Statistica Sinica

Supplement to Statistical inference with anchored Bayesian mixture of regressions models: An illustrative study of allometric data

Deborah Kunkel[1] and Mario Peruggia[2]

1. School of Mathematical & Statistical Sciences, Clemson University, Clemson, SC, USA

2. Department of Statistics, The Ohio State University, Columbus, OH, USA

1. Anchored EM algorithm

This section restates the steps of the anchored EM algorithm for the mixture of regressions model with additional details on its implementation.

Initialization. To initialize $\theta^0 = (\boldsymbol{\beta}^0, \sigma^0, \boldsymbol{\eta}^0)$, we recommend randomly partitioning the data into k groups and initializing $\boldsymbol{\beta}$ at the least-squares estimates calculated from each group. The residual variance σ^2 can be initialized at the estimate from one of these least-squares solutions. We initialize Δ at a large positive value (100 in this study) and set the tolerance to a small, positive value (1×10^{-5} in this study).

E-step. Calculate r_{ij}^t for $i = 1, \ldots, n$, $j = 1, \ldots, k$, where r_{ij} is the posterior probability that $S_i = j$ given $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma, \boldsymbol{\eta}$, and equals

$$r_{ij}^t \quad = \frac{\eta_j^t \, \phi\left(y_i; \boldsymbol{x}_i \boldsymbol{\beta}_j^t, \sigma^{2t}\right)}{\sum_{l=1}^k \eta_l^t \, \phi\left(y_i; \boldsymbol{x}_i \boldsymbol{\beta}_l^t, \sigma^{2t}\right)}. \tag{1.1}$$

where $\phi(\cdot; a, b)$ denotes the density function of a normal distribution with mean a and variance b. The distribution $q(\boldsymbol{s})$ is updated as

$$q^t(\boldsymbol{s}) \quad = \quad \prod_{i=1}^n q^t(s_i) \tag{1.2}$$

$$\quad = \quad \prod_{i=1}^n \left(r_{ij}^t\right)^{I(s_i = j)} \tag{1.3}$$

where $I(s_i = j)$ is an indicator function that equals 1 if $s_i = j$ and equals 0 otherwise.

Anchor step. For fixed values m_j, $j = 1, \ldots, k$, update the anchor points by finding $A^t = \cup_{j=1}^k A_j^t$ to maximize

$$\sum_{j=1}^k \sum_{i \in A_j} r_{ij}^t, \tag{1.4}$$

subject to $A_j \cap A_{j'} = \emptyset$ and $|A_j| = m_j$ for all $j \neq j'$. Then set

$$\widetilde{r}_{ij}^t = \begin{cases} r_{ij}^t & \text{if } i \notin A^t \\ 1 & \text{if } i \in A_j^t \\ 0 & \text{if } i \in A_{j'}^t, \quad j' \neq j. \end{cases} \tag{1.5}$$

The optimization step in 1.4 simply amounts to assigning to component j the m_j points with the highest posterior probability of allocation to component j given the current estimate of the model parameters, in situations where this does not anchor any observations to more than one component. If an observation would be anchored to more than one component, a situation that could occur if the m_j values are large relative to the sample size, an approximate solution may be used, or linear programming algorithms can produce an exact solution.

M-step. In the M-step, we update $\theta^t = (\boldsymbol{\beta}^t, \sigma^t, \boldsymbol{\eta}^t)$ to maximize $F(q^t, \theta)$. The objective function satisfies

$$F(q, \theta) \quad = E_q \log(p(\boldsymbol{\theta}, \boldsymbol{s}, \boldsymbol{y})) - E_q \log(q(\boldsymbol{s})), \tag{1.6}$$

with the second term constant with respect to θ. The M-step is thus derived by maximizing the following with respect to θ:

$$\begin{aligned}
F^*(q^t, \theta) \quad &= \quad E_q \log(p(\boldsymbol{\theta}, \boldsymbol{s}, \boldsymbol{y})) && (1.7)\\
&= \quad E_q \log(f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{s}, \boldsymbol{\theta})) + \log(p(\boldsymbol{\beta}) + \log(p(\boldsymbol{\sigma^2})) + \log(p(\boldsymbol{\eta})) && (1.8)\\
&= \quad \sum_{j=1}^{k} \sum_{i=1}^{n} \widetilde{r}_{ij}^{t} \log(\phi(y_i; \boldsymbol{x}_i \boldsymbol{\beta}_j, \sigma^2)) + \log(p(\boldsymbol{\beta}) + \log(p(\boldsymbol{\sigma^2})) + \log(p(\boldsymbol{\eta})) && (1.9)\\
&= \quad \frac{n}{2} \log(\sigma^{-2}) - \frac{1}{2} \sum_{j=1}^{k} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_j)' \boldsymbol{R}_j^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_j) - \\
&\quad\quad \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)' V^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_\beta) + \\
&\quad\quad (a-1) \log\left(\sigma^{-2}\right) - b\sigma^{-2} + (\alpha - 1) \sum_{j=1}^{k} \log(\eta_j) + c. && (1.10)
\end{aligned}$$

In the expression above, c represents terms that are constant with respect to θ and $\boldsymbol{R}_j^t$ is an $n \times n$ diagonal matrix whose i-th diagonal element is $\widetilde{r}_{ij}^{t}$. The update steps are as follows:

$$\begin{aligned}
\boldsymbol{\beta}_j^t &= \left(\boldsymbol{X}'\boldsymbol{R}_j^t\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{R}_j^t\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{\mu}_\beta\right), \;\; j = 1, \ldots, k, \\
\left(\sigma^{-2}\right)^t &= \frac{a + n/2 - 1}{b + .5\sum_{j=1}^k \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_j^t\right)'\boldsymbol{R}_j^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_j^t)}, \\
\eta_j^t &= \frac{\sum_{i=1}^n r_{ij}^t + \alpha - 1}{\sum_{l=1}^k \sum_{i=1}^n r_{il}^t + \alpha - 1}, \quad j = 1, \ldots, k. \qquad (1.11)
\end{aligned}$$

Monitoring convergence. After the expectation, anchoring, and maximization steps, t is increased by 1 and Δ is updated until its improvement does not exceed the tolerance. To update Δ, we calculate $F^*(q^t, \theta^t) - F^*(q^{t-1}, \theta^{t-1})$. The function $F^*(q, \theta)$ is given by 1.10 above. Convergence may instead be monitored using $F(q^t, \theta^t) - F(q^{t-1}, \theta^{t-1})$ by adding the term $E_q \log(q^{t-1}(\boldsymbol{s})) - E_q \log(q^t(\boldsymbol{s}))$ to Δ.

Initialization and convergence. We have found the algorithm to be sensitive to initial values and prone to visit local maxima, especially when components are not well-separated. We thus recommend running the algorithm several times (at least 15-20) and selecting the solution with the largest ending value of $F^*(q, \theta)$. We used 50 runs in the data analysis and simulations in this study. For the 3-component mixture models considered in this study, we have found that it is typical for about half of the sequences to converge to the same solution.

In the analysis and simulations of this manuscript, we set the tolerance for the algorithm to be 1×10^{-5} and typically saw convergence in less than 200 iterations. If the algorithm has not converged after a very large number of iterations, it should be stopped and re-started from new random starting points. We have found that, more often, the algorithm may converge in 3 or 4 iterations to a “poor” solution, such as one in which two components have identical parameters. If this behavior is observed in the final solution, the algorithm should be re-started from new random starting points; however, we have found that this situation is prevented by using several runs and retaining only the best solution.

