Estimating sexual size dimorphism in fossil species from posterior probability densities

Tomohiko Sasaki [a,1] (ID)**, Sileshi Semaw** [b,c]**, Michael J. Rogers** [d] (ID)**, Scott W. Simpson** [e] (ID)**, Yonas Beyene** [f]**, Berhane Asfaw** [g]**, Tim D. White** [b,h] (ID)**, and Gen Suwa** [i,1]

[a]The Kyoto University Museum, Kyoto University, Kyoto 606-8501, Japan; [b]Centro Nacional de Investigación Sobre la Evolución Humana 09002 Burgos, Spain; [c]Stone Age Institute, Gosport, IN 47408; [d]Department of Anthropology, Southern Connecticut State University, New Haven, CT 06515; [e]Department of Anatomy, Case Western Reserve University School of Medicine, Cleveland, OH 44106; [f]French Center for Ethiopian Studies, Addis Ababa, Ethiopia; [g]Rift Valley Research Service, Addis Ababa, Ethiopia; [h]Department of Integrative Biology, University of California, Berkeley, CA 94720-3140; and [i]The University Museum, The University of Tokyo, Tokyo 113-0033, Japan

Accurate characterization of sexual dimorphism is crucial in evolutionary biology because of its significance in understanding present and past adaptations involving reproductive and resource use strategies of species. However, inferring dimorphism in fossil assemblages is difficult, particularly with relatively low dimorphism. Commonly used methods of estimating dimorphism levels in fossils include the mean method, the binomial dimorphism index, and the coefficient of variation method. These methods have been reported to overestimate low levels of dimorphism, which is problematic when investigating issues such as canine size dimorphism in primates and its relation to reproductive strategies. Here, we introduce the posterior density peak (pdPeak) method that utilizes the Bayesian inference to provide posterior probability densities of dimorphism levels and within-sex variance. The highest posterior density point is termed the pdPeak. We investigated performance of the pdPeak method and made comparisons with the above-mentioned conventional methods via 1) computer-generated samples simulating a range of conditions and 2) application to canine crown-diameter datasets of extant known-sex anthropoids. Results showed that the pdPeak method is capable of unbiased estimates in a broader range of dimorphism levels than the other methods and uniquely provides reliable interval estimates. Although attention is required to its underestimation tendency when some of the distributional assumptions are violated, we demonstrate that the pdPeak method enables a more accurate dimorphism estimate at lower dimorphism levels than previously possible, which is important to illuminating human evolution.

sexual dimorphism | fossils | Bayesian estimate | mixture analysis | human evolution

S exual dimorphism across primates and in humans has been investigated to elucidate its evolutionary significance. Particular attention has been paid to body, skeletal, and canine size dimorphism. This is because of suggested relationships of these dimorphisms with ecological and sociobehavioral variables, especially in relation to reproductive behavior, and also because these parameters can be assessed in fossil assemblages (see refs. 1–3 for relatively recent overviews). Following the influential work of Clutton-Brock et al. (4), the ratio between male and female mean sizes (or its logarithm) (5, 6) has been the predominant parameter used in quantifying sexual size dimorphism, a simple, but fundamental, parameter in assessing adaptive strategies. In this paper, we refer to this measure as the "m/f ratio." Although this ratio is straightforwardly determinable in extant species and populations, this is not the case in fossils because the sex of a specimen is generally unknown.

In estimating the m/f ratio from fossil assemblages with no sex information, a commonly applied method is the "mean method" (here abbreviated the "MM") (7), which splits the sample into two subgroups, one above and the other below the sample mean. The ratio between the two subgroup means is considered the m/f ratio. In other words, the assumption is made that all specimens larger than the sample mean are males, and vice versa for females. This assumption would be reasonable only when the distributions of the two sexes overlap minimally, which is not necessarily the case in many features or taxa of interest. In estimating overall skeletal size dimorphism, a multivariate version of the MM has also been suggested (6). In this method, the m/f ratio is calculated as the geometric mean of the m/f ratios of multiple skeletal elements, thereby enabling dimorphism estimates from larger fossil sample sizes. However, caution is needed when partial skeletons are included in the analysis, because the dimorphism estimate can be weighted by relatively few individuals. This may result in a highly biased estimate, as in the case of the *Australopithecus afarensis* skeletal size dimorphism (6, 8).

Rather than accepting the assumption that the two sexes segregate at the mean, Lovejoy et al. (9) proposed a method that takes into account the unknown sex status of the specimens. Supposing a sample of N specimens, there are $N − 1$ possible splitting points between males and females and an equal number of possible m/f ratio values. Then, assuming equal chance of the two

Significance

The evolutionary study of sexual dimorphism has been hampered by the general inability to accurately estimate low levels of dimorphism in fossil assemblages with no information on sex membership. In previous estimates, accuracy decreased with lower dimorphism levels because the hidden distributions of the two sexes embedded in the sample were not adequately modeled. The posterior density peak (pdPeak) method that we developed in part solves this problem by estimating dimorphism and within-sex variance simultaneously and provides unbiased estimates of dimorphism at levels lower than previously possible. Thus, this method has potential to better address the evolutionary history of weakly dimorphic structures, such as the human canine, enabling a better understanding of dimorphisms and adaptive strategies among primate taxa.

ANTHROPOLOGY

sexes being represented in the fossil sample, the m/f ratio is calculated by weighing the $N - 1$ presumed male-to-female ratios with their binomial probabilities. This method was later termed the binomial dimorphism index (here, the "BDI") (10). When there is large size overlap between the sexes, this procedure is expected to correct for bias stemming from allocating sex via the sample mean. However, it shares with the MM the assumption that all males are larger than all females, and it has been shown (as with the MM) to overestimate weak skeletal size dimorphism, such as in chimpanzees (6, 10).

Another approach for estimating dimorphism is to use the empirical correlation observed between the m/f ratio and the coefficient of variation (CV; the SD divided by the mean). In traits that show a range of size dimorphism levels among taxa or populations, there is a strong tendency of a larger total CV to be associated with a larger m/f ratio. Using such a relationship, the m/f ratio of fossils can be estimated (11) via regression using known-sex extant samples (12–14) or simulated data (15, 16). However, as demonstrated in validation studies (15, 17), the CV method (hereafter, the "CVM") is susceptible to within-sex CV (hereafter, "wsxCV") levels and can overestimate, especially when sexual dimorphism is weak to moderate, as do the MM and BDI. This is because, when the two sex distributions substantially overlap, it is difficult to distinguish whether a large total CV stems from large sexual size dimorphism or large within-sex variance, the latter generally unknown in fossil assemblages.

Here, we introduce the "posterior density peak" (hereafter "pdPeak") method , a method of estimating the m/f ratio in fossil assemblages of unknown sex via a Bayesian mixture model. We model the background population from which the fossil sample is derived by three parameters, the male and female means and the common within-sex variance (all in log scale). Applying the Bayes theorem, we assign probabilities to the combinations of these three parameters that realize the sample distribution in hand. By this method, the distance between the two subgroups (dimorphism level) and within-subgroup variation can be evaluated simultaneously in terms of the fit of the model to a given sample distribution. In other words, "shape " of the sample distribution is considered when estimating size dimorphism. Fig. 1 illustrates how a population with the same total variation (combined-sex CV) can contain divergent latent distributions with a range of m/f ratios and within-sex variances, resulting in entirely different overall distributional shapes. It is clear that distribution shape is important in inferring distance (dimorphism) between subgroup means. We show that our method resolves the m/f ratio better than the other methods when male–female distributions overlap substantially.

