Statistical inference with anchored Bayesian mixture of regressions models:

A case study analysis of allometric data

Deborah Kunkel and Mario Peruggia

1. School of Mathematical & Statistical Sciences, Clemson University, Clemson, SC, USA
2. Department of Statistics, The Ohio State University, Columbus, OH, USA

Abstract

We present a case study in which we use a mixture of regressions model to improve on an ill-fitting simple linear regression model relating log brain mass to log body mass for 100 placental mammalian species. The slope of this regression model is of particular scientific interest because it corresponds to a constant that governs a hypothesized allometric power law that relates brain mass to body mass. A specific line of investigation is to determine whether the regression intercept and slope may vary across subgroups of related species.

We model these data using an anchored Bayesian mixture of regressions model, which modifies the specification of a standard Bayesian Gaussian mixture by pre-assigning small subsets of observations to given mixture components with probability one. These pre-classified observations (called anchor points) break the relabeling invariance typical of exchangeable model specifications (the so-called label-switching problem) that often prevents direct interpretation of component-specific parameters in mixture models. Anchoring can also be seen as a method for specifying weakly data-dependent prior distributions on the component-specific parameters. A careful choice of which observations to pre-classify to which mixture components is key to the specification of a well-fitting anchor model.

In the article we compare three distinct broad strategies for the selection of anchor points. The first strategy assumes that the underlying mixture of regressions model holds and assigns anchor points to different components so as to maximize the information about their labeling. The second strategy makes no assumption about the relationship between \( x \) and \( y \) and instead identifies anchor points using a bivariate
Gaussian mixture model that does not condition on body mass. The third strategy begins with the assumption that there is only one mixture regression component and identifies a clustering structure based on case-deletion importance sampling weights. The anchor points are then selected as typical representatives of the identified clusters. We compare the performance of the three strategies on the allometric data set and use auxiliary taxonomic information about the species to evaluate the model-based classifications that we can estimate from these models.

**Keywords**: Allometry, Case-deletion weights, Clustering, EM algorithm, Semi-supervised learning

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

1 **Introduction**

In the natural sciences, allometry studies the relationships between physical and physiological measurements taken on various animal species (Peters, 1983; Gayon, 2000). Of particular interest is to determine how other measurements may be affected by body mass. Examples include the relationships between body mass and brain mass, body mass and metabolic rate, body mass and gestation duration. It is often postulated that pairs \((x, y)\) of such measurements may be related via a power law of the form \(y = cx^b\), for some unknown constants \(c\) and \(b\), typically assumed to be positive. The estimation of the exponent \(b\) is often of primary scientific interest. On a logarithmic scale, the power law turns into the linear relationship \(\log y = (\log c) + b\log x\) and the investigative focus shifts toward the estimation of the slope of the regression line.

Given a set of \((x, y)\) pairs of traits measured on a variety of species, it is by now generally accepted that fitting a single linear regression model to the entire data set provides too crude a summary, especially when many species from different taxa and genetically diverse groups
are included in the data set (Jerison, 1955; Bennett and Harvey, 1985a,b). More refined approaches rely on the incorporation of evolutionary information (possibly inferred from a taxonomy) to perform an analysis based on models for derived quantities that can be treated as independent, rather than for the original measured traits that exhibit species-related dependencies. For example, this is the case for a popular type of analysis based on phylogenetically independent contrasts (Felsenstein, 1985; Garland Jr et al., 1992). MacEachern and Peruggia (2002) show that traditional Bayesian variance components models applied directly to allometric data for which taxonomic information is available can produce a good fit and yield easily interpretable inferences.

In this article we reanalyze the data that MacEachern and Peruggia (2002) used to illustrate their methods. The data comprise the body and brain mass measurements on 100 species of placental mammals originally reported by Sacher and Staffeldt (1974) as well as a taxonomy that assigns each species to an order and sub-order based on its morphological and physiological traits. For example, the African lion (Panthera Leo) is listed as belonging to the order Carnivora and sub-order Fissipeda and the chimpanzee (Pan Troglodytes) is listed as belonging to order Primates and sub-order Anthropoidea. In total, the data contain species that represent 13 orders and 19 sub-orders.

The data are shown in Figure 1. The left panel displays a scatterplot of the centered (i.e., mean adjusted) log body mass and log brain mass for the 100 mammalian species. The overlaid line is the least-squares fit from a naive simple linear regression model. The residuals for this model are shown in the right panel of Figure 1. This plot raises some concerns about model fit. First, there is a slight increase in residual variability as log body mass grows. Second, other features of the residuals can be traced back to the species orders (distinguished by plotting color): all Primates (red points) have positive residuals, while most Rodentia (blue points) have negative residuals. This structure points to the fact that the least squares line does not properly account for within-order similarities in the allometric relationship.

MacEachern and Peruggia (2002) present a detailed evaluation that uncovers the lack-of-
fit of this model, introducing Bayesian model diagnostic techniques based on case-deletion importance sampling weights (Geweke, 1989; Bradlow and Zaslavsky, 1997; Peruggia, 1997; Epifani et al., 2008; Thomas et al., 2018). They also present a Bayesian variance components model that includes additive random effects for orders and sub-orders. These random effects induce positive correlations between the residuals of species belonging to the same taxonomic groups and subgroups. The random effect adjustment effectively creates a separate regression line with its own intercept for the species within each subgroup. MacEachern and Peruggia (2002) show that this way of accounting for the taxonomic information significantly ameliorates the quality of the fit.

