Supporting information for: "The Mixed Cumulative Probit: A multivariate generalization of transition analysis that accommodates variation in the shape, spread, and structure of data" by Kyra Stull et al.

Contents

1 Algorithm Description

This section of the Supporting Information describes in detail the mixed cumulative probit model implemented in the R package 'yada'. By mixed we mean that response variables can be either continuous or ordinal. The response variables depend on a single, scalar independent variable, x. We begin by considering univariate responses for continuous and ordinal variables, then extend these to the general case. Next, we describe the Bayesian inference used to estimate the posterior density of x. This requires the formulation of a model for the prior probability of the variable x. Finally, we describe the cross-validation methodology we used for model selection (and present complete, associated results).

1.1 Motivating Application and Overview

Although this model and associated R code should have wide application, we had a specific application in mind when we developed it: subadult age estimation. The independent variable is x, age, and the dependent variables can be a mix of continuous skeletal traits (for example, femur length) and ordinal skeletal traits (for example, epiphyseal fusion). Ultimately, our goal is to predict posterior age given skeletal traits. This requires an additional set of Bayesian calculations that will not be required for every application.

1.2 Univariate continuous model

Let x be a scalar independent variable and let w be a scalar response variable distributed normally per

$$w \sim \mathcal{N}(h(x,\mathbf{c}),\psi^2(x,\boldsymbol{\kappa})), \tag{1}$$

where $\mathcal{N}$ denotes the normal distribution, $h(x, \mathbf{c})$ is the mean, $\mathbf{c}$ a vector that parameterizes the mean, $\psi(x, \boldsymbol{\kappa})$ the standard deviation (or noise), and $\boldsymbol{\kappa}$ a vector that parameterizes the noise. The likelihood of an observation (x, w) is

$$\lambda_w = \frac{1}{\sqrt{2\pi}\,\psi} \exp\left(-\frac{1}{2}\left[\frac{w - h(x)}{\psi}\right]^2\right). \tag{2}$$

The negative log-likelihood is

$$\eta_w = -\log \lambda_w = \log\sqrt{2\pi} + \log\psi + \frac{1}{2}\left[\frac{w-h}{\psi}\right]^2. \tag{3}$$

Define $S^{(h)}$ as the model specification for the mean and $S^{(\psi)}$ as the model specification for the noise, where the following parametric forms are allowed:

$$h(x, \mathbf{c}) = \left\{ c_2 x^{c_1} + c_3, \quad \text{if } S^{(h)} = 3. \right. \tag{4}$$

and

$$\psi(x, \boldsymbol{\kappa}) = \begin{cases} \kappa_1, & \text{if } S^{(\psi)} = 0. \\ \kappa_1\left[1 + \kappa_2 x\right], & \text{if } S^{(\psi)} = 1. \end{cases} \tag{5}$$

The reason indexing of $S^{(h)}$ starts at 3, rather than 0, is that the cases 0 through 2 are defined below for ordinal variables. The three noise cases are, respectively, constant (homoskedastic), linear (heteroskedastic), and hyperbolic (heteroskedastic).

The full parameter vector for the univariate continuous case is:

$$\boldsymbol{\theta}_w = \begin{bmatrix} \mathbf{c}^T & \boldsymbol{\kappa}^T \end{bmatrix}^T, \tag{6}$$

where T indicates a vector/matrix transpose, $\mathbf{c} = \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix}^T$, and the length of $\boldsymbol{\kappa}$ depends on the noise model. The yada function fit_pow_law does maximum likelihood estimation of this parameter vector, constraining all the parameters other than c_3 and κ_3 (for hyperbolic noise) to be positive.

1.3 Univariate ordinal model

Let v^* be a latent, scalar variable distributed normally per

$$v^* \sim \mathcal{N}(g(x, \mathbf{b}), \gamma^2(x, \boldsymbol{\beta})), \tag{7}$$

where $g(x, \mathbf{b})$ is the mean, $\mathbf{b}$ a vector that parameterizes the mean, $\gamma(x, \boldsymbol{\beta})$ the standard deviation (or noise), and $\boldsymbol{\beta}$ a vector that parameterizes the noise. The latent variable, v^*, is not directly observed; rather, what is observed is an ordinal response, v. Specifically, there are M ordered boundary parameters τ_m where

$m = 1, 2, \cdots, M$ and we adopt the conventions $\tau_0 = -\infty$ and $\tau_{M+1} = \infty$. The relationship between latent and observed responses is

$$v = \begin{cases} 0, & \text{if } -\infty < v^* \leq \tau_1. \\ m, & \text{if } \tau_m < v^* \leq \tau_{m+1}. \\ M, & \text{if } \tau_M < v^* \leq \infty. \end{cases} \tag{8}$$

The likelihood of the outcome (x, v) is

$$\lambda_v = \Phi\left(\frac{\tau_{v+1} - g}{\gamma}\right) - \Phi\left(\frac{\tau_v - g}{\gamma}\right), \tag{9}$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard univariate Gaussian density with a mean of 0 and a standard deviation of 1. As with the continuous case, we define the negative log-likelihood as $\eta_v = -\log \lambda_v$. We adopt the following set of specifications for the mean and noise:

$$g(x, \mathbf{b}) = \begin{cases} x^{b_1}, & \text{if } S^{(g)} = 0. \\ \log(x), & \text{if } S^{(g)} = 1. \\ x, & \text{if } S^{(g)} = 2. \end{cases} \tag{10}$$

and

$$\gamma(x, \boldsymbol{\beta}) = \begin{cases} \beta_1, & \text{if } S^{(\gamma)} = 0. \\ \beta_1 \left[1 + \beta_2 x\right], & \text{if } S^{(\gamma)} = 1. \end{cases} \tag{11}$$

These specifications result from identifiability considerations with the likelihood, Equation 9. Adopting the same parametric form as with the univariate continuous, $g(x) = b_2 x^{b_1} + b_3$, overspecifies the model. To demonstrate this, consider one of the interior terms in Equation 9,

$$\frac{\tau_m - b_2 x^{b_1} - b_3}{\gamma_0 \gamma}, \tag{12}$$

where we add the term λ_0 to account for overall re-scalings of the noise term. Simultaneous shifts of τ_m and b_3 yield the same model, as do simultaneous re-scalings of τ_m, b_2, and b_3 with λ_0. To ensure identifiability, we set $b_2 = 1$ and $b_3 = 0$. Thus, the mean function for a power law specification becomes $g(x, \mathbf{b}) = x^{b_1}$ and the interior term becomes

$$\frac{\tau_m - x^{b_1}}{\gamma}, \tag{13}$$

where there is now no longer a need to constrain the overall scaling of the noise, $\gamma(x, \boldsymbol{\beta})$. The second two specifications of the mean in Equation 10 are special cases of the first. The linear model, $S^{(g)} = 2$, is included in case the power law model, $S^{(g)} = 0$, overfits the data. The log model, $S^{(g)} = 2$, is included

because an identifiability problem exists for the power law model in the limit $b_1 \to 0$. Assuming $x \geq 0$, x^{b_1} becomes

$$\lim_{b_1 \to 0^+} x^{b_1} = \begin{cases} -\infty, & \text{if } x = 0. \\ 1 + b_1 \log(x), & \text{if } x > 0. \end{cases} \tag{14}$$

