomial
LOESS
Bayesian (mean)
Bayesian (mean), no known ages

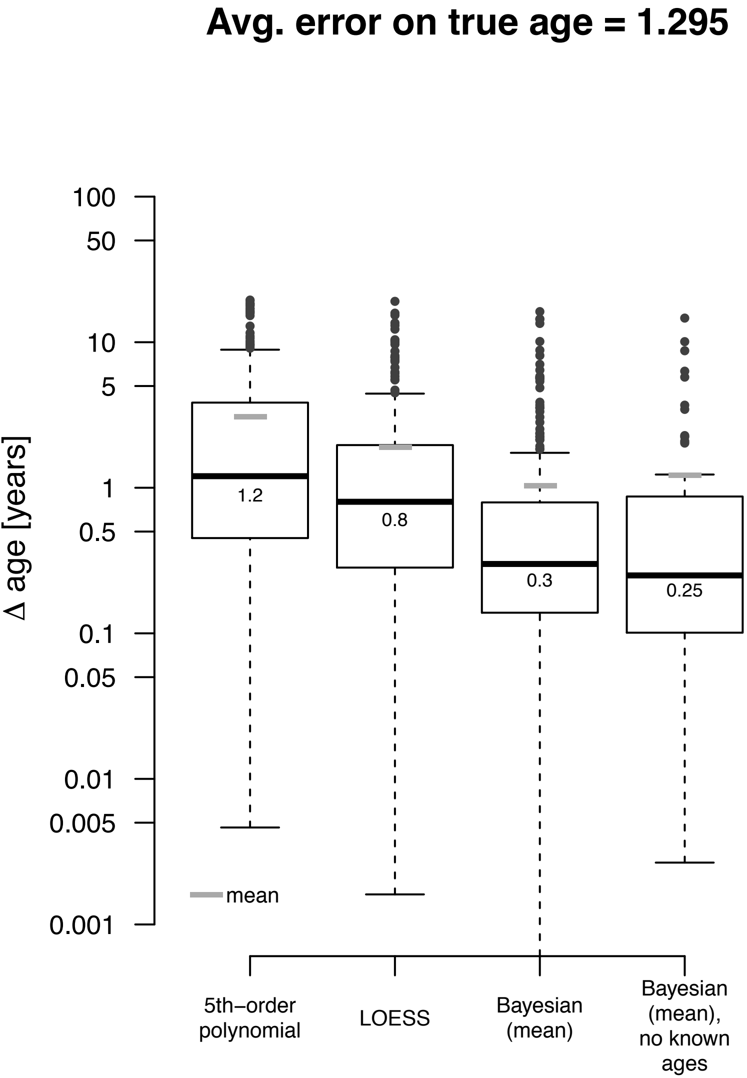

Avg. error on true age = 0.341
Δ age [years]
100
50
10
5
1
0.5
0.1
0.05
0.01
0.005
0.001
1.19
0.62
0.26
0.25
mean
5th−order polynomial
LOESS
Bayesian (mean)
Bayesian (mean), no known ages


Avg. error on true age = 0.506
Δ age [years]
100
50
10
5
1
0.5
0.1
0.05
0.01
0.005
0.001
1.18
0.66
0.28
0.31
mean
5th−order polynomial
LOESS
Bayesian (mean)
Bayesian (mean), no known ages