Suppose now that no information about the taxonomy were available. We might wish to remedy the lack-of-fit of the simple linear regression model, but a random effects model is not an option when the true group memberships are not known. In such situations, a reasonable modeling alternative is to assume that there exist finitely many subgroups of observations for which separate regression lines are appropriate, and yet it is unknown which observations to associate with which regression. This leads naturally to the formulation

Figure 1: Mammals data (left) with the estimated least-squares regression line and residuals (right) from the least-squares regression fit.
of a mixture of regressions model in which, conditional on unobserved group membership indicators, observations falling in separate groups follow separate regression models. Using these mammalian species data as the foundation for a case study, we consider the specification of a Bayesian mixture of regressions to account for unobserved heterogeneity in a data set. The fundamental difference between the mixture setting and the random effects setting considered by MacEachern and Peruggia (2002) is that, in the former, group memberships are latent and are the subject of inferential investigation. An additional difference is that we will allow for group-specific slopes in addition to group-specific intercepts.

A unique challenge arises when modeling with Bayesian mixture models: a typical Bayesian formulation would specify, a priori, a fully exchangeable mixture, yielding a fully exchangeable posterior. A fully exchangeable model is indifferent to any arbitrary relabeling of the mixture components implying that, a posteriori, each observation is equally likely to follow any of the postulated regression models. This phenomenon, often referred to as label switching, precludes any informative assignment of labels to the mixture components and prevents meaningful interpretation of the corresponding component specific parameter estimates. Kunkel and Peruggia (2018) introduce a modeling device called anchoring that breaks the labeling invariance of the exchangeable version of a Bayesian Gaussian mixture model. The idea is to identify small subsets of representative observations, called the anchor points, and allocate them to separate homogeneous groups before performing the analysis. Kunkel and Peruggia (2018) characterize the equivalence between this process and the specification of a weakly data-dependent prior for the component specific parameters. A shrewd selection of anchor points is essential for the successful implementation of the strategy. They propose two optimality criteria, one based on minimizing the entropy of the asymptotic distribution of class labeling and one based on a modified EM algorithm.

Motivated by the mixture of regressions model and its application to our allometric data, we extend the work of Kunkel and Peruggia (2018) in two main directions. First, we generalize the EM anchoring strategy from the case of multivariate Gaussian mixtures to the case of mixture of regressions models. Second, we introduce a new strategy for the selection
of anchor points that builds on the work on case-deletion analysis presented in MacEachern and Peruggia (2002) and Thomas et al. (2018). We compare this new strategy for the selection of anchor points (which we call CDW-reg) to the two EM anchoring strategies based on the bivariate Gaussian mixture model (which we call EM-MVN) and based on the mixture of regressions model (which we call EM-reg). The three strategies differ with respect to the model that they subsume (no regression structure, a single regression line, or several regression lines) and how they extract information from the data to determine which points to select as representative of the various mixture components (a principal components analysis of the log case-deletion weights or the EM algorithm).

In Section 2 we present the mixture of regressions model and its corresponding anchor model. In Section 3 we present the EM method for selecting anchor points under the mixture of regressions and multivariate Gaussian mixture models. In Section 4 we describe the CDW-reg method for choosing anchor points via clustering of case-deletion weights. The results of our analysis of the mammals data are presented in Section 5.

2 Background

In this section we give a formal definition of the mixture of regressions model and its anchored version that we will employ to analyze the mammals data.

2.1 Mixture of regressions model

The k-component mixture of linear regressions model specifies that the observations \((y_i, x_i), i = 1, \ldots, n,\) are drawn from \(k\) homogeneous subgroups within the population. The response \(y_i\) is a scalar and \(x_i \in \mathbb{R}^p\) is a row vector of predictors whose first element is equal to one. The data are denoted by \(y' = (y_1, \ldots, y_n)\) and by the \(n \times p\) matrix \(X\) with \(i\)-th row equal to \(x_i\). The mixture likelihood is

\[
f(y|X, \beta, \sigma^2, \eta) = \prod_{i=1}^{n} \sum_{j=1}^{k} \eta_j \phi(y_i; x_i \beta_j, \sigma^2),
\]
where \( \eta' = (\eta_1, \ldots, \eta_k) \) is a vector of mixture probabilities that satisfy \( \sum_{j=1}^k \eta_j = 1 \) and \( \phi(\cdot; a, b) \) denotes the density function of a normal distribution with mean \( a \) and variance \( b \), evaluated at its argument. The component-specific parameters in this model are \( \beta = (\beta_1, \ldots, \beta_k) \), where \( \beta_j \in \mathbb{R}^p \) is the regression coefficients associated with the \( j \)th component. This model can be written equivalently by introducing latent allocations \( s = (s_1, \ldots, s_n) \), where \( s_i = j \) indicates that observation \( i \) was generated from component \( j \). Conditional on \( s_i = j, i = 1, \ldots, n \), the mean of \( y_i \) is \( x_i \beta_j \) and the model is a random effects regression.

Introducing a probability distribution on \( s \) in which \( P(S_i = j) = \eta_j \), for \( j = 1, \ldots, k \), produces the mixture of regressions model in (1).

We specify the following exchangeable prior:

\[
\beta_j \mid \mu_\beta, V \sim N_p(\mu_\beta, V), \quad j = 1, \ldots, k,
\]

\[
\sigma^{-2} \mid a, b \sim \text{Gamma}(a, b),
\]

\[
\eta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k), \tag{2}
\]

where \( V \) is a diagonal matrix whose \( p \)th diagonal element is \( v_p \) and the Gamma distribution is parameterized to have mean \( a/b \). This type of exchangeable specification is a natural way of expressing prior ignorance about the relation between \( y \) and \( x \) within each group: we make no assumptions that induce different prior distributions on the \( \beta_j \). An unfortunate consequence of the exchangeable specification is that the posterior density of \( \beta \) is invariant to relabeling: for any permutation of the integers \( 1, \ldots, k \), denoted by \( \rho_q(1:k) \), \( p(\beta|y) \) is equal to \( p(\beta_{\rho_q(1:k)}|y) \) for any value of \( \beta \). In other words, the posterior density of a value of \( \beta \) is unchanged if the component labels are permuted. This produces \( k! \) symmetric regions in the posterior distribution of \( \beta \), each corresponding to one of the \( k! \) possible labelings of the mixture components, and the marginal distributions of the component specific parameters will all be identical. This is undesirable for several reasons (Jasra et al., 2005). Inferentially, it is impossible to use estimates of the labeled parameters \( \beta_1, \ldots, \beta_k \) to infer distinctions among features of different groups. Computationally, the label switching phenomenon hampers fitting the model by Markov chain Monte Carlo simulation.
In previous work, (Kunkel and Peruggia (2018)) we have proposed a new class of models call anchor models which pre-classify some observations in order to induce a non-exchangeable, data-dependent prior which can alleviate the inferential nuisances caused by this symmetry in the mixture model.

