Selection of a Representative Sample

Herbert K. H. Lee, Matthew Taddy, and Genetha A. Gray*

December 29, 2008

Abstract

Sometimes a larger dataset needs to be reduced to just a few points, and it is desirable that these points be representative of the whole dataset. If the future uses of these points are not fully specified in advance, standard decision-theoretic approaches will not work. We present here methodology for choosing a small representative sample based on a mixture modeling approach.

Key Words: Bayesian statistics; clustering; mixture model; data reduction.

1 Introduction

Data reduction has a wide variety of applications, and there exist a variety of suggested approaches (Kruskal, 1976; Späth, 1980; Jolliffe, 1986, for example). Here we explore methodology for data reduction in the form of choosing a small number of points which are representative of the dataset. We view this in the context of clustering, in that we want each selected point to represent a cluster of points in the original dataset.

This project originated from a collaboration with scientists at Sandia National Laboratories who were studying circuit devices via both physical experiments and computer simulation. Specifically, the overall goal of this project was to both "calibrate" and "validate"

*H. K. H. Lee is Professor, Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064, USA; M. Taddy is Assistant Professor, Graduate School of Business, University of Chicago; G. A. Gray is Technical Staff Member, Sandia National Laboratories, Livermore, California. This work was partially supported by Sandia grant 673400.

radiation-aware models of electrical circuit devices. In order to satisfy the requirements of both processes, the data must be divided such that a small representative sample of the physical data is used in the calibration process, with the remaining points held out for validation of the model.

In this context, the process of calibration uses experimental observations of a response to inform upon uncertain input parameters of a computational simulation (Kennedy and O'Hagan, 2001; Campbell, 2002; Trucano et al., 2006, for example). In other words, relevant model parameters are adjusted so that the agreement between the code calculations in the simulator best match the observed behavior illustrated by the experimental data. A calibrated model both shows improved predictive capabilities and minimizes the information lost by using a numerical model instead of the actual system. Experimental data is a critical component of this process. Inadequate data can render calibration useless and lead to faulty model predictions. Furthermore, data sufficiency is not merely dependent on quantity, and uncertainties in the model are not necessarily decreased by calibrating to a larger data set (Moore and Doherty, 2005). Instead, the quality, relevancy, and accuracy of the data must also be considered.

In contrast, the validation process is applied in order to quantify the degree to which a computational model is an accurate representation of the real world phenomena it seeks to represent (Trucano et al., 2002). Validation is critical to ensure some level of predictive capability of the simulation codes so that these codes can be subsequently used to study or predict situations for which experimental data are unavailable due to environmental or economic limitations (Oberkampf et al., 2003). It should be noted that model validation is a major effort regularly undertaken by numerous government agencies and private companies.

Satisfying validation benchmarks requires a significant amount of relevant data.

The problem of separating an experimental data set for calibration and validation is similar to a common challenge encountered in machine learning—how to split up the data into training versus testing sets. In the training phase of classification, the classifier uses a subset of the data to "learn" how to classify objects. Similarly, in the calibration process, the computational model "learns" how to best set inherent model parameters using the given experimental data. Then, in the validation phase, the remaining data set is used to "test" the predictive capabilities of the calibrated model just as is done in the testing phase of classification. Considering this analogy, it should be clear that the calibration (or training) data set must be representative of the entire data set. Moreover, adequate data must remain to substantiate validation results. Note that this problem of selecting a calibration data set should be solved using a robust, repeatable method in order to lend credibility to the validation (Trucano et al., 2002).

We note that this problem requires the selection of points in the dataset, as the data have already been collected and there may not be an opportunity to collect data at new locations. Because some aspects of our problem are classified, some of the information in the dataset available to us is actually summary information, and thus raw-level data that might be of interest to the scientists will only be available from points in the original experiment. So while points outside the dataset may be more statistically representative of clusters in the data, experimental information is not available for them, and we must restrict ourselves to actual datapoints.

An obvious approach to the problem of selecting a calibration data set is to frame it as one of an optimal decision, choosing the points to maximize a measure of utility, as in Müller et al.