We present the pdPeak method and evaluate its performance by 1) using computer-generated samples from simulated populations and 2) applying the method to a large dataset of extant anthropoid canine crown metrics of known sex. We compare the pdPeak method with the other most commonly used methods, the MM, BDI, and CVM. Lastly, we apply the pdPeak method to an actual sample of fossil canines to explore its potential in human evolutionary studies.

In evaluating the pdPeak method, we focus on canine crown size for simplicity. The canine crown diameter m/f ratios seen in the extant great apes and modern humans are summarized in *SI Appendix,* Table S1. In extant great apes, the m/f ratio ranges from ∼1.2 to ∼1.5. The most dimorphic is the gorilla, with maximum canine crown diameter m/f ratios of ∼1.4 to 1.5. The same ratios of chimpanzees and orangutans range from ∼1.2 to 1.4, with the least dimorphism seen in the bonobo lower canine (13, 18–22). To the contrary, in humans, the m/f ratio of canine crown diameters varies among populations, predominantly between 1.03 and ∼1.10 (23). Based on this, we focus on

discerning m/f ratios of the 1.1 to 1.2 interval, the range bridging the extant great ape and human conditions.

The pdPeak Method

The pdPeak method provides an estimate of the m/f ratio together with that of within-sex variation, here represented by the wsxCV. We outline the method below and provide further details in *SI Appendix, SI Text*. The metric data were log-transformed, and a mixture model of two homoscedastic (equal variance) normal distributions was assumed. Log-transformation reduces skewness of the size distributions and tends to equalize male and female variances. The model also assumes a population with the two sexes mixed in equal proportions, i.e., an equal probability for the two sexes to be fossilized and included in the analysis.

In this model, situations are specified by three parameters: μ_m for population male mean, μ_f for population female mean ($\mu_m \geq \mu_f$), and σ for population SD common to the sexes, all in log scale (population parameters are italicized). The parameters of interest, the *m/f ratio* and *wsxCV*, mathematically relate to μ_m, μ_f, and σ by the following equations:

$$m/f\,ratio = \exp\left[\mu_m - \mu_f\right], \qquad \textbf{[a]}$$

and

$$wsxCV = \sqrt{\exp[\sigma^2] - 1}. \qquad \textbf{[b]}$$

Assuming independent uniform prior probability distributions for the three parameters and independent prior probabilities of 0.5 for sexes of all specimens, the Bayes theorem yields the joint posterior probability distribution of:

$$P\left(\mu_m, \mu_f, \sigma, \mathbf{S}|\mathbf{d}\right) = \frac{1}{\sigma^N} \exp\left[-\frac{1}{2\sigma^2}\{\mathbf{S} \cdot (\mathbf{d} - \mu_f)^2 \right.$$
$$\left. + (1 - \mathbf{S}) \cdot (\mathbf{d} - \mu_m)^2\}\right]/C, \qquad \textbf{[f1]}$$

where $\mathbf{S}$ is a row vector of sex assignments for all specimens, in which 0 stands for male and 1 for female; $\mathbf{d}$ is a column vector of log-transformed measurements whose length is the sample size N; the denominator C is a normalizing constant calculated as the integral of the numerator with regard to all the variables, μ_m, μ_f, σ, and $\mathbf{S}$. We outline the details of the methods and derivation of **[f1]** in *SI Appendix, SI Text*. The theories and procedures of the mixture analyses used in this method can be found in McLachlan and Peel (24).

The joint probability function **[f1]** was further processed to obtain bivariate/marginal posterior distributions, either with the Markov chain Monte Carlo sampling procedure or by direct calculation through integrals of the function **[f1]**, depending on the sample size. With either procedure, first, a bivariate posterior probability density distribution of $\mu_m - \mu_f$ and σ is obtained, enabling assessment of the joint probability of the *m/f ratio* and *wsxCV*. Then, the marginal posterior density distributions of $\mu_m - \mu_f$ and σ were separately obtained by collapsing the bivariate density on either variable.

The pdPeak is determined by finding the point of highest density in the marginal distribution. Then, the value is back-transformed to original scale either from $\mu_m - \mu_f$ to m/f ratio by Equation **a** or from σ to wsxCV by **[b]**. Interval estimates, or credible intervals in terms of the Bayesian inference, were also obtained from the marginal distributions. The credible interval defined here is the highest posterior density interval (HDI), which, in the case of the 95% HDI, is the interval(s) enclosed by points whose probability density and above integrates to 0.95 (the shortest interval). Likewise, the 68% HDI integrates to 0.68 (corresponding to the ±1 SD range of a normal

www.manaraa.com

combined-sex CV (10%) common to all plots

A wsxCV = 8%   B wsxCV = 7%   C wsxCV = 6%   D wsxCV = 5%

E   F   G   H

true m/f ratio = 1.13   true m/f ratio = 1.15   true m/f ratio = 1.17   true m/f ratio = 1.19
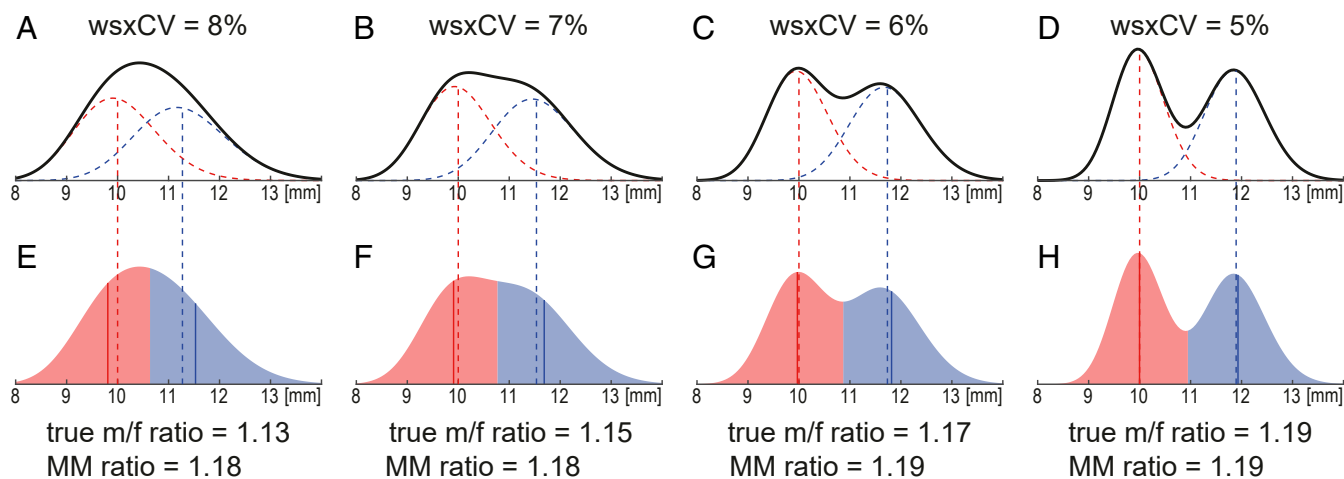MM ratio = 1.18   MM ratio = 1.18   MM ratio = 1.19   MM ratio = 1.19