The identifiabilty problem arises because, for small b_1, overall rescalings of the variables τ_m and b_1 with the mean function γ yield the same model. An identifiable specification for the special case $b_1 \to 0$ merely has the mean function equaling $\log(x)$, which is why this is one of the cases in Equation 11. In fact, previous work that uses cumulative probit models has often assumed one of the two preceding special cases, $g = x$ (*e.g.*, Chapter 8 of Jackman, 2009) or $g = \log(x)$ (*e.g.*, Konigsberg, 2015). We use K-fold cross-validation of single variable models to select parametric specifications of the mean and noise (see below).

The full parameter vector is

$$\boldsymbol{\theta}_v = \begin{bmatrix} \mathbf{b}^T & \boldsymbol{\tau}^T & \boldsymbol{\beta}^T \end{bmatrix}^T, \tag{15}$$

where $\boldsymbol{\tau}^T = \begin{bmatrix} \tau_1 & \cdots & \tau_M \end{bmatrix}^T$ and both $\mathbf{b}$ and $\boldsymbol{\beta}$ may have length zero. The yada function fit_pow_law_ord does maximum likelihood estimation of this parameter vector, constraining all the parameters other than β_3 (for hyperbolic noise) to be zero.

1.4 Full Mixed Model

Let x be a (scalar) independent variable and let $\mathbf{v}^*$ and $\mathbf{w}$ be, respectively, a vector of latent ordinal response variables indexed by $j = 1, 2, \cdots J$ and a vector of directly observed continuous variables indexed by $k = 1, 2, \cdots, K$. We assume the overall latent response vector $\mathbf{y}^{*T} = \begin{bmatrix} \mathbf{v}^{*T} & \mathbf{w}^T \end{bmatrix}^T$ is distributed normally per

$$\mathbf{y}^* \sim \mathcal{N}(\mathbf{f}(x, \mathbf{a}), \Sigma(x, \boldsymbol{\alpha}, \mathbf{z})), \tag{16}$$

where $\mathbf{f}$ is the vector of means, $\mathbf{a}$ is a vector that parameterizes the mean, Σ is the covariance matrix, $\boldsymbol{\alpha}$ is a vector that parameterizes the diagonal terms of Σ, and $\mathbf{z}$ is a vector that parameterizes the correlation terms of Σ. The correlation terms are ρ_{il} and the off-diagonal elements of the covariance matrix are $\Sigma_{il} = \rho_{il}\sigma_i\sigma_l$, where σ_i is the standard deviation of variable i (the diagonal terms of the covariance matrix are $\Sigma_{ii} = \sigma_i^2$).

The vector of means $\mathbf{f}^T = \begin{bmatrix} \mathbf{g}^T & \mathbf{h}^T \end{bmatrix}^T$ consists of the vector of means for ordinal variables, $\mathbf{g}(x, \mathbf{b})$, and the vector of means for continuous variables, $\mathbf{h}(x, \mathbf{c})$, where the parameter vector for the means consists of an ordinal and a continuous component, $\mathbf{a}^T = \begin{bmatrix} \mathbf{b}^T & \mathbf{c}^T \end{bmatrix}^T$. In turn, $\mathbf{b}$ and $\mathbf{c}$ consist of separate specifications for individual variables, which we denote by $\mathbf{b}^{(j)}$ and $\mathbf{c}^{(k)}$ (e.g., $\mathbf{b}^T = \begin{bmatrix} (\mathbf{b}^{(1)})^T & \cdots & (\mathbf{b}^{(J)})^T \end{bmatrix}^T$). We index vectors such as $\mathbf{a}$ that consist of both an ordinal and continuous component with i, adopting the following convention:

$$i = \begin{cases} j, & \text{if } i \leq J. \\ J + k, & \text{if } i > J. \end{cases} \tag{17}$$

We allow the following parametric specifications of the mean:

$$f_i(x,\mathbf{a}^{(i)}) = \begin{cases} x^{a_1^{(i)}}, & \text{if } S_i^{(f)} = 0. \\ \log(x), & \text{if } S_i^{(f)} = 1. \\ x, & \text{if } S_i^{(f)} = 2. \\ a_2^{(i)} x^{a_1^{(i)}} + a_3^{(i)}, & \text{if } S_i^{(f)} = 3. \end{cases} \tag{18}$$

where $S_i^{(f)}$ is a variable that specifies the mean model to be used for each variable. For ordinal variables ($i \leq J$) we define $S_i^{(g)} = S_i^{(f)}$. For continuous variables ($i > J$) we define $S_{i-J}^{(h)} = S_i^{(f)}$. We allow the following specifications of the standard deviations:

$$\sigma_i(x,\boldsymbol{\alpha}^{(i)}) = \begin{cases} \alpha_1^{(i)}, & \text{if } S_i^{(\sigma)} = 0. \\ \alpha_1^{(i)} \left[1 + \alpha_2^{(i)} x\right], & \text{if } S_i^{(\sigma)} = 1. \end{cases} \tag{19}$$

where $S_i^{(\sigma)}$ is a variable that specifies the noise model to be used for each variable. The vector of standard deviations, $\boldsymbol{\sigma}^T = \begin{bmatrix}\boldsymbol{\gamma}^T & \boldsymbol{\psi}^T\end{bmatrix}^T$, consists of the vector of standard deviations for ordinal variables, $\boldsymbol{\gamma}$, and the vector of standard deviations for continuous variables, $\boldsymbol{\psi}$. For ordinal variables ($i \leq J$) we define $S_i^{(\gamma)} = S_i^{(\sigma)}$. For continuous variables ($i > J$) we define $S_{i-J}^{(\psi)} = S_i^{(\sigma)}$.