(2004). However, in this case, we did not have access to the computer simulators, because of national security classification. Not having access to simulators is actually not unusual as many codes are proprietary or not widely distributed. Moreover, the computational cost of many simulators can be prohibitive to utilizing a formal utility function. Furthermore, our collaborators did not want to specify a particular utility, as they wanted to be able to use the representative sample for multiple possible future goals, some of which they might decide based on future analyses. In fact, specifying a specific utility is ineffective in the case where the development of the simulator is ongoing because when a model is updated, the calibration must be repeated. Therefore, to increase the usefulness of this exercise, we were asked to base our representative set solely on the experimental data. This would eliminate the need to select a representative set each time a change was implemented in the simulator. This request completely eliminated any approach using an explicit evaluation of a formal utility function.

One related approach in the literature is data squashing (Dumouchel et al., 1999; Madigan et al., 2002; Owen, 2003), where the idea is to create a relatively small pseudo-dataset that when appropriately weighted, is representative of the original large dataset for a variety of uses. The pseudo-dataset is constructed by segmenting the data and then matching moments or using using likelihood-based methods to ensure that the properties of the reduced set match those of the larger set. However, the pseudo-datapoints generated typically are not points in the original dataset, and thus the methodology does not apply directly to our problem. One could imagine a re-derivation that restricts the pseudo-datapoints to be real datapoints, but we do not pursue that idea herein.

Instead, we turn to a more intuitive approach, that of mixture modeling, both in the context of clustering (Fraley and Raftery, 2002) and density estimation (Roeder and Wasserman, 1997). We fit a mixture of Gaussians, using an informative prior to help separate the components. We restrict the centers to be at observed datapoints, and then declare the fitted centers to be the representative sample. In Section 2 we describe our motivating example data. Section 3 details our selection methodology. Results are given in Section 4.

2 Circuit Device Experiments

This problem arose from an extended study at Sandia National Laboratories on photocurrent radiation aware models for electrical devices. The interest is in understanding current output as a function of the intensity of a pulse of gamma radiation applied to the devices. The current output is characterized by the peak amplitude reached during the experiment. The range of possible radiation doses is somewhat limited at any given testing facility, so experiments were run at three different facilities to span a broader range of possible dose rates. In principle, the relationship between dose rate and peak amplitude should not depend on the facility. Experimental runs were done at three different temperature settings. In principle, results could depend on temperature, and in fact separate computer simulation models are used for each temperature. Further details on the physical experiment are available in Gray et al. (2007).

The data for three of the devices is presented here. These three devices demonstrate the range of issues that arise in our analyses. We work with natural logarithm transformations of both the dose rate and the peak amplitude, as it helps linearize the problem. Figure 1 shows one device where there is a clear linear relationship between (log) dose rate and (log) peak amplitude, and no apparent effect from either temperature or facility. Figure 2 shows a device where the relationship is clearly linear, but the intercept appears to vary by facility. Again, temperature seems unimportant. Also, for facility 3, where the most experiments were run, there seem to be two parallel lines in the data. No information was available to us that could explain this phenomenon. Finally, Figure 3 shows a device where the relationship does not appear to follow any functional form, as there seem to be three separate branches on the right half of the plot. There is no clear effect from either temperature or facility. In summary, there appears to be no effect by temperature, despite an interest by the scientists in treating the temperatures separately. Typically there is no effect by facility, as expected by the experimenters, but occasionally the facilities do differ. And in some cases the relationship is clearly linear, but in others the relationship is not any functional form, as the same dose rate may correspond to several quite different possible outputs.