Fig. 1. Variation in shape of population distributions with a constant CV of 10%. (*A–D*) Hypothetical population (combined-sex) distributions with the same overall variation (CV of 10%) are plotted for several subgroup conditions. Log-normal distributions of males and females with a common wsxCV are mixed in equal proportions. Overall CV is fixed at 10% and the female mean at 10 mm. The wsxCV is set in four ways: 8%, 7%, 6%, and 5% (from left to right). Solid black curves indicate overall distributions, dashed curves are the latent within-sex distributions, red is for female, and blue is for male. Vertical dashed lines indicate within-sex means. The same distributions are shown in *E–H* segregated to subpopulations by the mean. Vertical solid lines indicate the presumed sex means of the mean method (MM). The true male mean/female mean (m/f ratio), as well as the MM ratio (mean of presumed males divided by that of the presumed females), is shown below each wsxCV condition. Note that the MM increasingly overestimates as the male and female distributions increasingly overlap (from right to left). The Rd values (see *Application to Actual Cases*), i.e., the distance between means in within-sex SD units, are 1.6, 2.2, 2.9, and 3.8, respectively in *A*, *B*, *C*, and *D*. Note also that, under the same overall CV, the true m/f ratio can vary substantially depending on wsxCV.

distribution). Then, the interval boundaries were back-transformed to original scale by either Equation **a** or **b**.

Computations were performed by using MATLAB (R2019b; MathWorks). The script codes are available at https://github.com/sxdm/pdPeak. This is a package of MATLAB scripts that computes the pdPeak estimates and related statistics given a user's metric data input. Instructions for using the program are included in the package (also in *SI Appendix*, *SI Text*, *Instructions for using the MATLAB program "pdPeak"*).

"Concordant" Simulations

Performance of the pdPeak method was assessed by using computationally generated (simulated) samples under a variety of conditions. The simulated samples were generated based on a model of two homoscedastic normal distributions in log scale, i.e., two log-normal distributions with the same CV in original scale, mixed in the same proportion. The distribution with the larger population mean was assumed to be male. The female population mean was arbitrarily set at a constant value, 10 mm, and the other conditions—i.e., sample size, wsxCV, and the male population mean—were set as follows. Note that these predetermined values (except for sample size) refer to that of the source populations from which the simulated samples were generated and are not to be confused with the actually generated sample distributions and statistics. Sample size was set in three ways: $N = 10$, 15, and 30; male population mean was set so that the m/f ratio ranged from 1.0 to 1.3 in 0.03 increments; and wsxCV was set either to 5% or 8%. Extant anthropoids have canine wsxCVs mostly between 5% and 8% with a range of <3% to >10% (Fig. 2 and Dataset S1). To summarize, sample size was set in 3 ways, m/f ratio in 11 ways, and wsxCV in 2 ways (a total of 66 conditions). Under each condition, 2,000 sets of simulated data were generated. In some of the simulated datasets, instances of uniform sex were possible because of small *N*, but they were not discarded. The pdPeak, MM, BDI, and CVM estimates of dimorphism were calculated in each

dataset. For the CVM, the "Plavcan (1994) equation" (15) was used.

The results of the concordant simulations are shown in Fig. 3 and Dataset S2. All methods tended to overestimate the m/f ratio when it was low (<1.1). In all methods, overestimation
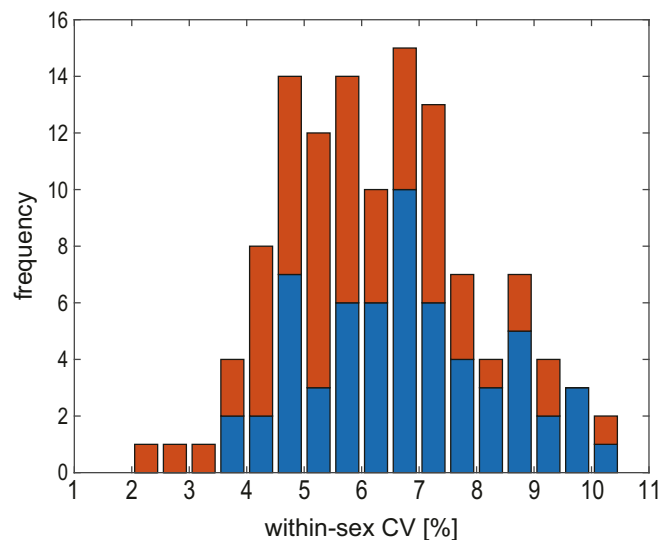

Fig. 2. Distribution of wsxCVs of canine crown diameters in extant anthropoids. Blue histograms are for male and red for female. The upper canines ($N = 29$) and lower canines ($N = 31$) of both sexes are plotted together ($N = 120$). The CVs are those of species/subspecies samples with 15 or more specimens available for each sex. The correction suggested by Sokal and Braumann (28) was used in calculating the CVs. Data are from Plavcan (29) and Suwa et al. (21). See *SI Appendix*, *SI Text* and Dataset S1 for further details and source information. Out of 120 samples, a total of 29 samples have a wsxCV lower than 5%, and 20 samples have a wsxCV higher than 8%.
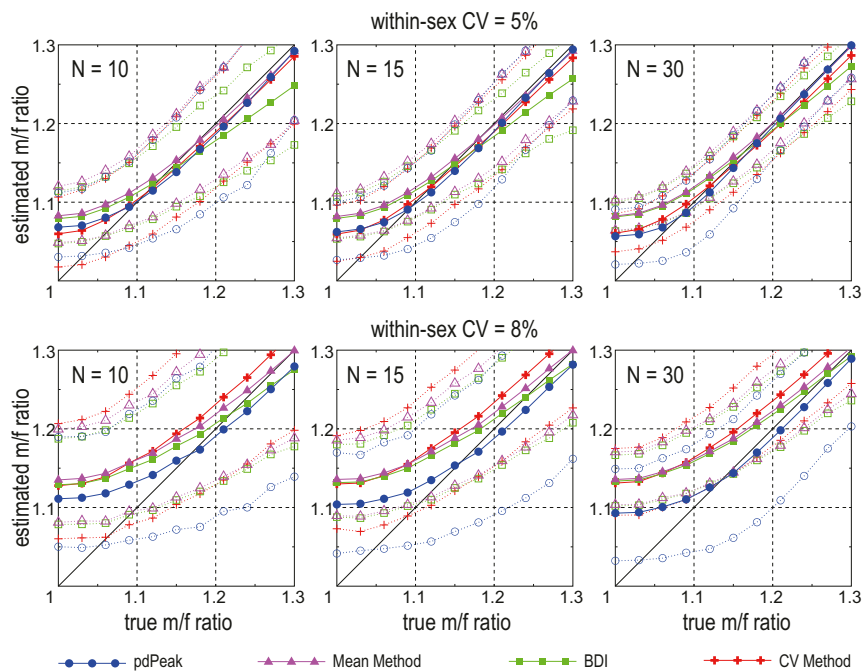
Fig. 3. Simulation results under conditions concordant to assumptions. The pdPeak method is evaluated and compared with the MM, BDI, and the CVM. The CVM is that of Plavcan (15). Three conditions were set for sample size *N*: 10 (left plots), 15 (middle plots), and 30 (right plots). Two conditions were set for wsxCV: 5% (*Upper*) and 8% (*Lower*). This resulted in six combinations. Under each combination, the m/f ratio was set in 0.3 increments indicated by the markers, and 2,000 samples were generated for each m/f ratio in each condition. True m/f ratio is on the *x* axis, and estimated values are on the *y* axis. Solid lines indicate the mean, and dotted lines are the 5th- and 95th-percentile values of the 2,000 samples. The pdPeak was the only method capable of estimating m/f ratios between 1.1 and 1.2 with little bias when wsxCV was 8%. Note that, with regard to the pdPeak, MM, and BDI estimates, the wsxCV 8% results correspond to enlarged portions of the wsxCV 5% plots. This is because these methods are directly affected by the distance between the two means relative to within-sex variance. This can be expressed by alternative expressions of the same measure, the relative distance (Rd) (difference of means divided by female SD) or the Rm/f ratio ("m/f ratio minus 1" divided by wsxCV) (see *Applications to Actual Cases*). For example, with an m/f ratio of 1.30 and wsxCV of 8%, the Rm/f ratio is 3.75, which corresponds to an m/f ratio of ~1.19 and a wsxCV of 5%. Thus, the wsxCV 8% plots largely correspond to the lower ~60% subset (m/f ratio of 1.0 to 1.19) of the wsxCV 5% plots. This is not the case with the CVM, which depends on a regression relationship between the m/f ratio and combined-sex CV.

