Classification in Nuclear Forensics with Quantile Comparisons

Charles Weber and Ken Dayman
Oak Ridge National Laboratory

Presented at the
INMM Nuclear Materials Science, Processing and Signature Discovery Workshop
Richland, Washington
May 1-2, 2018
Nuclear Forensics Involves
Measurements and Inverse Calculations

- Measurements evaluate current conditions and observables
- Inverse calculations describe processes, reactions, and interactions that occurred in order to produce the given sample measurements
  - Simple—hand calculation
  - Complex—difficult model and inverse calculations
  - Processes not characterized well—data analytics
Quantile Comparisons (QC): Development of Method in Progress

• Applied to NMIP Pu database (2014-16)
• Used in NA-22 Reactor Venture (2015-17)
• Developed further under DHS NTNFC (2016-17)
  – Primary sponsor for current work
• Quantiles: median, quartile, 90th percentile
Quantile Comparisons (QC): Development of Method in Progress

- Applied to NMIP Pu database (2014-16)
- Used in NA-22 Reactor Venture (2015-17)
- Developed further under DHS NTNFC (2016-17)
  - Primary sponsor for current work
- Quantiles: median, quartile, 90\textsuperscript{th} percentile
  - Implies ranking (for one dimension)
    
    \[
    ^{235}\text{U} / ^{238}\text{U} = \{3.9, 4.1, 4.4, 4.9\} \text{ (\%)}
    \]
Quantile Comparisons (QC): Development of Method in Progress

- Applied to NMIP Pu database (2014-16)
- Used in NA-22 Reactor Venture (2015-17)
- Developed further under DHS NTNFC (2016-17)
  - Primary sponsor for current work
- Quantiles: median, quartile, 90\textsuperscript{th} percentile
  - Implies ranking (for one dimension)
    \[
    \frac{^{235}\textnormal{U}}{^{238}\textnormal{U}} = \{3.9, 4.1, 4.4, 4.9\} \text{ (%)}
    \]
  - Interpretation for multiple dimensions not apparent
    \[
    \left(\frac{^{235}\textnormal{U}}{^{238}\textnormal{U}}, \frac{^{240}\textnormal{Pu}}{^{239}\textnormal{Pu}}\right) = \{(4,8), (5,7)\} \text{ (%)}
    \]
Quantile Comparisons

**Quantiles for Multivariate Problems**

- Most data contain multiple variables:
  \[ X = (x_1, \ldots, x_d)^T \quad Y = (y_1, \ldots, y_d)^T \]
  - Single-variable notion of quantile not extendable
  - Samples with multiple specimens \( X_i, Y_i \), compare the two distributions

- Method of Dhar and Chaudhuri
  - Calculate Spatial Rank:
  \[ U_j = \frac{1}{N} \sum_{i=1}^{N} \frac{(X_j - X_i)}{|X_j - X_i|} \]
Quantiles for Multivariate Problems

- Most data contain multiple variables:
  \[ X = (x_1, \ldots, x_d)^T \quad Y = (y_1, \ldots, y_d)^T \]
  - Single-variable notion of quantile not extendable
  - Samples with multiple specimens \( X_i, Y_i \), compare the two distributions

- Method of Dhar and Chaudhuri
  - Calculate Spatial Rank:
    \[ U_j = \frac{1}{N} \sum_{i=1}^{N} \frac{(X_j - X_i)}{|X_j - X_i|} \]
  - Inverse Spatial Rank in \( Y \) distribution:
    - What specimen in the \( Y \) distribution would produce \( U_j \)?
Quantiles for Multivariate Problems

• Most data contain multiple variables: 
  \[ X = (x_1, \ldots, x_d)^T \quad Y = (y_1, \ldots, y_d)^T \]
  – Single-variable notion of quantile not extendable
  – Samples with multiple specimens \( X_i, Y_i \), compare the two distributions