2.2 Anchor models

To specify an anchor model, we select a small number of points to be assigned a priori to particular components of the mixture model. The model is defined using $k$ index sets $A_1, \ldots, A_k$, where $A_j$ contains the indices of the observations to be pre-labeled (or “anchored”) to component $j$. The number of points in $A_j$, denoted by $m_j$, is chosen ahead of time and each observation is anchored to at most one component; i.e., $A_j \cap A_{j'} = \emptyset$ for $j \neq j'$. Given the anchor points, $A = \cup_{j=1}^k A_j$, the anchored version of the model (1) has likelihood

$$f_A(y|X, \beta, \sigma^2, \eta) = \prod_{i=1}^n f_A(y_i|x_i, \beta, \sigma^2, \eta),$$

with

$$f_A(y_i|x_i, \beta, \sigma^2, \eta) = \begin{cases} 
\sum_{j=1}^k \eta_j \phi(y_i;x_i;\beta_j;\sigma^2), & i \notin A, \\
\phi(y_i;x_i;\beta_j;\sigma^2), & i \in A_j, \ j = 1, \ldots, k.
\end{cases} \quad (3)$$

Similarly, if the latent allocation variables are used, we have the equivalent representation

$$f_A(y|X, \beta, \sigma^2, \eta) = \prod_{i=1}^n \sum_{s_i=1}^k P_A(S_i = s_i)f(y_i|s_i, x_i, \beta, \sigma^2, \eta),$$

where $f(y_i|s_i, x_i, \beta, \sigma^2, \eta) = \phi(y_i;x_i;\beta_{s_i};\sigma^2)$, $P_A(S_i = j) = \eta_j I(i \notin A) + I(i \in A_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, k$, and $I(\cdot)$ is an indicator function that equals 1 when its argument is true and zero otherwise.

An anchored mixture model can be regarded as a hybrid between a random effects model, in which the class membership is known for all observations, and a pure mixture model, in which the class membership is unknown for all observations. It can be shown that the pre-labeled anchor points in the sets $A_1, \ldots, A_k$ eliminate the prior exchangeability of the model that leads to label-switching if at least $k - 1$ components have at least one anchor point each. Kunkel and Peruggia (2018) have shown through a simulation study that the price to be paid for an erroneous attribution of class membership can be large in terms of out-of-sample predictive performance. Their general recommendation is then to confine the anchor points
to small subsets of optimally chosen observations that can well-characterize the underlying
groups and are sufficiently informative to effectively eliminate the labeling invariance. The
anchor model induces a data-dependent prior, informed only by the anchor points, and so it is
crucial to select anchor points that provide useful information about the component-specific
parameters. Kunkel and Peruggia (2018) propose two strategies for specifying anchor models,
one based on minimizing the entropy of the asymptotic distribution of class labeling and one
based on a modified EM algorithm. The method based on the modified EM algorithm is the
basis for two of the models used in our case study.

3 Anchoring with the EM algorithm

Kunkel and Peruggia (2018) describe the anchored expectation-maximization (EM) algo-
rithm, a modified version of the EM algorithm for maximum a posteriori estimation of
anchored mixture models. For a generic vector $\gamma = (\gamma_1, \ldots, \gamma_k)$ of the component-specific
model parameters, the standard EM algorithm for mixture models uses the latent variable
representation of the model and iterates between an E-step, which estimates a probability
distribution $p(s)$ on the latent allocations conditional on the current estimate of $\gamma$, and an
M-step, which updates $\gamma$ to maximize the expected (with respect to $p(s)$) joint posterior
density of $\gamma$ and $s$. Instead, the E-step of the modified algorithm estimates a probability
distribution $q(s)$ that corresponds to an anchor model for some choice of $A = \bigcup_{j=1}^{k} A_j$, that
is, a distribution for which each $A_j$ specifies $m_j$ observations allocated to component $j$ with
probability 1. The sets $A_j$ are chosen to make $q(\cdot)$ as close as possible to the unanchored dis-
tribution $p(s)$. This method seeks to estimate an anchoring structure that produces a close
approximation to the unanchored posterior distribution of $\gamma$ near one of its local modes. The
steps of this algorithm are outlined below for a generic parameter $\gamma$ and mixture probability
distribution $p(y_i|\gamma) = \sum_{j=1}^{k} \eta_j f(y_i|\gamma_j)$. 
### 3.1 General anchored EM

**0:** Initialize $\gamma^0, \eta^0$ and repeat the following steps until convergence.

**E step:** Calculate the following distribution on the latent allocations: $p(s|\gamma^{t-1}, \eta^{t-1}) = \prod_{i=1}^{n} p(s_i|\gamma^{t-1}, \eta^{t-1})$, where

$$p(s_i = j|\gamma^{t-1}, \eta^{t-1}) = r_{ij}^t = \frac{\eta_{j}^{t-1} f(y_i|\gamma_j^{t-1})}{\sum_{l=1}^{k} \eta_l^{t-1} f(y_i|\gamma_l^{t-1})}. \quad (4)$$

**Anchor step:** Find $A^t = \bigcup_{j=1}^{k} A^t_j$ to maximize $\sum_{j=1}^{k} \sum_{i \in A_j} r_{ij}^t$ subject to $A_j \cap A_{j'} = \emptyset$ for all $j \neq j'$. Calculate the distribution $q^t(s|\gamma^{t-1}, \eta^{t-1}) = \prod_{i=1}^{n} q^t(s_i|\gamma^{t-1}, \eta^{t-1})$, where

$$q^t(s_i = j|\gamma^{t-1}, \eta^{t-1}) = \tilde{r}_{ij}^t = \begin{cases} r_{ij}^t, & i \notin A^t, \\ 1, & i \in A^t_j, \\ 0, & i \in A^t_{j'}, j \neq j'. \end{cases} \quad (5)$$

**M step:** Update $\gamma^t$ as $\arg \max_{\gamma} E_{q^t} \log(p(\gamma, \eta, s|y))$, where the expectation is taken with respect to the distribution $q^t$ defined in the Anchor step.

