Supplementary material for “Manual hierarchical clustering of regional geochemical data using a Bayesian finite mixture model”

Karl J. Ellefsena, David B. Smithb

aU.S. Geological Survey, MS 964, Box 25046, Denver, Colorado, USA, [ellefsen@usgs.gov](mailto:ellefsen@usgs.gov), 303-236-7032

bretired from U.S. Geological Survey, MS 973, Box 25046, Denver, Colorado, USA, [dbsmith13@gmail.com](mailto:dbsmith13@gmail.com)

Corresponding author: K.J. Ellefsen

Table S1. Univariate statistics for the measured concentrations. The lower limits of determination (LLD) are published in Smith et al. (2009). Leaders (---) for interquartile range indicate that the value could not be computed because the LLD exceeded the 0.25 quantile. Leaders for median absolute deviation and robust coefficient of variation indicate that the values could not be computed because too many concentrations were below the LLD: the criterion was that the LLD exceeded the 0.05 quantile.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Element | Lower limit of determination  (mg/kg) | Percentage of samples less than the lower limit of determination | Minimum (mg/kg) | 0.05 quantile (mg/kg) | 0.25 quantile (mg/kg) | Median (mg/kg) | 0.75 quantile (mg/kg) | 0.95 quantile (mg/kg) | Maximum (mg/kg) | Interquartile range (mg/kg) | Median absolute deviation (mg/kg) | Robust coefficient of variation (%) |
| Al | 100 | 0 | 11200 | 34300 | 46600 | 55400 | 63500 | 77600 | 99800 | 16900 | 12600 | 22.7 |
| Ca | 100 | 0 | 1100 | 3700 | 7100 | 12300 | 23800 | 55900 | 219000 | 16600 | 9790 | 79.6 |
| Fe | 100 | 0 | 2700 | 7300 | 15500 | 21400 | 27700 | 40500 | 94200 | 12200 | 9040 | 42.3 |
| K | 100 | 0 | 3800 | 14200 | 19300 | 22900 | 26500 | 33400 | 58600 | 7200 | 5340 | 23.3 |
| Mg | 100 | 0 | 300 | 1500 | 3700 | 5700 | 7900 | 13300 | 37700 | 4200 | 3110 | 54.6 |
| Na | 100 | 0 | 700 | 3400 | 6900 | 9050 | 12900 | 20900 | 40500 | 6030 | 4080 | 45.1 |
| S | 100 | 3.02 | <100 | 100 | 200 | 300 | 400 | 1000 | 95400 | 200 | 148 | 49.4 |
| Ti | 100 | 0 | 300 | 900 | 1500 | 2000 | 2500 | 3900 | 12100 | 1000 | 741 | 37.1 |
| Ag | 1 | 98 | <1 | <1 | <1 | <1 | <1 | <1 | 35 | --- | --- | --- |
| As | 1 | 0.417 | <1 | 2 | 3 | 5 | 7 | 11 | 126 | 4 | 2.97 | 59.3 |
| Ba | 5 | 0 | 155 | 424 | 606 | 719 | 814 | 1040 | 4660 | 208 | 153 | 21.2 |
| Be | 0.1 | 0 | 0.4 | 0.9 | 1.3 | 1.6 | 1.9 | 2.5 | 6.4 | 0.6 | 0.445 | 27.8 |
| Bi | 0.04 | 1.15 | <0.04 | 0.06 | 0.12 | 0.18 | 0.24 | 0.43 | 14.6 | 0.12 | 0.089 | 49.4 |
