The Gini Unbound: Analyzing Class Inequality with Model-Based Clustering

Tim F. Liao

University of Illinois

Paper Prepared for the International Conference in Memory of Two Eminent Social Scientists: C. Gini and M. O. Lorenz, Their Impact in the XX-th Century Development of Probability, Statistics and Economics, Università degli Studi di Siena, 23-26 May 2005

Please treat as copy-righted material

Send Correspondence to: Tim F Liao, Department of Sociology, University of Illinois, Urbana, IL 61801, USA. Email: tfliao@uiuc.edu
The Gini Unbound: Analyzing Class Inequality with Model-Based Clustering

Tim F. Liao

University of Illinois

Abstract

For students of social and economic inequality, the most widely used measure is no other than the Gini index (or Gini inequality ratio). Whereas some other measures of inequality possess certain useful characteristics, such as the straightforward decomposability of the generalized entropy measures, the Gini index has remained the most popular, at least in part due to its ease in interpretation. However, the Gini index has a limitation in measuring inequality. It is less sensitive to how the population is stratified than how individual values differ. The twin purposes of this paper are to explain the limitation and to propose a model-based method—latent class/clustering analysis for understanding and measuring inequality. The latent cluster approach has the major advantages of being able to identify potential “classes” of individuals who share similar levels of income (or another attribute) and to assess the fit to the empirical data of alternative models of different assumptions and varying number of latent classes. We distinguish class inequality from individual inequality, the type that is better measured by the Gini. Once the classes are estimated, the number of estimated classes obtained from the best fitting model facilitates the decomposition of the Gini index into individual and class inequality. Class inequality is then measured by a stratification index based on the decomposition of the overall Gini into between-class and within-class inequality components. Therefore, the Gini index is extended and assisted by model-based clustering to measure class inequality, thereby realizing its potential applicability.
The Gini Unbound: Analyzing Class Inequality with Model-Based Clustering

Introduction

For students of social and economic inequality, the most widely used measure is no other than the Gini index (or Gini inequality ratio). The index or ratio is named after the Italian statistician Corrado Gini (1912), who first proposed this insightful measure of inequality. Together with its associated Lorenz curve, attributed to Lorenz (1905), the Gini has seen extremely widespread application in the social sciences. We are now approaching the centennial year of the Lorenz original publication, and there have been numerous alternative measures proposed, including the class of generalized entropy measures. Whereas the alternative measures of inequality may possess certain useful characteristics, such as the straightforward decomposability of the generalized entropy measures, the Gini index has remained the most popular, at least in part due to its ease for interpretation.

Research on the decomposition of the Gini has also gained momentum in recent years (Dagum 1997a, 1997b, 1998; Milanovic and Yitzhaki 2002; Mussard, Terraza, and Seyte 2003; Yao and Liu 1996; Yao 1999; Yitzhaki 1994).

However, the Gini index has a consequential limitation in measuring inequality that has not been previously discussed. It is less sensitive to how the distribution is stratified than to how individual values differ. The twin purposes of this paper are to explain this limitation and to propose an alternative, model-based method—latent class/clustering analysis to extend the Gini index for understanding and measuring stratified inequality.

The paper has six sections. The section following the introduction discusses the limitation of the Gini index in measuring inequality. I next present a form of latent class or clustering analysis that can not only avoid the limitation but also capture features of
stratification. A real-world example of income data is then examined to illustrate the proposed method and to compare it with the Gini index. In the section to follow I apply Gini decomposition by using cluster classification schemes suggested by the model-based method. The decomposition results are then used to form a stratification index for measuring the amount class inequality relative to the total amount of inequality. A few concluding remarks are offered in the final section.

**Gini’s Limitation in Measuring Inequality**

Let us use $y_i$ to indicate a random distribution such as income, and let us begin with a very simple simulated distribution of only six $y_i$ data points where $i=1\ldots6$. Suppose that we have a first distribution that contains the following observations:

1: [1, 1, 1, 4, 4, 4]

where the values are simply certain units of income. The Gini ratio for this distribution is 0.300000, and the value of the coefficient is not affected by sample size as long as the nature of inequality remains the same. Namely, repeating the 1s and the 4s a large number of times would result in an identical Gini value of 0.300000. This observation, however, will not hold for the case of complete equality, i.e., all observations having the same number of income units. But for our purposes here that point is irrelevant.