2. Simulation study

We conducted a simulation study to evaluate the performance of the three anchoring methods: A-EM, CDW-cov, and CDW-cor.

2.1 Data generation.

Data sets were generated under mixtures of simple linear regressions models with $k = 3$ components. To mimic the mammals data studied in the main text, we used sample sizes of $n = 100$ for all settings. Six settings were used to generate the data: A1, A2, B1, B2, C1, and C2. Settings A, B, and C used different values of $\boldsymbol{\beta}$, inducing different relationships among the true regression lines. Figures 1, 4, and 7 show the true regression lines for each case. One additional setting with a multiple regression model is considered in Section 2.5.

For each value of $\boldsymbol{\beta}$, we used two values of σ. These values were chosen so that the ratio

$$\frac{\sum_{j=1}^{k} \eta_j (\boldsymbol{\beta}_j - E_s(\boldsymbol{\beta}))^2}{\sum_{j=1}^{k} \eta_j (\boldsymbol{\beta}_j - E_s(\boldsymbol{\beta}))^2 + \sigma^2}, \tag{2.12}$$

which reflects the ratio of expected across-cluster variability to total variability, was set to be 0.95 (settings A1, B1, and C1) and 0.8 (settings A2, B2, and C2).

For each setting, 100 data sets were generated at random. The predictor variable, x, was simulated from a standard normal distribution and centered and rescaled. Latent allocations were sampled with $\eta_1 = \eta_2 = \eta_3 = 1/3$, and data were sampled conditional on the latent allocations.

As we did in the analysis of the mammals data performed in the main text, $m = 3$ anchor points per component were selected using the A-EM, CDW-cov, and CDW-cor methods. Each anchor model was fit using a Gibbs sampler and posterior means of the model parameters were estimated. The hyperparameters were fixed at the following values: $a = 5$, $b = 1$, $v_0 = 1$, $v_1 = 3$, $\alpha = 1.5$, $\boldsymbol{\mu}_\beta = (\bar{y}, 0)'$, where $\bar{y}$ is the sample mean of the simulated response variable. After fitting the models, posterior means were computed for the model parameters, and maximum a posteriori estimates were obtained for the latent allocation.

In addition to the three anchor models, we analyzed each data set using a traditional mixture of regressions model with no anchor points. Post-hoc relabeling was used to relabel the posterior samples. This procedure involves first fitting a mixture of regressions model with no anchor points and then applying the likelihood-based relabeling method strategy of Stephens (2000) as implemented in the R package `label.switching` (Papastamoulis, 2016). The package does not automatically relabel the sampled allocation vectors, so estimated allocations were not computed for this method.

CDW implementation details. When implementing the CDW methods, anchor points were chosen automatically using k means on the rows of the $\widehat{\boldsymbol{C}}$ and $\widehat{\boldsymbol{R}}$ matrices, and then k means was run a second time to find sub-clusters and their centroids. In order to automate the procedure, we did not make PCA displays. We recommend including this step in analysis of real data, however, to evaluate graphically the features of the selected anchor points.

For some simulated data sets, the CDW method occasionally estimated initial clusters with too few observations to select 3 anchor points per component. This was more typical of the CDW-cov method, due to k means identifying a small cluster of points with large variances of the log case deletion weights. In these situations, we still selected three anchor points per component by artificially adding points to the too-small cluster. The points that were closest to the centroid in Euclidean distance were artificially added until each cluster contained at least 5 points. This decision was made to facilitate the automation of the procedures. In practice, if this occurs, it may be appropriate to instead modify the model to require fewer anchor points for the affected component(s).

Evaluation

We evaluated the methods' performance on the simulated data sets by measuring estimation accuracy and clustering accuracy.

Squared estimation error. Monte Carlo estimates of the posterior means were used as the primary estimates of each model parameter. We denote by $\widehat{\theta}_d$ the estimated posterior mean of a parameter θ from simulated data set d.

To calculate error, we first relabeled the posterior means to minimize the error in estimating $\boldsymbol{\beta}$. Relabeling is not typically necessary in an anchor model; however, this step ensured that it was possible to compare estimated parameters to the true values that generated the data. For each possible relabeling of the component-specific parameters, the error was calculated as the sum of the relative squared distances, $\left((\widehat{\theta}_d - \theta_d)/\theta_d\right)^2$, of the posterior means from their true values. The relabeling that minimizes this value was chosen as the final parameter estimate.

After achieving the optimal relabeling of parameter estimates, we calculated the mean squared error separately for the intercept parameters, slope parameters, mixture weights, and residual variance. The mean squared error for vector-valued $\boldsymbol{\theta}_d$ was calculated as

$$\frac{1}{D}\sum_{d=1}^{D}(\boldsymbol{\theta}_d - \boldsymbol{\theta})^T(\boldsymbol{\theta}_d - \boldsymbol{\theta}).$$

Clustering accuracy. We also assessed the accuracy of the maximum a posteriori estimates of $\boldsymbol{s}$ from each anchor model. To measure accuracy, we calculated the Rand index between the estimated allocation to the true allocation that generated the data. The Rand index measures the similarity of two clustering structures estimated from the same data (Rand, 1971). Numbers close to 1 indicate that the anchor model's clustering is similar to the allocation that in truth generated the data. Numbers close to 0 indicate a dissimilar grouping of the observations.

As a benchmark, we also estimated an allocation using likelihood-based classification probabilities evaluated at the true values of the model parameters:

$$s_i^{oracle} = \max_j P(S_i = j|y_i, \boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2), \quad i = 1, \ldots, n \tag{2.13}$$

where

$$P(s_i = j|y_i, \boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2) \propto \eta_j \phi(y_i; x_i\beta_j, \sigma^2). \tag{2.14}$$

We refer to the allocation vector estimated using 2.13 as the "oracle" allocation.

2.2 Results: Setting A

Figure 1 shows the regression lines used to simulated data in settings A1 and A2 (left panel). The center and right panels show examples of simulated data sets from these two settings.

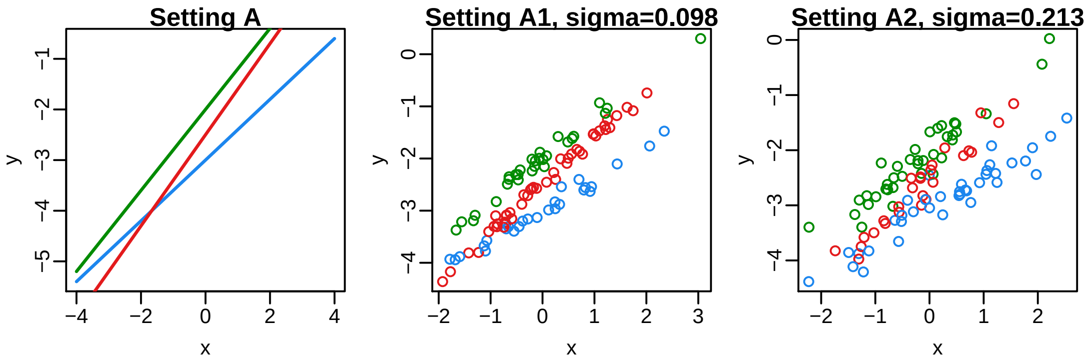

Figure 1: True regression lines and representative simulated data sets for Setting A. Plotting symbols are colored by their true allocation.