| Cd | 0.1 | 5.62 | <0.1 | <0.1 | 0.2 | 0.3 | 0.4 | 1 | 20.2 | 0.2 | --- | --- |
| Ce | 0.05 | 0 | 14.6 | 35.1 | 52.3 | 63.9 | 76.7 | 112 | 350 | 24.4 | 17.9 | 28.1 |
| Co | 0.1 | 0 | 0.7 | 2.1 | 5.2 | 7.3 | 9.4 | 13.8 | 43.3 | 4.2 | 3.11 | 42.7 |
| Cr | 1 | 0 | 3 | 11 | 20 | 29 | 39 | 61 | 141 | 19 | 14.8 | 51.1 |
| Cs | 5 | 76.6 | <5 | <5 | <5 | <5 | <5 | 7 | 22 | --- | --- | --- |
| Cu | 0.5 | 0 | 1.7 | 4.7 | 10.7 | 15.6 | 20.8 | 32.2 | 464 | 10.1 | 7.56 | 48.5 |
| Ga | 0.05 | 0 | 2.76 | 7.52 | 10.4 | 12.9 | 15.1 | 18.5 | 26.5 | 4.72 | 3.48 | 27 |
| In | 0.02 | 14.1 | <0.02 | <0.02 | 0.03 | 0.04 | 0.05 | 0.08 | 13.8 | 0.02 | --- | --- |
| La | 0.5 | 0 | 7.5 | 18.6 | 27.1 | 33.1 | 38.9 | 54.4 | 176 | 11.8 | 8.75 | 26.4 |
| Li | 1 | 0 | 4 | 8 | 16 | 21 | 28 | 42 | 259 | 12 | 8.9 | 42.4 |
| Mn | 5 | 0 | 75 | 175 | 316 | 431 | 608 | 1060 | 3460 | 292 | 196 | 45.6 |
| Mo | 0.05 | 0 | 0.15 | 0.32 | 0.64 | 0.97 | 1.44 | 3.49 | 23.4 | 0.803 | 0.563 | 58.1 |
| Nb | 0.1 | 0 | 1.5 | 4.4 | 7.1 | 9.1 | 11.3 | 18.9 | 69.9 | 4.2 | 3.11 | 34.2 |
| Ni | 0.5 | 0 | 2 | 4.2 | 9.28 | 13.8 | 19.1 | 30.7 | 149 | 9.85 | 7.26 | 52.6 |
| P | 50 | 0.521 | <50 | 250 | 450 | 650 | 860 | 1260 | 2840 | 410 | 297 | 45.6 |
| Pb | 0.5 | 0 | 5.5 | 13.8 | 18.8 | 21.7 | 26.5 | 58.3 | 3600 | 7.72 | 5.34 | 24.6 |
| Rb | 0.2 | 0 | 22.1 | 55.8 | 77.3 | 91.6 | 105 | 143 | 312 | 27.7 | 20.8 | 22.7 |
| Sb | 0.05 | 0.208 | <0.05 | 0.19 | 0.34 | 0.5 | 0.72 | 1.39 | 66.7 | 0.38 | 0.267 | 53.4 |
| Sc | 0.1 | 0 | 0.9 | 2.3 | 5 | 6.8 | 8.6 | 12.2 | 33.9 | 3.6 | 2.67 | 39.2 |
| Sn | 0.1 | 0 | 0.2 | 0.6 | 1.1 | 1.4 | 1.7 | 2.7 | 125 | 0.6 | 0.445 | 31.8 |
| Sr | 0.5 | 0 | 37.5 | 91.3 | 134 | 181 | 254 | 516 | 1430 | 120 | 78.6 | 43.4 |
| Te | 0.1 | 93.6 | <0.1 | <0.1 | <0.1 | <0.1 | <0.1 | 0.1 | 4.3 | --- | --- | --- |
| Th | 0.2 | 0 | 2.2 | 5.1 | 8 | 10 | 12.6 | 19.6 | 82.6 | 4.6 | 3.41 | 34.1 |
| Tl | 0.1 | 0 | 0.1 | 0.3 | 0.5 | 0.6 | 0.7 | 0.905 | 3.1 | 0.2 | 0.148 | 24.7 |
| U | 0.1 | 0 | 0.6 | 1.2 | 1.9 | 2.4 | 3 | 4.6 | 18.6 | 1.1 | 0.741 | 30.9 |
| V | 1 | 0 | 4 | 20 | 40 | 59 | 80 | 123 | 373 | 40 | 29.7 | 50.3 |
| W | 0.1 | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1 | 1.8 | 14 | 0.4 | 0.297 | 37.1 |
| Y | 0.1 | 0 | 3.8 | 8.6 | 12.8 | 16.2 | 19.3 | 25.8 | 169 | 6.52 | 4.82 | 29.8 |
| Zn | 1 | 0 | 10 | 24 | 47 | 66 | 90 | 158 | 7000 | 43 | 32.6 | 49.4 |
| Se | 0.2 | 45.3 | <0.2 | <0.2 | <0.2 | 0.2 | 0.4 | 1 | 11.3 | --- | --- | --- |
| Hg | 0.02 | 69 | <0.02 | <0.02 | <0.02 | <0.02 | 0.02 | 0.04 | 0.4 | --- | --- | --- |