bias and percentile width (width between the 5th and 95th percentiles) were both greater with a larger wsxCV (compare the wsxCV 8% vs. 5% results). This reflects the difficulty of estimating the m/f ratio when overlap of male and female distributions increases with larger wsxCV (15, 17).

The degree of overestimation (and underestimation) bias differs by method (Fig. 3). Accuracy (as opposed to precision, which refers to uncertainty level) is expressed by how close the means follow the $y = x$ diagonal. In the 5% wsxCV simulations, the pdPeak and CVM showed better accuracy than the other methods in the lower m/f ratio range. In the higher m/f ratio range, the differences among methods were not substantial, except that BDI tended to underestimate in small samples. In the 8% wsxCV simulations, pdPeak was by far the most accurate in the 1.1 to 1.2 m/f ratio interval, although precision was lower compared with the other methods. The CVM was the least accurate among the methods, tending to substantially overestimate. With the pdPeak, accuracy increased with sample size. This was not the case with the MM, CVM, and BDI; overestimation bias did not noticeably change with sample size in either the 5% or the 8% wsxCV simulations.

"Disconcordant" Simulations

Actual samples may deviate from the pdPeak method assumptions. Guided by the empirical distributions seen in the extant anthropoid datasets (Dataset S1), we investigated the effects of deviations from assumptions regarding eight conditions. We exaggerated the conditions so that the possible effects can be easily seen (see *SI Appendix, SI Text* for how the simulation data were generated).

1) Heteroscedastic ($M\sigma > F\sigma$): Male SD (in log scale) is twice that of female.
2) Heteroscedastic ($M\sigma < F\sigma$): Male SD (in log scale) is half that of female.
3) Unidirectional skew (positive): Both sex distributions are positively skewed.
4) Unidirectional skew (negative): Both sex distributions are negatively skewed.
5) Tail-to-tail skew: Male and female distributions are negatively and positively skewed, respectively.
6) Head-to-head skew: Male and female distributions are positively and negatively skewed, respectively.
7) Unbalanced sex (M > F): Male population to female population ratio set at 7:3.
8) Unbalanced sex (M < F): Male population to female population ratio set at 3:7.

As in the concordant condition simulations, the m/f ratio was set from 1.0 to 1.3 (with 0.03 increments), and wsxCV was set at 5% or 8%. The sample size was set at $N = 15$. Note that conditions 7 and 8 concern the source population and do not apply to the actually generated samples. Presetting the source population to a skewed sex ratio simulates bias stemming from demography and/or taphonomy that differentially affects males and females of the source population and would bias the sex ratios of the actual samples.

The results of the disconcordant simulations are summarized in Fig. 4 and *SI Appendix, SI Text*. As a whole, the effects of deviations from assumptions were found to be limited and largely of inconsequence. However, heteroscedasticity, mildly noted in extant anthropoid upper canines (*SI Appendix, SI*
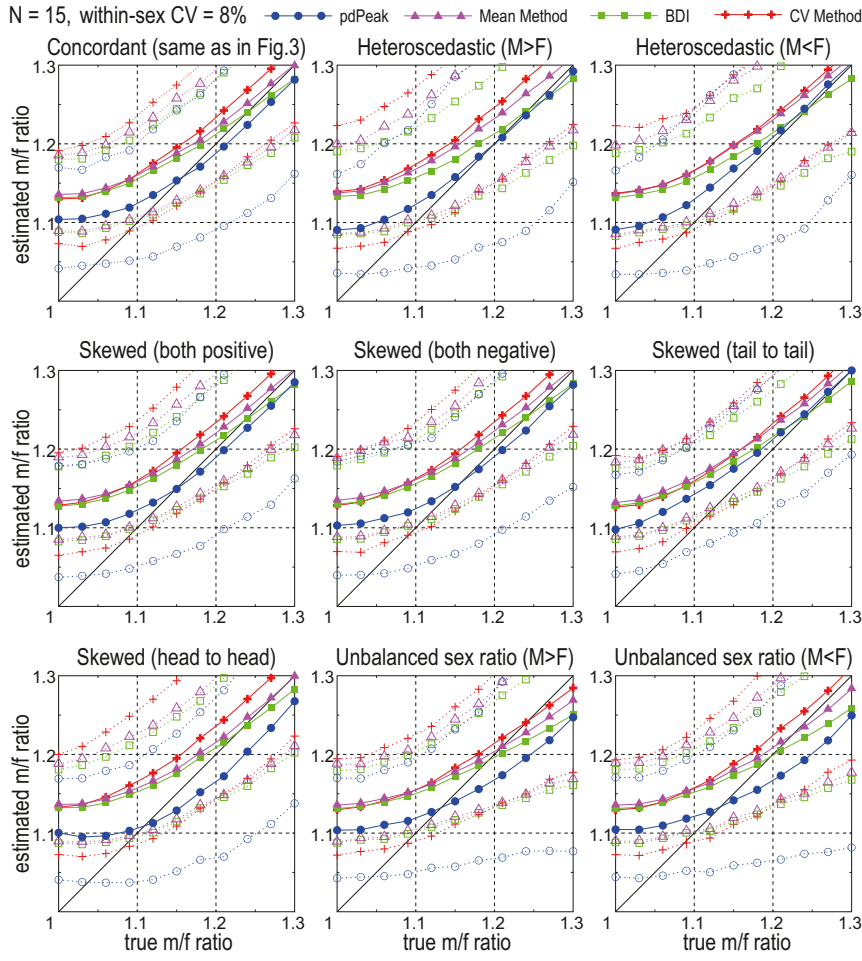
www.manaraa.com

Fig. 4. Simulation results under conditions disconcordant to assumptions. Simulation tests were conducted on background populations that deviate from assumptions in eight ways: two heteroscedastic conditions (Mσ > Fσ and Mσ < Fσ); four skewed distributions (both positive; both negative; male negative and female positive [tail to tail]; and male positive and female negative [head to head]); and two unbalanced sex ratios (M:F = 7:3 and 3:7). N = 15, wsxCV = 8% (average wsxCV used when heteroscedastic). *Top Left* reproduces the concordant condition to facilitate comparisons. See Fig. 3 for explanation of axes, lines, and symbols. The effects of deviation from assumptions are generally not conspicuous, except in the head-to-head and unbalanced sex ratio conditions.