The plots in Figure 2 summarize the clustering and estimation performance of the four methods for setting A1. The left panel shows boxplots of the Rand index for the three anchor models and the oracle allocation. Among the anchor models, the CDW-cor method has the highest median clustering accuracy and, for a few data sets, achieves values close to those typical of the oracle allocation. The A-EM and CDW-cov models perform similarly to each other, with more variability in the EM values.

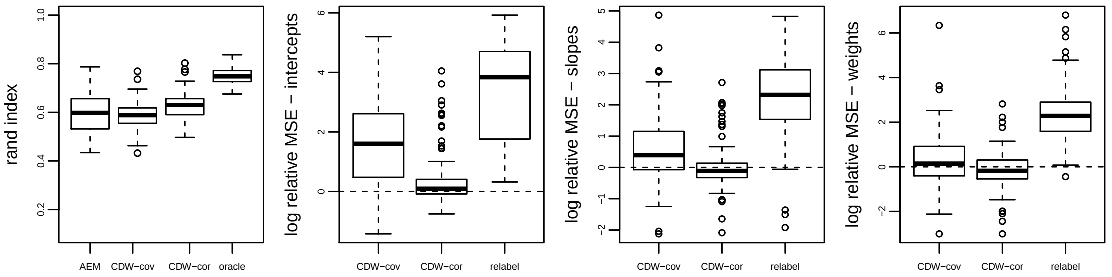

Figure 2: Results from setting A1: Rand index (left) and MSE relative to anchored EM for the intercept parameters (left center), slope parameters (right center), and mixture weights (right).

The right panels of Figure 2 show boxplots of the log MSE relative to the A-EM model under Setting A1. Values greater than zero indicate poorer performance than the A-EM model and values less than zero indicate better performance. The left center plot shows that the CDW-cov and relabeling method typically have much higher error than A-EM, while CDW-cor outperforms A-EM about 25% of the time. In estimating the slope (right center panel), CDW-cor performs comparably to A-EM, with CDW-cov and relabeling again resulting in larger error. All of the anchor models show similar performance in estimating the mixture weights, $\boldsymbol{\eta}$.

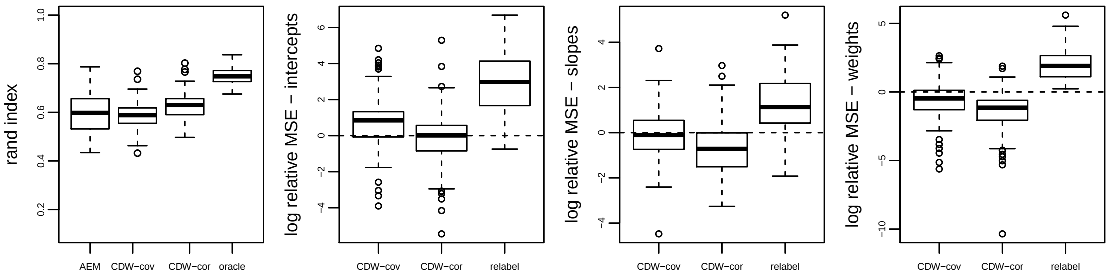

Figure 3: Results from setting A2: Rand index (left) and MSE relative to anchored EM for the intercept parameters (left center), slope parameters (right center), and mixture weights (right).

Figure 3 shows the same summaries for setting A2, where the residual variability is higher. In this more difficult case, the methods perform similarly in classification accuracy, with CDW-cor again tending to be the best. In this setting, CDW-cor improves in estimation accuracy relative to A-EM, and in fact out-performs A-EM when estimating the slopes. The CDW-cov also exhibits better relative performance, but still has poorer estimation accuracy than A-EM. As in the previous setting, the relabeling method provides the least accuracy.

The average posterior means of each parameter are shown in Table 1 for settings A1 and A2. These values indicate that the relabeling method tends to estimate one component with very low weight. This component is estimated to have a shallow slope and intercept between that of the other two components. An explanation for this is that the method identifies some red points between the green and blue lines, as seen in Figure 1, as belonging to a distinct component, but fails to accurately detect the linear pattern followed by the red points across a wider range of x-values without prior information from the anchor points.

Table 1: Average posterior means for each model under setting A1 (top) and A2 (bottom). Values are the posterior means for each parameter, averaged over all data sets. Values in parentheses are estimated standard errors.

Setting A1

	Component 1			Component 2			Component 3			
	β_0	β_1	η	β_0	β_1	η	β_0	β_1	η	σ^2
true	-3	0.6	0.3333	-2.5	0.9	0.3333	-2	0.8	0.3333	0.009591
AEM	-2.948 (0.05)	0.577 (0.04)	0.343 (0.05)	-2.485 (0.06)	0.932 (0.07)	0.321 (0.07)	-2.062 (0.03)	0.769 (0.03)	0.335 (0.06)	0.049 (0.01)
CDW-cov	-2.862 (0.14)	0.613 (0.13)	0.323 (0.06)	-2.553 (0.23)	0.807 (0.16)	0.304 (0.05)	-2.152 (0.10)	0.808 (0.07)	0.373 (0.07)	0.064 (0.01)
CDW-cor	-2.936 (0.08)	0.589 (0.06)	0.338 (0.05)	-2.494 (0.10)	0.905 (0.10)	0.311 (0.06)	-2.083 (0.07)	0.770 (0.06)	0.351 (0.06)	0.051 (0.01)
relabel	-2.910 (0.04)	0.566 (0.03)	0.377 (0.06)	-2.460 (0.04)	0.601 (0.17)	0.168 (0.13)	-2.195 (0.09)	0.785 (0.10)	0.455 (0.12)	0.069 (0.01)

Setting A2

	Component 1			Component 2			Component 3			
	β_0	β_1	η	β_0	β_1	η	β_0	β_1	η	σ^2
true	-3	0.6	0.3333	-2.5	0.9	0.3333	-2	0.8	0.3333	0.04556
AEM	-2.920 (0.07)	0.558 (0.08)	0.318 (0.08)	-2.479 (0.15)	0.874 (0.20)	0.350 (0.09)	-2.096 (0.10)	0.765 (0.11)	0.332 (0.11)	0.095 (0.01)
CDW-cov	-2.842 (0.16)	0.602 (0.14)	0.309 (0.07)	-2.541 (0.20)	0.783 (0.20)	0.320 (0.07)	-2.182 (0.15)	0.810 (0.10)	0.371 (0.08)	0.106 (0.02)
CDW-cor	-2.917 (0.11)	0.608 (0.08)	0.329 (0.05)	-2.467 (0.14)	0.828 (0.14)	0.324 (0.05)	-2.118 (0.09)	0.797 (0.07)	0.347 (0.05)	0.095 (0.01)
relabel	-2.841 (0.09)	0.569 (0.06)	0.361 (0.08)	-2.468 (0.05)	0.491 (0.14)	0.113 (0.10)	-2.256 (0.08)	0.830 (0.10)	0.526 (0.12)	0.125 (0.02)

2.3 Results: Setting B

Settings B1 and B2 generated data from three parallel lines, as shown in Figure 4.