# Sampling the posterior probability density function (pdf)

The finite mixture model is coded in the Stan probabilistic programming language (Stan Development Team, 2015); samples of its posterior pdf are obtained using the Hamilitonian Monte Carlo method (Neal, 2011; Gelman et al., 2014, p. 300–305), which is implemented within Stan (Hoffman and Gelman, 2013).

The posterior pdfs for finite mixture models usually have multiple modes (Carreira-Perpinan and Williams, 2003; Marin et al., 2005; and Edelsbrunner et al., 2012). It is important to find these modes, or at least as many of them as is practically possible. To this end, multiple Monte Carlo chains with random starting points are computed. There is no simple rule for selecting a suitable number of chains; our experience is that 35 chains may be adequate.

A simple way to summarize the results from these multiple chains is to compute point statistics for the model parameters. For the finite mixture model, the number of model parameters is where is the number of principal components that are processed. Because , the number of model parameters is Analyzing point statistics for each chain would be overwhelming, so only are selected: , , , , and . In addition, the point statistics are computed for the natural logarithm of the likelihood function , which is called “log-likelihood.” The log-likelihood is defined as

|  |  |
| --- | --- |
|  | (S-1) |

where is the number of field samples and represents the probability density for vector . The point statistics are chosen to be the median and the 95% credible interval because they succinctly summarize a distribution. These point statistics for the selected model parameters and the log-likelihood are plotted in Figure S1.

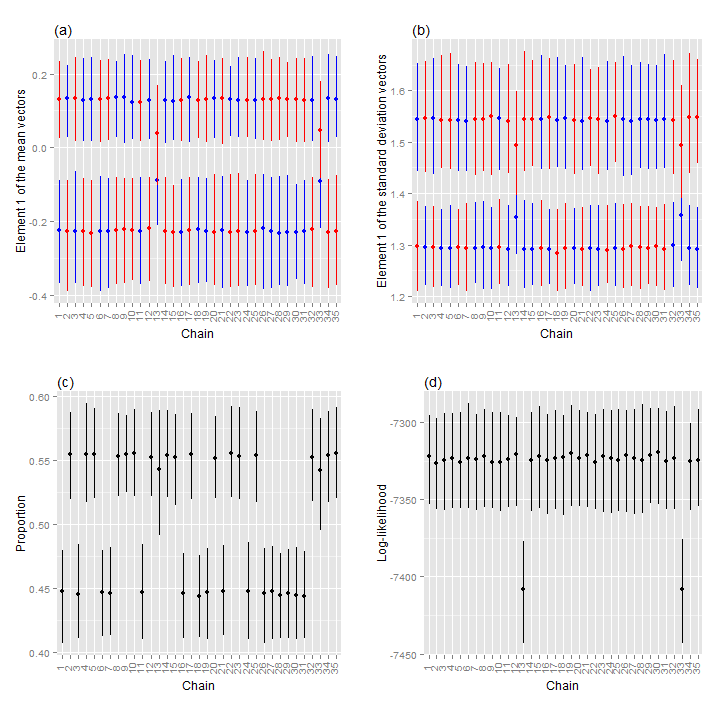


Figure S1. Point statistics for (a) element 1 of the two mean vectors, and , (b) element 1 of the two standard deviation vectors, and , (c) proportion, , and (d) the log-likelihood, . A dot represents the median, and the vertical line represents the 95% credible interval. Blue and red represent pdfs 1 and 2, respectively.

Examine the point statistics for the log-likelihood (Figure S1d). The medians are either approximately or , which suggests that two modes in the posterior pdf were sampled. We are unaware of any point statistics that are suitable for characterizing multi-modal distributions, so we analyze the modes individually. It is tempting to select just that mode having the highest log-likelihood and to neglect the other modes because the high log-likelihood indicates a relatively good fit between the model and the data. However, our experience is that this approach is incorrect—all modes should be investigated. This issue is discussed further in section 3.4 (Interpretation to check model).