Text), would result in a somewhat lower precision, and reciprocal skewness and an unbalanced sex ratio were seen to produce noticeable effects to the pdPeak estimates. In the tail-to-tail condition, the pdPeak tended to overestimate when wsxCV was 8% (Fig. 4) (but less than in the other methods), although this was minimal when wsxCV was 5% (*SI Appendix,* Fig. S1). Conversely, in the head-to-head condition, the pdPeak tended to underestimate and with a slightly lower precision than in the concordant cases. In the unbalanced sex ratio simulations (both M > F and M < F), the pdPeak estimates tended to underestimate the m/f ratio in the higher m/f ratio range, accompanied by a substantial reduction in precision. The estimated means of the other methods also tended to be lower in the unbalanced sex ratio simulations and with a reduction of precision, but not as substantially as seen in the pdPeak.

Application to Actual Cases
We applied the pdPeak and the other methods to 110 extant anthropoid taxonomic samples of canines with maximum crown diameter m/f ratios of less than 1.3 (*SI Appendix, SI Text* and Dataset S3). This dataset includes 45 species/subspecies samples of upper canines and 65 of lower canines. In this analysis, samples with a total size of five or more (regardless of sex) were used. Using these extant anthropoid samples, the m/f ratio

estimates of the pdPeak method and those derived from the MM, BDI, and CVM were compared to the actual sample m/f ratios. Because these are actual samples that variably deviate from assumptions (see above and *SI Appendix, SI Text*), they function as empirical checks on the simulation results.

Results are shown in the upper row of Fig. 5 and numerically summarized in Dataset S3. In samples of $10 \leq N < 30$, and especially $30 \leq N$, the pdPeak estimates clustered around the $y = x$ diagonal, with a tendency to overestimate only when the sample m/f ratio was lower than ~1.1. Note that the pdPeak 95% credible intervals encompass all the sample m/f ratios and broadly match the sample CIs (*SI Appendix,* Dataset S3), regardless of sample sizes, supporting the reliability of the pdPeak credible intervals. Deviations from the $y = x$ diagonal are explained by the combination of both sample and estimated m/f ratio uncertainties. The latter is larger in cases of larger wsxCV, which appears to be a dominating factor of the dispersion around the diagonal. When the sample m/f ratio was >1.1 and $N > 10$, there were only six cases in which deviation from the diagonal was greater than 0.04 (Dataset S3). Four of these can be explained by relatively large uncertainties stemming from large wsxCVs (>7% to close to 9%). The remaining two cases can be explained by a combination of a moderately high wsxCV (~6.5%) and either heteroscedasticity or a head-to-

Sasaki et al.
Estimating sexual size dimorphism in fossil species from posterior probability densities

PNAS | 5 of 9
https://doi.org/10.1073/pnas.2113943118
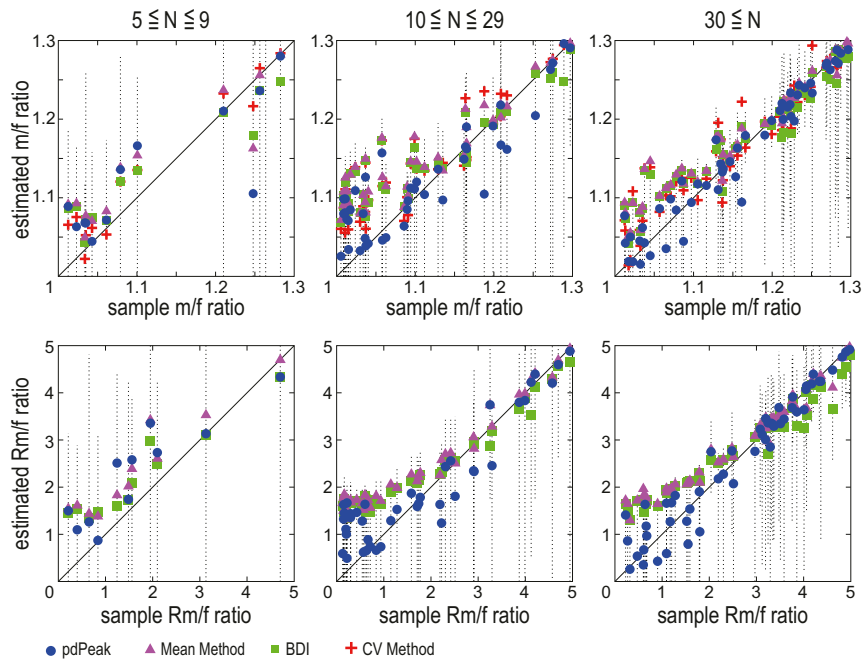
www.manaraa.com

Fig. 5. Applications to extant anthropoid canine samples of known sex. The panels summarize results of the pdPeak and other methods applied to 110 extant anthropoid samples. The samples varied in size (*N*) from 5 to 89 and were separated into three categories. See *SI Appendix*, *SI Text* and Dataset S3 for sample and data information. *Upper* row shows the m/f ratio estimates (*y* axis) against the actual sample values (*x* axis). Note that the pdPeak estimates cluster around the *y = x* diagonal, especially when the sample size is large, whereas the other methods tend to overestimate when the m/f ratio is <1.2. The few cases of relatively large underestimation of the pdPeak estimate involve samples with large sample wsxCVs and, hence, with relatively larger uncertainties (see *Applications to Actual Cases*). *Lower* row shows the estimated and actual sample values of the quantity "m/f ratio minus 1" divided by sample wsxCV (male and female average), termed the Rm/f ratio (see *Applications to Actual Cases*). The *y* axis deviation from the *y = x* diagonal expresses difference between estimated and actual m/f ratios standardized by wsxCV. *Lower* row shows a pattern that conforms to what is expected from the simulation tests (Fig. 3); the pdPeak estimates are unbiased when the Rm/f ratio is >∼1.5, while MM and BDI overestimate when the Rm/f ratio is <∼2.5. The CVM is not shown in the *Lower* row because there are no general relationships between the CVM and the Rm/f ratio. The interrupted vertical lines indicate the 95% credible intervals of the pdPeak estimates. Note that the 95% credible intervals encompass the sample values in all cases. Note also that the pdPeak credible intervals broadly match the sample m/f ratio CIs of standard statistics (Dataset S3), but with a tendency to be wider (more conservative). From this, it is seen that the pdPeak credible intervals are reliable, even with small sample sizes.

head situation, the latter potentially contributing to the observed underestimation. The MM and BDI appear to overestimate when the sample m/f ratio is ∼1.15 or lower. The CVM of the present study also tends to overestimate when the sample m/f ratio is lower than ∼1.15.