• Method of Dhar and Chaudhuri
  – Calculate Spatial Rank:
    \[ U_j = \frac{1}{N} \sum_{i=1}^{N} \frac{(X_j - X_i)}{|X_j - X_i|} \]
  – Inverse Spatial Rank in \( Y \) distribution: 
    \[ U_j = \frac{1}{M} \sum_{i=1}^{M} \frac{(\bar{X}_j - Y_i)}{|\bar{X}_j - Y_i|} \]
    • What specimen in the \( Y \) distribution would produce \( U_j \)?
    • Convex nonlinear optimization—unique solution exists for \( \bar{X}_j \)
Quantiles for Multivariate Problems

• Most data contain multiple variables:
  \( X = (x_1, \ldots, x_d)^T \quad Y = (y_1, \ldots, y_d)^T \)
  
  – Single-variable notion of quantile not extendable
  
  – Samples with multiple specimens \( X_i, Y_i \), compare the two distributions

• Method of Dhar and Chaudhuri
  
  – Calculate Spatial Rank:
    \[
    U_j = \frac{1}{N} \sum_{i=1, i \neq j}^{N} \frac{(X_j - X_i)}{|X_j - X_i|}
    \]
  
  – Inverse Spatial Rank in \( Y \) distribution:
    \[
    U_j = \frac{1}{M} \sum_{i=1}^{M} \frac{(\bar{X}_j - Y_i)}{|\bar{X}_j - Y_i|}
    \]
  
    • What specimen in the \( Y \) distribution would produce \( U_j \)?
  
    • Convex nonlinear optimization—unique solution exists
  
  – Sum-of-squared-error measures deviation in the two distributions:
    \[
    S_{XY} = \sum_{i=1}^{N} (\bar{X}_i - X_i)^T (\bar{X}_i - X_i)
    \]
Quantiles for Multivariate Problems

- Most data contain multiple variables: 
  $$X = (x_1, \ldots, x_d)^T, \quad Y = (y_1, \ldots, y_d)^T$$
  - Single-variable notion of quantile not extendable
  - Samples with multiple specimen \(X_i, Y_i\), compare the two distributions

- Method of Dhar and Chaudhuri
  - Calculate Spatial Rank: 
    $$U_j = \frac{1}{N} \sum_{i=1}^{N} \frac{(X_j - X_i)}{|X_j - X_i|}$$
  - Inverse Spatial Rank in \(Y\) distribution: 
    $$U_j = \frac{1}{M} \sum_{i=1}^{M} \frac{(\tilde{X}_j - Y_i)}{|\tilde{X}_j - Y_i|}$$
    - What specimen in the \(Y\) distribution would produce \(U_j\)?
    - Convex nonlinear optimization—unique solution exists
  - Sum-of-squared-error measures deviation in the two distributions: 
    Same analysis for \(Y\) distribution
    $$S_{XY} = \sum_{i=1}^{N} (\tilde{X}_i - X_i)^T (\tilde{X}_i - X_i) + \sum_{i=1}^{M} (\tilde{Y}_i - Y_i)^T (\tilde{Y}_i - Y_i)$$
Classification Using Quantile Comparisons

• $X_i$ represents an unknown test sample
• Several $Y_i$ drawn from known classes: $Y_i^a, Y_i^b, Y_i^c, ...$
• Calculate scores for comparisons: $S_{XY^a}, S_{XY^b}, S_{XY^c}, ...$
  – Lowest score is the likely class
• Better: many repetitions by drawing many samples from known classes
  – Compare average scores: $\bar{S}_{XY^a}, \bar{S}_{XY^b}, \bar{S}_{XY^c}, ...$
• Probability of misclassification: $Pr(\bar{S}_{XY^a} - \bar{S}_{XY^b} > 0)$
  – Knowing Class $Y^a$ is correct
  – Student’s $t$ Distribution
Example: Burnup of Fuel Samples

- Gas-cooled reactor, 3-D core model
  - 990 depletion zones, 19 time points=19 burn classes