Examine the point statistics for proportion for chains 1 to 4 (Figure S1c). These four chains are associated with one mode (Figure S1d), which we call the “first mode.” Nonetheless, is not a single value. For chains 1 and 3, is approximately , so must be approximately . In contrast, for chains 2 and 4, is approximately , so must be approximately . In effect, the labels for the two pdfs in the finite mixture model have been switched; this label switching is apparent also in the other point statistics (Figures S1a and S1b). Such label switching is common and normal in finite mixture models (Marin et al., 2005; Gelman et al., 2014, p. 533–536). Although there are algorithms to account for this switching (Stephens, 2000), they are unnecessary for this application because the label switching is easy to identify.The pdfs for chains 2 and 4 can be switched manually so that they conform to the pdfs for chains 1 and 3. Now, examine the point statistics for proportion for chains 13 and 33 (Figure S1c). These two chains are associated with one mode (Figure S1d), which we call the “second mode.” In this case, label switching has not occurred, so the pdfs must not be switched.

It is very important that each set of chains be checked to ensure that the Monte Carlo sampling has converged. For this application, convergence means that the samples in a set of chains are representative of the associated mode. Well-established procedures for assessing convergence are described in Gelman et al. (2014, p. 281–288); so the description is not repeated here.

# Checking the fit of the model to the data

One part of model checking is evaluating how well the model fits the data (Gelman et al., 2014, p. 141). For this application, the model is the finite mixture model, and the data are the principal components. These checks are performed for model parameters , , , , , and and require corresponding statistics calculated from the principal components. The difficulty in calculating these statistics is that the principal components are not divided into two groups corresponding to pdfs 1 and 2. Nonetheless, suitable statistics may be estimated—the key idea is to use the conditional probabilities as weights. To calculate the statistics associated with pdf 1, the principal components are weighted by the median of the samples of . These weighted principal components are used to calculate a sample mean vector and a sample covariance matrix , using the formulas presented in Price (1972). Then the sample covariance matrix is decomposed into a sample standard deviation vector and sample correlation matrix . A similar procedure is used to calculate the sample statistics associated with pdf 2, namely , , and .

For the second mode, sample statistics , , , and and model parameters , , , and are compared in Figure S2. For each vector element of a model parameter, the Monte Carlo samples constitute a univariate distribution, which is represented by its median and its 95% credible interval in Figure S2. All sample statistics are within their associated 95% credible intervals. The sample standard deviations are consistently lower than the associated distributional medians because of the weighting.

The relation between a sample statistic and the associated distribution may be summarized by the posterior predictive p-value (Gelman and others, 2014, p. 146). It is defined as the probability that a model parameter could be more extreme than the associated sample statistic. In mathematical terms, where is a generic notation for the model parameter and is a generic notation for the sample statistic. When is in the left tail of the distribution for , the p-value is close to 1. This situation can confuse the interpretation of the p-value, so the mathematical definition is modified slightly: (Gelman and others, 2014, p. 148). Consequently, the calculated p-value is always less than 0.5, and it may be interpreted in the standard way. The p-values, which are printed in Figure S2, are moderate to large.