Because performance of the methods is affected by the distance between the two sex means relative to within-sex variance, we standardized the m/f ratio by the wsxCV. This was done by defining:

$$Rm/f \ ratio = (m/f \ ratio - 1)/wsxCV,$$

which is mathematically equivalent to:

$$Rd = (M - F)/sd_f,$$

where *M* and *F* stand for male and female population means and sd_f for female SD. Whereas *Rd* is a familiar measure of relative distance (the separation between male and female means relative to within-sex variance), the *Rm/f ratio* is useful in evaluating the simulations and actual dataset tests of the present study. The sample Rm/f ratio was calculated by substituting the denominator with the mean of male and female sample wsxCVs.

The results are shown in the lower row of Fig. 5 and numerically summarized in Dataset S3. The pdPeak method is seen to yield unbiased m/f ratios to a condition of the Rm/f ratio as low as ∼1.5. This corresponds to the simulation results that showed unbiased pdPeak m/f ratios of ∼1.075 when wsxCV was 5% (Rm/f ratio of ∼1.5) when sample size was *N* = 30. To the

contrary, MM and BDI methods start to overestimate at an Rm/f ratio of ∼2.5 and consistently so when the Rm/f ratio is <2. These latter observations also conform to the simulation results.

Finally, we investigated the relation between MM and BDI by examining differences between their m/f ratio estimates. Our results confirmed that BDI tends to correct for overestimation in 67/74 cases when the m/f ratio is <1.2 by a modest amount of 0.005 on average (Dataset S3).

Application to a Fossil Assemblage

The pdPeak and the other methods were applied to *Ardipithecus ramidus* from the African Pliocene. Dental remains of that species have been reported from the Middle Awash (21, 25, 26) and the Gona (27) study areas, Ethiopia. The crown diameters of a total of 11 lower canines were available for analysis (*SI Appendix*, Table S2). The Gona (*N* = 5) and the Middle Awash (*N* = 6) samples do not differ significantly in lower canine maximum diameter (*t* test, *P* = 0.403; basic statistics shown in *SI Appendix*, Table S3). Further details of the *Ar. ramidus* sample and a fuller investigation of canine sexual dimorphism in *Ardipithecus*, *Australopithecus*, and *Homo* will be presented elsewhere. The results of the pdPeak analysis are shown in Fig. 6. The pdPeak m/f ratio estimates are 1.13 in both the Gona and Middle Awash subsamples and in the combined sample. The estimated wsxCV is ∼6 to 7%, at midrange of the observed extant anthropoid values. Although the pdPeak m/f ratio estimates of the Gona and Middle Awash subsamples are similar
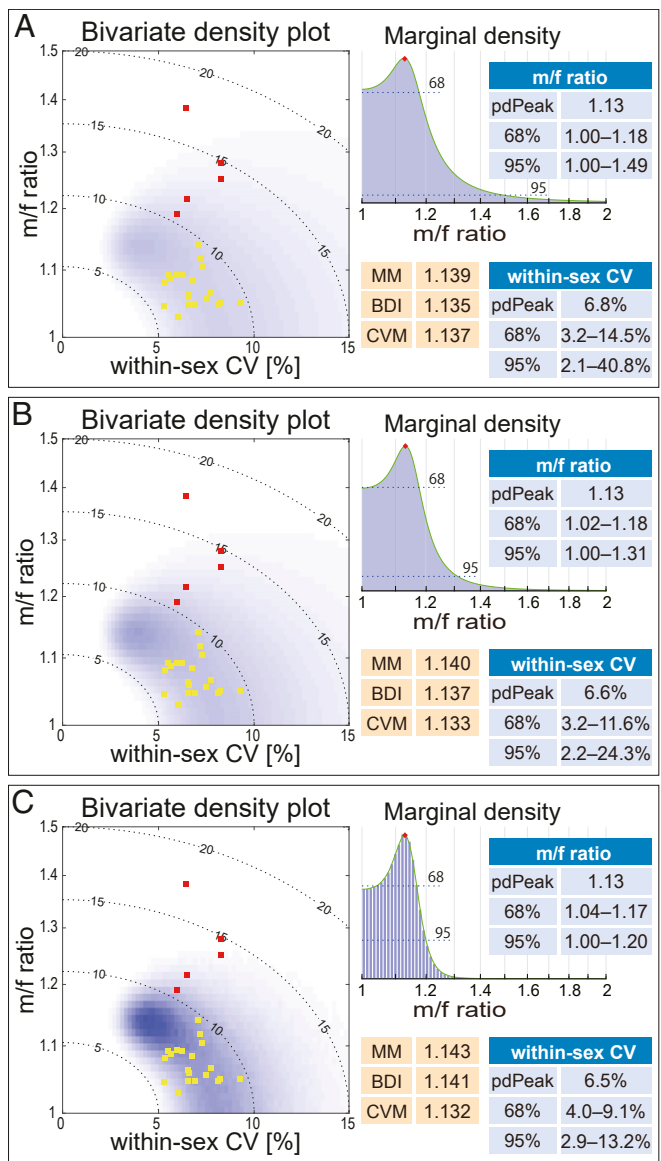
www.manaraa.com

Fig. 6. The pdPeak method applied to the lower canine crown diameter of *Ar. ramidus*. Three runs were performed on the maximum canine diameter of different sample compositions. (*A*) Gona specimens (*N* = 5). (*B*) Middle Awash specimens (*N* = 6). (*C*) Both combined (*N* = 11). In each panel, *Left* shows the bivariate probability distribution of wsxCV (*x* axis) and the m/f ratio (*y* axis). Probability densities were obtained with the log-transformed data, and the axis labels were back-transformed to original scale (*SI Appendix, SI Text*). Dotted contour lines are the combined-sex CV levels. The red and yellow squares show distribution of the extant great ape and modern human reference data (*SI Appendix*, Table S1; the weighted wsxCVs of the males and females were used). In *Right* is the marginal posterior density plot of the logarithm of the m/f ratio, with a red diamond marker at its peak and dotted horizontal lines at the 68% and 95% credible interval levels.

to the combined sample, their uncertainty levels are high (large credible intervals). When the two subsamples are combined, the uncertainty level is lower and enables probability assessments. The *Ar. ramidus* m/f ratio is estimated to between 1.04 and 1.17 at *P* = 0.68, and <1.20 at *P* = 0.95. The probability of having an m/f ratio as high as seen in the least dimorphic extant great ape example (the lower canine of the bonobo with an m/f ratio of 1.19) is *P* = 0.068. The MM, BDI, and CVM methods yielded similar m/f ratio estimates, corroborating the pdPeak estimate.

Discussion

In this paper, we introduced the pdPeak method, a method that models dimorphism in relation to latent subgroup distributions, as opposed to methods that only evaluate gross variation (such as the CVM) or methods that partition the observed variation without considering overlap between sexes (the MM and BDI). However, the validity of model assumptions needs confirmation, and the effects of deviations need to be understood.

With the extant anthropoid canine data, we confirmed that the assumptions of homoscedasticity and normality in log scale were largely not contradicted (*SI Appendix, SI Text* and Dataset S1). Indeed, we confirmed that log-transformation is generally effective in reducing skewness and difference in variance between the sexes. Although a weak tendency for male canines to be more variable than those of females remained after log-transformation in the upper canine crown diameter; results of the disconcordant simulations suggest that its effect would be little, although uncertainty levels may increase.