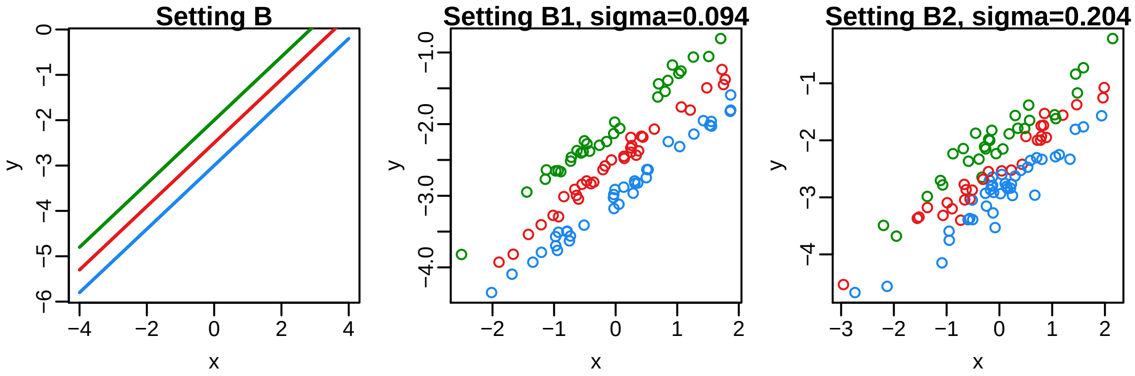

Figure 4: True regression lines and representative simulated data sets for Setting B. Plotting symbols are colored by their true allocation.

The performance summaries under setting B1 are shown in Figure 5. The typical values of the Rand index are highest for the CDW-cor models, and the MSE is lowest for the same method. Figure 6 shows that in Setting B2, the case with less separation, the advantage of CDW-cor in terms of MSE is even stronger. CDW-cov in this setting performs as well or better than anchored EM, in contrast to settings where models have differing slopes.

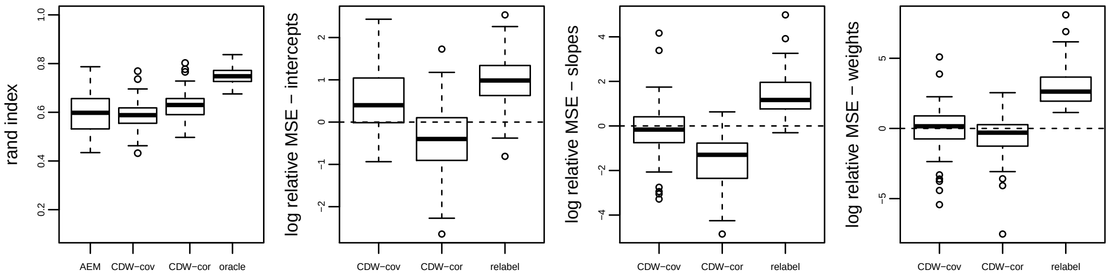

Figure 5: Results from setting B1: Rand index (left) and MSE relative to anchored EM for the intercept parameters (left center), slope parameters (right center), and mixture weights (right).

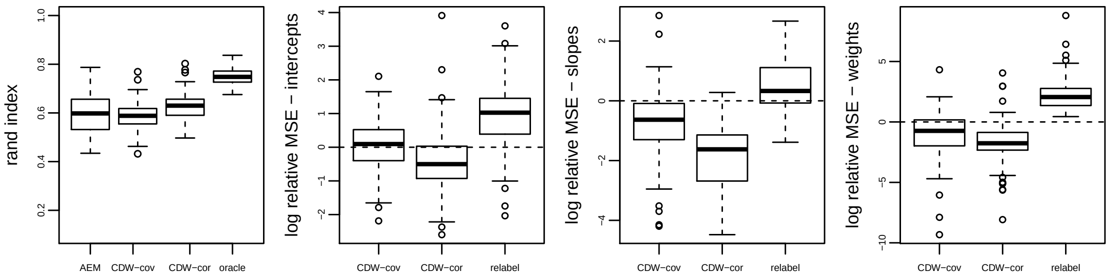

Figure 6: Results from setting B2: Rand index (left) and MSE relative to anchored EM for the intercept parameters (left center), slope parameters (right center), and mixture weights (right).

The estimated parameters in both settings are shown in Table 2. In all methods, although three distinct intercepts are estimated, the average posterior means of the high- and low-intercept lines are under- and over-estimated, respectively. This is expected behavior in a mixture model because uncertainty in classifications will lead to component means being shrunk towards each other.

Table 2: Average posterior means for each model under setting B1 (top) and B2 (bottom). Values are the posterior means for each parameter, averaged over all data sets. Values in parentheses are estimated standard errors.

Setting B1

	Component 1			Component 2			Component 3			
	β_0	β_1	η	β_0	β_1	η	β_0	β_1	η	σ^2
true	-3	0.7	0.3333	-2.5	0.7	0.3333	-2	0.7	0.3333	0.008772
AEM	-2.842 (0.08)	0.711 (0.10)	0.356 (0.07)	-2.501 (0.13)	0.672 (0.15)	0.296 (0.05)	-2.155 (0.08)	0.703 (0.09)	0.348 (0.07)	0.078 (0.01)
CDW-cov	-2.809 (0.08)	0.695 (0.08)	0.359 (0.08)	-2.496 (0.20)	0.692 (0.14)	0.289 (0.04)	-2.194 (0.10)	0.701 (0.08)	0.351 (0.08)	0.085 (0.01)
CDW-cor	-2.869 (0.05)	0.691 (0.05)	0.358 (0.05)	-2.498 (0.10)	0.707 (0.07)	0.289 (0.03)	-2.133 (0.05)	0.696 (0.05)	0.353 (0.05)	0.076 (0.01)
relabel	-2.709 (0.11)	0.685 (0.05)	0.476 (0.15)	-2.502 (0.03)	0.355 (0.05)	0.064 (0.02)	-2.293 (0.11)	0.674 (0.07)	0.459 (0.16)	0.115 (0.02)

Setting B2

	Component 1			Component 2			Component 3			
	β_0	β_1	η	β_0	β_1	η	β_0	β_1	η	σ^2
true	-3	0.7	0.3333	-2.5	0.7	0.3333	-2	0.7	0.3333	0.04167
AEM	-2.873 (0.09)	0.697 (0.13)	0.310 (0.08)	-2.529 (0.16)	0.657 (0.20)	0.346 (0.07)	-2.155 (0.08)	0.688 (0.13)	0.344 (0.09)	0.110 (0.01)
CDW-cov	-2.818 (0.16)	0.671 (0.10)	0.335 (0.07)	-2.485 (0.19)	0.711 (0.15)	0.325 (0.05)	-2.218 (0.15)	0.674 (0.12)	0.340 (0.07)	0.115 (0.02)
CDW-cor	-2.866 (0.10)	0.692 (0.07)	0.331 (0.04)	-2.517 (0.13)	0.671 (0.10)	0.317 (0.04)	-2.156 (0.12)	0.691 (0.07)	0.352 (0.05)	0.108 (0.01)
relabel	-2.662 (0.11)	0.660 (0.07)	0.449 (0.19)	-2.517 (0.05)	0.367 (0.07)	0.074 (0.06)	-2.352 (0.10)	0.671 (0.07)	0.477 (0.19)	0.158 (0.02)

2.4 Results: Setting C

Figure 7 shows the regression lines and sample data sets under Settings C1 and C2. The model is characterized by two parallel lines with a third line intersecting both. The points generated by setting C2 are particularly difficult to distinguish as being generated from different groups.

