Determination of Hidden Extrapolations via Gaussian Mixture Models

Ross Pivovar
BWXT Nuclear Energy, Inc. 109 Ramsey Place, Lynchburg, VA 24501
Email: rpivovar@bwxt.com

Testing and experimentation can, in general, be a costly or time consuming endeavor and, as such, companies are motivated to minimize costs. This often causes engineers to merge many different data sets, resulting in imbalanced multidimensional spaces of data. As a consequence, it becomes increasingly difficult to locate where hidden extrapolations have occurred when using these amorphous data sets. Thus, there is a need for a more rigorous method in extrapolation determination without requiring engineers to spend large of amounts of time parsing multidimensional data. A method relying on Gaussian Mixture Models for hidden extrapolation determination is presented in this paper.

I. INTRODUCTION

The nuclear industry is filled with data sets that are amalgamations of any and all available test data, ranging from data older than the 1970s through the present day. Many of the phenomena that impact safety calculations are quite difficult to accurately predict with mechanistic models, thereby requiring empirical models. Furthermore, new data is often difficult to obtain due to cost, time constraints, physical limitations, or safety.

Thus, over time, data sets grow into amorphous imbalanced shapes that require complex regressions/correlations. This becomes a problem particularly when developing models that rely on variables that are themselves functions of other measured variables, such as non-uniform power shape factors used in critical heat flux correlations, fluid qualities, enthalpies, etc.

Due to the multivariate nature of the phenomena of interest, when performing engineering calculations, it can be problematic for a human to know if they have accidently ventured outside the range of the available data and extrapolated into regions of unknown behavior. All approved NRC models rely on one dimensional applicability ranges (i.e. range of applicable mass fluxes, qualities, pressures, etc.).

This scenario has been referred to as a "hidden extrapolation." Naively, one may expect simply staying within the one dimensional range of available data prevents any extrapolation from being an issue. However, as illustrated in Fig. 1, it is quite possible to use an empirical model within the range of the data, yet still extrapolate outside the region of space occupied by the data.

The classical technique for hidden extrapolation detection is via the examination of the leverages, which are discussed in Section II. Unfortunately, this technique does not work well when the data set is amorphous and imbalanced. In Sections III and IV, this paper utilizes the more modern method of Gaussian Mixture Models to detect hidden extrapolations.

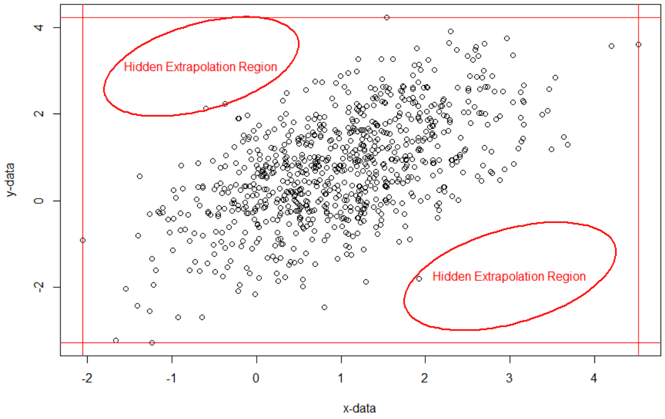

Fig. 1. Depiction of hidden extrapolations.

II. LEVERAGES

The leverages of a sample data set are defined in Eq. (1) (see Kutner et al.[1] for more details) in which X is an n × (1+p) matrix, where n is the number of data points and p are the independent variables with the first column equaling ones for the intercept. Conceptually, data leverages can be thought of as a "lever" or line drawn between a specific data point and the centroid of the data. The further the data point is from the centroid, the more influence this data point will have over a regression fit to the data set (i.e. greater leverage).

$$h = diag(X_i'(X'X)^{-1}X_i) \qquad (1)$$

As recommended by Kutner et al.[1], a data point (i.e. a new 1 × (1+p) vector X_{new}) of interest is likely a hidden extrapolation if the leverage of the new data point (Eq. (2)) is well outside the range of the leverages calculated in Eq. (1).

$$h_{new} = (X'_{new}(X'X)^{-1}X_{new}) \qquad (2)$$

The general idea of using leverages for extrapolation detection is quite useful since it reduces data of any number of dimensions down to one dimension, which is easily understandable by a human.