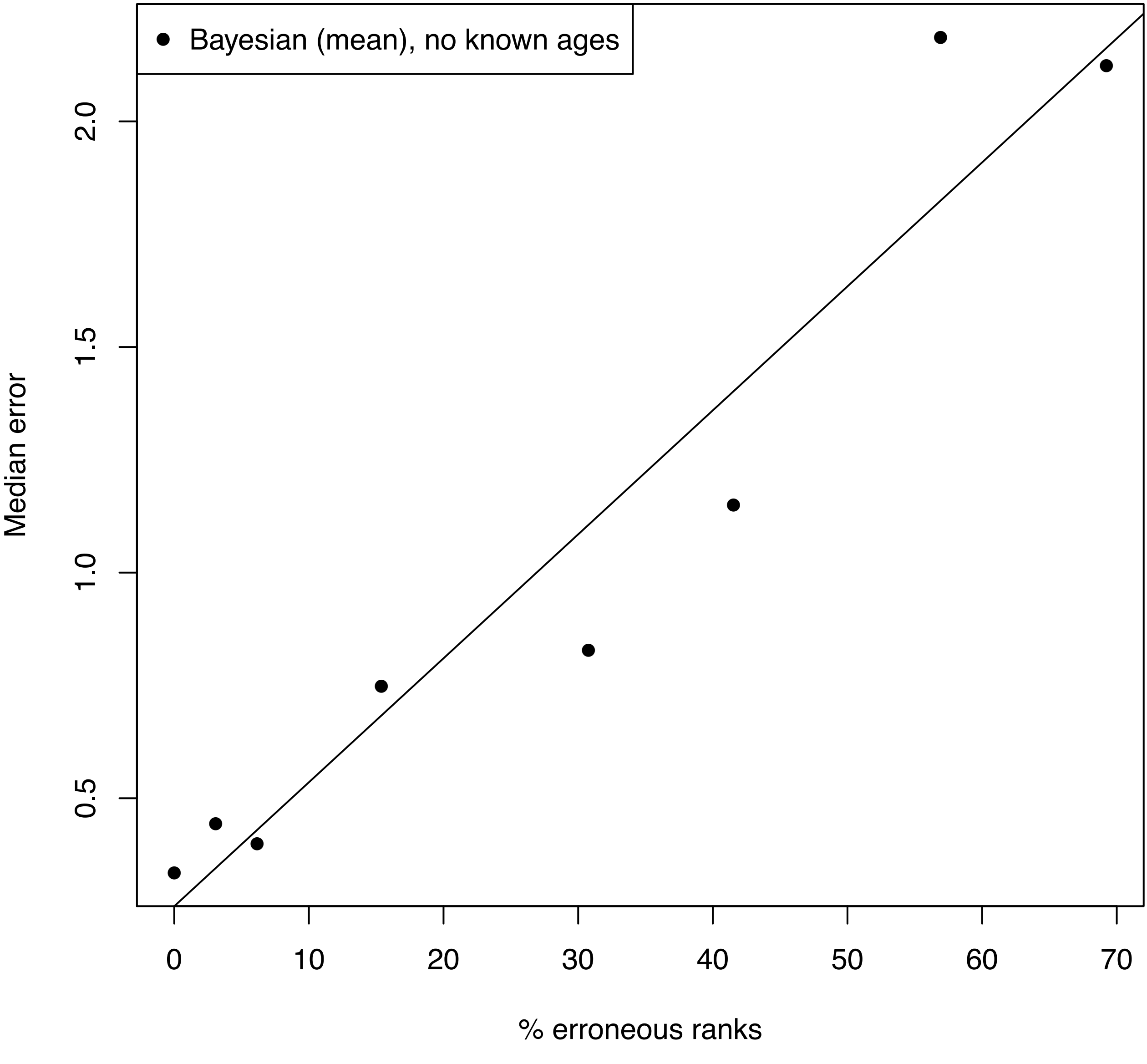

Avg. error on true age = 0.594
Δ age [years]
100
50
10
5
1
0.5
0.1
0.05
0.01
0.005
0.001
1.15
0.57
0.28
0.29
mean
5th−order polynomial
LOESS
Bayesian (mean)
Bayesian (mean), no known ages