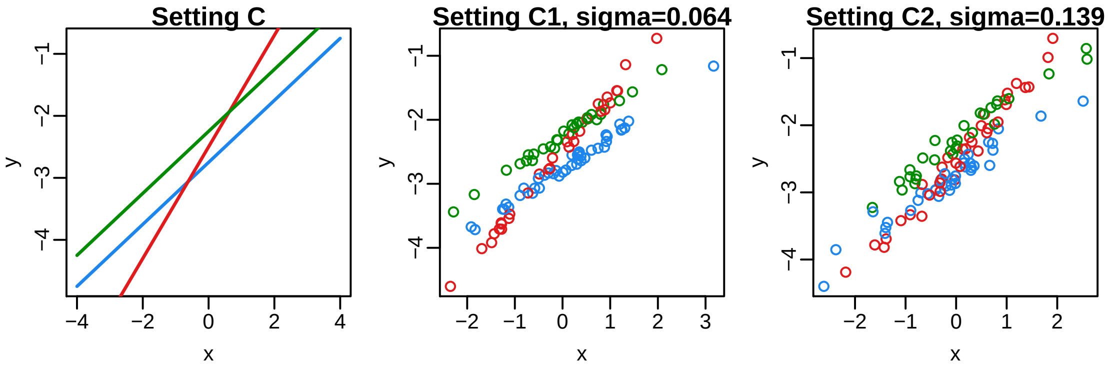

Figure 7: True regression lines and representative simulated data sets for Setting C. Plotting symbols are colored by their true allocation.

The estimation accuracy of the four methods is shown in the right panels of Figures 8 and 9 for settings C1 and C2, respectively. In setting C1, the anchored EM has the strongest accuracy in parameter estimation, while the CDW-cor method has the highest clustering accuracy.

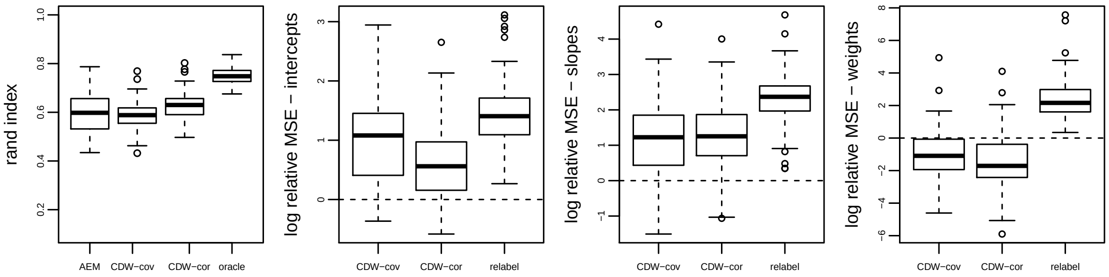

Figure 8: Results from setting C1: Rand index (left) and MSE relative to anchored EM for the intercept parameters (left center), slope parameters (right center), and mixture weights (right).

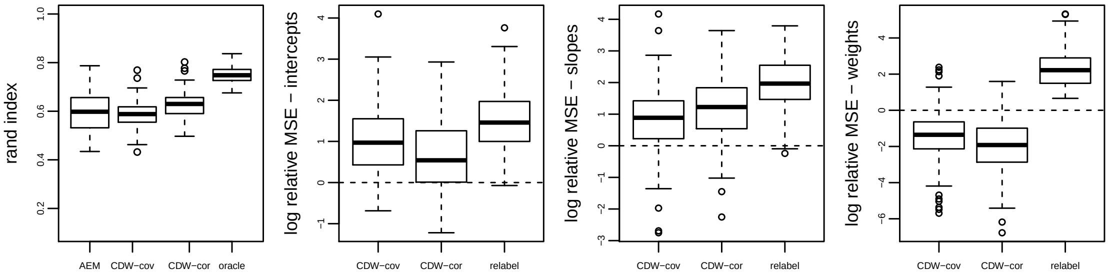

Figure 9: Results from setting C2: Rand index (left) and MSE relative to anchored EM for the intercept parameters (left center), slope parameters (right center), and mixture weights (right).

A very similar pattern is seen in setting C2, as shown in Figure 9. The estimates in Table 3 indicate that there is particularly high error in estimating the steep slope of the intersecting line under the two CDW methods. In addition, both CDW methods have average intercepts that are closer together than the true values. The anchored EM also results in some shrinkage of the intercept estimates towards each other, but to a lesser degree.

Table 3: Average posterior means for each model under setting C1 (top) and C2 (bottom). Values are the posterior means for each parameter, averaged over all data sets. Values in parentheses are estimated standard errors.

Setting C1

	Component 1			Component 2			Component 3			
	β_0	β_1	η	β_0	β_1	η	β_0	β_1	η	σ^2
true	-2.75	0.5	0.3333	-2.5	0.9	0.3333	-2.25	0.5	0.3333	0.004064
AEM	-2.633 (0.04)	0.494 (0.06)	0.281 (0.05)	-2.500 (0.03)	0.825 (0.07)	0.423 (0.09)	-2.360 (0.04)	0.491 (0.06)	0.297 (0.06)	0.054 (0.00)
CDW-cov	-2.561 (0.06)	0.614 (0.16)	0.335 (0.06)	-2.495 (0.05)	0.640 (0.19)	0.330 (0.07)	-2.434 (0.06)	0.618 (0.14)	0.336 (0.05)	0.060 (0.01)
CDW-cor	-2.597 (0.06)	0.574 (0.12)	0.315 (0.04)	-2.498 (0.05)	0.693 (0.15)	0.358 (0.05)	-2.401 (0.05)	0.592 (0.13)	0.327 (0.04)	0.059 (0.01)
relabel	-2.515 (0.03)	0.559 (0.21)	0.404 (0.22)	-2.499 (0.03)	0.388 (0.22)	0.237 (0.26)	-2.476 (0.03)	0.510 (0.19)	0.360 (0.20)	0.069 (0.01)

Setting C2

	Component 1			Component 2			Component 3			
	β_0	β_1	η	β_0	β_1	η	β_0	β_1	η	σ^2
true	-2.75	0.5	0.3333	-2.5	0.9	0.3333	-2.25	0.5	0.3333	0.01931
AEM	-2.642 (0.05)	0.503 (0.13)	0.285 (0.07)	-2.503 (0.04)	0.793 (0.13)	0.404 (0.10)	-2.363 (0.05)	0.512 (0.10)	0.312 (0.07)	0.068 (0.01)
CDW-cov	-2.567 (0.06)	0.623 (0.15)	0.343 (0.05)	-2.502 (0.06)	0.609 (0.20)	0.325 (0.08)	-2.430 (0.06)	0.620 (0.17)	0.332 (0.06)	0.075 (0.01)
CDW-cor	-2.618 (0.05)	0.606 (0.12)	0.318 (0.04)	-2.496 (0.05)	0.662 (0.13)	0.356 (0.04)	-2.392 (0.06)	0.606 (0.13)	0.326 (0.04)	0.074 (0.01)
relabel	-2.527 (0.03)	0.535 (0.20)	0.382 (0.23)	-2.502 (0.03)	0.384 (0.18)	0.201 (0.24)	-2.478 (0.04)	0.548 (0.20)	0.417 (0.24)	0.084 (0.01)

2.5 Setting D: three predictors

In addition, we considered one multiple regression case in which three numeric predictors were used. We generated three numeric predictors, x_1, x_2, x_3 with the following correlation matrix:

$$\begin{bmatrix} 1 & 0.8 & 0.05 \\ 0.8 & 1 & -0.10 \\ 0.05 & -0.10 & 1 \end{bmatrix} \tag{2.15}$$

The regression coefficients for each component were:

$$\boldsymbol{\beta}_1 = (-3.0, 0.7, -1.6, 0.2)'; \;\; \boldsymbol{\beta}_2 = (-2.5, 0.9, -1.6, -0.2)'; \boldsymbol{\beta}_3 = (-2.0, 0.5, -1.6, 0.0)'.$$

The residual standard deviation, σ, was set to be 0.224.