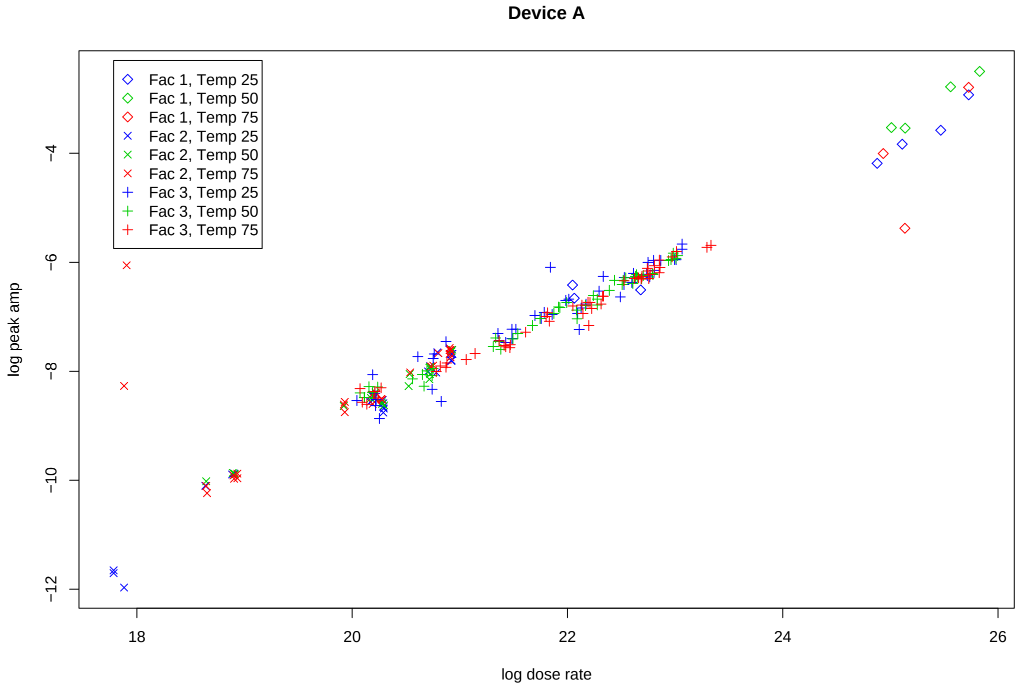

Figure 1: Electrical Output versus Dose Rate for Device A

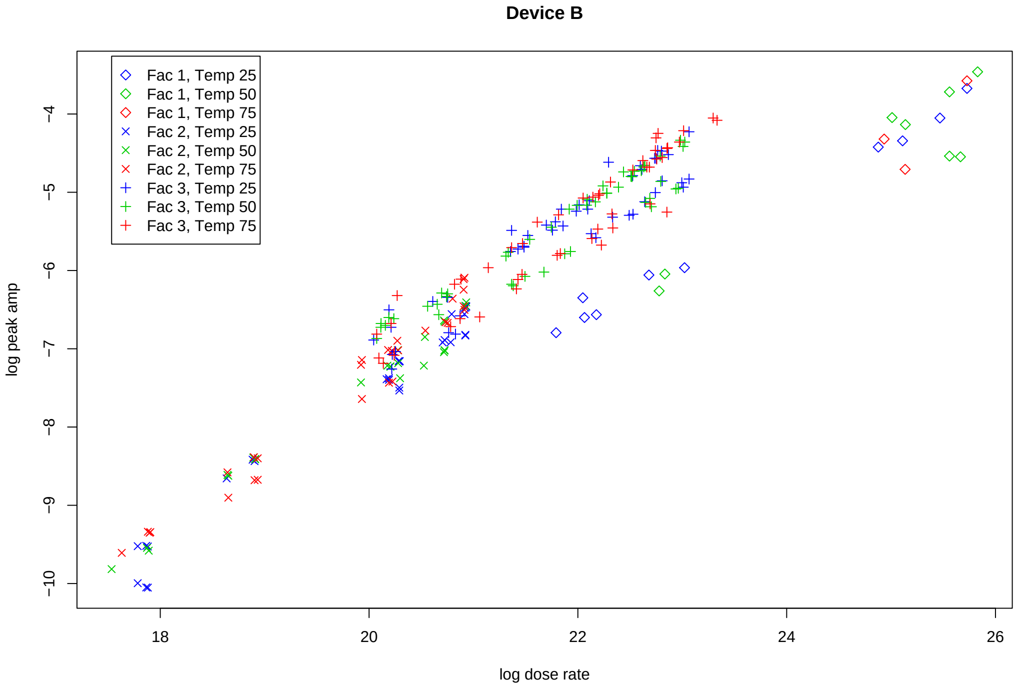

Figure 2: Electrical Output versus Dose Rate for Device B

3 Selection Methodology

Most clustering techniques view the formation of clusters as the only goal (Hartigan, 1975; Arabie et al., 1996; Duda et al., 2001, for example). Here we could first form clusters, then select one point from each cluster. However, it would make more sense to form the clusters with the intention of choosing a point from each, and thus take this into account during the clustering process. Thus we take the approach of fitting a mixture of normal densities to the data, but we restrict the centers of the distributions to the set of datapoints, and thus the fitted centers are the required sample. This approach is similar to standard mixture model or model-based clustering (Fraley and Raftery, 2002), but with the restriction on the cluster means directly addressing our need for finding representative points. This approach can also be thought of in the context of density estimation (Roeder and Wasserman, 1997), where the