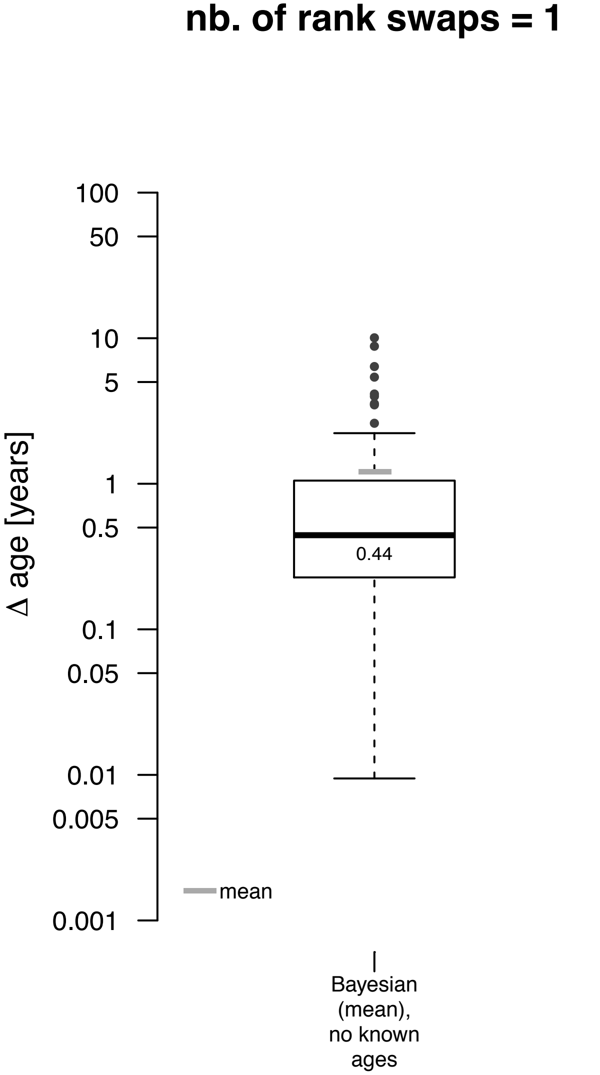

Avg. error on true age = 0.84
Δ age [years]
100
50
10
5
1
0.5
0.1
0.05
0.01
0.005
0.001
1.4
0.66
0.31
0.32
mean
5th–order polynomial
LOESS
Bayesian (mean)
Bayesian (mean), no known ages

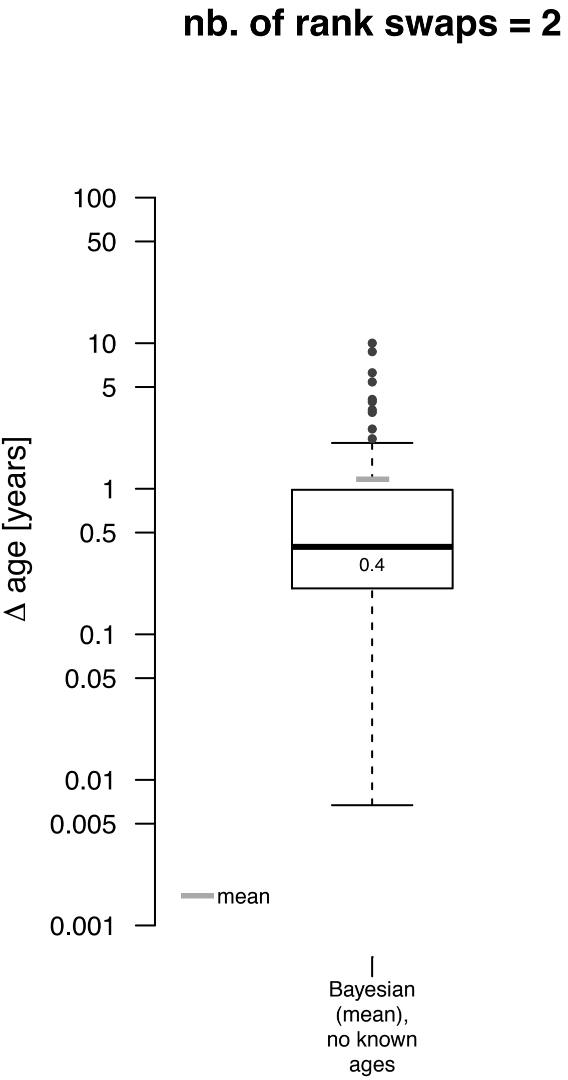

Avg. error on true age = 1.026
Δ age [years]
100
50
10
5
1
0.5
0.1
0.05
0.01
0.005
0.001
1.87
0.72
0.26
0.31
mean
5th−order polynomial
LOESS
Bayesian (mean)
Bayesian (mean), no known ages