Unfortunately the major shortcoming of this method is the assumption of the hyper-ellipsoid. Reviewing only the maximum and minimum leverages is an excellent metric for hidden extrapolation detection only if the multivariate data exhibits a shape similar to a multivariate normal distribution. As described in Section I, such data is quite often not available in the nuclear industry.

III. BRIEF DESCRIPTION OF GAUSSIAN MIXTURE MODELS

The key difference between using leverages and Gaussian Mixture Models (GMM) for extrapolation detection is the assumption of many multivariate normal distributions instead of one. A GMM is a model-based clustering algorithm in which the data is modeled as a mixture of many different normal (Gaussian) distributions. The available data is broken down into subgroups that resemble hyper-ellipsoids.

A GMM was specifically chosen to provide a statistical model of the data. The well known kmeans[2] clustering was not used due to the expectation of heterogeneous covariance structures. The myriad of hierarchical clustering methods (the most common are discussed in James et al.[3]) were not used as they do not readily provide a model for inferences. Other model based clustering such as self-organizing maps[4] were not used because they provided a less direct means of analyzing the statistical significance of sparse regions.

Within the context of this paper, the R package mclust[5] was used to fit the GMM to our hypothetical example data. The details of the mclust package are described in the technical report by Fraley et al.[6]. The two most relevant equations to accomplish the objective are Eq. (3) (the joint distribution of the data), which is the definition of the mixture model, and Eq. (4), which is a standard multivariate normal distribution (see Appendix of Dempster et al.[7] for additional GMM details).

$$\prod_{i=1}^{n}\sum_{k=1}^{G}\tau_k\phi_k(x_i|\mu_k\,\Sigma_\mathrm{k}) \qquad (3)$$

$$\phi_k(x_i|\mu_k\,\Sigma_\mathrm{k}) = (2\pi)^{-\frac{p}{2}}|\Sigma_\mathrm{k}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x_i-\mu_k)^T{\Sigma_\mathrm{k}}^{-1}(x_i-\mu_k)\right\} \quad (4)$$

GMMs work by assuming G number of multivariate normal distributions for a data set of size n and p dimensions (e.g. $p = 3$ if analyzing data with three variables such as pressure, mass flux, and enthalpy). Then the expectation-maximization (EM) algorithm (originally developed by Dempster et al.[7]) is used to maximize the joint probability given in Eq. (3) by finding the best distribution parameters τ_k, μ_k, Σ_k for the assumed number of G distributions.

These steps are then repeated for various numbers of G distributions, in which the Bayesian Information Criterion (BIC), a model selection criteria that attempts to provide guidance on the most accurate model that does not overfit the sample data (see Schwarz and Gideon[8]) is recorded for each value of G. The value of G that results in the largest BIC is considered to be the optimal number of Gaussian (normal) distributions to fit the available data while minimizing the chance of overfitting the data.

III.A. COMMENT ON GEOMETRIC HULLS

A problem tackled in geometric computation is the determination of the "data hull," which is defined by drawing a multidimensional shell around the bounds of the data. Geometric hulls may seem to answer the objective of understanding the bounds of the data, but it assumes that all regions of data are of equal density, which is typically not true for empirical data. Hence, geometric hulls will not differentiate between the bulk of the data and outliers and sparse regions of data.

There are also conceivable scenarios in which the hull will be too conservative in that it may assume a fluid dynamic system could make non-physical jumps. For these reasons, geometric hulls were not further investigated for hidden extrapolation detection.

IV. APPLICATION OF GAUSSIAN MIXTURE MODELS

To fully demonstrate the utility of a GMM, a random amorphous data set is generated, shown in Fig. 2. While the data provided in Fig. 2 is synthetic, it does represent a data shape that is reasonably realistic for the highly controlled experiments performed in the nuclear industry. Included in Fig. 2 is a GMM with three groups. As shown in Fig. 3, the mclust package automatically runs a series of groups to find the best model based on BIC that fits the data. In addition, the package tests many different EM models used during the fitting of the data. These models make various assumptions like equal variance, unequal variance, equal-volume spherical, etc. A GMM with three groups (i.e. multivariate normal distributions) was found to best fit the sample data.