For the correlation terms of the covariance matrix, ρ_{il}, we allow for the possibility that groups of variables behave identically, where the groups are indexed by $r = 1 \cdots R$. Let $S_i^{(z)}$ give the group, r, that each variable belongs to. There are two types of correlations: intra-group correlations, which only apply for groups with more than one member (non-singleton groups), and inter-group correlations, of which there are $\binom{R}{2}$. The reason the intra-group correlations do not apply to singleton groups is that off diagonal terms must be between two different variables (or, somewhat differently, variables always correlate perfectly with themselves, which is already accounted for by the diagonal terms). The vector $\mathbf{z}^T = \begin{bmatrix}(\mathbf{z}^{(ns)})^T & (\mathbf{z}^{(cr)})^T\end{bmatrix}^T$ consists of a vector of correlations for non-singleton groups, $\mathbf{z}^{(ns)}$, and a vector of inter-group correlations, $\mathbf{z}^{(cr)}$. The ordering of $\mathbf{z}^{(ns)}$ is the same as r, except that singleton groups are excluded. The ordering of $\mathbf{z}^{(cr)}$ is such that the correlation term linking groups r and r', where $r < r'$, is given by the location of the pair (r, r') in the lexical ordering $(1,2),(1,3),\cdots,(R-1,R)$. The R package yada supports variables being assigned to no group, $S_i^{(z)} = \emptyset$, by setting the group assignment to NA. For such variables, the associated correlation term is zero, $\rho_{il} = 0$. For all other variables, the correlation term equals the pertinent element of $\mathbf{z}$ given the group assignments specified by the vector $\mathbf{S}^{(z)}$. All constraints on variables are as before for the univariate cases with the addition that the elements in $\mathbf{z}$ must be between -1 and $+1$.

To make the preceding assumptions clear, consider the groupings we chose for the conditionally dependent model used in the second cross-validation step: $\mathbf{S}^{(\sigma)} = \begin{bmatrix}1 & 2 & 3 & 3 & 4 & 4\end{bmatrix}^T$. That is, variables $i = 1$ and $i = 2$ belong in singleton groups by themselves, variables $i = 3$ and $i = 4$ belong together in a non-singleton group, and variables $i = 5$ and $i = 6$ belong together in another non-singleton group. The correlation coefficients and elements of $\mathbf{z}$ are related per $z_1 = \rho_{34}$, $z_2 = \rho_{56}$, $z_3 = \rho_{12}$, $z_4 = \rho_{13} = \rho_{14}$, $z_5 = \rho_{15} = \rho_{16}$,

$z_6 = \rho_{23} = \rho_{24}$, $z_7 = \rho_{25} = \rho_{26}$, and $z_8 = \rho_{35} = \rho_{36} = \rho_{45} = \rho_{46}$. The full set of relationships in matrix form is:

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} & \rho_{16} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} & \rho_{25} & \rho_{26} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} & \rho_{35} & \rho_{36} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 & \rho_{45} & \rho_{46} \\ \rho_{15} & \rho_{25} & \rho_{35} & \rho_{45} & 1 & \rho_{56} \\ \rho_{16} & \rho_{26} & \rho_{36} & \rho_{46} & \rho_{56} & 1 \end{bmatrix} = \begin{bmatrix} 1 & z_3 & z_4 & z_4 & z_5 & z_5 \\ z_3 & 1 & z_6 & z_6 & z_7 & z_7 \\ z_4 & z_6 & 1 & z_1 & z_8 & z_8 \\ z_4 & z_6 & z_1 & 1 & z_8 & z_8 \\ z_5 & z_7 & z_8 & z_8 & 1 & z_2 \\ z_5 & z_7 & z_8 & z_8 & z_2 & 1 \end{bmatrix} \tag{20}$$

Observed ordinal responses: The latent variables in the vector $\mathbf{v}^*$ are not directly observed; rather, what is observed for each variable is an ordinal response v_j that depends on a set of boundary parameters. Specifically, for each variable j there are M_j ordered boundary parameters $\tau_m^{(j)}$ where $m_j = 1, 2, \cdots, M_j$ and we adopt the conventions $\tau_0^{(j)} = -\infty$ and $\tau_{M+1}^{(j)} = \infty$. The relationship between latent and observed responses is

$$v_j = \begin{cases} 0, & \text{if } -\infty < v_j^* \leq \tau_1^{(j)}. \\ m_j, & \text{if } \tau_{m_j}^{(j)} < v_j^* \leq \tau_{m_j+1}^{(j)}. \\ M_j, & \text{if } \tau_{M_j}^{(j)} < v_j^* \leq \infty. \end{cases} \tag{21}$$

The full parameter vector is

$$\boldsymbol{\theta}_y = \begin{bmatrix} \mathbf{a}^T & \boldsymbol{\tau}^T & \boldsymbol{\alpha}^T & \boldsymbol{z}^T \end{bmatrix}^T. \tag{22}$$

Likelihood: The likelihood of observing the pair $(x, \mathbf{y})$ is

$$\lambda_y = \int_{\tau_{v_1}^{(1)}}^{\tau_{v_1+1}^{(1)}} \cdots \int_{\tau_{v_J}^{(J)}}^{\tau_{v_J+1}^{(J)}} \frac{\exp(-0.5 \begin{bmatrix} (\mathbf{v}^* - \mathbf{g})^T & (\mathbf{w} - \mathbf{h})^T \end{bmatrix} \Sigma^{-1} \begin{bmatrix} (\mathbf{v}^* - \mathbf{g})^T & (\mathbf{w} - \mathbf{h})^T \end{bmatrix}^T)}{(2\pi)^{\frac{J+K}{2}} |\Sigma|^{\frac{1}{2}}} d\mathbf{v}^*. \tag{23}$$

The vector $\mathbf{w}$ is directly observed, and the full multivariate density inside the integral in Equation 23 can be separated into a component that depends on $\mathbf{w}$, which can be moved outside the integral, and a new, conditional density that depends only on $\mathbf{v}^*$. The conditional mean and covariance matrix are

$$\bar{\mathbf{g}} = \mathbf{g} - \Sigma_{vw} \Sigma_{ww}^{-1} (\mathbf{w} - \mathbf{h}) \tag{24}$$

and

$$\bar{\Sigma}_{vv} = \Sigma_{vv} - \Sigma_{vw} \Sigma_{ww}^{-1} \Sigma_{vw}^T, \tag{25}$$

where Σ_{vv} is the upper left block of Σ with dimensions J by J, Σ_{vw} is the upper right block of Σ with dimensions J by K, and Σ_{ww} is the lower right block of Σ with dimensions K by K (and similarly for $\bar{\Sigma}$). Using these definitions and moving the conditional mean into the integration limits yields

$$\lambda_y = \frac{\exp(-0.5(\mathbf{w} - \mathbf{h})^T \Sigma_{ww}^{-1} (\mathbf{w} - \mathbf{h}))}{(2\pi)^{\frac{K}{2}} |\Sigma_{ww}|} \int_{\tau_{v_1}^{(1)} - \bar{g}_1}^{\tau_{v_1+1}^{(1)} - \bar{g}_1} \cdots \int_{\tau_{v_J}^{(J)} - \bar{g}_J}^{\tau_{v_J+1}^{(J)} - \bar{g}_J} \frac{\exp(-0.5 \mathbf{v}^{*T} \bar{\Sigma}_{vv}^{-1} \mathbf{v}^*)}{(2\pi)^{\frac{J}{2}} |\bar{\Sigma}_{vv}|^{\frac{1}{2}}} d\mathbf{v}^*. \tag{26}$$

Equation 26 is a multivariate Gaussian integral of dimension J over a rectangular domain, which can be calculated numerically with the function dmvnorm in the R package mvtnorm.