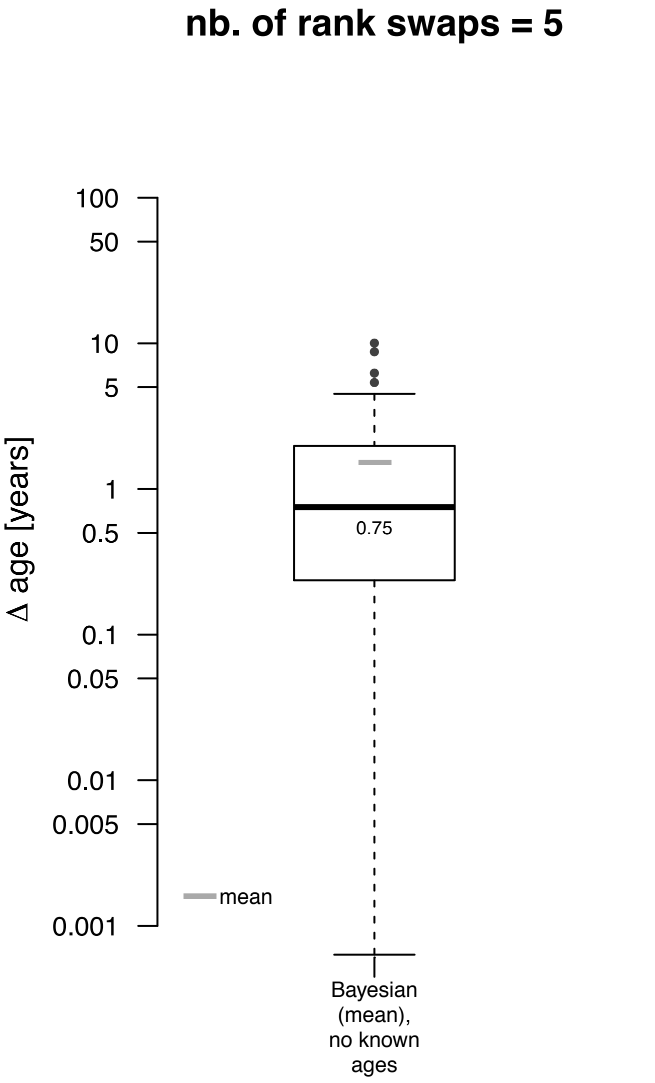

Avg. error on true age = 1.295
Δ age [years]
100
50
10
5
1
0.5
0.1
0.05
0.01
0.005
0.001
1.2
0.8
0.3
0.25
mean
5th−order polynomial
LOESS
Bayesian (mean)
Bayesian (mean), no known ages

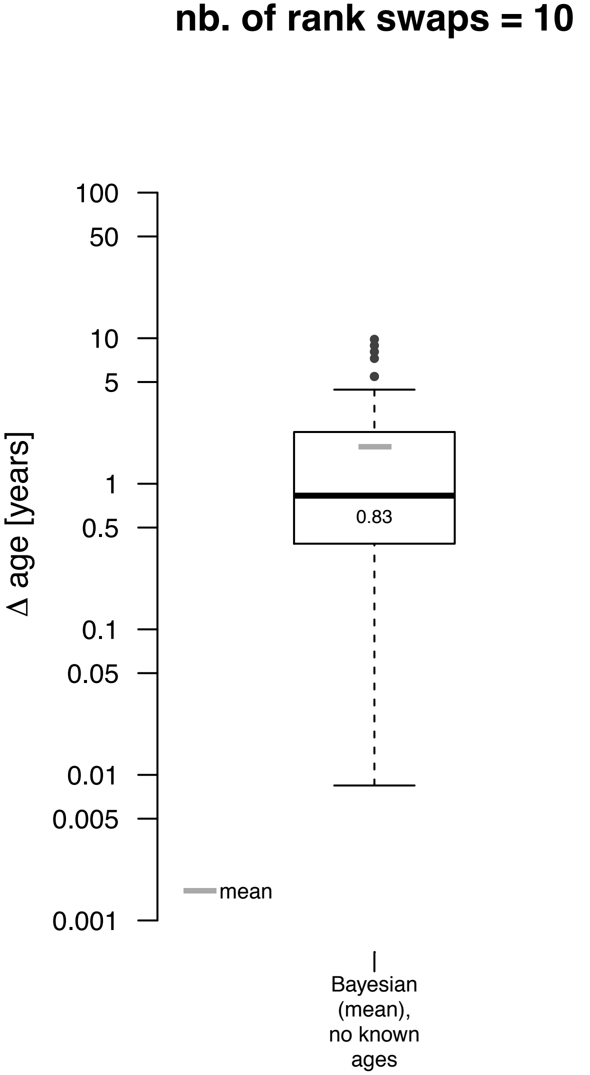
Supplementary Figure S5. *Estimation robustness to error in ranking order.* We repeated validation from Figures 1 and 2, however, introduced different amounts of error in the ranking order (all errors we introduce are consistent with the age brackets). As changing the ranking order would require to adjust the age of the individuals to reflect the altered ranking order, we focus on the performance of our Bayesian method when no ages are considered known. This prevents that the effects of errors in ranking order and age (see Supplementary Figure S4) are conflated. Panel A summarizes the results showing the medians of the differences between estimated and actual ages of the individuals, Panel B gives the corresponding distributions in form of boxplots.