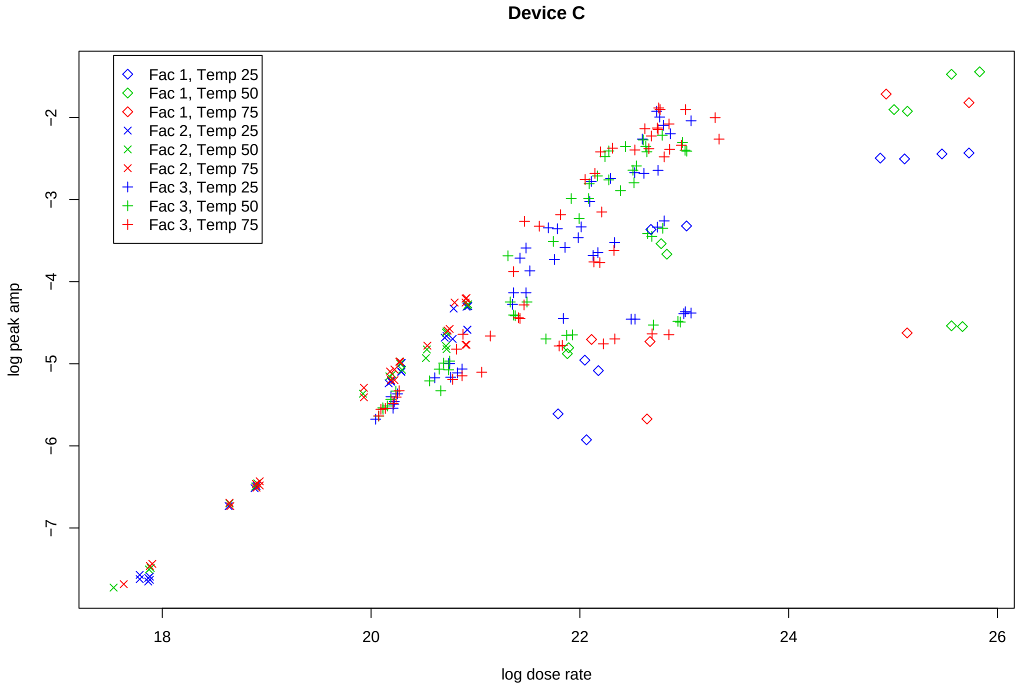

Figure 3: Electrical Output versus Dose Rate for Device C

density of the whole dataset is now being represented by distributions centered at a small set of observed datapoints.

Let k be the desired number of representative points. We fit a k-component mixture of Gaussians, restricted so that the centers of the components are data locations. Each component is allowed a different mean $\boldsymbol{\mu_j}$ and covariance $\boldsymbol{\Sigma_j}$, as well as a different weight p_j, $j = 1, \ldots, k$. We denote the observed data by (X_i, Y_i), $i = 1, \ldots, n$, where X is the explanatory variable or variables and Y is the response. In our case, X and Y are univariate (the log dose rates and the log peak amplitude respectively), and thus $\boldsymbol{\mu_j} = (\mu_1, \mu_2)$ and $\boldsymbol{\Sigma_j}$ are bivariate, although this setup clearly generalizes to multivariate X. The full likelihood for the model is:

$$f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{p}|\mathbf{x}, \mathbf{y}) = \left[\prod_{j=1}^{k} I_{\{\boldsymbol{\mu_j}=(x_{r_j}, y_{r_j})\}}\right] \sum_{j=1}^{k} p_j (2\pi)^{-n/2} \left|\boldsymbol{\Sigma_j}\right|^{-1/2}$$

$$* \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\left[\begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \mu_{1j} \\ \mu_{2j} \end{pmatrix}\right]^T \boldsymbol{\Sigma_j}^{-1} \left[\begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \mu_{1j} \\ \mu_{2j} \end{pmatrix}\right]\right)\right\}$$

where $\sum_{j=1}^{k} p_j = 1$, $I_{\{\}}$ is the indicator function that takes value one when its subscript set is true and zero otherwise, and $r_j \in \{1, \ldots, n\}$ for $j = 1, \ldots, k$. Thus the indicator functions specify in the likelihood that each component center must be a datapoint.