Missing data: If a variable is missing for either an ordinal or continuous variable, it can be marginalized out of the likelihood calculation by integrating over the interval $-\infty$ to ∞ for the pertinent variable. This is equivalent to simply removing the variable from the calculation by subsetting $\mathbf{y}$, $\mathbf{f}$, and Σ to remove the variable.

A special case for ordinal variables with $\log(x)$: For a mean specification $S_j^{(g)} = 1$ with $g_j = \log(x)$ there is a special case that must be handled; if $x = 0$, the mean tends to $-\infty$ and the shifted integration limits $\tau_{v_j}^{(j)} - g_j$ and $\tau_{v_j+1}^{(j)} - g_j$ tend to ∞. This yields an invalid model unless $v_j = 0$, for which the integration is on the interval $-\infty$ to ∞. Therefore, a $\log x$ specification for ordinal variable j can only be used if all observations for which $x = 0$ have associated ordinal responses $v_j = 0$. For such cases, the variable can be treated as missing since integrating from $-\infty$ to ∞ is simply marginalizing over the variable. A $\log(x)$ specification of the mean cannot be used for continuous variables if $x = 0$ for any observations.

1.5 Bayesian Age Estimation

Ultimately, our goal is to estimate the posterior probability distribution over x (for individuals of unknown age in our example application of subadult age estimation) given a vector of new ordinal and/or continuous measurements, $\tilde{\mathbf{y}}$. In the preceding section, we describe a model for the likelihood λ_y (Equation 26). Although one could in principle specify priors over the model parameters and sample from the posteriors, doing so in practice with an algorithm such as Hamiltonian MCMC is challenging since the gradient of λ_y must be calculated numerically, and doing so is computationally expensive. Indeed, merely solving the maximum likelihood problem to maximize λ_y is challenging (see below). For this reason, in estimating the posterior age distribution we assume only a single parameter vector is available (the vector that maximizes the likelihood, λ_y).

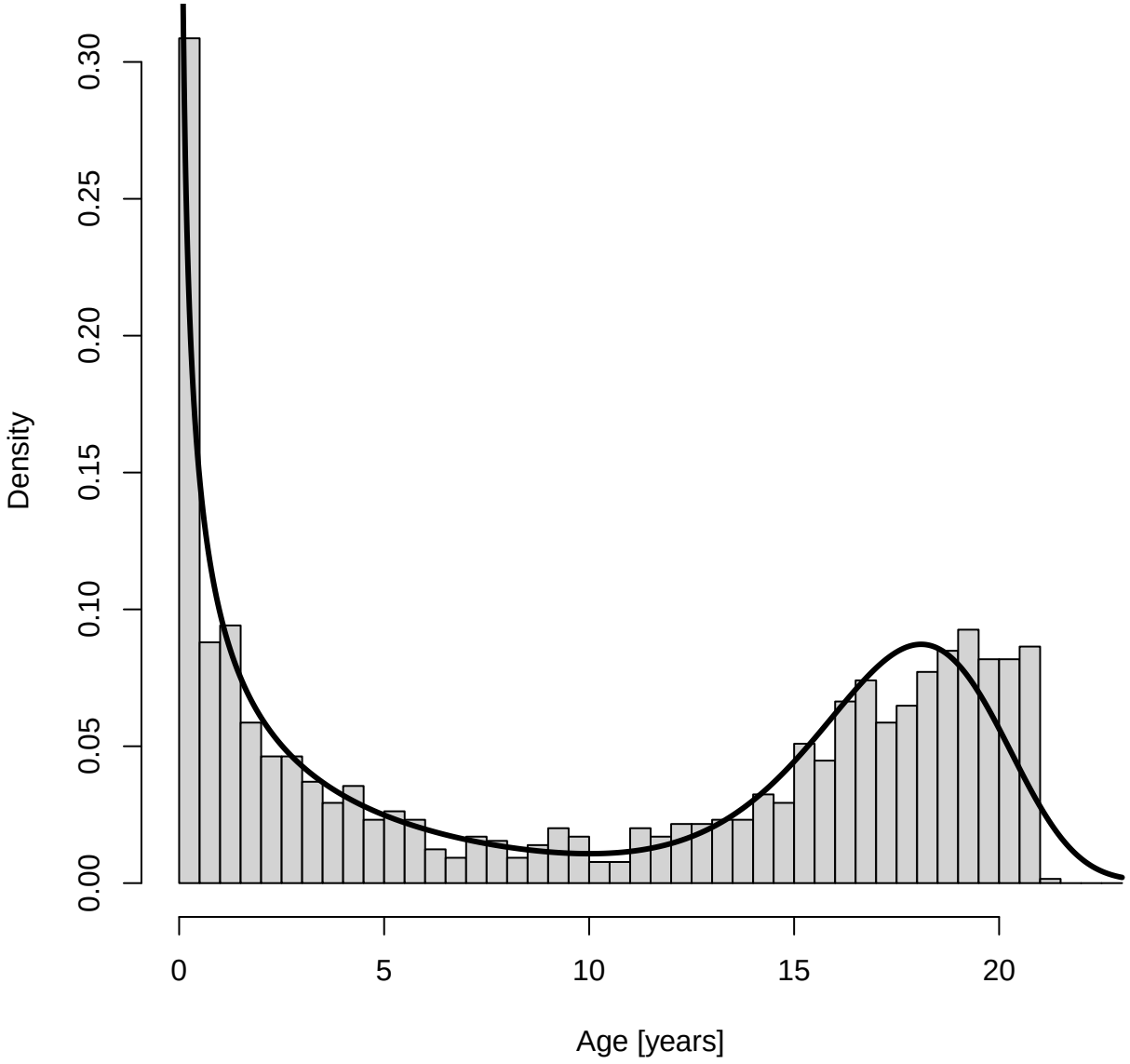
Let $\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\theta}_x^T & \boldsymbol{\theta}_y^T\end{bmatrix}^T$ be the full parameter vector used in inference, where $\boldsymbol{\theta}_y$ is as in the preceding section and $\boldsymbol{\theta}_x$ is a vector that parameterizes the prior distribution of x, $p(x|\boldsymbol{\theta}_x)$. The posterior distribution of x for the vector of observations $\mathbf{y}$ is (applying Bayes' theorem)

$$p(x|\boldsymbol{\theta},\mathbf{y}) = \frac{p(\mathbf{y}|x,\boldsymbol{\theta})p(x|\boldsymbol{\theta})}{p(\boldsymbol{\theta},\mathbf{y})} = \frac{p(\mathbf{y}|x,\boldsymbol{\theta}_y)p(x|\boldsymbol{\theta}_x)}{p(\boldsymbol{\theta},\mathbf{y})}, \tag{27}$$

where we use the fact that y depends on x and $\boldsymbol{\theta}$ only through $\boldsymbol{\theta}_y$ (and, similarly, x depends on $\boldsymbol{\theta}$ only through $\boldsymbol{\theta}_x$), and $p(\mathbf{y}|x,\boldsymbol{\theta}_y)$ is the likelihood, λ_y, described in the preceding section. The denominator in Equation 27 is a normalization of the numerator to ensure that probabilities integrate to 1. Hence, the posterior density can be calculated by calculating the numerator as a function of x, and normalizing it such that integrating over x on the relevant domain yields 1. To parameterize the prior over x, we use an offset mixture of Weibulls with three mixture components (however, other specifications of the prior are supported by yada, including a uniform prior). Figure 1 shows this mixture of Weibulls fit along with the histogram of know ages used to create the fit.