Figure 10 summarizes the performance of each of the methods. For this setting, the EM and CDW methods tend to have similar MSEs for the coefficients corresponding to the numeric predictors. The error is somewhat higher for CDW-cov in estimating the intercept. As in the other settings, we see the highest accuracy in clustering from CDW-cor and the lowest from CDW-cov. The parameter estimates are given in Table 4. All methods typically produce accurate estimates of β_2, which does not differ across the components. The β_1 coefficient, associated with the predictor x_1 which is highly correlated with x_2, is also estimated with accuracy. The intercepts, as in the simpler models, tend to shrink together with the CDW-cov method exhibiting this behavior to the greatest degree.

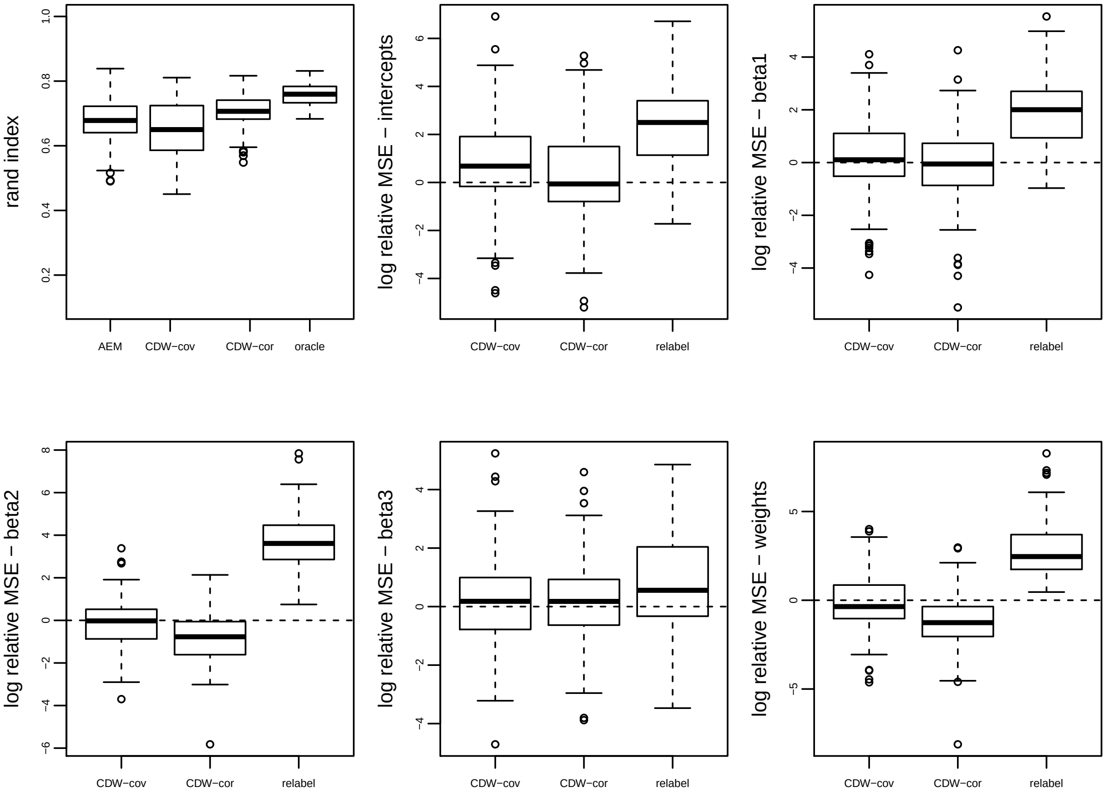

Figure 10: Results from setting D: Rand index (top left) and MSE relative to anchored EM for the regression coefficients and mixture weights.

Table 4: Average posterior means for each model under setting D. Values are the posterior means for each parameter, averaged over all data sets. Values in parentheses are estimated standard errors.

	Setting D															
	Component 1					Component 2					Component 3					
	β_0	β_1	β_2	β_3	η	β_0	β_1	β_2	β_3	η	β_0	β_1	β_2	β_3	η	σ^2
true	-3	0.7	-1.6	0.2	0.3333	-2.5	0.9	-1.6	-0.2	0.3333	-2	0.5	-1.6	0	0.3333	0.05
AEM	-2.86 (0.19)	0.68 (0.18)	-1.56 (0.18)	0.24 (0.10)	0.32 (0.10)	-2.47 (0.19)	0.80 (0.22)	-1.59 (0.17)	-0.14 (0.14)	0.38 (0.10)	-2.18 (0.24)	0.46 (0.24)	-1.55 (0.21)	0.02 (0.14)	0.30 (0.09)	0.09 (0.02)
CDW-cov	-2.76 (0.24)	0.68 (0.21)	-1.60 (0.19)	0.21 (0.11)	0.30 (0.08)	-2.47 (0.19)	0.77 (0.18)	-1.60 (0.14)	-0.12 (0.16)	0.38 (0.09)	-2.27 (0.24)	0.52 (0.24)	-1.56 (0.20)	0.00 (0.12)	0.32 (0.08)	0.11 (0.02)
CDW-cor	-2.83 (0.23)	0.69 (0.17)	-1.58 (0.14)	0.17 (0.12)	0.33 (0.05)	-2.45 (0.23)	0.76 (0.17)	-1.59 (0.11)	-0.10 (0.12)	0.35 (0.05)	-2.22 (0.27)	0.57 (0.18)	-1.59 (0.12)	-0.01 (0.09)	0.32 (0.05)	0.10 (0.02)
relabel	-2.60 (0.25)	0.45 (0.29)	-1.17 (0.50)	0.12 (0.10)	0.33 (0.21)	-2.42 (0.23)	0.63 (0.21)	-1.52 (0.24)	-0.04 (0.11)	0.60 (0.21)	-2.50 (0.07)	0.06 (0.16)	-0.24 (0.36)	0.02 (0.04)	0.06 (0.16)	0.17 (0.05)

3. Sensitivity analysis

In this section, we assess the sensitivity of our results to the number of mixture components, the strength of prior assumptions, and to the number of anchor points.

3.1 Number of mixture components

Here, we show the results of fitting a mixture with $k = 2$ components to the mammals data. The main text discussed that in selecting the number of components, two or three components are preferred. Figure 11 shows the anchor points and estimated regression lines resulting from the two-component model.

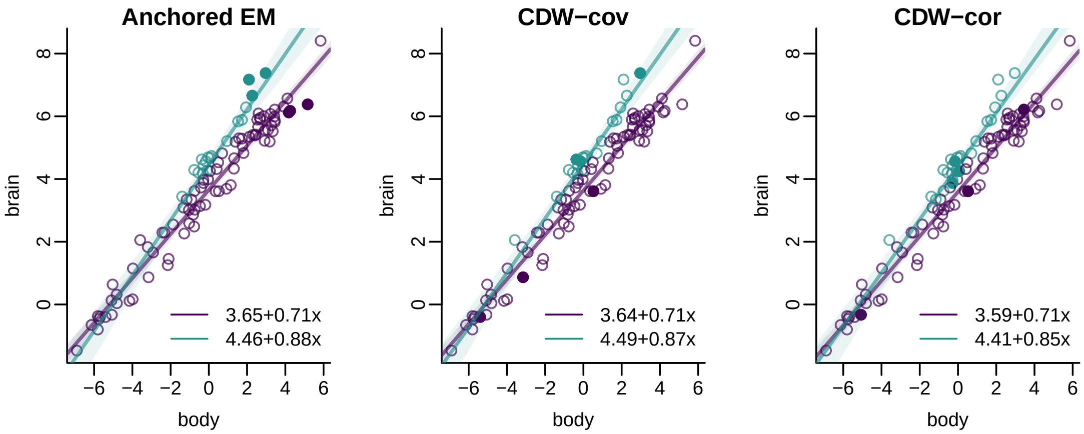

Figure 11: Anchor points and estimated regression lines for a model with two components.