In order for the chosen centers to be representative of the whole dataset, we want them to not only be close to the centers of density, but also to be at least somewhat representative of the range of the data. In our case, we know that our collaborators plan to use the points for calibration of computer code, so it could be useful if the selected points are reasonably spread out across the range of values observed. If the data are spread evenly, this is not an issue. However, the data may be unevenly distributed, and thus we want to incorporate our desire for some spread in the model. We do this through the Bayesian paradigm, specifying a prior for the component means that spreads them out. We can also use a prior for the component weights that shrinks them towards equality at $1/k$, thus incorporating a desire that each selected point represents somewhat equal portions of the dataset. Of course, clumps in the data will mean that exactly equal portions are undesirable, so the Bayesian approach gives us an explicit, transparent, and coherent mechanism for balancing observed data features with modeling intentions. Thus, we take the prior for each μ_{hj} to be to a normal distribution centered $j/(k+1)$ of the way from the minimum observed value to the maximum observed value for each dimension $h \in \{x, y\}$. The choice of the covariance matrix for this normal prior specifies the degree to which we want separation in the selected points versus allowing the data to completely dictate the cluster centers. The prior for the weights, $\mathbf{p}$, is a Dirichlet distribution with all k parameters equal to the number of total observations divided by the number of representative points, i.e., n/k. The prior for each of the diagonal (variance) components of $\mathbf{\Sigma}$ is inverse-gamma, with different parameters for each element to reflect for the possibly different ranges of the X and Y data. The prior for the off-diagonal (correlation) components of $\mathbf{\Sigma}$ is uniform on $[0, 1]$ for the correlation (scaled by the diagonal elements to get the covariance matrix).

We fit the model with Markov chain Monte Carlo to find the maximum a posteriori (MAP) parameter values, and use the MAP component centers, $\boldsymbol{\mu}$, as the representative sample. Metropolis-Hastings proposals are used for all of the parameters, because the summation in the likelihood makes Gibbs sampling difficult. For $\mathbf{\Sigma}$, a multivariate random-walk style proposal is used to propose an update for the whole matrix in a single step, with standard Metropolis-Hastings accept/reject procedures. For updating the mixing proportions, the proposal $\mathbf{p}^*$ is drawn from a Dirichlet distribution centered at the current values $\mathbf{p}_l$, i.e., $\mathbf{p}^* \sim Dir\left(p_{l1}, \ldots, p_{lk}\right)$. For updating the center points $\boldsymbol{\mu}$, it is helpful to sort the data, so that the proposals for moving the centers can be done with respect to the ordered data. Each $\boldsymbol{\mu_j}$ is updated individually, with a proposal to move it with equal probability to the closest datapoint in each direction (in a one-dimensional representation, this would be one point on each side; in a two-dimensional problem, one could move along coordinate axes, or one could reduce the problem to one dimension by moving along the first principal component). In some problems, better mixing may be achieved by allowing a move to several points away.

As with most mixture model problems, the posterior is rather multi-modal, and one can easily become stuck in local modes. Thus it is helpful to do a number of runs with different starting values. We have run each model at least ten times and taken the result of highest posterior probability.

4 Results

Our collaborators asked for $k = 6$ points for each temperature level for each device. We included data from all facilities together, and did not label the points within the selection algorithm, on the premise that if the facilities are interchangeable as they were supposed to be, then the points could be selected without regard to the facility. However, if the facilities did turn out to be different, then we would expect separate clusters to form for different facilities, and that is exactly what we see in those cases. Thus the issue of the different facilities is automatically accounted for by the clustering.

We show here three examples of the clusters and representative points that were found. In each case, the log dose rates are on the X axis and the log peak amplitude responses are on the Y axis. The ellipses are the contours at one standard deviation for each of the normal mixture components.