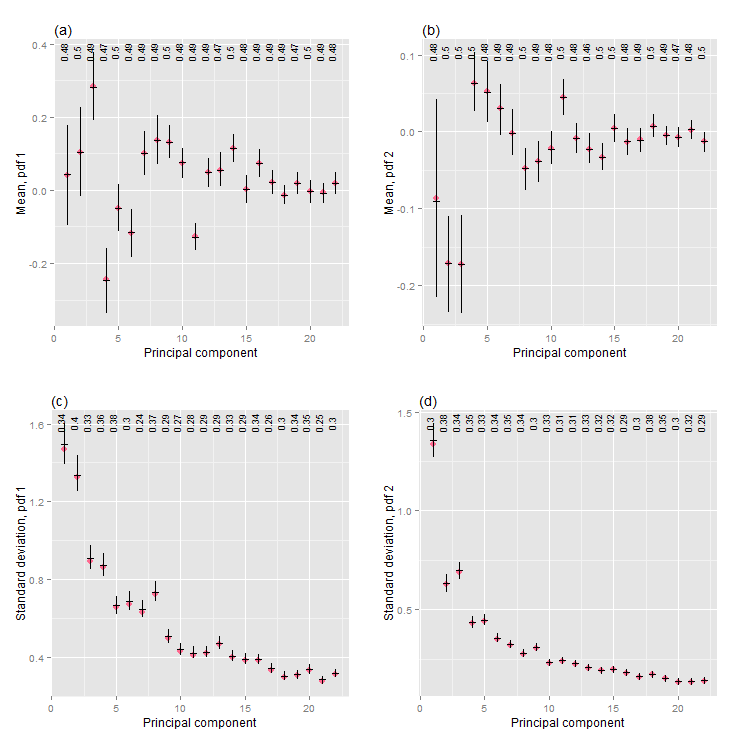


Figure S2. Model check, for the second mode, of (a and b) the mean vectors and (c and d) the standard deviation vectors. A red dot represents a sample statistic. A horizontal black line represents the median of associated Monte Carlo samples, and the vertical line represents its 95% credible interval. The numbers along the top of each plot are the posterior predictive p-values.

A model check of the correlation matrices could be like the model check of the mean and standard deviation vectors (Figure S2). However, such a model check would require 462 individual comparisons for this set of principal components because there are 231 unique elements in each correlation matrix. Such a large number of comparisons would be overwhelming, so a different model check is used and is demonstrated for the second mode. One aspect of the model check involves a matrix that combines parts of two other matrices (Figures S3a and S3c). One part is the upper triangle of the correlation matrix that is calculated from the principal components. The other part is the lower triangle of the correlation matrix that is the median of its Monte Carlo samples. Consequently, corresponding elements in the upper and lower triangles should be compared to one another. For both pdfs, corresponding elements appear the same. Some readers may expect that all correlations should be zero because they pertain to principal components. Indeed, the correlations between the principal components for all field samples are zero. However, the correlations between the principal components for groups of field samples (namely, clusters) are not. The other aspect of the model check is a graphical display of the posterior predictive p-values for the off-diagonal elements of the correlation matrix (Figures S3b and S3d). Because of the symmetry in the correlation matrix, only the p-values for one triangle must be displayed, and the upper triangle is chosen. The color scale ranges from the smallest p-value to 0.5, which is the largest possible p-value. For both pdfs, the p-values are moderate to large.

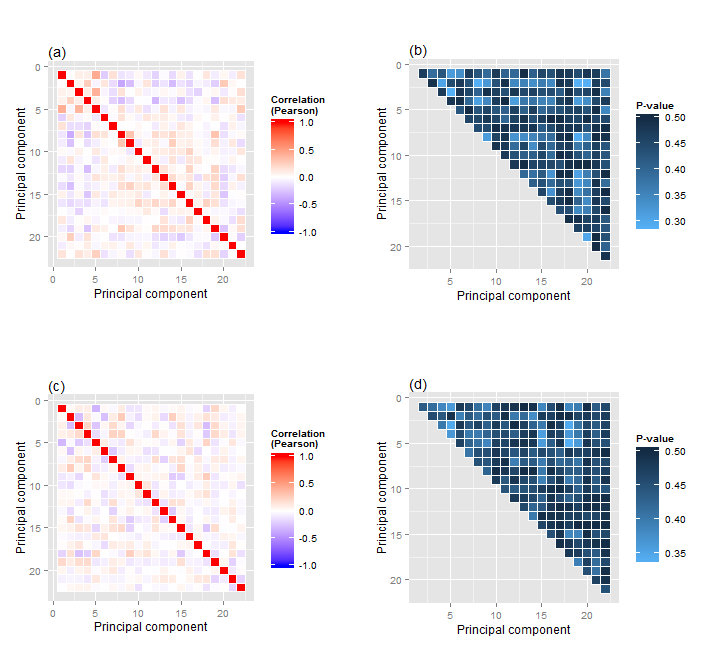


Figure S3. Model check, for the second mode, of (a and b) the correlation matrix for pdf 1 and (c and d) the correlation matrix for pdf 2. In (a) and (c), the matrix comprises the upper triangle of the correlation matrix that is calculated from the principal components and the lower triangle of the correlation matrix that is the median of its Monte Carlo samples. In (b) and (d), the upper triangle comprises the posterior predictive p-values.