2 Cross-validation of univariate models

Equations 4, 5, 10, and 11 provide alternative specifications of the mean and noise functions for continuous and ordinal variables. We utilize 4-fold cross-validation to choose these parameterizations. The function



Supporting Figure 1. Histogram of known ages (as a density plot, not a frequency plot). The number of observations used to create the histogram is $N = 1296$. The solid black line shows the maximum-likelihood fit for the offset mixture of Weibulls, which is used for the prior in Bayesian inference.

crossval_univariate_models() in the R package 'yada' makes these cross-validation choices. These choices fundamentally stem from the following rule: (1) prefer the model with the lowest out-of-sample negative log-likelihood values unless (2) one or more models are within a user-specified tolerance ("cand_tol") of each other in their out-of-sample negative log-likelihood values (we consider those within 0.05 of the best model), at which point the simplest model is preferred. In addition, some models could not be assessed for

one of the following reasons: (a) at least one of the folds failed to fit successfully; (b) the log_ord models could not be fit (see discussion above on when a log_ord model can be fit); (c) the scaling exponent is close to zero, which implies an identifiability problem (we require that the scaling coefficient ("scale_exp_min") is greater than or equal to 0.01); and (d) The heteroskedastic noise term, β_2, is too large, which implies that the noise at x=0 tends to zero (we require $\beta_2 \leq 5$, "beta2_max"); this issue with β_2 could be addressed by adopting a linear noise model where the intercept is forced to be zero, and a new model could be added, but preliminary work suggests there are some complications with fitting such models, so such a noise model has not yet been added to yada. We summarize the final cross-validation choices, including the best-fit parameter vector for each preferred model with the 6 variables in Table 1. Table 1 in the main text provides detailed cross-validation reports for all 6 variables. These choices were used in the conditionally dependent model described in the next section.

3 Conditionally dependent models

As described above, we allow for the possibility that, for the correlation terms of the covariance matrix, groups of variables behave identically. Our prior expert knowledge allows us to sensibly group variables. We chose to train a four group model with the following groups: (1) All Epiphyseal Fusion variables, (2) All Ossification variables, (3) all Dental variables, and (4) all Continuous variables. These four groups are directly referenced in the model specification as 'cdep_groups' by the multivariate model-fitting function, fit_multivariate() in 'yada'. Response variable parameterization is initialized using the cross-validated univariate model fits (see Table 1), whereas correlation terms are initially set at zero. This initial parameter vector is considered the "conditionally independent" case. Final response variable parameters are included in Table 1.

Table 1. Final model parameters, $\boldsymbol{\theta}_y$, for the conditionally independent and conditionally dependant models. Univariate parameter notation is provided in parentheses. The ordering of $\boldsymbol{\theta}_y$ in the table differs from that in yada and in Equation 22. In particular, the rows of the table are organized by variable, whereas in Equation 22 the ordering is $\boldsymbol{\theta}_y = \left[\mathbf{a}^T \quad \boldsymbol{\tau}^T \quad \boldsymbol{\alpha}^T \quad \boldsymbol{z}^T\right]^T$.

Response Variable	Mean Specification	Noise Specification	Model Parameter	cindep Value ($\boldsymbol{\theta}_y$)	cdep Value ($\boldsymbol{\theta}_y$)
Humerus Medial Epicondyle (HME_EF)	Power Law Ordinal	Linear Positive Intercept	a_1 (b_1)	1.33	1.33
			τ_1 (τ_1)	7.88	7.57
			τ_2 (τ_2)	27.50	27.55
			τ_3 (τ_3)	28.66	28.66
			τ_4 (τ_4)	30.12	30.00
			τ_5 (τ_5)	31.65	31.46
			τ_6 (τ_6)	35.68	35.45
			α_1 (β_1)	2.02	2.22
			α_2 (β_2)	0.17	0.16
Tarsal Count (TC_Oss)	Power Law Ordinal	Linear Positive Intercept	a_2 (b_1)	0.47	0.45
			τ_7 (τ_1)	0.15	0.18
			τ_8 (τ_2)	0.47	0.49
			τ_9 (τ_3)	1.15	1.15
			τ_{10} (τ_4)	1.34	1.33
			τ_{11} (τ_5)	1.53	1.51
			α_3 (β_1)	0.25	0.24
			α_4 (β_2)	0.08	0.06
Maxillary First Molar (max_M1)	Power Law Ordinal	Linear Positive Intercept	a_3 (b_1)	0.54	0.54
			τ_{12} (τ_1)	0.65	0.64
			τ_{13} (τ_2)	0.81	0.82
			τ_{14} (τ_3)	1.12	1.13
			τ_{15} (τ_4)	1.33	1.34
			τ_{16} (τ_5)	1.57	1.59
			τ_{17} (τ_6)	1.77	1.76
			τ_{18} (τ_7)	2.01	2.01
			τ_{19} (τ_8)	2.44	2.41
			τ_{20} (τ_9)	2.67	2.63
			τ_{21} (τ_{10})	3.09	3.06
			τ_{22} (τ_1)	3.23	3.20
			β_5 (α_1)	0.13	0.14
			β_6 (α_2)	0.11	0.11

Mandibular Lateral Incisor (man_I2)	Power Law Ordinal	Constant	a_4 (b_1) τ_{23} (τ_1) τ_{24} (τ_2) τ_{25} (τ_3) τ_{26} (τ_4) τ_{27} (τ_5) τ_{28} (τ_6) τ_{29} (τ_7) τ_{30} (τ_8) τ_{31} (τ_9) τ_{32} (τ_{10}) τ_{33} (τ_{11}) α_7 (β_1)	0.27 0.93 1.00 1.05 1.15 1.31 1.39 1.52 1.63 1.72 1.79 1.86 0.07	0.27 0.93 1.00 1.06 1.15 1.30 1.39 1.52 1.62 1.71 1.78 1.85 0.08
Femur Diaphyseal Length (FDL)	Power Law	Linear Positive Intercept	a_5 (c_1) a_6 (c_2) a_7 (c_3) α_8 (κ_1) α_9 (κ_2)	0.61 65.94 64.88 7.70 0.20	0.62 65.44 65.20 6.50 0.28
Radius Diaphyseal Length (RDL)	Power Law	Linear Positive Intercept	a_8 (c_1) a_9 (c_2) a_{10} (c_3) α_{10} (κ_1) α_{11} (κ_2)	0.59 32.85 47.05 4.82 0.18	0.60 32.27 47.45 4.11 0.23