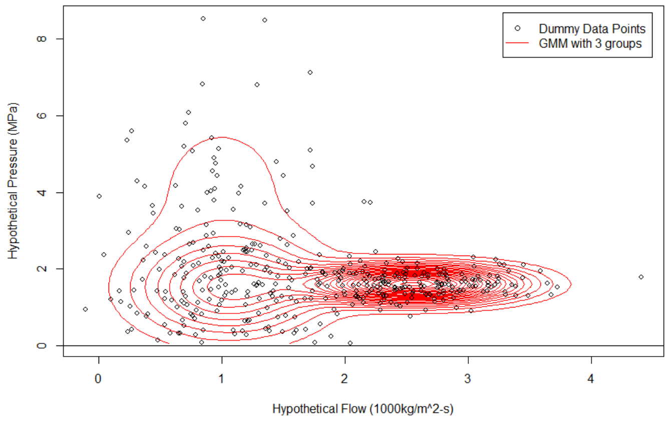

Fig. 2. The best GMM with three groups overlaid with a hypothetical data set.

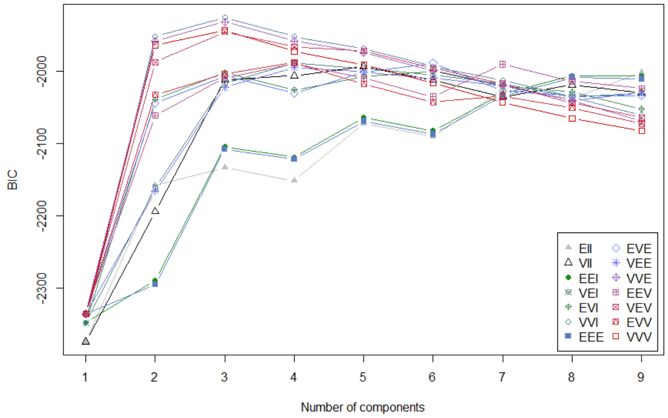

Fig. 3. Various EM models tested for different numbers of multivariate normal distributions for the Gaussian mixture model.

Now that a probability density function that models the data has been determined, the probability of a new vector actually coming from the multi-dimensional space can be assessed. If the probability associated with the new vector is extremely low, then either 1) we have no data in this subregion or 2) the new vector is in a very sparse region of data.

The problem of integrating a mixture density function must be solved, which will require numerical techniques. Monte Carlo importance sampling provides a robust and well-defined method for obtaining fast results that can be inserted into any engineering code. The Monte Carlo sampling portion can take some time to compute, but these results can be saved for later use and enable virtually instantaneous integration.

There are many references that can provide in-depth theory for Monte Carlo importance sampling, but the following discussion describes specifically how the GMM can be integrated. Eq. (5) provides the integral for a two dimensional example in which the probability that the new data point (c_1,c_2) could have come from the distribution of our tested region of data. Note that two dimensions were used only to demonstrate the approach and this GMM method can be used for any number of dimensions.

$$P(x > c_1, y > c_2) = \int_{c_1}^{\infty} \int_{c_2}^{\infty} \sum_{k=1}^{G} \tau_k \phi_k(x, y|\mu_k\, \Sigma_\mathrm{k})\, dydx \qquad (5)$$

The importance sampling version of Eq. (5) is then given in Eq. (6), where m is the Monte Carlo sample size and x is sampled from one of the G multivariate normal distributions in which the probability of sampling from each of the distributions is given by τ as calculated by the EM algorithm. In other words, $x^{(j)} \sim N(\mu_k, \Sigma_\mathrm{k})$, where there is a τ_k probability of selecting the k distribution of the GMM.

$$P(x > c_1, y > c_2) = \frac{\sum_{j=1}^{m} \sum_{k=1}^{G} \tau_k \phi_k(x^{(j)} > c_1, y^{(j)} > c_2 | \mu_k\, \Sigma_\mathrm{k})}{\sum_{j=1}^{m} \sum_{k=1}^{G} \tau_k \phi_k(x^{(j)}, y^{(j)} | \mu_k\, \Sigma_\mathrm{k})} \qquad (6)$$

Thus, with a fairly small number of samples (much less than 10^5) Eq. (5) can be solved numerically and the desired probability obtained, which takes a few minutes to calculate. Each repeat use of the generated Monte Carlo data does not require resampling and only needs the numerator of Eq. (6) to be summed up across the desired region of space, which takes less than a second to calculate. Hence, any engineering code on a modern computer should be able to make online calculations for extrapolation determination.