In the case of the canine crown diameter, deviation from normality appears to rarely occur, judging from the low frequencies of samples that tested positive in the extant anthropoid dataset (*SI Appendix, SI Text* and Dataset S1). In such cases, the disconcordant simulations caution that head-to-head skewness and unbalanced sex ratio have potential to cause notable effects. Under either condition, the pdPeak method would tend to underestimate the m/f ratio. Thus, when applying the pdPeak method, the possibility of such deviations from assumptions needs to be considered. However, it should also be noted that our simulations were exaggerated. Indeed, none of the extant anthropoid canine samples that we examined exhibited (at the *P* < 0.05 level of significance) the strong head-to-head skew that we used in the simulations. Also, the sex ratio of 7:3 (or 3:7) that we used in the simulations is best considered an unrealistic scenario (17), and we found its effect (underestimation of the m/f ratio) to be modest (∼0.03) in the 1.1 to 1.2 m/f ratio interval. It is difficult or perhaps impossible to detect whether latent within-sex distributions are skewed with the available small sample sizes of the fossils of interest. Whole-sample skewness may be used as an indication of an unbalanced sex ratio, although distinguishing this from heteroscedasticity and/or within-sex skewness is generally not possible. Thus, any a priori information would also deserve attention. For example, expected behaviors and taphonomic situations may, at times, validate an a priori unbalanced population sex ratio.

Resolving the 1.1 to 1.2 m/f ratio range of dimorphism is important in understanding the evolution of human canine sexual dimorphism. Our simulations showed that the pdPeak method is capable of unbiased estimates of the m/f ratio as low as ∼1.1 with a sample size of ≥10 if wsxCV is 5% or with a sample size of ≥30 if wsxCV is 8%. The other methods also yield comparable m/f ratio estimates when wsxCV is 5%, but this is not the case if wsxCV is 8%, regardless of sample size (Fig. 3). With canine crown diameters, a wsxCV of 8% is not exceptionally large among extant anthropoids. In fossil assemblages, even larger wsxCVs are possible because of the temporal depth factor that may effectively commingle subgroups with different central tendencies. With fossils, preservation may also necessitate inclusion of estimated values, thereby adding error variance and elevating the wsxCV.

As is the case with the other methods, with realistic sample sizes (*N* ≤ 30), the pdPeak method cannot accurately estimate the weakest levels of dimorphisms (overestimation bias in low m/f ratios). However, bias is less in the pdPeak than in the other methods, especially when wsxCV is large. Both the simulation results and the analysis of extant anthropoid canines demonstrated that, with a sample size of ∼30, the pdPeak m/f

Sasaki et al.
Estimating sexual size dimorphism in fossil species from posterior probability densities

PNAS | 7 of 9
https://doi.org/10.1073/pnas.2113943118

www.manaraa.com

ratio estimates are unbiased, even when the Rd and the Rm/f ratio are as low as 1.5. This represents a condition of substantial overlap between the male and female distributions, wherein the male and female means are separated by only 1.5 female SD units (slightly more overlap than in Fig. 1A).

Another important characteristic of the pdPeak method is that both accuracy and precision of the m/f ratio estimate increase with sample size, which is not the case with the other methods. This can be understood by the fact that the shape of the population distribution becomes more certain as sample size increases. Additionally, whereas the BDI tends to underestimate when sexual dimorphism is large, this is generally not the case with the pdPeak. Thus, the pdPeak method accurately estimates dimorphism across a broader range of the m/f ratio than the other methods.

Compared with the pdPeak method, our simulations show that the MM begins to overestimate at an m/f ratio of ~1.2 when wsxCV is 5% and ~1.3 when wsxCV is 8%, which corresponds to a distance between the two means of about 4 units of within-sex variation, i.e., an Rd and Rm/f ratio of 4 (Fig. 3). Populations with an Rd of 4 are clearly bimodal (Fig. 1D). Our simulations also show that overestimation bias of the MM is not substantial if Rd (and the Rm/f ratio) is at least 2 to 2.5, which corresponds to when bimodality starts to be expressed (Fig. 1 B and C).

A limitation of the pdPeak method is its relatively high uncertainty levels. To evaluate the performance of pdPeak combining both accuracy and uncertainty, we calculated the SE (17), defined as $\sqrt{\sum(\text{estimate} - \text{true})^2/n}$, where n is the number of simulations (Dataset S2). In the concordant simulations, in the lower m/f ratio range, the SE of the pdPeak method is smaller than in the other methods due to higher accuracy. However, when the m/f ratio is large, the pdPeak SE is larger than that of the other methods due to greater uncertainty. Because the SE of the pdPeak m/f ratio estimates is slightly larger than that of the MM when Rd is >4 (Dataset S2), both MM and pdPeak estimates should be considered if clear bimodality is observed.

It was observed that the BDI estimates are consistently lower than those of the MM (Fig. 3 and Dataset S2), particularly with smaller sample sizes and higher wsxCV. With the extant anthropoid canine samples that we examined (m/f ratios < 1.20), the difference between MM and BDI estimates averaged 0.005, for the most part compensating for overestimation bias. However, in the higher m/f ratio range, BDI tends to underestimate, whereas MM is less biased, consistent with previous evaluations (17). This is because the BDI is calculated by considering all possible male–female splitting points. Underestimation occurs in the BDI when Rd is greater than ~3 (depending on sample size), in parallel to diminishing overestimation in MM.

The CVM examined in this study is that of Plavcan (15), which was used in investigating canine dimorphism in *Australopithecus* (16). This formula was derived from a regression analysis of computationally generated samples under a condition of wsxCV of 5.5%. As originally pointed out (15), the accuracy of this method is only as good as the appropriateness of the reference sample for the sample being investigated. Our simulations show this in exhibiting high accuracy when wsxCV is 5%, but a considerably lower accuracy throughout the m/f ratio range when wsxCV is 8% (Fig. 3). The susceptibility of the CVM to wsxCV levels has also been documented by others (15, 17).

Finally, application to an actual fossil case illuminates the usefulness of the pdPeak method, even with small sample sizes. The credible intervals became narrower as the sample size increased. This enabled probabilistic hypothesis testing, such that the *Ar. ramidus* lower canine was demonstrated unlikely to

be as dimorphic as in extant great apes. Although the MM, BDI, and CVM methods yielded similar m/f ratio estimates, these methods do not provide reliable interval estimates for the population m/f ratio. While bootstrap CIs can be calculated by resampling, this would not account for bias inherent in the estimation method. For example, a bootstrap interval for the MM estimate will never encompass an m/f ratio of 1.0, even if the true value is 1.0. This is in contradistinction to the credible interval of the pdPeak method, which we found to reliably include the true population value within its intervals in the extant anthropoid canine samples.

Summary and Conclusions

Accurately measuring sexual dimorphism levels is important because of their significance in understanding how males and females differentially respond or relate to sociobehavioral variables and ecological signals as a part of the species' overall adaptive strategy.