Correlation terms for cdep model

Group Comparison Type	cdep Groups	Group 1 Variables	Group 2 Variables	Model Parameter	cdep Value
Within	Dental	max_M1	man_I2	z_1	0.65
Within	Long bones	FDL	RDL	z_2	0.86
Between	Epiphyseal fusion - Ossification	HME_EF	TC_Oss	z_3	0.42
Between	Epiphyseal fusion - Dental	HME_EF	max_M1 / man_I2	z_4	0.66
Between	Epiphyseal fusion - Long bones	HME_EF	FDL / RDL	z_5	0.68
Between	Ossification - Dental	TC_Oss	max_M1 / man_I2	z_6	0.22
Between	Ossification - Long bones	TC_Oss	FDL / RDL	z_7	0.34
Between	Dental - Long bones	max_M1 / man_I	FDL / RDL	z_8	0.32

4 Example of turning a continuous variable into an ordinal one

One aspect both our model and other models within the broader family of transition analysis models is that they are latent trait models (Konigsberg, 2015). In particular, we assume that for each ordinal response variable there is a corresponding unobserved, latent response. For ordinal variables, this constitutes a model assumption. However, for continuous response variables an actual response is measured, and some insight can be gained from the exercise of turning a continuous variable into an ordinal one.

The top sub-plot of Figure 2 shows Femuar Diaphyseal Length (FDL) as a function of known age; this is identical to the top sub-plot of Figure 1 in the main text, except for the horizontal grey bars and associated labels. The grey bars mark the 25%, 50%, and 75% quantiles of the FDL values, which we used to convert the continuous FDL measurements into ordinal responses. In particular, all observations below the 25% quantile were given the ordinal response $v = 0$, all observations between the 25% and 50% quantiles were given the ordinal response $v = 1$, etc. The middle sub-plot shows these ordinal responses as a function of known age, and the bottom sub-plot visualizes the ordinal fit for category $v = 2$ using binned probabilities.

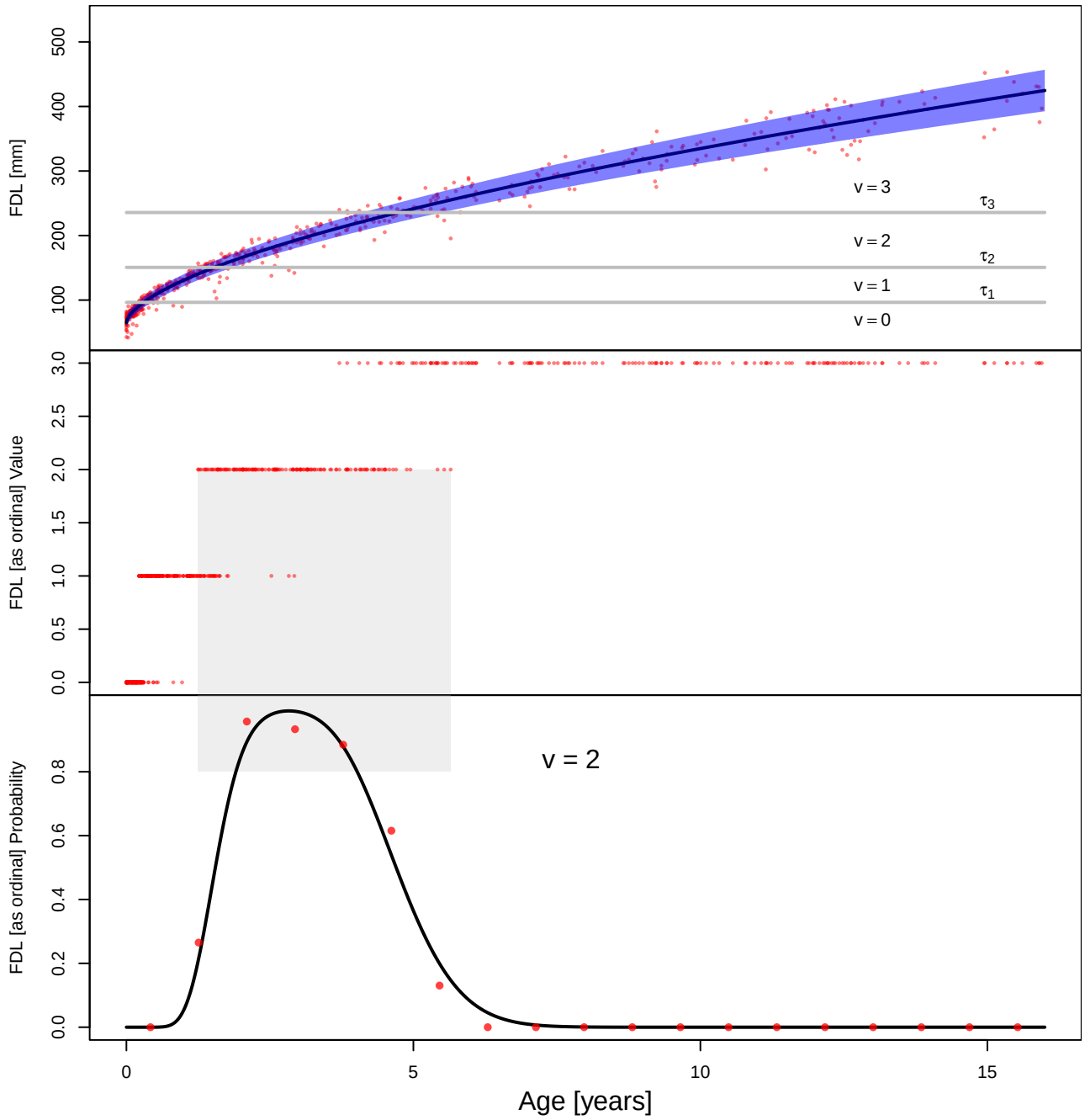
We provide this example, of turning a continuous variable into an ordinal variable, for two reasons: one conceptual and one practical. Conceptually, we expect it will be useful for some reasons to have a tangible example of how the latent space "works". Of course, for actual ordinal variables, there is no continuous response that can actually be observed; the existence of a latent response is a useful modeling assumption. Nevertheless, the top sub-plot of Figure 2 illustrates what such a latent variable might look like.