As an example using a dummy data set, if a regression model based on our sample data set were to be used at a mass flux and pressure of (1.5, 2.0), there is a 0.395 probability that this statepoint came from the tested region of data (i.e. it is very likely). However, using a mass flux and pressure of (2.0, 5.0), there is a 6.2×10^{-8} probability that this statepoint came from the tested region of data (i.e. extremely unlikely).

As a point of comparison to the leverages discussed in Section II, the leverages for a grid of possible statepoints was compared against the criteria discussed in Section II. Fig. 4 displays the results of this calculation, demonstrating the inadequacy of leverages for our imbalanced data set.

Fig. 5 uses the same grid of data with the GMM approach in which a fairly accurate assessment of which statepoints are likely hidden extrapolations can be seen.

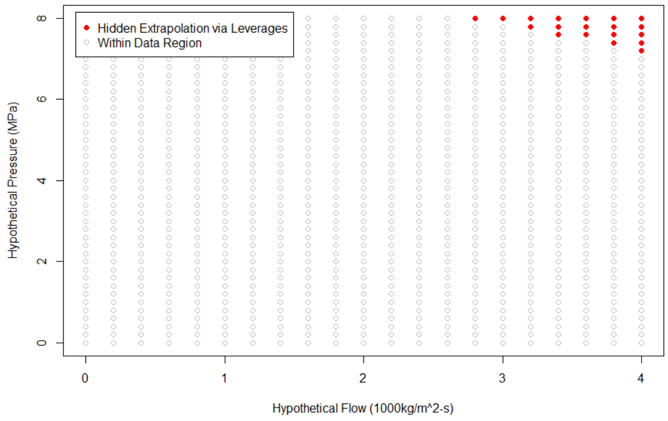

Fig. 4. Hidden extrapolation detection using the leverage approach.

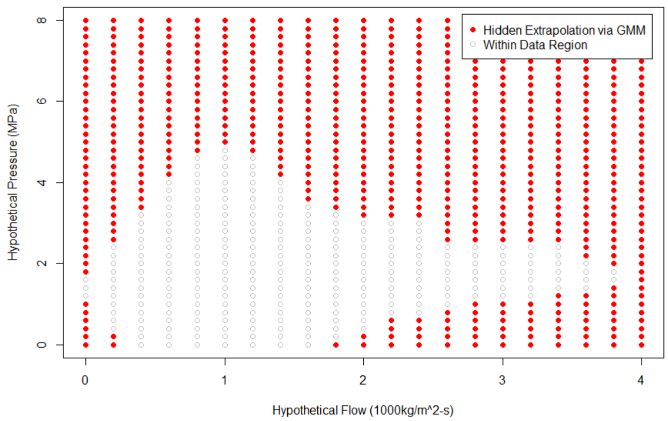

Fig. 5. Hidden extrapolation detection using the GMM approach.

V. GENERAL GUIDANCE FOR GMM EXTRAPOLATION DETERMINATION

1. The inequalities for the calculation of the Eq. (6) probabilities need to be directed away from the centroid of the data to yield a meaningful value. Hence, the minimum of 2^p combinations of inequalities need to be included in the algorithm that checks for extrapolations for each new vector. It is recommended a computer algorithm is setup to perform this calculation as data sets with greater than three dimensions become exceedingly difficult to know where the centroid is based only on visual inspection.

   Example: if we have five measured variables $X_1, X_2 \dots X_5$, the minimum probability, i.e. the solution of Eq. (6) that is pointing away from the centroid is one of the 32 inequality combinations:

$$P(X_1 > c_1, X_2 > c_2, \dots X_5 > c_5)$$
$$P(X_1 < c_1, X_2 > c_2, \dots X_5 > c_5)$$
$$P(X_1 > c_1, X_2 < c_2, \dots X_5 > c_5)$$
$$\vdots$$