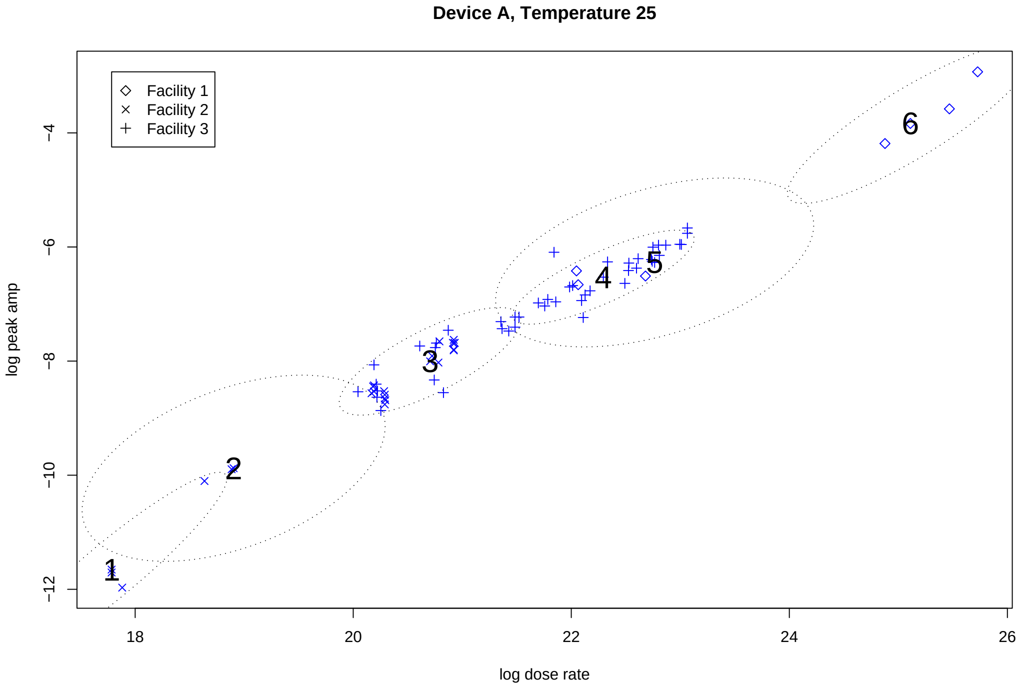

Figure 4: Representative Samples and Clusters for Device A

Figure 4 shows device A at temperature 25, where the data are fairly linear. The chosen points are fairly spread out, although points 4 and 5 are relatively close to each other, which takes into account the larger amount of data in that region. Clusters 4 and 5 overlap, because of both the linear structure and the concentration of the data. Note that cluster 3 includes data from two different facilities, and that facilities do appear to be interchangeable for this device. For all of the mixture components, the fitted covariance matrices show a strong correlation, as would be expected for linearly-related data.

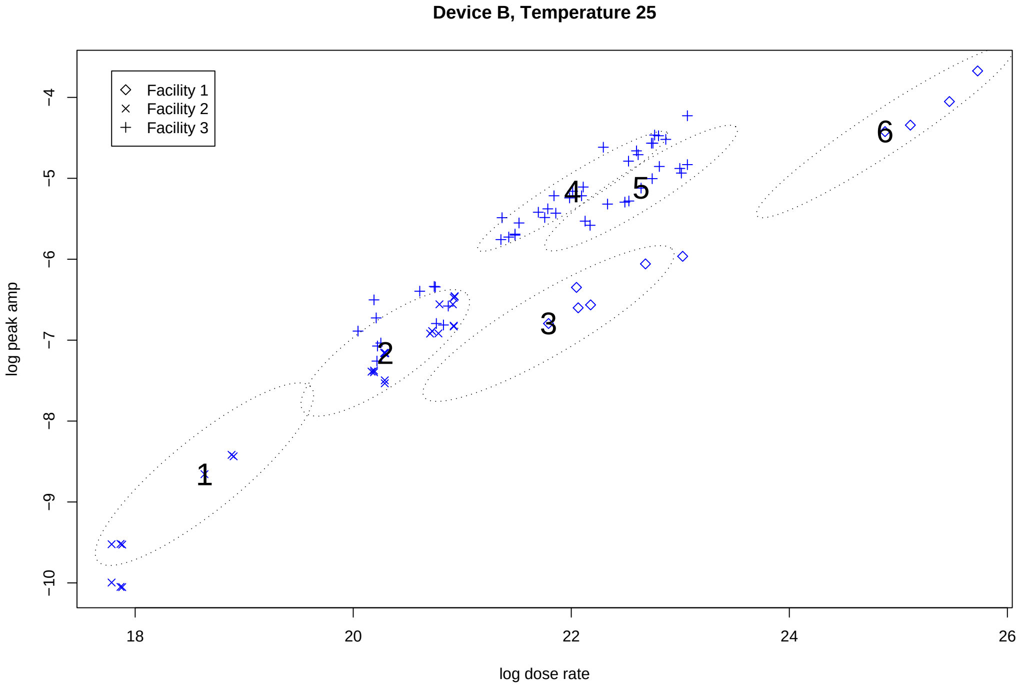

Figure 5: Representative Samples and Clusters for Device B

Figure 5 shows device B at temperature 25. While cluster 2 encompasses both facilities 2 and 3, clusters 3 and 6 separate facility 1 from facility 2 in clusters 4 and 5. Also note that clusters 4 and 5 pick up the parallel line fits that were seen in Figure 2. Thus the selected points do clearly represent the key aspects of this dataset.