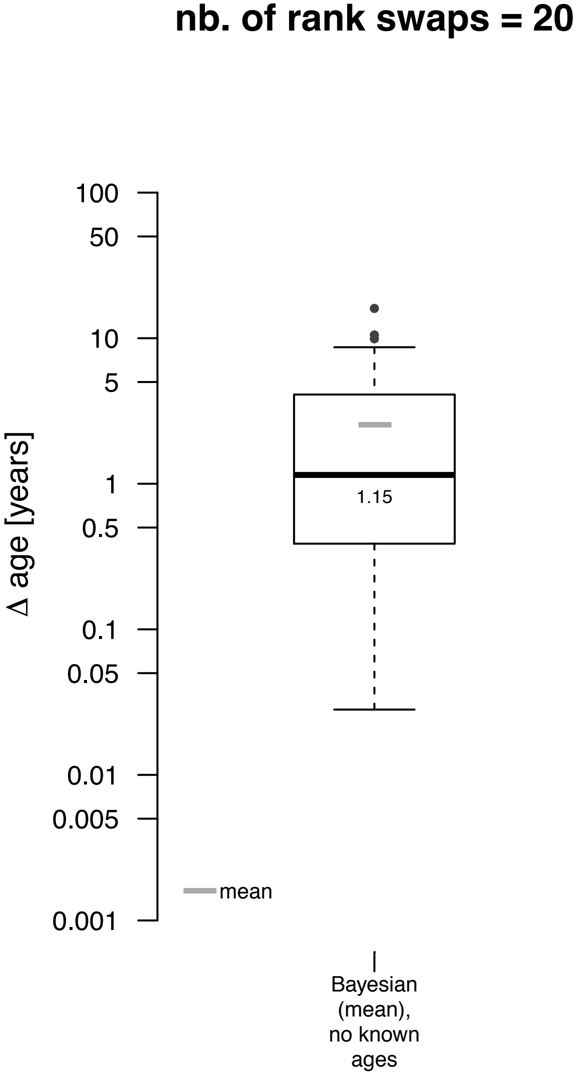

● Bayesian (mean), no known ages
Median error
0.5
1.0
1.5
2.0
0
10
20
30
40
50
60
70
% erroneous ranks


nb. of rank swaps = 1
Δ age [years]
100
50
10
5
1
0.5
0.1
0.05
0.01
0.005
0.001
0.44
mean
Bayesian (mean), no known ages