Compared to the 3-component model presented in the main text, the estimated teal component captures the species that constitute a group with a higher slope than the others, including some of the large primates. This sub-group is similar to Component 3 identified in the main text, although, in the two-component fit, the estimated regression line is somewhat less steep and has a smaller intercept than Component 3 in the main text. The second group, shown in purple in Figure 11, has an estimated slope of 0.71 for all anchor models, which is similar to the slopes of Components 1 and 2 in the main text. The estimated intercepts of this purple group range from 3.59 to 3.65, which falls between the intercepts estimated for Components 1 and 2 in the main text. The similarities between the lines estimated in the $k = 2$ and $k = 3$ models indicate that both models are able to partition the species into similar subgroups, and that the 3-component model further refines the grouping by capturing differences in average brain size.

Sensitivity to hyperparameters

The main text presents results under the following prior specification:

$$\beta_j \sim N_2((3.5, 0.6)', \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix}) \tag{3.16}$$

$$\sigma^{-2} \sim \text{Gamma}(\text{shape=5}, \text{rate=1}) \tag{3.17}$$

We will assess the effect of changing these hyperparameters on conclusions from the analysis of the mammals data. Table 5 gives the posterior means and 90% credible intervals under the original hyperparameters for reference.

The anchored EM method includes update steps that depend on the prior hyperparameters. The CDW methods select anchor points based on a preliminary simple linear regression fit, which may also be affected by specification of the hyperparameters. Because of this, in assessing sensitivity to the hyperparameters, we do not hold the anchor points fixed, but re-select them for each case.

Table 5: Posterior means and (90% credible intervals) of the regression coefficients for the mammals data. Estimates are conditional on the hyperparameters used in the main text.

	Component 1		Component 2		Component 3	
	β_0	β_1	β_0	β_1	β_0	β_1
AEM	3.39 (3.18, 3.60)	0.697 (0.65, 0.74)	3.97 (3.72, 4.24)	0.712 (0.64, 0.77)	4.56 (4.35, 4.76)	0.911 (0.83, 1.02)
CDW-cov	3.43 (3.22, 3.65)	0.695 (0.65, 0.74)	4.00 (3.75, 4.26)	0.691 (0.59, 0.76)	4.53 (4.31, 4.73)	0.915 (0.83, 1.02)
CDW-cor	3.47 (3.21, 3.83)	0.694 (0.61, 0.75)	3.83 (3.54, 4.08)	0.724 (0.67, 0.78)	4.52 (4.32, 4.70)	0.891 (0.81, 0.99)

Sensitivity to V. We first consider the sensitivity of the results to the strength of prior information on $\boldsymbol{\beta}$ by considering prior variances of 4 and 2 on the intercepts and slopes, respectively. These variances are larger than those specified in the main text analyses by a factor of 4. The prior means were left unchanged.

Table 6 gives the posterior means and credible intervals for the regression coefficients. The estimates are very similar to those in Table 5 under the A-EM and CDW-cor method. For the CDW-cov method, the estimates for the steepest regression line (labeled Component 3) are similar to their original values. For the two shallower lines (Component 1 and 2), the posterior credible interval of the estimated intercepts overlap to a much greater degree under this less-informative prior.

Table 6: Posterior means and (90% credible intervals) of the regression coefficients with $v_0 = 4$, $v_1 = 2$.

	Component 1		Component 2		Component 3	
	β_0	β_1	β_0	β_1	β_0	β_1
AEM	3.39 (3.16, 3.60)	0.699 (0.66, 0.74)	3.98 (3.73, 4.25)	0.712 (0.64, 0.77)	4.58 (4.37, 4.78)	0.910 (0.82, 1.02)
CDW-cov	3.55 (3.20, 3.96)	0.732 (0.67, 0.81)	3.79 (3.52, 4.07)	0.674 (0.58, 0.74)	4.56 (4.32, 4.77)	0.895 (0.79, 1.03)
CDW-cor	3.43 (3.19, 3.75)	0.696 (0.63, 0.75)	3.87 (3.61, 4.09)	0.726 (0.68, 0.78)	4.55 (4.36, 4.73)	0.897 (0.81, 1.00)

Sensitivity to a, b. We next consider the sensitivity of results to the strength of prior information on σ^{-2} by setting the gamma shape and rate to be $a = 0.5$ and $b = 0.1$, respectively. This specification leaves the prior mean of σ^{-2} unchanged, but increases the prior variance from 5 to 50. Table 7 shows the estimated posterior means under this weaker prior. The slope of Component 2 estimated by CDW-cor is smaller than its estimate under the original hyperparameters and the slope of Component 1 is larger. The credible intervals of these parameters nonetheless overlap with the original estimates.

Table 7: Posterior means and (90% credible intervals) of the regression coefficients with $a = 0.5$, $b = 0.1$.

	Component 1		Component 2		Component 3	
	β_0	β_1	β_0	β_1	β_0	β_1
AEM	3.35 (3.17, 3.55)	0.701 (0.66, 0.74)	3.98 (3.76, 4.21)	0.709 (0.64, 0.77)	4.58 (4.40, 4.75)	0.903 (0.83, 0.99)
CDW-cov	3.40 (3.15, 3.95)	0.687 (0.54, 0.74)	3.87 (3.61, 4.06)	0.723 (0.67, 0.77)	4.58 (4.40, 4.75)	0.889 (0.82, 0.97)
CDW-cor	3.48 (3.19, 3.78)	0.735 (0.67, 0.81)	3.82 (3.44, 4.23)	0.648 (0.49, 0.74)	4.50 (4.29, 4.68)	0.865 (0.79, 0.93)

3.2 Sensitivity to m

While selection of the number of anchor points is not a focus of our study, we considered the estimates resulting from using the proposed anchoring procedures with fewer anchor points. Intuitively, at least two anchor points per component should be needed to pin down the component-specific linear regression. Specification of a single anchor point per component, while sufficient to avoid label switching, may not be enough to provide accurate modeling. The table below displays posterior means for the three anchoring methods with 1 and 2 points per component.

The top panel of Table 8 shows the estimated regression coefficients under the anchor model with one point per component. Compared to the fit with $m = 3$, the credible intervals are much wider, which is a natural consequence of a model with weaker prior information. In these models, all methods have identified a line with a steep slope and large intercept, arbitrarily labeled Component 3, although the estimated slopes for this line are slightly smaller in the CDW models than their original fits. For all methods, the credible intervals for the regression parameters of Components 1 and 2 overlap substantially, particularly those estimated by the CDW-cov and A-EM methods.

With two anchor points per component, the results summarized in the bottom panel of Table 8 show that there is a clearer separation of Component 1 and 2, with the former having a small intercept with credible intervals that do not overlap with those of the intercepts for the other components under the A-EM and CDW-cor methods. As in the $m = 1$ case, the estimates for Component 3 are similar to those obtained with the original analysis.