Finally, Figure 6 shows device C at temperature 75, where the relationship between dose

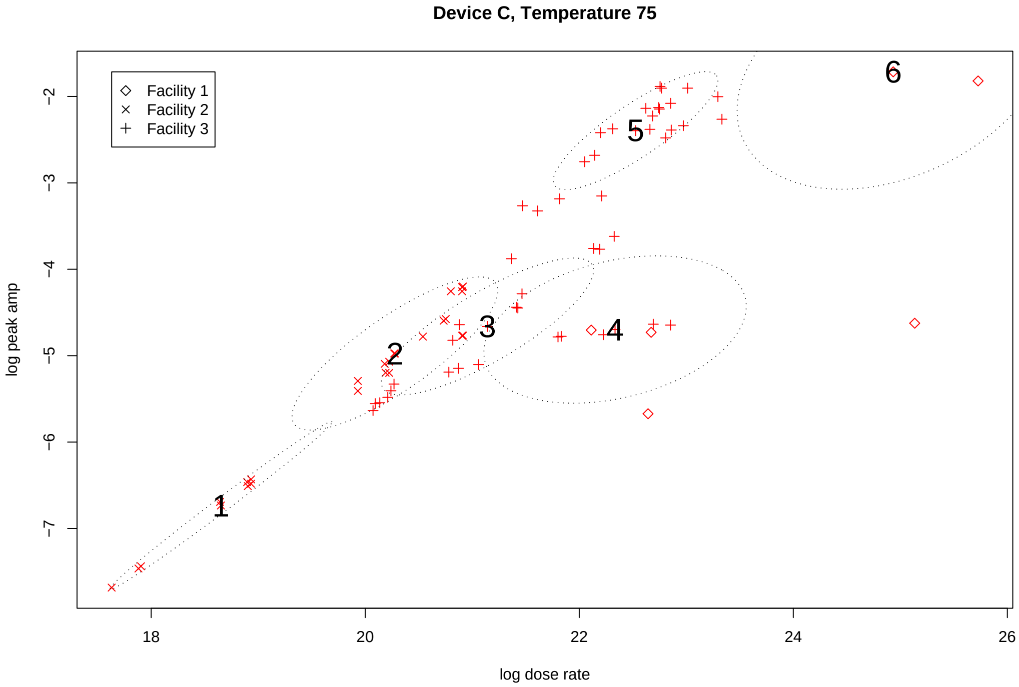

Figure 6: Representative Samples and Clusters for Device C

and response is much less clear. Cluster 1 has a strong correlation, encoding a tight linear relationship at low dose rates. Clusters 2, 3, and 5 represent points in the main linear trend, although there is more variability around the trend than in the previous examples. Cluster 4 models the lower branch of points away from the main trend, which includes observations from both facilities 1 and 3. Cluster 6 models points from facility 1 which are on an upper branch away from the main trend.

5 Conclusions

The mixture model approach allows us to cluster the data with the ultimate goal of using the cluster centers as points representative of the clusters and of the dataset as a whole. Working within the Bayesian paradigm allows us to formally encode desirable properties of representation, such as spread across the domain and balance in cluster sizes. Our formulation provides a general solution, not tied to any specific utility function, and thus applicable when a utility function cannot be formally specified. Hence, these sets are still useful if the model is updated and needs to be re-calibrated.

Our examples showed that the mixture approach can effectively segment the data in both the case of a strong functional relationship, as well as in the case where the relationship is non-functional. It can automatically account for latent variables when appropriate (e.g., facility in Figure 5, or the separation of clusters 4 and 5 in Figure 5 where the reason for the separation is not known to us). We also note that our collaborators used the resulting data sets to successfully perform calibrations. Prior to this work, the calibration sets were chosen using expert opinion about which experiments might provide the most useful information. The calibration results with the sets chosen by the method described here led to an improved calibration process (Gray et al., 2008).