The Anchor step in the algorithm above amounts to assigning to component $j$ the points with the highest posterior probability of allocation to component $j$ given the current estimate of the model parameters, as long as this does not anchor any observations to more than one component. If that happens, an approximate solution may be used or linear programming algorithms can produce an exact solution if necessary. We have found the algorithm to be sensitive to the selection of a starting point and prone to visit local maxima, especially when modeling groups that are not well-separated. We thus recommend running the algorithm several times and selecting the solution with the largest ending value of $E_{q^t} \log(p(\gamma, \eta, s|y))$.

This algorithm is applied to a specific model by substituting the appropriate $j$-th component density for $f(y_i|\gamma_j^{t-1})$ in Equation (4) of the E step above and deriving appropriate updates in the M step. The M step updates the values of the component-specific param-
eters to maximize $\sum_{i=1}^{n} \sum_{j=1}^{k} \tilde{r}_{ij} \log(f(y|\gamma, s)) + \log(p(\gamma)) + \log(p(\eta))$, analogously to the standard EM algorithm for maximum a posteriori estimation. In the next subsections we outline the steps of this algorithm for two mixture models that we will use in our analysis of the mammals data.

### 3.2 EM-reg

The EM-Reg anchoring method applies the anchored EM algorithm to the mixture of regressions model in (3). The E step requires updating the $r_{ij}^t$ as

$$ r_{ij}^t = \frac{\eta_{ij}^{t-1} \phi(y_i; x_i \beta_j^{t-1}, (\sigma^2_j)^{t-1})}{\sum_{i=1}^{k} \eta_{ij}^{t-1} \phi(y_i; x_i \beta_j^{t-1}, (\sigma^2_j)^{t-1})}, $$

and the M step updates of $\beta^t$, $\sigma^t$, $\eta^t$ are

$$ \beta_j^t = (X'R_j^t X + V^{-1})^{-1}(X'R_j^t y + V^{-1} \mu_\beta), \quad j = 1, \ldots, k, $$

$$ (\sigma^2_j)^t = \frac{a + n/2 - 1}{b + .5 \sum_{j=1}^{k} (y - X \beta_j^t)' R_j^t (y - X \beta_j^t)}, $$

$$ \eta_j^t = \frac{\sum_{i=1}^{n} r_{ij}^t + \alpha - 1}{\sum_{i=1}^{n} \sum_{i=1}^{n} r_{il}^t + \alpha - 1}, \quad j = 1, \ldots, k, $$

where $R_j^t$ is an $n \times n$ diagonal matrix whose $i$-th diagonal element is $\tilde{r}_{ij}^t$.

### 3.3 EM-MVN

We now introduce a second anchored EM scheme, EM-MVN, which we will apply to data of the type described in Section 2.1. The EM-MVN scheme assumes, in the anchoring stage, that the vectors $z_i = (y_i, x_{2i}, \ldots, x_{pi})'$ arise from a mixture of multivariate normal distributions whose $j$-th component has mean $\theta_j \in \mathbb{R}^p$ and covariance matrix $\Sigma_j$. The elements of $z_i$ are the response $y_i$ and the $p - 1$ predictors $x_{2i}, \ldots, x_{pi}$ excluding the first element of $x_i$ equal to one. The distribution of $z_i$ under the $p$-variate Gaussian mixture model
is \[ \sum_{j=1}^{k} \eta_j \phi_p(z_i; \theta_j, \Sigma_j), \]
where \( \phi_p(z_i; \theta_j, \Sigma_j) \) denotes the \( p \)-variate Gaussian density function with mean \( \theta_j \) and covariance \( \Sigma_j \), evaluated at its argument.

The M step is derived assuming a conjugate normal-Wishart prior on \( (\theta_j, \Sigma_j) \), in which
\( \theta_j | \mu, \kappa, \Sigma_j \sim N_p(\mu, \kappa^{-1} \Sigma_j) \) and \( \Sigma_j^{-1} | \nu, W \sim \text{Wishart}_p(\nu, W) \), a Wishart distribution with prior scale \( W \) and degrees of freedom \( \nu \). The updates for \( \theta_j \) and \( \Sigma_j \) are given by:

\[
\theta^t_j = \frac{\kappa^{-1} \mu + \sum_i r^t_{ij} z_i}{\kappa^{-1} + \sum_{i=1}^n r^t_{ij}}, \quad j = 1, \ldots, k, \tag{10}
\]

\[
\Sigma^t_j = \frac{W^{-1} + \sum_i r^t_{ij}(z_i - \theta^t_j)(z_i - \theta^t_j)' + \kappa^{-1}(\mu - \theta^t_j)(\mu - \theta^t_j)'}{\nu - p + \sum_{i=1}^n r^t_{ij}}, \quad j = 1, \ldots, k. \tag{11}
\]

The update for \( \eta_j \) is unchanged from (9).

Multivariate Gaussian mixture models are frequently used for classification and are a natural default choice for selecting anchor points. In contrast to the EM-reg method described in the previous subsection, this selection method does not condition on \( X \), nor does it assume that an underlying mixture of regressions holds. In practice, the two methods select very different anchor points. This owes to the fact that the selection is determined by the allocation probabilities \( r_{ij} \), which, under the MVN model, will be large for observations near the \( p \)-variate mean \( \theta_j \) of component \( j \). For the EM-reg method, however, the magnitude of \( r_{ij} \) is driven by the observation’s proximity to the regression line determined by both \( \beta_j \) and the \( x \)-values. This fundamental difference noticeably affects the resulting inferences, as we will see in the sequel.