Practically, we can illustrate one of the benefits of using continuous variables: by their very nature they offer more precise estimates. Conversely, ordinal variables, which possess a constrained set of values, do not necessarily offer the most precise age estimates, even though some scholars prefer them since they are associated with more canalized age indicators (Cardoso, 2007; Conceição and Cardoso, 2011). The loss of precision in going from a continuous FDL response to an ordinal one with four categories can be quantified using information theory. Rather than using the KL divergence as in the main text, we use the mutual information as in Stull et al., 2021. With the KL divergence, one must specify a known response vector. With the mutual information, one can treat the response vector as a random variable, and ask how much information is gained, on average, by knowing the response. We condition on age in making this calculation to yield the mutual information (information gain) as a function of age (alternatively, one could average across the prior probability to yield an age-independent measure). To calculate the mutual information, the data must be represented or modeled probabilistically, which can be done by either fitting the data or discretizing the continuous variables to directly calculate the mutual information (and, optionally, binning observations by age). In Stull et al., 2021, we did the latter. Here, we do the former.

To make this precise, let x_0 be the baseline age to be conditioned. As before, we represent the prior by $p(x|\boldsymbol{\theta}_x)$ and the posterior by $p(x|\boldsymbol{\theta},\mathbf{y})$. Now let the response vector, $\mathbf{y}$, depend on the baseline age using the preceding likelihood equations. The mutual information is

$$I = \mathbb{E}_{y|x_0}\{KL(p(x|\boldsymbol{\theta},\mathbf{y})|p(x|\boldsymbol{\theta}_x))\}, \tag{28}$$

where $\mathbb{E}_v$ represents an expectation over the variable $\mathbf{y}$. For univariate ordinal models Equation 28 is a sum for which $\mathbf{v}$ takes on discrete values (for ordinal variables, we use $\mathbf{v}$ rather than $\mathbf{y}$ for the response



Supporting Figure 2. Top: femur diaphyseal length (FDL) vs known age with a heteroskedastic maximum likelihood fit. Red dots are observations, the black line is the mean response, and the blue shaded region marks the noise bounds. Middle: FDL as an ordinal variable vs known age. Bottom: Probability of observing the ordinal category m = 2 as a function of age for FDL as an ordinal variable. The grey band that extends from the middle to bottom plot marks the range of ages for which v = 2 is observed in the data. The black curve is the predicted probability for a heteroskedastic fit with a power law mean response. The red dots are the actually observed proportions in the underlying data, which are calculated by binning observations by known age value and calculating the proportion of observations in each bin for which v = 2.

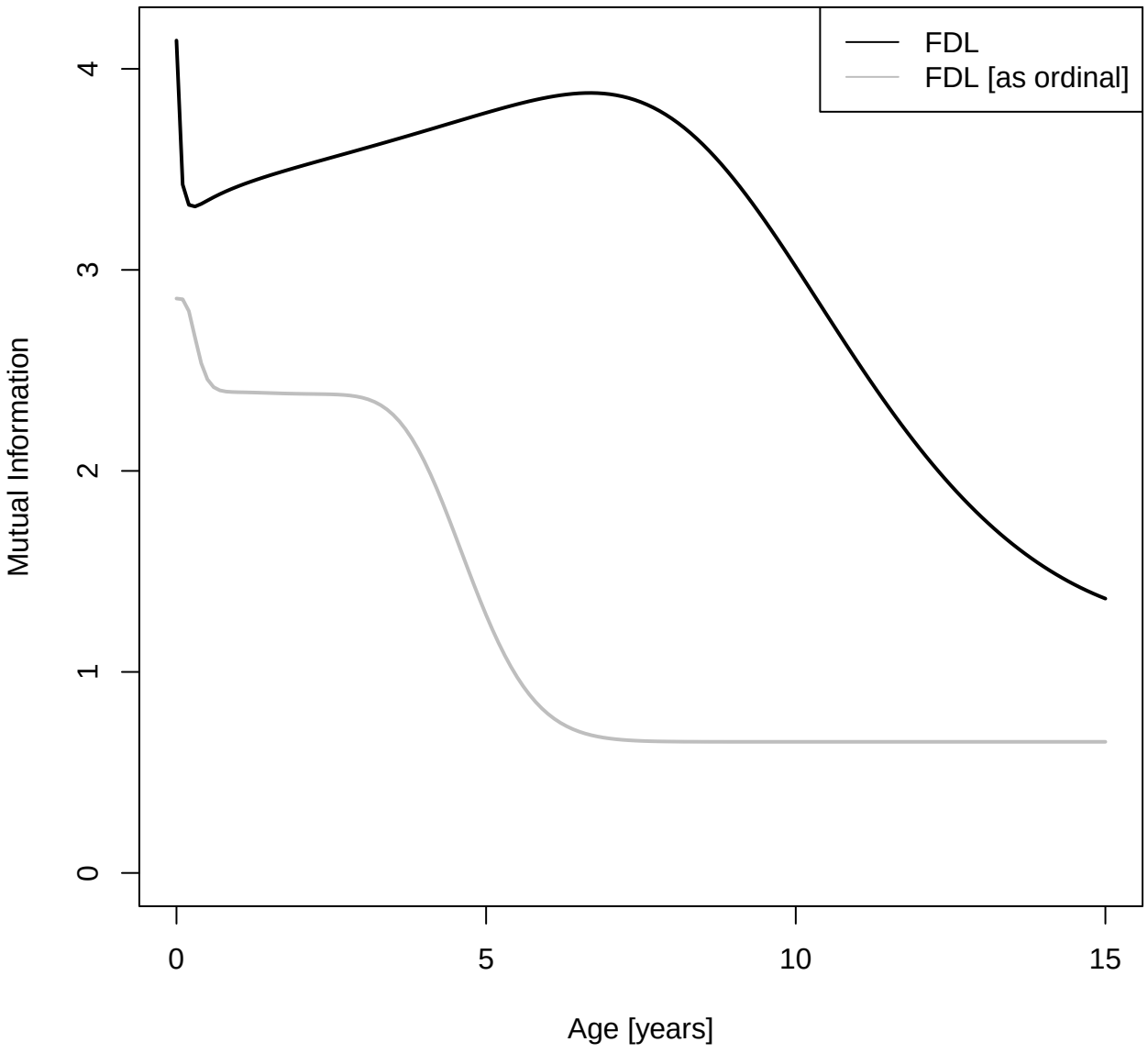
vector) with probabilities given by Equation 9,

$$I = \sum_{m=0}^{M} p(v = m|x_0) KL(p(x|\boldsymbol{\theta}, v = m)|p(x|\boldsymbol{\theta}_x)), \tag{29}$$

where

$$p(v = m|x_0) = \Phi(\frac{\tau_{m+1} - g(x_0, \mathbf{b})}{\gamma(x_0, \boldsymbol{\beta})}) - \Phi(\frac{\tau_m - g(x_0, \mathbf{b})}{\gamma(x_0, \boldsymbol{\beta})}). \tag{30}$$

A similar set of equations holds for univariate continuous models, except that the expectation is an integral rather than a sum and the conditional probabilities come from Equation 2. Figure 3 plots the mutual information as a function of age (x_0) for both the FDL and FDL as an ordinal variable. Regardless of age, the ordinal representation necessarily yields a less informative model, and amount of information lost can be quite substantial, ranging from 16% to 83% of the total information available.