   It is acknowledged that this method loses utility for ultra-high dimension datasets (e.g. greater than 20 dimensions).
2. Standard inferencing can be used as a first pass criterion, e.g. a probability greater than 0.05 can be said to come from the tested region of data with 95% confidence.
3. Probabilities less than a typical level of significance are not necessarily extrapolations, e.g. a probability of 0.0005 could be in a very sparse data region. It should also be expected that using regression models (such as Critical Heat Flux models) in these sparse data regions will exhibit appreciable increases in uncertainty of the predicted values. Thus, Eq. (6) probabilities that are less than the significance level but are greater than a user defined cutoff value (5×10^{-5} was used for the example data in this paper) might not be extrapolations, but should be used with care due to low data density. The exact cutoff value for declaration of an extrapolation was determined via trial and error through inspection of the available data. Engineers may need to calibrate the cutoff criteria to their particular data set.
4. Eq. (6) probabilities less than the cutoff value are very likely extrapolations and engineers should take great care in examining this region based on available data.

VI. EXAMPLE CASE WITH REAL DATA

The publically available EPRI Critical Heat Flux (CHF) data provided by Fighetti and Reddy[9] is an excellent and quite representative dataset often used or referenced in the nuclear industry. The Pressurized Water Reactor (PWR) data for Westinghouse designs was arbitrarily selected for this example.

This dataset provides the added complexity of data points being taken in clusters around specific pressure/mass flux combinations. Clusters of data points in CHF testing is not unusual as it limits stress on the experimental apparatus.

For this study, simply relying on the BIC to select the best structure of the GMM did not work well since the clustered data misleads the GMM algorithm into assuming the clusters are independent distributions with a very small probability of data existing between clusters. However, we know that the data points are covariates and we would expect interpolated points between clusters to exhibit similar behavior to the surrounding clusters of data. Thus the best model can be assessed manually. The ellipsoidal covariance structure with equal volume, shape, and orientation was selected based on visual trial and error (i.e. the 'EEE' model option in the mclust[5] R package). The selected data with the resulting fitted GMM are illustrated in Fig. 6.

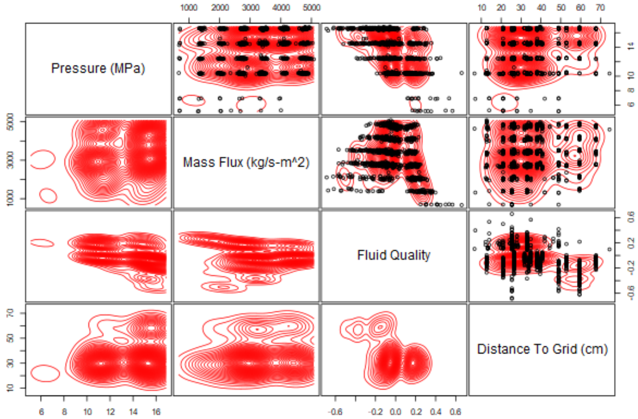

Fig. 6. Pair wise comparison of the EPRI data with a GMM fit to this dataset.

Fig. 6. clearly demonstrates this dataset includes many pockets of missing data. In the scenario that the NRC approved a CHF correlation for use with this particular dataset, the standard review procedure would always request justification of sparse regions or missing regions of data. For demonstration purposes, assume that this dataset is approved as is presented in Fig. 6. The NRC "applicability range" would be given as:

```
Pressure (MPa)             5.13    16.86
Mass Flux (kg/s-m^2)      670.6    5076.6
Fluid Quality           -0.6872    0.6593
Distance To Grid (cm)       6.6     74.51
```

However it would be quite difficult to ensure any transient simulation does not unintentionally venture into a missing data region. An example transient forcing function that stays within the bounds of the applicability range is given in Fig. 7 with a constant distance to grid of 35.6cm. As an added challenge to demonstrate the utility of a GMM, the transient ranges given in Fig. 7 were setup to stay within the bulk of the data provided in Fig. 6 (i.e. an attempt has been made to stay out of sparse data regions). Hence, similar to a real life situation, the engineer would attempt to stay within the bounds of the data.

After the GMM has been fit, a possible cutoff probability needs to be determined. A total of 50,000 samples were taken from the GMM for use with summations in Eq. (6). The exact number of the samples depends on the complexity of the fitted GMM and should be increased until the probabilities converge.