### 3.4 EM anchoring on the mammals data

For the analysis of the mammals data presented in the remainder of the paper, we used the two EM methods outlined in the previous sections to specify anchor models with \( k = 3 \) mixture components and \( m = 3 \) anchor points in each component. The choice of this number of components is motivated by a desire to restrain model complexity while retaining sufficient flexibility to capture salient morphological differences between groups of species. The choice of this number of anchor points per component aims at imposing enough data-
dependent prior information so as to control the label switching problem. We set the prior hyperparameters in (2) for EM-reg to be $a = 5$, $b = 1$, $\mu_\beta = (3.5, 0.6)^T$, and $v_0 = 1$, $v_1 = 0.5$. For EM-MVN we set $\mu = \bar{z}$, the sample mean vector of the $z_i$, $\kappa = 1.25$, $W = 1.5I_2$, where $I_2$ denotes the $2 \times 2$ identity matrix, and $\nu = 2$. The left and center panels of Figure 2 show the anchor points selected by EM-reg and EM-MVN, respectively.

The EM-reg method chooses points that provide prior information about the distinct linear relationships for each component. The anchor points for the green and blue groups identify approximately parallel lines that differ only in intercept. The red points suggest a line with a steeper slope than that of the other two groups. None of the EM-reg anchor points have extreme $x$ or $y$ values while they fall almost exactly in straight-line patterns. This is in contrast to the choices operated by method CDW-reg (described in Section 4) which selected more influential points to be among the anchors, as shown in the right panel of Figure 2.
The middle panel of Figure 2 shows the EM-MVN anchor points. For these data, the EM-MVN method selects tight clusters of anchor points that are close to each other in both the $x$ and $y$ directions. The selected anchor points do not suggest any distinct features among the regression lines in our target model because they show almost no variability in the $x$ direction; this is a consequence of using a selection method that does not condition on $x$. It is also interesting to observe that this method, favoring anchor points that fall on the extreme ends of the observed data, ends up selecting some of the same high leverage points identified by CDW-reg.

4 Anchoring with case deletion weights

In a Bayesian context, case-deletion analysis quantifies the influence of an individual observation on the overall analysis by comparing the posterior distribution conditional on the entire data set and the posterior distribution conditional on the reduced data set obtained by omitting the observation under consideration. Case-deletion analysis is an effective tool for identifying influential observations in Bayesian models (Bradlow and Zaslavsky, 1997; MacEachern and Peruggia, 2002). It also provides an avenue to assess the similarity of the impact of observations on the inferential conclusions and, as a byproduct, a means to cluster the observations on the basis of a similarity measure derived from the specified model. We now derive a strategy for selecting the anchor points of a Bayesian Gaussian mixture model as “typical” representatives of the clusters identified via a preliminary case-deletion analysis based on a model which assume the existence of a single mixture component.

Consider a set of observations $\mathbf{y} = (y_1, \ldots, y_n)$ following a model with density $f(\mathbf{y}|\theta)$ conditional on a set of parameters $\theta$ having prior distribution $\pi(\theta)$. The posterior distribution of $\theta$ given $\mathbf{y}$ is proportional to the joint distribution of $\mathbf{y}$ and $\theta$ which is $q(\mathbf{y}, \theta) = f(\mathbf{y}|\theta)\pi(\theta)$. Similarly, denoting by $\mathbf{y}_{-i}$ the reduced data set obtained by deleting observation $i$, the posterior distribution of $\theta$ given $\mathbf{y}_{-i}$ is proportional to $q_{-i}(\mathbf{y}_{-i}, \theta) = f(\mathbf{y}_{-i}|\theta)\pi(\theta)$.

The ratio between the case-deleted and full posterior reacts to the influence of the deleted case on the inferential conclusions. For this reason it is useful to understand the behavior of
the random variables

\[ w_i(\theta) = \frac{q_i(y_i, \theta)}{q(y, \theta)}, \quad i = 1, \ldots, n, \] (12)

when \( \theta \) follows the posterior distribution conditional of the entire data set. In practice, we can compute the normalized empirical versions of the ratios in (12) using a sample \( \theta_1, \ldots, \theta_M \) from (approximately) the posterior distribution of \( \theta \) given \( y \). These quantities, known as case-deletion importance sampling weights, are given by

\[ \bar{w}_i(\theta_m) = \frac{w_i(\theta_m)}{\sum_{\ell=1}^{M} w_i(\theta_\ell)}, \quad i = 1, \ldots, n; \quad m = 1, \ldots, M. \]

An important practical use of the normalized case-deletion weights is to reweight empirical averages based on samples from the full posterior to approximate expectations with respect to the case-deleted posterior (Geweke, 1989). The behavior of the resulting weighted estimator is affected by the theoretical properties of the unnormalized weights in (12). For example, a fundamental result in Geweke (1989) requires that the unnormalized weights have finite variance with respect to the full posterior for a central limit theorem for the weighted estimator to hold.