The model checks of the mean vectors, the standard deviation vectors, and the correlation matrices (Figures S2 and S3) show that the model fits the data for the second mode. The model also fits the data for the first mode (not shown).

# Sensitivity analysis

The distribution of the model parameters in the finite mixture model are specified by prior pdfs (section 3.2—Finite mixture model). These prior pdfs are characterized by parameters that are chosen by us, based upon our exploratory analysis of the principal components (section 3.1—Preprocessing and analysis). For example, the prior pdf of (the first element of the mean vector for the first pdf) is a normal pdf for which the mean is and the variance is . The prior pdf parameters are chosen, not estimated from the data, so the question arises, “How sensitive are the model parameters to the prior pdf parameters?” This question is addressed with a sensitivity analysis.

A sensitivity analysis is performed by systematically changing each prior pdf parameter and then observing how it affects the model parameters in the finite mixture model. The analysis is conducted for data subset 2 (section 3.6—Manual hierarchical clustering)—both the number of field samples (486) and the number of principal components (19) are smaller than the corresponding quantities for the entire data set (959 and 22, respectively), so the analysis is quicker. Nonetheless, for 19 principal components, there are 419 model parameters, making a comprehensive sensitivity analysis overwhelming. So, the analysis is restricted both to selected model parameters and to important prior pdf parameters.

First consider the prior pdf for the elements of the mean vectors (section 3.2): Recall that it is a normal pdf, for which the parameters are the mean and the variance (or equivalently, the standard deviation). Of these two parameters, the standard deviation is deemed more important, so a sensitivity analysis is conducted only for it. The standard deviation is set to 1, 2, 3, 4, and 5, and for each value the Monte Carlo sampling is performed. (The value used for the clustering is 3 (section 3.2).) Although the changes in the standard deviations affect all model parameters, the largest effects are in the elements of the mean vectors. Of the 19 elements for each mean vector, only the first elements and are analyzed. For the five values of the standard deviation, the distributions for and appear the same (Figure S4a).

The sensitivity analysis for the other prior pdf parameters is similar to the previously described procedure, so only a few remarks are necessary. For the prior pdf for the elements of the standard deviation vectors, the scale is analyzed for values of 1, 2, 3, 4, and 5. (The value used for the clustering is 3 (section 3.2).) For these five values of the scale, the distributions for and appear the same (Figure S4b). For the prior pdf for the elements of the correlation matrices, the shape parameter is analyzed for values of 1.1, 1.5, 2, 2.5, and 3. (The value used for the clustering is 2 (section 3.2).) For these five values of the shape parameter, the distributions for and (namely element of correlation matrices and ) appear the same (Figure S4c). For the prior pdf for the proportion, the shape parameter is analyzed for values of 1.1, 2, 3, 4, and 5. (The value used for the clustering is 4 (section 3.2).) For these five values of the shape parameter, the distributions for appear almost the same (Figure S4d)—from 1.1 to 5, the changes in the distributions are insignificant.

Based upon the results of this sensitivity analysis, we infer that, if physically reasonable values are assigned to the prior pdf parameters, then these parameters do not significantly affect the estimated model parameters.