An initial test point is considered to ensure the Monte Carlo integration is behaving as expected. Reviewing the probability of a random point coming from pressure < 13.8MPa, mass flux > 2030kg/s-m^2, any quality, and any distance to grid results in a probability of 0.347, which agrees with expectations (i.e. a large value). Next, as a calibration test point, a data point is selected from a region with very sparse data (e.g. pressures < 8.62 MPa, any mass flux, any quality, and any distance to grid), which results in a probability of 0.0025. This region includes a very small amount of data and thus it would be highly advised to review any data points that result in probabilities of approximately equal magnitude. Probabilities much less than this calibration value are very likely hidden extrapolations.

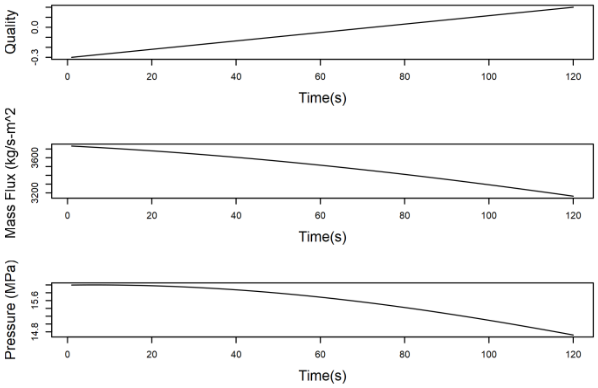

Fig. 7. Example variable transient simulation for demonstration of GMM utility

The calibrated GMM can now be tested with an example transient scenario in which the mass flux is decreasing simultaneously with pressure, and quality is increasing. The resulting probabilities of this transient coming from the approved database are shown in Fig. 8. The horizontal dashed line is drawn at half the calibration probability to ensure we are not being overly aggressive with our hunt for hidden extrapolations. Despite attempting to stay within the available dataset, simultaneously modifying three variables resulted in almost half the transient existing outside the approved database. The first half of the transients is within the available pressure and mass flux, but in a hidden quality extrapolation. The latter half of the transient is in a hidden extrapolation region due to the quality and distance to grid.

Thus, as a follow-up it would be prudent to review these newly identified areas and determine if any non-conservatism may exist in the resulting safety assessment.

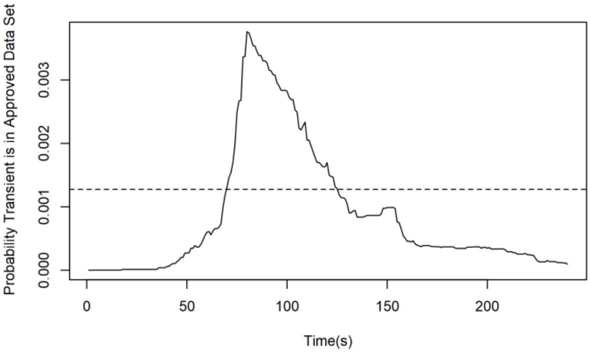

Fig. 8. Resulting probability of a transient coming from a dataset included in the approved database

ACKNOWLEDGMENTS

None.

REFERENCES

1. KUTNER, NACHTSHEIM, NETER, LI, *Applied Linear Statistical Models*, 5th ed, (2005).
2. HARTIGAN, J. A. and WONG, M. A. Algorithm AS 136: "A K-means clustering algorithm". *Applied Statistics*, **28**, 100–108 (1979).
3. JAMES, WITTEN, HASTIE, TIBSHIRANI, *An Introduction to Statistical Learning, (*2015).
4. KOHONEN, "The Self-Organizing Map", *Proceedings of the IEEE* **78** (9): 1464-1480 (1990).
5. SCRUCCA L., FOP M., MURPHY T. B. and RAFTERY A. E. "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models". *The R Journal 8/1*, pp. 205-233 (2017).
6. FRALEY, RAFTERY, MURPHY, SCRUCCA, MCLUST "Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation", *Technical Report no. 597*, Department of Statistics, University of Washington, (2012).
7. DEMPSTER, LAIRD, RUBIN, "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society*, Series B. **39** (1): 1–38. (1977).
8. SCHWARZ, GIDEON, "Estimating the dimension of a model", *Annals of Statistics*, **6** (2): 461–464 (1978).
9. FIGHETTI and REDDY, "Parametric Study of CHF Data Volume 1: Compilation of Rod bundle CHF Data Available at the Columbia University heat Transfer Facility", *EPRI NP-2609*, **1** Project 813 (1982).