In addition, the theoretical variability of the \( w_i(\theta) \) and the sample variability of the \( w_i(\theta_m) \) and \( \bar{w}_i(\theta_m) \), \( m = 1, \ldots, M \), are indicators of the influence of observation \( i \) on the posterior distribution, with higher variability indicating larger influence Bradlow and Zaslavsky (1997); Peruggia (1997); Epifani et al. (2008). The covariance matrix (with respect to the full posterior distribution of \( \theta \)) of the log weights, \( C = [C_{ij}] = [\text{Cov}(\log w_i(\theta), \log w_j(\theta))] \), is a particularly useful quantity: \( C_{ij} \) can be interpreted as summarizing the degree of similarity between the influence of deletion of case \( i \) and case \( j \). Further, for models of conditional independence, Thomas et al. (2018) detail the existence of a close relationship between \( C \) and measures of influence based on infinitesimal geometric perturbations of the multiplicative contributions of each observation to the overall likelihood. In the perturbed likelihood, the multiplicative factor corresponding to each observation is raised to a power \( \omega_i, i = 1 \ldots, n \). The original likelihood is recovered by setting \( \omega_i = 1, i = 1 \ldots n \). Thomas et al. (2018)
show that, in a neighborhood of \( \omega = (1, \ldots, 1)' \), \( C \) characterizes the curvature of the \( n \)-dimensional surface representing the Kullback-Leibler divergence of the posterior based on the original likelihood from the posterior based on the geometrically perturbed likelihood. Thus, \( C \) contains information about the directions in \( R^n \) along which the Kullback-Leibler divergence surface changes more rapidly in response to geometric likelihood perturbations. This insight can be used to assess the directional influence of cases.

As an exploratory tool for assessing such influence, Thomas et al. (2018) recommend to compute the sample covariance matrix \( \hat{C} \) based on the sample \( \theta_1, \ldots, \theta_M \) and to perform a principal component analysis (PCA) based on an eigendecomposition of the estimated matrix. Typically, the first several components explain most of the observed variability in the log weights and well-summarize the main directions of influence. A PCA display consisting of a scatterplot of the first two or three normalized eigenvectors helps to reveal structure in the data: points with high loadings in one or more components are particularly influential and points with similar loadings in all components have similar influence.

We propose leveraging these ideas to form a strategy for choosing anchor points based on the case-deletion weights, which we call CDW-reg. First, we fit a simple model that does not accommodate latent heterogeneity in the data. To specify an appropriate mixture model that improves on this naive model, we identify clusters based on a PCA summary of the log case-deletion weights. Thereby we uncover homogeneous subsets of observations that have similar directional influence on the posterior distribution and for which separate regression structures are possibly needed. Our anchor points are chosen to be representatives of clusters in the PCA display.

### 4.1 CDW-reg anchoring on the mammals data

To apply the proposed technique to the analysis of the mammals data, we began by fitting a Bayesian simple linear regression model. This is the same model found to be inadequate by MacEachern and Peruggia (2002). The left panel of Figure 3 shows the PCA display based on the first three eigenvalues of the eigendecomposition of the sample covariance matrix \( \hat{C} \) of
Figure 3: PCA display based on the first three eigenvalues from the eigendecomposition of the sample covariance matrix of the log case-deletion weights (left) and resulting k means clustering with highlighted selected anchor points (right).

the empirical case-deletion log weights for the 100 species. The weights were computed from an MCMC run of length 10,000 implemented using the JAGS software (Plummer, 2003).

Again postulating three mixture components, we ran the R implementation of the k means clustering algorithm (Hartigan and Wong, 1979) identifying the red, green, and blue clusters in the PCA display. The scatterplot in the right panel of Figure 3 shows how the identified PCA clusters map back into observation space. The clustering appears to be responding primarily to the size and sign of the residuals from the fitted simple linear regression line. The red cases tend to have larger positive residuals, the blue cases tend to have larger negative residuals and the green cases are in the middle. As an exploratory step, we superimposed to the three clusters the least squares regression lines estimated from the points in the cluster. The blue and green lines are essentially parallel, while the red line has a slightly larger slope. There is a hint of non-constant variance in the scatterplot and the differences in slopes are conceivably driven by the reaction of the PCA induced clustering to this feature of the data (in addition to the size of the residuals).

As a next step, we identified three representative observations within each PCA cluster to be used as anchor points in the anchored Bayesian mixture of regressions model. To do
so, for each of the three original clusters, we again ran the k means clustering algorithms to identify three PCA sub-clusters. A representative point from each sub-cluster was chosen as the point nearest to the sub-cluster centroid. The selected nine points (three for each cluster) are drawn as solid dots in the PCA display and scatterplot of Figure 3. The same nine points are also displayed in the right panel of Figure 2, where the orders to which the nine species belong are also identified.

When observed in the scatterplot, the selected anchor points can be seen to be adequately spaced out in the $x$ dimension, a feature that will provide estimation stability for the Bayesian mixture of regressions model. We also see that, within each cluster, the set of anchor points comprises some points with apparently high influence (points with large positive and negative residuals in the red and blue groups, respectively, and one point with a very small $x$ value in the green group) and points that fall very near the least-squares lines. This feature follows from the way in which k means decides to allocate member species to the various sub-clusters, with some sub-clusters comprising mostly “usual” observations and other, typically smaller, sub-clusters comprising mostly “unusual” observations.

5 Analysis of the mammals data

We now present the inferential results obtained by fitting the mixture of regressions model to the mammals data under the three sets of anchor points. One byproduct of this analysis will be to identify groups of mammalian species that are similar with respect to the brain/body mass relationship. Our model fit uses the same hyperparameters specified in the EM-reg selection method: $a = 5$, $b = 1$, $\mu_\beta = (3.5, 0.6)'$, and $v_0 = 1, v_1 = .5$. For each anchor model we used the JAGS software (Plummer, 2003) to obtain 20,000 posterior samples of the model parameters via MCMC simulation and assessed convergence using trace plots.
5.1 Parameter and cluster estimates

We used the posterior means of the model parameters to estimate component regression lines for each anchor model. (These are the component-specific, posterior predictive regression lines.) These lines are shown in Figure 4, with Components 1, 2, and 3 drawn in red, green, and blue, respectively. We also estimated group membership for each species by finding a maximum a posteriori estimate of its latent allocation, $s_i$. Using the posterior samples, $s_i^m$, $m = 1, \ldots, M$, the Monte Carlo estimate of the MAP allocation is $\hat{s}_i = \max_j \sum_{m=1}^M I(s_i^m = j)$. Each data point in Figure 4 is color-coded according to its corresponding value of $\hat{s}_i$.

The anchored points, whose group assignments are assumed to be known, are shown as solid symbols to distinguish them from the remaining observations, whose group assignments are estimated.