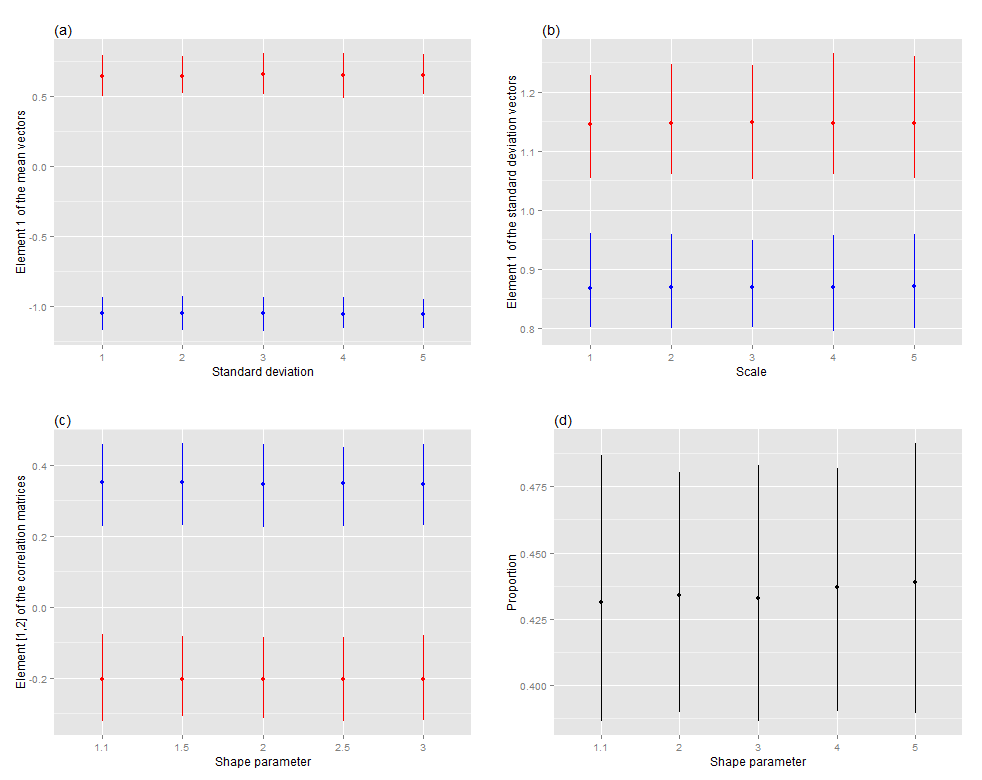


Figure S4. Sensitivity analysis for (a) the standard deviation in the prior pdf for elements and of the mean vectors, (b) the scale in the prior pdf for elements and of the standard deviation vectors, (c) the shape parameter in the prior pdf for elements and of the correlation matrices, and (d) the shape parameter in the prior pdf for proportion . A dot represents the median of a distribution, and the associated vertical line represents the 95% credible interval. Blue and red represent pdfs 1 and 2, respectively.

# Miscellaneous issues

Most regional geochemical surveys will have many clusters, and the obvious way to account for these clusters is to increase the number of pdfs in the finite mixture model. Because the correct number is unknown, the number itself may be cast as a random variable in the Bayesian formulation, and the posterior pdf may be sampled using the reversible jump Markov chain Monte Carlo method (Richardson and Green, 1997). These investigators demonstrate that this method works successfully for univariate distributions. However, we are unaware of any demonstration showing that the method works successfully for multivariate distributions with many variates, which is a common situation for geochemical surveys.

A practical difficulty with the method developed by Richardson and Green (1997) is that, if there are pdfs in the finite mixture model, then the number of modes is of order (Marin et al., 2005). For example, if there are pdfs (which is appropriate for the Colorado geochemical data), then the number of modes is of order . If there are pdfs (which might be appropriate for a continental scale survey), then the number of modes is of order . With current sampling methodology, thorough investigation of the posterior pdf is very difficult when the number of modes is of order , and it is impossible when the number is of order .

Because of these difficulties, we use the method presented in section 3 (Clustering procedure). An advantage of this method is that using just two pdfs in the finite mixture model minimizes the number of modes in the posterior pdf. Consequently, the chances of finding all of the modes increase. A strong assumption underlying this method is that, at each level of the hierarchy, two pdfs in the finite mixture model are enough to represent the distribution of the data. Our experience heretofore is that two pdfs are enough; the model fits both the data and our independent knowledge of the geology and geochemistry (sections 3.4 and 3.6).

An important issue is the number times that the data are partitioned. It is desirable to make as many partitions as possible because the resulting clusters will show geologic and geochemical process at ever smaller spatial scales. We observe that, after some number of partitions, the Monte Carlo sampling does not converge. A reason for this lack of convergence is that, with each partition, the number of field samples decreases, so there are fewer field samples to estimate the parameters in the finite mixture model. Another reason is that, after some number of partitions, the remaining geochemical data are roughly homogenous, so the data should not be partitioned further. This issue requires additional study.

Instead of sampling the posterior pdf, its modes may be located using an optimization algorithm. The model parameters associated with the mode are then used to analyze the geochemical data. An advantage of this approach is that Monte Carlo sampling, which is difficult, is not used. Another advantage is that optimization requires much less computer time and memory than sampling requires. Because of these advantages, optimization appears very attractive.