<table>
<thead>
<tr>
<th>Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burn time (days)</td>
<td>0</td>
<td>15</td>
<td>46</td>
<td>76</td>
<td>106</td>
<td>137</td>
<td>167</td>
<td>198</td>
<td>228</td>
<td>259</td>
<td>289</td>
<td>319</td>
<td>350</td>
<td>395</td>
<td>456</td>
<td>517</td>
<td>578</td>
<td>639</td>
<td>700</td>
</tr>
</tbody>
</table>

- Considerable variation as irradiation time increases
Fuel Samples—Classification Procedure

- A sample consists of multiple points (usually 20)
  - Each point is a vector of isotopic ratios: $^{234}\text{U}$, $^{235}\text{U}$, $^{236}\text{U}$, $^{238}\text{Pu}$, $^{240}\text{Pu}$, $^{241}\text{Pu}$, $^{242}\text{Pu}$, $^{134}\text{Cs}$, $^{133}\text{Cs}$
  - Denominators are $^{238}\text{U}$, $^{239}\text{Pu}$, $^{137}\text{Cs}$

- A Test Sample is chosen randomly from the inventories at one burn time
  - Pretend time is unknown
  - Samples are drawn from inventories at various burn times, including the time of the test sample
  - Each sample compared with test sample using QC method
  - Method with lowest score is likely class
  - Repetition for many samples from known classes
**Burn Time Classification Example**

<table>
<thead>
<tr>
<th>Class</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burn time (days)</td>
<td>198</td>
<td>228</td>
<td>259</td>
<td>289</td>
<td>319</td>
<td>350</td>
</tr>
</tbody>
</table>

- Test sample drawn from Class 10 (pretend class not known)
- Samples drawn from known classes 8, 9, 10, 11, 12, 13 (10 repetitions)
- Comparison of known Class 10 clearly superior to other classes
- Would classify Test Sample in Class 10
Repeat: Another Test Sample from Class 10

• Repeat tests against known classes 8, 9, 10, 11, 12, 13

• Calculate average scores for multiple tests
Repeat: Another Test Sample from Class 10

- Repeat tests against known classes 8, 9, 10, 11, 12, 13
- Calculate average scores for multiple tests
- Class 11 is best!!!
- Test sample exhibit unusual skewness?
Repeat: Another Test Sample from Class 10

- Repeat tests against known classes 8, 9, 10, 11, 12, 13
- Calculate average scores for multiple tests
- Class 11 is best!!!
- Test sample exhibit unusual skewness?
- Further repeats with many additional test samples from Class 10: all gave the correct class
  - Only one incorrect of 10 Test Samples
Probability of Misclassification

- Using average scores $\bar{S}_{10.10}, \bar{S}_{10.11}, \ldots$

- Given Test Sample from class 10, calculate $\Pr(\bar{S}_{10.11} - \bar{S}_{10.10} > 0)$

- Student’s t distribution: $T = \frac{\bar{S}_{10.11} - \bar{S}_{10.10}}{\sqrt{(\sigma^2_{10.11} + \sigma^2_{10.10})/N}}$
  
  - Average scores use $N$ points $\to N - 1$ degrees of freedom
  
  - Assumes $\bar{S}_{10.10}, \bar{S}_{10.11}$ are Normal
    
    - Approximately true for large enough $N$
    - Error for non-normality only 5-10 %

- Misclassification probability: $\Pr(t < T)$
Misclassification Using T Statistic

- Case 2 shows high likelihood of misclassification
- Most cases indicate misclassification improbable
Summary and Conclusions

• Quantile Comparisons is a multivariate analog to one-dimensional quantiles
• Used to compare sets of samples as if they represent statistical distributions
• Example problem in predicting irradiation time (or burnup) for irradiated fuel samples
• Probability of misclassification can be quantified
• Can render a “none-of-the-above” classification decision (demonstrated in previous work)