All three methods have identified Component 1 (shown in red) to be a subgroup whose slope is considerably steeper than those of the other groups, representing a subset of species whose brain masses show an unusually large increase as body masses increase. The estimated regression lines are similar across the methods, with EM-reg estimating the lowest slope.
(0.88) and EM-MVN estimating the highest (0.93). The three methods assign many of the
same species to Component 1, identifying a string of points in the upper-right end of the
scatterplot to be those with the unusually steep regression line. The only difference is that
the EM-reg model assigns four additional points to Component 1, two of which are its anchor
points. These points, which appear in the lower left portion of the left panel of Figure 4,
are assigned to Component 2 under the other two anchor models. These species have low
body masses compared to the other species allocated to Component 1, and they fall in an
area where the estimated regression lines for Components 1 and 2 overlap. In fact, one of
the EM-reg anchor points (triangle) falls almost exactly on the intersection of the two lines;
it is only its selection as an anchor point that ensures it is assigned to Component 1.

Component 3, plotted in blue in Figure 4, represents, broadly speaking, species with
brains that are small relative to species of the same size. The estimated regression lines
are similar for the CDW-reg and EM-reg models. Both models estimate that this group
has a small intercept and a slope of about 0.7, which is similar to that of Component 2
(plotted in green). Under EM-reg, 41 of the 100 species are assigned to this component, and
these species have a wide range of body masses (x-values). The CDW-reg model assigns
fewer species to this group (31), but these species share the characteristics of having small
brains relative to their body masses and to span a wide range of body masses. The features of
Component 3 are quite different under the EM-MVN model; the regression line has a shallow
slope and large intercept, causing the line to move away from the observed data for small
values of x. The estimated regression line under EM-MVN is predominantly representative
of the x-y relationship among the six species that EM-MVN assigns to Component 3. These
species form a small cluster of observations with large body masses on the far right end of
the scatterplot. They have small brains relative to their body mass and are also assigned to
Component 3 under CDW-reg and EM-reg.

Component 2, with regression lines drawn in green, represents the species with moderate
brain masses relative to their body masses. Under the EM-MVN model, this group is
estimated to describe the majority of species; the model has identified two relatively small
subsets of species that differ meaningfully from this main group. The two other methods have
assigned comparatively fewer observations to this component. The EM-reg model estimates
a Component 2 line with a larger intercept and shallower slope than the other methods.
This difference can be attributed to the influence of the EM-reg anchor points, which fall
in the center of the scatterplot; the green line fits these points quite well, as well as many
other points in this region. For the other two methods, both of which have anchor points
with low \( x \)-values, the line fits the data well in the low-\( x \) range, and less well in the middle
of the scatterplot. In particular, there is a string of points corresponding to log body masses
between \(-1\) and \(1\) which all fall above the EM-MVN and CDW-reg lines. This suggests
that additional mixture components might be needed to adequately describe the points not
allocated to Components 1 and 2 under these models.

In all three models, the features of the mixture components and the resulting classifi-
cations have inherited characteristics of the chosen anchor points. The EM-MVN method
selects anchor points unconditional on \( x \), which favors points on the periphery of the scatter-
plot that are representative of bivariate elliptical clusters. The effect of these anchor points
is that the inferred regression line moves to accommodate these points and it need not reflect
the \( x-y \) relationship in other areas of the sample space. A very different behavior is seen
in the EM-reg method: here we have conditioned not only on \( x \), but on a model specifying
three distinct regression lines. The selected anchor points are highly representative of three
distinct lines and the resulting inference produces estimates that closely fit the lines to the
anchor points.

5.2 Comparison to known taxonomy

The estimated component assignments \( \hat{s} \) 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 \( \hat{s} \) have similar estimated regression slopes, and if there is a correspondence between \( \hat{s} \)
Figure 5: Data for the species in orders Primates, Artiodactyla, and Rodentia color-coded according to their model-based $\hat{s}$. The rows correspond to the three anchor models.

and the species' true orders, it may indicate that certain taxonomic groups have distinct body mass/brain mass relationships. Figure 5 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 1 and 2. Given body mass, most large-brained species are assigned to Component 1 and most of those with smaller brains are assigned to Component 2. Component 1 also contains the three species of the Cetacea order (not pictured) under all
three model fits. Interestingly, all three models assign the same seven species of Primates to Component 2. Three of these species belong to the sub-order Prosimii (all other primates in the data set belong to the sub-order Anthropoidea). 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 two extreme Primate species in the $x$ dimension are the Gorilla in the upper end, with a brain smaller than expected and the Hepale Leucocephala in the lower end. The latter species’ brain is actually large given its body mass, but it is classified as a Component 2 species because all three models do not use Component 1 to describe Primate species in the lower $x$ range.

The clustering among the Rodentia highlights some additional interesting features of the model fit. Two species are estimated to belong to the large-brained Component 1 under the EM-reg method, in part because one of the anchor points for Component 1 was from the Rodentia order. Although these estimated groups have no well-defined meaning, we may view these red Rodentia species as something of a misclassification, since Component 1 primarily describes the Primates and Cetacea species under the other two models. A partial explanation for this is that larger body masses are a characterizing feature of these species, but the EM-reg method chooses anchor points conditional on $x$ and thus is less likely to identify groups that are characterized by differences in $x$-values.

6 Discussion

When we specify a finite mixture model, we assume the existence of distinct subgroups in the population. The group membership of any individual observation is unknown and the mixture likelihood combines with the prior specifications (including the prior group membership probabilities) to determine both the estimated features of the mixture components and the posterior probabilities of group membership. A random effects model similarly allows inference on group-specific features, but requires auxiliary, deterministic information indicating which observations are to be grouped together, and the similarities among these observations drive the estimated features of the various groups. In an anchored mixture model, the
anchored observations affect the model fit in the same way as labeled observations in a random effects model affect inferences on their groups: they inform the distinct features of their mixture components. These features then influence the probabilities of group membership of the remaining unanchored observations. Thus, the similarities among a component’s anchor points are a key driver of what types of groups the mixture model identifies.