We developed a method of estimating sexual dimorphism levels in fossils based on Bayesian posterior probability densities. We named this method the pdPeak method, which simultaneously estimates the population m/f ratio (ratio of male to female means) and within-sex variance in samples of unknown sex membership. Our aim was to develop a method that enables better m/f ratio estimates in the low to moderate levels of dimorphisms—for example, below an m/f ratio of 1.2—crucial in elucidating primate and human canine evolution. Validation analyses demonstrated that the strengths of this method compared to the other methods are:

1) The pdPeak method is capable of unbiased estimates of a broader range of dimorphisms. It provides estimates with little or no biases in the high m/f ratio ranges and continues to do so to an m/f ratio of ~1.1 or lower, whereas the MM and BDI provide unbiased or little-biased estimates only to an m/f ratio level of ~1.15 to 1.2. Compared with the CVM, the pdPeak method maintains accuracy across a range of within-sex variance, which is not the case with the CVM (accuracy is tied to confined levels of sample wsxCV).
2) In the pdPeak method, both accuracy and precision increases with sample size, which is not the case with the other methods. The limits of estimating the m/f ratio by the pdPeak method are as summarized above with the currently realistic sample sizes ($N \leq$ ~30), but an accurate estimate of even lower dimorphism levels is potentially possible with larger sample sizes.
3) Statistical assessments such as hypothesis testing are possible by utilizing the posterior probability distributions, i.e., the credible intervals of Bayesian inference.

On the other hand, some limitations to the pdPeak method are as follows: First, precision of the pdPeak m/f ratio estimates tends to be lower than that of the MM, BDI, and CVM (especially if heteroscedastic), resulting in a slightly higher SE (mean error) in the higher range of the m/f ratio. Next, the pdPeak method tends to underestimate the m/f ratio if the sample involves a "head-to-head" skew (male and female distributions that are positively and negatively skewed, respectively) and/or if the sex ratio is substantially unbalanced. The possibility of these complications needs to be considered, although we found that such cases are rare in anthropoid canines. We recommend that the sample be examined with respect to skewness in the data (which can indicate either heteroscedasticity, within-sex skew, or unbalanced sex representation) and for any relevant expectations from contextual information. We also recommend that the results of the other methods are considered together with the pdPeak estimates.

www.manaraa.com

1. J. M. Plavcan, Understanding dimorphism as a function of changes in male and female traits. *Evol. Anthropol.* **20**, 143–155 (2011).
2. J. M. Plavcan, Implications of male and female contributions to sexual size dimorphism for inferring behavior in the hominin fossil record. *Int. J. Primatol.* **33**, 1364–1381 (2012).
3. J. M. Plavcan, Sexual size dimorphism, canine dimorphism, and male-male competition in primates. *Hum. Nat.* **23**, 45–67 (2012).
4. T. H. Clutton-Brock, P. H. Harvey, B. Rudder, Sexual dimorphism, socionomic sex ratio and body weight in primates. *Nature* **269**, 797–800 (1977).
5. R. J. Smith, Statistics of sexual size dimorphism. *J. Hum. Evol.* **36**, 423–459 (1999).
6. A. D. Gordon, D. J. Green, B. G. Richmond, Strong postcranial size dimorphism in *Australopithecus afarensis*: Results from two new resampling methods for multivariate data sets with missing data. *Am. J. Phys. Anthropol.* **135**, 311–328 (2008).
7. L. R. Godfrey, S. K. Lyon, M. R. Sutherland, Sexual dimorphism in large-bodied primates: The case of the subfossil lemurs. *Am. J. Phys. Anthropol.* **90**, 315–334 (1993).
8. P. L. Reno, C. O. Lovejoy, From Lucy to Kadanuumuu: Balanced analyses of *Australopithecus afarensis* assemblages confirm only moderate skeletal dimorphism. *PeerJ* **3**, e925 (2015).
9. C. O. Lovejoy, K. F. Kern, S. W. Simpson, R. S. Meindl, "A new method for estimation of skeletal dimorphism in fossil samples with an application to *Australopithecus afarensis*" in *Hominidae*, G. Giacobini, Ed. (Jaka Book, Milano, Italy, 1989), pp. 103–108.
10. P. L. Reno, R. S. Meindl, M. A. McCollum, C. O. Lovejoy, Sexual dimorphism in *Australopithecus afarensis* was similar to that of modern humans. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9404–9409 (2003).
11. J. G. Fleagle, R. F. Kay, E. L. Simons, Sexual dimorphism in early anthropoids. *Nature* **287**, 328–330 (1980).
12. R. F. Kay, Sexual dimorphism in Ramapithecinae. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 209–212 (1982).
13. W. Leutenegger, B. Shell, Variability and sexual dimorphism in canine size of *Australopithecus* and extant hominoids. *J. Hum. Evol.* **16**, 359–367 (1987).
14. C. A. Lockwood, B. G. Richmond, W. L. Jungers, W. H. Kimbel, Randomization procedures and sexual dimorphism in *Australopithecus afarensis*. *J. Hum. Evol.* **31**, 537–548 (1996).
15. J. M. Plavcan, Comparison of four simple methods for estimating sexual dimorphism in fossils. *Am. J. Phys. Anthropol.* **94**, 465–476 (1994).
16. J. M. Plavcan, C. P. van Schaik, Interpreting hominid behavior on the basis of sexual dimorphism. *J. Hum. Evol.* **32**, 345–374 (1997).
17. K. Kościński, S. Pietraszewski, Methods to estimate sexual dimorphism from unsexed samples: A test with computer-generated samples. *Anthropol. Rev.* **67**, 33–55 (2004).
18. A. J. Almquist, Sexual differences in the anterior dentition in African primates. *Am. J. Phys. Anthropol.* **40**, 359–367 (1974).
19. R. F. Kay, *Sivapithecus simonsi*, a new species of Miocene hominoid, with comments on the phylogenetic status of Ramapithecinae. *Int. J. Primatol.* **3**, 113–173 (1982).
20. J. M. Plavcan, C. P. van Schaik, Intrasexual competition and canine dimorphism in anthropoid primates. *Am. J. Phys. Anthropol.* **87**, 461–477 (1992).
21. G. Suwa *et al.*, Paleobiological implications of the *Ardipithecus ramidus* dentition. *Science* **326**, 94–99 (2009).
22. F. K. Manthi, J. M. Plavcan, C. V. Ward, New hominin fossils from Kanapoi, Kenya, and the mosaic evolution of canine teeth in early hominins. *S. Afr. J. Sci.* **108**, 1–9 (2012).
23. J. A. Kieser, *Human Adult Odontometrics* (Cambridge University Press, Cambridge, UK, 1990).
24. G. J. McLachlan, D. Peel, *Finite Mixture Models* (John Wiley & Sons, New York, NY, 2000).
25. T. D. White, G. Suwa, B. Asfaw, *Australopithecus ramidus*, a new species of early hominid from Aramis, Ethiopia. *Nature* **371**, 306–312 (1994).
26. T. D. White, C. O. Lovejoy, B. Asfaw, J. P. Carlson, G. Suwa, Neither chimpanzee nor human, *Ardipithecus* reveals the surprising ancestry of both. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 4877–4884 (2015).
27. S. Semaw *et al.*, Early Pliocene hominids from Gona, Ethiopia. *Nature* **433**, 301–305 (2005).
28. R. R. Sokal, C. A. Braumann, Significance tests for coefficients of variation and variability profiles. *Syst. Zool.* **29**, 50–66 (1980).
29. J. M. Plavcan, "Sexual dimorphism in the dentition of extant anthropoid primates," PhD dissertation, Duke University, Durham, NC (1990).

ANTHROPOLOGY

www.manaraa.com