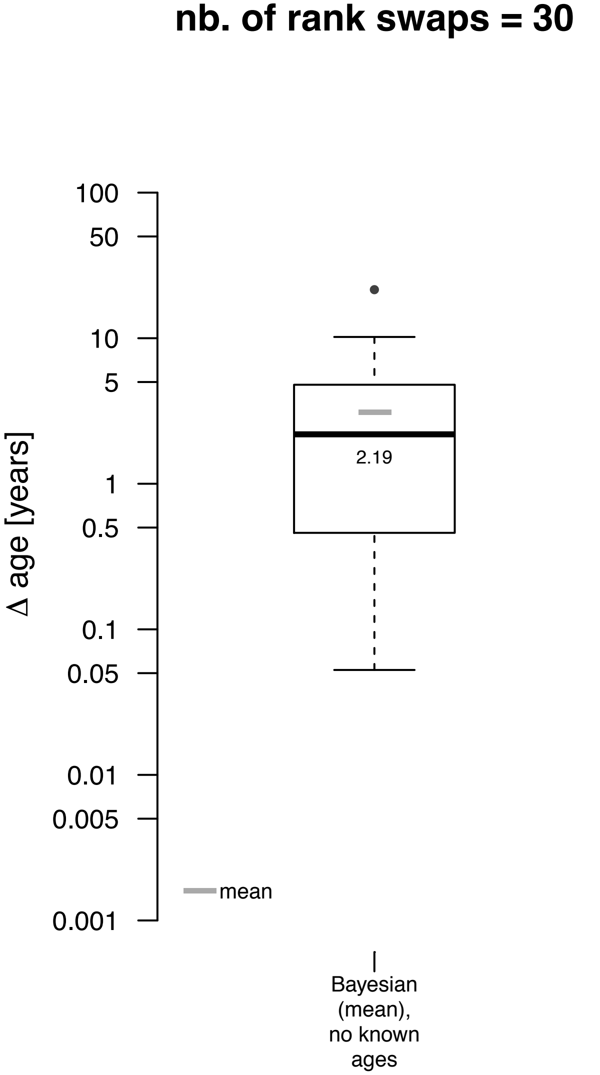

nb. of rank swaps = 2
Δ age [years]
100
50
10
5
1
0.5
0.1
0.05
0.01
0.005
0.001
0.4
mean
Bayesian (mean), no known ages


nb. of rank swaps = 5

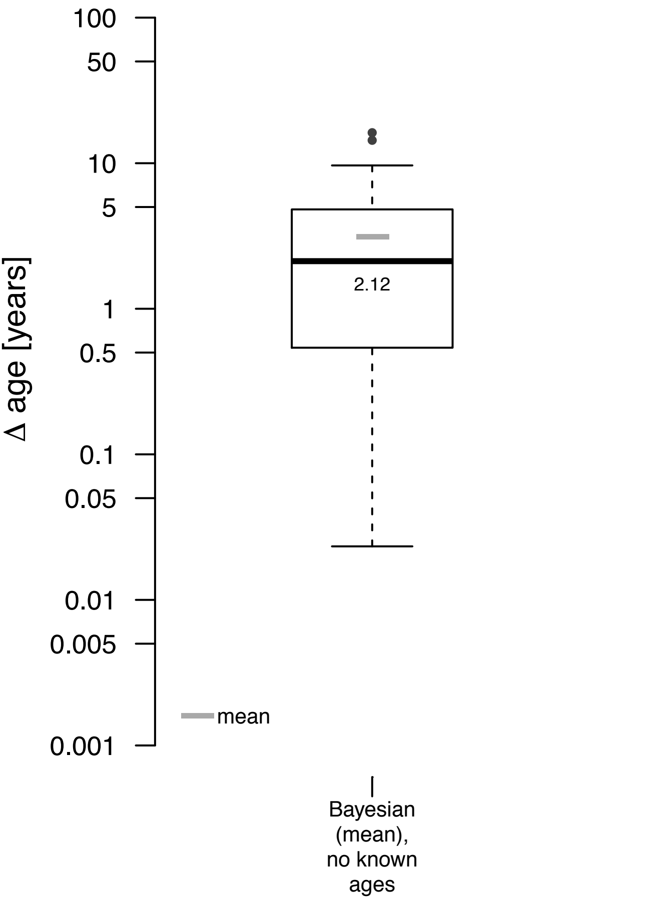

nb. of rank swaps = 10
Δ age [years]
100
50
10
5
1
0.5
0.1
0.05
0.01
0.005
0.001
0.83
mean
Bayesian
(mean),
no known
ages