In our case study, the EM-reg method chose anchor points to fall along straight lines and its accompanying mixture model defined groups based on proximity to these lines. On the other hand, the EM-MVN method selected anchor points to form tight clusters within the bivariate scatterplot, inducing a mixture model whose components are more apt to represent groups that are similar in both the $x$ and $y$ dimensions. The CDW-reg method selected anchor points that were representative of groups of observations exerting similar influence on the posterior inferences from a naive model. Which method we prefer is somewhat situation-dependent. For example, if we truly believe the mixture model we have specified, a method such as EM-reg, which subsumes this model at the anchoring stage, is perfectly appropriate. However, in our case study this perspective was restrictive in its inability to detect differences in groups that depend on $x$.

The anchoring of EM-MVN assumes a completely different model and, in some sense, produces a mixture of regressions model that is more descriptive of clusters in bivariate space than of groups of similar $x$-$y$ relationships. The resulting inferences lack cohesiveness because of the mismatch between the anchoring method and the model of analysis. The CDW-reg method falls between these two extremes; the anchor points are selected using the principle that a mixture model is appropriate because the simple model is inadequate; the mixture components are identified as groups exhibiting similar misfit and the estimated model is consistent with the mixture of regressions but flexible in accommodating unexpected features in the data. This flexibility suggests that the CDW-reg method has potential as an effective strategy for specifying a wide variety of anchored mixture models. Its reliance on clustering performed in the derived space determined by the PCA eigendecomposition of the log case-deletion weights enables the method to select anchors by honing in on features that
affect directly the quality of model fit.

Partly, the choice to fix the number of mixture components to three in the interest of modeling parsimony limits the ability of the mixture model to adapt to finer structure in the data. On the other hand, adding more components and companion anchor points would make the reliance of the use of a data-dependent prior too substantial.

The known taxonomic structure of species in our case study provides insight about the model-based classifications produced by the three anchor models which extends beyond the particular data we analyzed. In our data, the groups of species from the same order exhibit distinct characteristics that might be primarily revealed by similarities in brain or body mass (the Artiodactyla species, for example, all have above-average body mass), by similarities in the change in brain mass as body mass changes, or both (as we see for the large primates). The EM-MVN model is useful in identifying groups of the former type because it selects anchor points that correspond to bivariate clusters in the data. However, this model is less flexible in identifying homogeneous groups that span a wide range of $x$ values. The EM-reg model is particularly well-suited for identifying groups of the latter type because its anchor points are selected using the mixture of regressions, but it ignores differences in $x$ which provide valuable information for classification in this setting, where body mass is a distinguishing features of some of the taxonomic groups.

Thoughtful application of anchor models requires understanding of these differing characteristics. The field of semi-supervised learning has addressed many relevant questions regarding the combination of labeled and unlabeled data, such as selecting some observations to artificially pre-label (Yarowsky, 1995) and constraining certain observations to the same group in an effort to learn a distance metric that discerns similarity among future observations (Shental et al., 2002). We think that integration of our methods with these well-studied problems is a promising direction for future research.

In sum, to select a method for finding anchor points, it is important to consider which types of distinguishing features the underlying model uses to identify similar observations and to determine how such features become relevant to answer the scientific questions of
interest.

References

Bennett, P. and Harvey, P. (1985a). Relative brain size and ecology in birds. *Journal of Zoology*, 207(2):151–169.

Bennett, P. M. and Harvey, P. H. (1985b). Brain size, development and metabolism in birds and mammals. *Journal of Zoology*, 207(4):491–509.

Bradlow, E. T. and Zaslavsky, A. M. (1997). Case influence analysis in Bayesian inference. *Journal of Computational and Graphical Statistics*, 6(3):314–331.

Epifani, I., MacEachern, S. N., Peruggia, M., et al. (2008). Case-deletion importance sampling estimators: Central limit theorems and related results. *Electronic Journal of Statistics*, 2:774–806.

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15.

Garland Jr, T., Harvey, P. H., and Ives, A. R. (1992). Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic biology*, 41(1):18–32.

Gayon, J. (2000). History of the Concept of Allometry. *American Zoologist*, 40(5):748–758.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, pages 50–67.
Jerison, H. J. (1955). Brain to body ratios and the evolution of intelligence. *Science*, 121(3144):447–449.

Kunkel, D. and Peruggia, M. (2018). Anchored Bayesian Gaussian Mixture Models. *arXiv e-prints*.

MacEachern, S. N. and Peruggia, M. (2002). Bayesian tools for EDA and model building: A brainy study. In *Case Studies in Bayesian Statistics*, pages 345–362. Springer.

Peruggia, M. (1997). On the variability of case-deletion importance sampling weights in the Bayesian linear model. *Journal of the American Statistical Association*, 92(437):199–207.

Peters, R. H. (1983). *The ecological implications of body size*. Cambridge University Press.

Plummer, M. (2003). Jags: A program for analysis of Bayesian graphical models using Gibbs sampling.

Sacher, G. A. and Staffeldt, E. F. (1974). Relation of gestation time to brain weight for placental mammals: implications for the theory of vertebrate growth. *The American Naturalist*, 108(963):593–615.

Shental, N., Hertz, T., Weinshall, D., and Pavel, M. (2002). Adjustment learning and relevant component analysis. In Heyden, A., Sparr, G., Nielsen, M., and Johansen, P., editors, *Computer Vision — ECCV 2002*, pages 776–790, Berlin, Heidelberg. Springer Berlin Heidelberg.

Thomas, Z. M., MacEachern, S. N., and Peruggia, M. (2018). Reconciling curvature and importance sampling based procedures for summarizing case influence in Bayesian models. *Journal of the American Statistical Association*, 113(524):1669–1683.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, ACL ’95*, pages 189–196, Stroudsburg, PA, USA. Association for Computational Linguistics.