Next, we consider a second distribution, which is identical to Distribution 1 except for the last three observations, whose holding of income are all increased by one unit:

2: [1, 1, 1, 5, 5, 5]

The Gini index for the distribution is 0.333333, an 11% increase over the Gini of the first distribution. Now let us consider a third distribution:
3: [1, 1, 2.5, 2.5, 4, 4]

This time the two middle observations have an income level that is halfway between the first two and the last two observations in Distribution 1. The third distribution has a Gini value of 0.266667, an 11% decrease from the Gini for the first distribution. A comparison of the three simple distributions reveals something rather fundamental, and rather disturbing. Either Distribution 2 or 3 represents the same amount of departure from Distribution 1 in the value of its Gini index, but for understanding their distributions and measuring inequality, are Distributions 2 and 3 similarly different from Distribution 1? Of course not. Distribution 2 is almost the same as Distribution 1, other than cases 4, 5, and 6 are a bit richer. Unlike Distribution 1 (or 2), that contains two (social) classes, Distribution 3 has three (social) classes, and has different features of stratification and inequality from the other two distributions. To push the point a bit further, let us double the values of cases 4, 5, and 6 in Distribution 1, resulting in:

4: [1, 1, 1, 8, 8, 8]

The new distribution has a Gini index of 0.388889, a 30% increase over that of Distribution 1 even though the form of inequality is the same.

To further illustrate the relative insensitivity of the Gini to the shape of inequality or stratification, let us consider two more distributions:

5: [1, 2.3, 3.6, 4.9, 6.2, 7.5]

which has a Gini coefficient of 0.297386, and

6: [1, 1, 3, 3, 5, 5]

which has a Gini index of 0.296296. In Distribution 5, each observation after the first is 1.3 income units higher than the previous case, showing no clear patterns of class formation.
and only an even distribution over the scale. The distribution can be viewed as all cases belonging to one class or as they belonging to six individual classes. In Distribution 6, the cases form three distinctive classes. Interestingly, the three distributions of 1, 5, and 6 have an almost identical Gini value. (By fine-tuning the values in Distributions 5 and 6, we can obtain an exact match for the value of 0.300000 from Distribution 1, but that is not necessary here to show the point.) However, the form of inequality is drastically different across the three distributions in question.

Why is the Gini index so insensitive to changes in the form of inequality? To answer the question, let us review how the Gini is calculated. The Gini index is related to the Lorenz curve as twice the area between the 45-degree line and the Lorenz curve, and can be formally written as below (see, e.g., Chotikapanich and Griffiths 2001).

Let $\pi = F(y_i)$ indicate the distribution for $y_i$, and let $\eta = F_1(y_i)$ represent the corresponding first moment distribution function. The relation between $\eta$ and $\pi$, defined for $0 \leq y_i < \infty$, is the Lorenz curve, and relation can be denoted by $\eta = L(\pi)$. The Gini index can then be defined accordingly:

$$G = 1 - 2 \int_0^1 L(\pi)d\pi \quad (1)$$

And can also be written as

$$G = -1 + \frac{2}{\mu} \int_0^y yF(y)f(y)dy \quad (2)$$

There exits numerous computational formulae for implementing the Gini calculation. Perhaps the most revealing one is that used by Dagum (1997a, 1997b, 1998) and Mussard, Terraza, and Seyte (2003):
All the various computation formulae will give you the same results. However, (3) demonstrates that what is really measured by the Gini index is a weighted average of pairwise differences between the individual cases in the sample. It is the overall individual mean differences that matter, whether or not these individuals may fall into classes or clusters. Although recent research has shown that Gini’s mean difference is a superior measure of variability for non-normal distribution (Yitzhaki 2003), the Gini does not capture well the clustering nature of the data, a more important point for studying social stratification.