This approach is clearly generalizable to the case where the number of points is not specified ahead of time, but where one has an approximate idea of how many points are desired. A prior can be placed over the number of mixture components, which would then be treated as another unknown. Fully Bayesian inference extends naturally in this setting.

It is worth noting that the points chosen are dependent on the entire dataset, and thus dependent on the points in the validation sample. While traditionally one may select randomly from the data to ensure independence between the training and test sets, that requires a larger training sample. By selecting the points carefully, we can represent the dataset with a smaller training sample, at the cost of dependence with the test set.

Our methods give intuitively reasonable results for the two-dimensional motivating example. We expect that they will also perform well in higher dimensions, when the results may not be easily visualizable and automated methods are more critical. Finally, we note that the data-centered clustering concept may generalize to other sorts of problems, such as stratified sampling.

Acknowledgments

The authors would like to thank the editor and two reviewers for their helpful suggestions that have improved this paper.

References

Arabie, P., Hubert, L. J. and De Soete, G. (eds.) (1996) *Clustering and Classification.* Singapore: World Scientific.

Campbell, K. (2002) A brief survey of statistical model calibration ideas. *Tech. Rep. LA-UR-02-3157*, Los Alamos National Lab.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001) *Pattern Classification.* New York: John Wiley and Sons.

Dumouchel, W., Volinsky, C., Johnson, T., Cortes, C. and Pregibon, D. (1999) Squashing flat files flatter. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, 6–15. AAAI Press.

Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611–631.

Gray, G. A., Martinez-Canales, M., Lam, C., Owens, B. E., Hembree, C., Beutler, D. and Coverdale, C. (2007) Designing dedicated experiments to support validation and calibration activities for the qualification of weapons electronics. In *Proceedings of the 14th NECDC*. Also available as Sandia National Labs Technical Report SAND2007-0553C.

Gray, G. A., Taddy, M., Griffin, J. D., Martinez-Canales, M. and Lee, H. K. H. (2008) Hybrid optimization: A tool for model calibration. *Tech. Rep. SAND2008-0145J*, Sandia National Labs, Livermore, CA.

Hartigan, J. (1975) *Clustering Algorithms*. New York: John Wiley and Sons.

Jolliffe, I. T. (1986) *Principal Component Analysis*. New York: Springer-Verlag.

Kennedy, M. C. and O'Hagan, A. (2001) Bayesian calibration of computer models. *J. Royal Statistical Society*, **63**, 425–464.

Kruskal, J. B. (1976) The relationship between multi-dimensional scaling and clustering. In *Classification and Clustering: Proceedings of an Advanced Seminar Conducted by the Mathematics Research Center, the University of Wisconsin-Madison, May 3-5, 1976* (ed. J. V. Ryzin), 7–44. New York: Academic Press.

Madigan, D., Raghavan, I., Dumouchel, W., Nason, M., Posse, C. and Ridgeway, G. (2002) Likelihood-based data squashing: a modeling approach to instance construction. *Data Mining and Knowledge Discovery*, **6**, 2002.

Moore, C. and Doherty, J. (2005) Role of calibration in reducing model predictive error. *Water Resources Res.*, **41**.

Müller, P., Sansó, B. and de Iorio, M. (2004) Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*, **99**, 788–798.

Oberkampf, W. L., Trucano, T. G. and Hirsch, C. (2003) Verification, validation, and predictive capability. *Tech. Rep. SAND2003-3769*, Sandia National Labs, Albuquerque, NM.

Owen, A. (2003) Data squashing by empirical likelihood. *Data Mining and Knowledge Discovery*, **7**, 101–113.

Roeder, K. and Wasserman, L. (1997) Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894–902.

Späth, H. (1980) *Cluster Analysis Algorithms for Data Reduction and Classification of Objects.* New York: John Wiley & Sons.

Trucano, T., Swiler, L., Igusa, T., Oberkampf, W. and Pilch, M. (2006) Calibration, validation, and sensitivity analysis: what's what. *Reliability Engineering and System Safety*, **91**.

Trucano, T. G., Pilch, M. and Oberkampf, W. L. (2002) General concepts for experimental validation of ASCI code applications. *Tech. Rep. SAND2002-0341*, Sandia National Labs, Albuquerque, NM.