Table 8: Posterior means and (90% credible intervals) of the regression coefficients with $m = 1$ (top) and $m = 2$ (bottom).

			$m = 1$			
	Component 1		Component 2		Component 3	
	β_0	β_1	β_0	β_1	β_0	β_1
AEM	3.68 (3.22, 4.43)	0.635 (0.41, 0.73)	3.84 (3.52, 4.27)	0.725 (0.66, 0.80)	4.50 (4.22, 4.76)	0.917 (0.80, 1.10)
CDW-cov	3.75 (3.22, 4.53)	0.773 (0.67, 1.01)	3.77 (3.40, 4.22)	0.679 (0.52, 0.75)	4.46 (4.00, 4.75)	0.835 (0.52, 1.00)
CDW-cor	3.56 (3.24, 3.99)	0.711 (0.63, 0.79)	3.99 (3.49, 4.65)	0.739 (0.61, 0.93)	4.40 (3.83, 4.70)	0.852 (0.62, 1.02)
			$m = 2$			
AEM	3.42 (3.17, 3.66)	0.697 (0.65, 0.74)	3.97 (3.69, 4.28)	0.714 (0.64, 0.78)	4.55 (4.31, 4.77)	0.914 (0.82, 1.05)
CDW-cov	3.58 (3.20, 4.22)	0.660 (0.46, 0.75)	3.80 (3.53, 4.08)	0.724 (0.67, 0.77)	4.53 (4.29, 4.75)	0.899 (0.80, 1.02)
CDW-cor	3.48 (3.24, 3.70)	0.711 (0.67, 0.76)	4.05 (3.61, 4.61)	0.735 (0.58, 0.93)	4.37 (3.91, 4.64)	0.820 (0.63, 0.94)

4. Model-based clustering and known taxonomy

The estimated component assignments $\widehat{\boldsymbol{s}}$ give a model-based grouping of the species which ignores the additional data on the species' taxonomic orders and suborders. A comparison of these groups with the true taxonomy of the species can shed light on the allometric questions posed at the beginning of this article: the species assigned to the same component by $\widehat{\boldsymbol{s}}$ have similar estimated regression slopes, and if there is a correspondence between $\widehat{\boldsymbol{s}}$ and the species' true orders, it may indicate that certain taxonomic groups have distinct body mass/brain mass relationships. Figure 12 displays the data from species in three of the 13 orders, color-coded according to the model-based cluster assignment. The displayed orders are Primates, Artiodactyla, and Rodentia, which account for 21, 21, and 24 of the 100 species in the data, respectively.

The first column of the figure shows the Primate species, which, for all three models, are split between Components 2 and 3. For a given body mass, most larger-brained species are assigned to Component 3 and most of those with smaller brains are assigned to Component 2. Component 3 also contains the three species of the Cetacea order (not pictured) under all three model fits. Two primate species are assigned to Component 3 by the CDW-cor method and to Component 2 by the other two models: Hapale Leucocephala and Macaca Maurus, both of the sub-order Anthropoidea. The Hapale Leucocephala is the primate with the smallest body, but this species' brain is actually large given its body mass. Component 2 also contains all three primate species of the Prosimii sub-order. So, the mixture model is sensitive to this aspect of the taxonomic classification, recognizing that the Prosimii species have small brains given their body masses.

The clustering among the Rodentia differs most across the three anchor models.

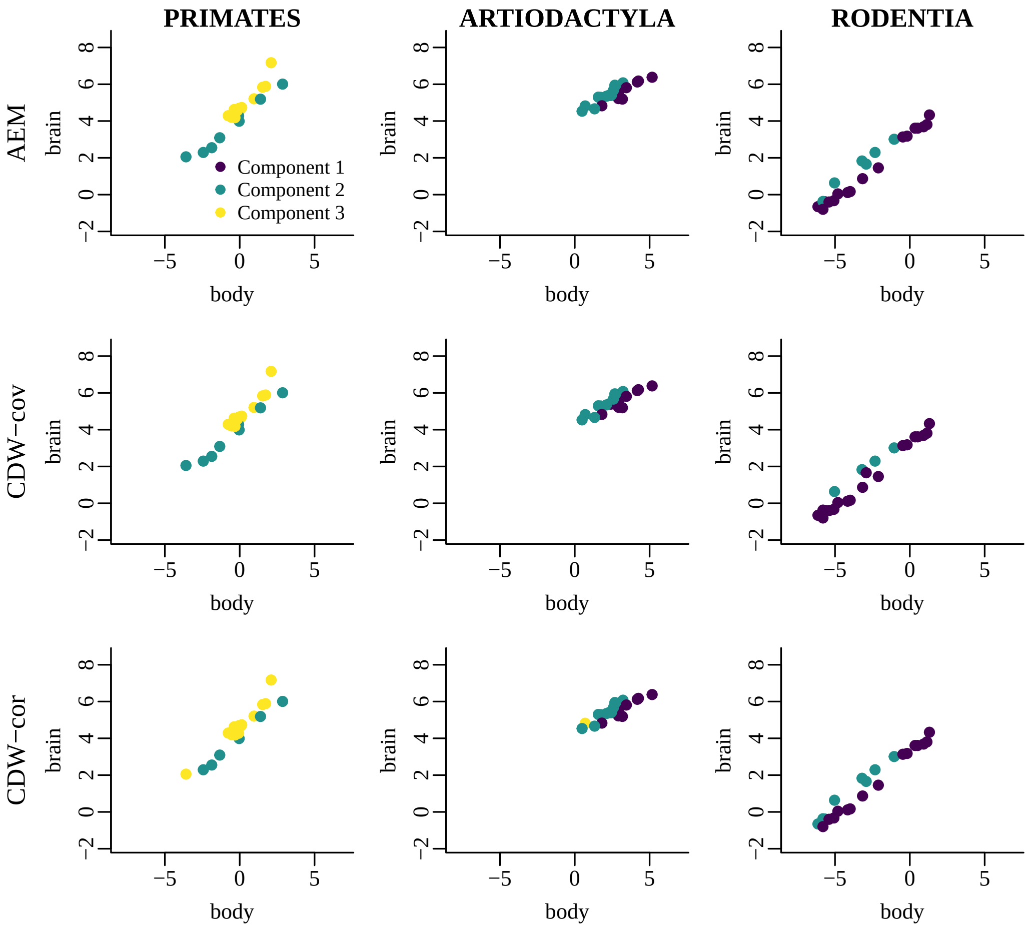

Figure 12: Data for the species in orders Primates, Artiodactyla, and Rodentia color-coded according to their model-based $\widehat{\boldsymbol{s}}$. The rows correspond to the three anchor models. The color coding used to distinguish the three mixture components matches that used in Figure 4 of the main article.

The CDW-cov model assigns only four of 24 Rodentia to Component 2, while the CDW-cor method assigns 9. All of the largest-bodied Rodentia species, seen as the far right points colored in blue, are assigned to Component 1 by all of the models. Interestingly, one species that falls next to these points in the plot, the Myocastor Coypus, is assigned to Component 2 by all models, in spite of its similar body size to the neighboring points. This indicates a sensitivity of all models to its slight decrease in brain size compared to the adjacent species, whose body sizes are very similar.

Acknowledgments

This material is based on work supported by the National Science Foundation under grant no. SES-1424481 and SES-1921523.

References

Papastamoulis, P. (2016). label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software, Code Snippets 69*, 1–24.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 62*, 795–809.