Supporting Figure 3. Mutual information (mean information gain) as a function age for FDL and FDL as an ordinal variable (both heteroskedastic and with power law response specifications). Turning the continuous variable into an ordinal one necessarily entails a loss of information regardless of age.

5 Credible intervals for univariate ordinal responses

A common representation of age estimation using ordinal variables are tables that provide the point estimate and confidence intervals as age ranges for each corresponding stage. The MCP provides a credible interval in lieu of a confidence interval, as the former is used to describe an interval over a Bayesian probability distribution and the latter is based in a Frequentist framework (Kruschke, 2021). Table 2 provides a summary of the ordinal response variables used for the main article to provide analogous results for comparison with other age estimation publications. We utilize the 2.5% and 97.5% values of the quantile of the posterior distribution (that is, an equal-tailed interval) rather than the highest density interval.

Table 2. Point estimates and 95% credible intervals of age in years for each stage by ordinal response variable

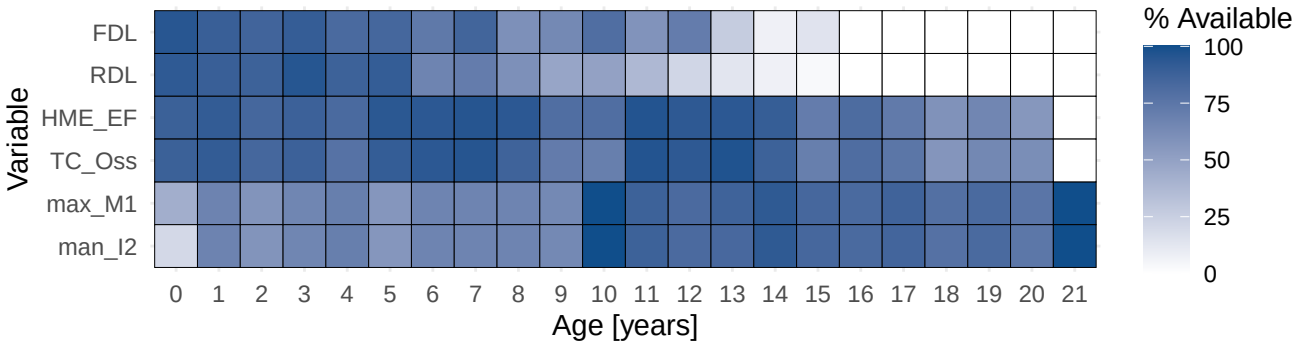
Response Variable	Stage ($m =$)	Point Estimate	95% Credible Interval
HME_EF	0	1.41	0.01 - 5.25
	1	8.54	2.91 - 16.39
	2	13.95	9.45 - 18.43
	3	14.38	9.96 - 18.70
	4	14.82	10.57 - 18.92
	5	15.53	11.47 - 19.42
	6	18.08	13.97 - 21.83
TC_Oss	0	0.05	0.00 - 0.21
	1	0.21	0.00 - 0.80
	2	0.79	0.05 - 2.36
	3	1.71	0.50 - 3.83
	4	2.28	0.78 - 4.85
	5	14.30	2.35 - 21.35
max_M1	0	0.18	0.00 - 0.60
	1	0.57	0.23 - 1.04
	2	0.96	0.45 - 1.64
	3	1.48	0.87 - 2.27
	4	2.02	1.26 - 3.00
	5	2.63	1.75 - 3.75
	6	3.31	2.25 - 4.66
	7	4.45	3.02 - 6.33
	8	5.76	4.18 - 7.86
	9	7.24	5.14 - 10.28
	10	8.82	6.45 - 12.68
	11	17.02	9.99 - 21.55
man_I2	0	0.29	0.00 - 0.91
	1	0.90	0.47 - 1.48
	2	1.13	0.62 - 1.82
	3	1.48	0.83 - 2.35
	4	2.21	1.25 - 3.47
	5	3.10	1.96 - 4.54
	6	4.13	2.63 - 6.01
	7	5.53	3.71 - 7.77
	8	6.93	4.80 - 9.58
	9	8.31	5.84 - 11.53
	10	9.95	6.92 - 13.86
	11	17.23	11.25 - 21.58

6 Missing data visualization

Observational datasets, as are common in biological anthropology, often have large frequencies of missing data. Data may be missing due from taphonomic processes (*e.g.*, damaged or not recovered), trauma (*e.g.*, preventing the observation of a feature), or processes of growth and development (*e.g.*, complete fusion of long bones or plateauing at the highest developmental stage). Many statistical models require complete datasets, unlike the MCP, which removes a missing variable during likelihood calculations (see **Missing data** discussion within Section 1 above). To demonstrate the pattern of missing data commonly found in growth and development and observational data, a visual representation of data availability (the opposite of missing data) by age is provided in Figure 4. When evaluating the percentage of missing data over all ages by response variable, diaphyseal lengths present with the most amount of missing data, followed by dental development, and finally epiphyseal fusion and ossification (see Table 3 below).

Table 3. Summary of missing data percentages by response variable over entire data.

Response Variable	% Missing Data
HME_EF	20.80
TC_Oss	21.11
max_M1	30.14
man_I2	35.00
FDL	53.99
RDL	57.02



Supporting Figure 4. Percentage of data availability for each of the six response variables is presented over chronological age (0-21 years-old). This figure demonstrates the pattern of missing data that are common barriers to other subadult age estimation methods in biological anthropology.

Supporting Information References

Cardoso, H. (2007). Environmental effects on skeletal versus dental developement: Using a documented subadult skeletal sample to test a basic assumption in human osteological research. *American Journal of Physical Anthropology*, *132*, 223–233.

Conceição, E., & Cardoso, H. (2011). Environmental effects on skeletal versus dental development ii: Further testing of a basic assumption in human osteological research. *American Journal of Physical Anthropology*, *144*, 463–470.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. John Wiley & Sons.

Konigsberg, L. W. (2015). Multivariate cumulative probit for age estimation using ordinal categorical data. *Annals of Human Biology*, *42*(4), 368–378. https://doi.org/10.3109/03014460.2015.1045430

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, *5*(10), 1282–1291.

Stull, K. E., Corron, L. K., & Price, M. H. (2021). Subadult age estimation variables: Exploring their varying roles across ontogeny (B. Algee-Hewitt & J. Kim, Eds.). In B. Algee-Hewitt & J. Kim (Eds.), *Remodeling forensic skeletal age*. Elsevier.