nb. of rank swaps = 20

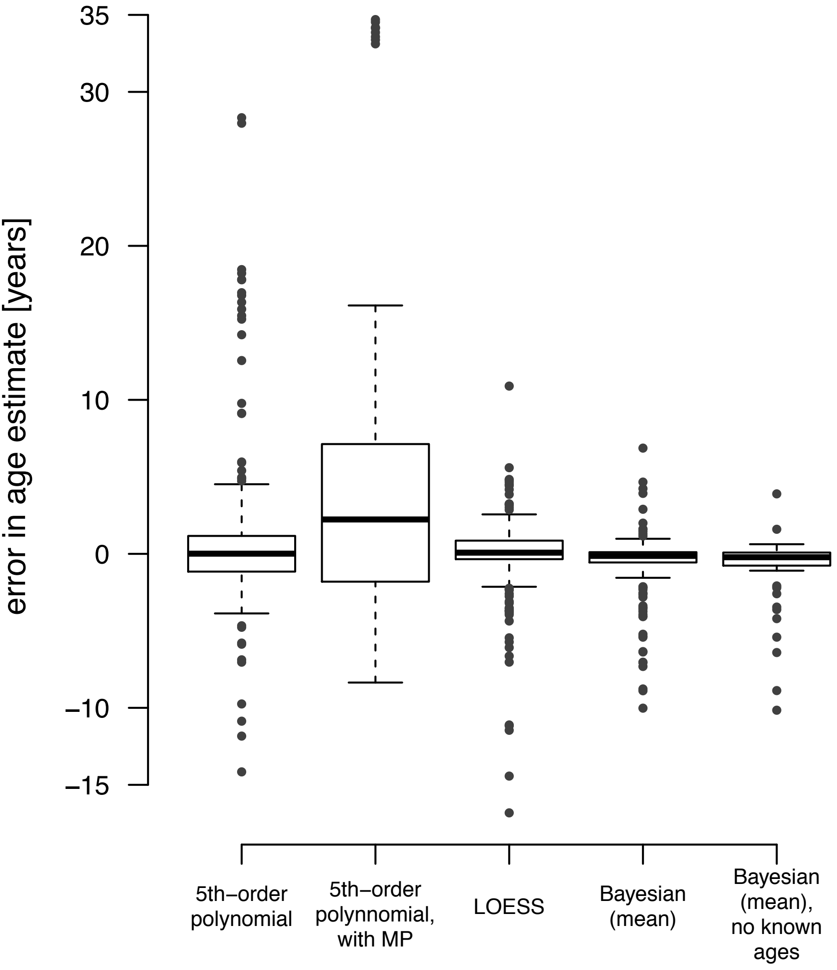


nb. of rank swaps = 30



nb. of rank swaps = 50

Supplementary Figure S6. *Raw values behind Figure 1.* We show the same distributions as in Figure 1 in the main text, however, without showing absolute differences and with a y-axis in natural scale.


error in age estimate [years]
35
30
20
10
0
−10
−15
5th−order polynomial
5th−order polynnomial, with MP
LOESS
Bayesian (mean)
Bayesian (mean), no known ages

SUP. TABLES

Supplementary Table S1. *Numerical values corresponding to absolute differences between actual and estimated ages shown in Figure 1 of the main text.* The minimum, 25th percentile, median, mean, 75th percentile and maximum given in the last row (total) directly correspond to the boxplots plotted in Figure 1. The remaining rows provide more detail as the results are split by age cohort. Bold red values indicate worst, bold black best performance. See legend of Figure 1 and explanation of the benchmarking procedure in the main text for further information. Note that photographs in the Headland database (13) were taken in different years (between 1972 and 2010), and all ages and age estimates were therefore adjusted to the present day (2015). Hence, the youngest age is 15 explaining why the 10-20 cohort is the first row.
Abbreviations: minimum (min.), maximum (max.), percentile (per.), standard deviation (sd.), mid-point (MP)

Age Cohort	Sample Size	Statis-tic	5th-order polynomial	5th-order polynomial, with MP	LOESS	Gibbs (mean)	Gibbs (mean), no known ages
10-20	10	min.	0.06	**1.17**	0.03	**0.00**	0.04
		25th per.	0.71	**6.71**	0.12	**0.08**	0.15
		median	1.64	**7.92**	**0.18**	**0.18**	0.24
		mean (sd.)	4.71 (6.30)	**7.96** (3.20)	**0.25** (0.20)	**0.25** (0.27)	0.28 (0.23)
		75th per.	3.91	**8.88**	**0.29**	0.31	0.32
		max.	**18.46**	14.55	**0.84**	1.12	0.85
20-45	40	min.	0.01	**0.02**	0.01	**0.00**	**0.00**
		25th per.	0.23	**1.70**	0.20	**0.11**	0.12
		median	0.67	**4.32**	0.57	**0.25**	0.26
		mean (sd.)	1.13 (1.19)	**4.25** (2.73)	0.94 (1.11)	**0.45** (0.58)	0.47 (0.56)
		75th per.	1.65	**6.23**	1.19	**0.49**	0.53
		max.	5.41	**10.37**	5.59	3.38	**2.59**
45+	15	min.	0.08	0.07	0.14	**0.01**	**0.21**
		25th per.	1.64	**1.81**	1.25	**0.57**	0.88
		median	3.42	2.89	2.71	**1.03**	**3.46**
		mean (sd.)	5.38 (6.01)	**8.02** (11.03)	4.09 (4.05)	**2.57** (2.68)	3.45 (3.15)
		75th per.	5.94	**11.83**	5.28	**4.06**	4.81
		max.	28.32	**34.70**	16.82	**10.02**	10.15
Total	65	min.	0.01	**0.02**	0.01	**0.00**	**0.00**
		25th per.	0.38	**2.05**	0.21	**0.12**	0.18
		median	1.16	**4.39**	0.64	**0.29**	0.33
		mean (sd.)	2.66 (4.35)	**5.69** (6.09)	1.47 (2.36)	**0.91** (1.64)	1.13 (2.01)
		75th per.	2.90	**7.36**	1.57	**0.80**	0.90
		max.	28.32	**34.70**	16.82	**10.02**	10.15