It is worth asking why optimization is not used here. First, the likelihood function (Eq. 1) sometimes can be become very large if the data, or parts of them, are fit very well by the finite mixture model (MacKay, 2003, p. 306; Marin et al., 2005). This phenomenon may occur near the boundary of the space for the model parameters where, for example, the standard deviations can be very small, almost . Such standard deviations as well as the other associated model parameters are not physically reasonable. This problem may be mitigated somewhat by using a highly informative prior pdf for the standard deviations (Gelman et al., 2014, p. 313–315). Second, the mode depends upon the basis for the model parameters (MacKay, 2003, p. 306; Druilhet and Marin, 2007). That is, if the model parameters undergo a non-linear transformation, then the modes for the original and transformed parameters will not correspond to one another. Third, the Hessian associated with a mode may provide reliable information about the uncertainty in the model parameters, only if the shape of the distribution is well approximated by a normal distribution (Gelman et al., 2014, p. 84 and 88). These three problems do not occur with sampling of the posterior pdf, so only sampling is used here.

# References for supplementary materials

Carreira-Perpiñán, M.Á, and Williams, C.K.I., 2003, On the number of modes of a Gaussian mixture *in* Griffin, L.D., and Lillholm, M., eds., Scale Space Methods in Computer Vision: Springer-Verlag, p. 625–640.

Druilhet, P. and Marin, J.M., 2007, Invariant HPD credible sets and MAP estimators: Bayesian Analysis, vol. 2, no. 4, p. 681–692.

Edelsbrunner, H., Fasy, B.T., and Rote, G., 2012, Add Isotropic Gaussian Kernels at Own Risk–More and More Resiliant Modes in Higher Dimensions: ACM Symposium on Computational Geometry (SoCG 2012), June 17–20 2012, Chapel Hill, North Carolina, USA.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B., 2014, Bayesian data analysis (3rd ed.), CRC Press.

Hoffman, M., and Gelman, A., 2013, The no-U-turn sampler–Adaptively setting path lengths in Hamiltonian Monte Carlo, available on line at http://arxiv.org/abs/1111.4246 (last accessed November 2015).

MacKay, D.J.C., 2003, Information theory, inference, and learning algorithms: Cambridge University Press.

Marin, J.M., Mengersen, K. and Robert, C.P., 2005, Bayesian modelling and inference on mixtures of distributions, *in* Dey, D. and Rao, C.R., eds., Handbook of statistics 25, Elsevier-Sciences.

Neal, R., 2011, MCMC for using Hamiltonian dynamics, *in* Brooks, S., Gelman, A., Jones, G., and Xiao-Li, M., eds., Handbook of Markov Chain Monte Carlo, Chapman & Hall, p. 113–162.

Price, G.R., 1972, Extension of covariance selection mathematics: [Annals of Human Genetics, vol. 35, p 485–490.,](http://www.dynamics.org/Altenberg/LIBRARY/REPRINTS/Price_extension_AnnHumGenetLond.1972.pdf) available at <http://www.dynamics.org/Altenberg/LIBRARY/REPRINTS/Price_extension_AnnHumGenetLond.1972.pdf>, last accessed December 2015.

Richardson, S., Green, P.J., 1997, On Bayesian analysis of mixtures of an unknown number of components: Journal of the Royal Statistical Society, Series B (Methodological), vol. 59, no. 4, p. 731–792 (including comments and response).

Smith, D.B., Woodruff, L.G., O’Leary, R.M., Cannon, W.F., Garrett, R.G., Kilburn, J.E., Goldhaber, M.B., 2009. Pilot studies for the North American Soil Geochemical Landscapes Project – Site selection, sampling protocols, analytical methods, and quality control protocols. Applied Geochemistry, vol 24, p. 1357–1368, http://dx.doi.org/10.1016/j.apgeochem.2009.04.008.

Stan Development Team, 2015, Stan Modeling Language - User’s Guide and Reference Manual, Version 2.8.0, available on line at http://mc-stan.org/ (last accessed October 2015).

Stephens, M., 2000, Dealing with label switching in mixture models: Journal of the Royal Statistical Society, B, vol. 62, part 4, p. 795–809.