Supplementary Table S2. *Kolmogorov-Smirnov p-values and Bayes factors for all pairwise comparisons of error distributions shown in Figure 1.* BFs greater than three are considered positive evidence, above 150 as strong evidence. *Abbreviations:* mid-point (MP), Bayes factor (BF)

	5th-order polynomial	**5th-order polynomial, with MP**	**LOESS**	**Gibbs (mean)**

5th-order polynomial, with MP	*p*=1.554312e-15;
	BF=4.128145e+20			

LOESS	*p*=0.0004320986;
	BF= 29.39566
		p=1.776357e-15;
		BF= 7.426503e+38		

Gibbs (mean)	*p*=3.108624e-15;
	BF=2.81064e+12
		p=1.554312e-15;
		BF=2.04265e+63
			p=4.486276e-06;
			BF=1377.745	

Gibbs (mean), no known ages	*p*=9.447281e-06;
	BF=419.9913
		p=7.771561e-16;
		BF=6.354463e+27
			p=0.02777288;
			BF=1.054143
				p=0.6081314;
				BF=0.2328565

Supplementary Table S3. *Average difference between upper and lower bound of the age bracket and number of accurately known ages for different age cohorts of the Palanan Agta*. For the purposes of this table, the mean value of the upper and lower bound was considered an individual's age and used for grouping into cohorts. Number of exact birth dates and birth dates accurate within +/- 1 year are also displayed.

Age Cohort	Sample Size	Average Difference	Number of Exact Birthdates	Percentage of Exact Birthdates	Number of Birthdates +/- 1 year	Percentage of Birthdates +/- 1 year
<1	20	0.16	15	75%	20	100%
1-5	103	1.73	30	29.13%	67	65.05%
5-10	103	3	19	18.45%	33	32.04%
10-20	116	4.1	13	11.21%	33	28.45%
20-45	164	9.47	18	10.98%	26	15.85%
45+	81	18.56	3	3.7%	12	14.81%
Total	587	6.85	98	16.7%	191	32.54%

REFERENCES

1. Kaplan H, Hill J, Lancaster J, Hurtado A M, Hill K I M, Lancaster J, Hurtado A M (2000) A theory of human life history evolution: diet, intelligence, and longevity. *Evolutionary Anthropology 9*:156–185.

2. Hill K, Hurtado A M (1996) *Aché Life History: The Ecology and Demography of a Foraging People* (Aldine de Gruyter, New Haven).

3. Walsh B (2004) Markov Chain Monte Carlo and Gibbs Sampling. *Lecture Notes for EEB*:1–24.

4. Headland T N, Headland J D, Uehara R T (2011) *Agta Demographic Database: Chronicle of a hunter-gatherer community in transition* (SIL Language and Culture Documentation and Description, 2).

5. Python Software Foundation. (2016) Python Language Reference.

6. R Core Team (2012) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

7. Cleveland W S, Grosse E, Shyu W M (1992) Local Regression Models. *Statistical Models in S,* eds Chambers J M, Hastie J J (Wadsworth & Brooks, Pacific Grove, CA), pp 309–376.

8. Gelman A, Rubin DB (1992) Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science 7*(4):457–511.

9. Raftery AE, Lewis SM (1995) The number of iterations, convergence diagnostics and generic Metropolis algorithms. In: Gilks WR, Spiegelhalter DJ, Richardson S, eds. *Practical Markov Chain Monte Carlo*. London, UK: Chapman and Hall; pp 1–15.