\[
G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |y_j - y_i|}{2n^2 \mu}
\]  

(3)

Model-Based Cluster Analysis

Cluster analysis can be viewed as a way to group similar objects, with unknown number of groups whose forms (i.e., cluster parameters) are also unknown (Kaufman and Rousseeuw 1990). Similarly, Everitt (1993) views cluster analysis as obtaining a useful division of objects into classes, whose numbers and properties are to be determined. These ideas convey the essence of analysis of social inequality, where the researchers seeks an understanding of the groups or social classes within which individuals are more similar than across these classes, judged by certain attributes such as income.

A recent development takes the approach of model-based clustering, which specifies a statistical model for the population from which the sample under study is assumed to have come. Model-based clustering has a number of advantages (Vermunt and Magidson 2002): First, the choice of the cluster criterion that is used to minimize the
within-cluster variation and maximize the between-cluster variation is less arbitrary in model-based clustering than in conventional cluster analysis; second, model-based clustering is flexible in allowing various simple and complex distributional forms for the observed variables within the clusters; furthermore, no scaling decisions have to be made about the observed variables in model-based clustering while in conventional cluster analysis scaling is always an issue.

Model-based clustering also allows the observed variables to be continuous or categorical (i.e., nominal or ordinal) because clusters can be treated as latent classes, thereby the method can be seen as latent class analysis. Here we consider only continuous observed variables in our model specifications because our concern in this paper is income. The basic model-based clustering specification takes the form

$$ f(y_i | \theta) = \sum_{k=1}^{K} \pi_k f_k(y_i | \theta_k), \quad (4) $$

where $y_i$ represents an individual’s scores on a set of observed variables, $K$ denotes the number of clusters, $\pi_k$ designates the prior probability of a case’s belonging to cluster $k$ (or the size of cluster $k$), and $\theta$ defines the model parameters (Vermunt and Magidson 2002). Equation (4) specifies the distribution of $y_i$ given the model parameter $\theta$ as a mixture of cluster-specific densities, $f(y_i | \theta_k)$.

Equivalently, we may express the model in (4) in its likelihood form (Fraley and Raftery 2002),

$$ L(\theta, \pi_k | y_i) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(y_i | \theta_k), \quad (5) $$
where most commonly \( f(y_i \mid \theta_k) \) is the multivariate normal (Gaussian) density \( \phi_k \), parameterized by its mean \( \mu_k \) and covariance matrix \( \Sigma_k \). Banfield and Raftery (1993) proposed parameterizing the cluster-specific covariance matrices \( \Sigma_k \) by eigenvalue decomposition:

\[
\Sigma_k = \lambda_k D_k A_k D_k^T ,
\]

(6)

where \( D_k \) is the orthogonal matrix of eigenvectors, \( A_k \) is a diagonal matrix whose elements are proportional to the eigenvalues, \( \lambda_k \) is an associated scalar of proportionality. More specifically, \( \lambda_k = |\Sigma_k|^{1/d} \), where \( d \) is the number of indicators, and \( A_k \) is scaled such that \( |A_k| = 1 \). The three parameters offer a nice interpretation: \( D_k \) describes the orientation of the \( k \)th cluster in the mixture, \( A_k \) its shape, and \( \lambda_k \) its volume. Put differently, if a latent class or cluster is viewed as a group or cluster of points in a multidimensional space, the volume is the size of the cluster, and the orientation and shape parameters indicate whether the cluster mixture is spherical, diagonal, or ellipsoidal. For example, the \( k \)th cluster will be roughly spherical if the largest and the smallest eigenvalues of \( \Sigma_k \) are of the same magnitude.

The combination of these parameter specifications determines the specific statistical model to fit. For example, for one dimensional data such income distributions, there are only two models to estimate, E for equal variance and V for varying variance. For multidimensional data that involve multiple variables, things are more complex. For example, VEI denotes a model in which the volumes of the clusters are varying (V), the shapes of all the clusters are equal (E), and the orientation is of the identity (I). For further details, see Fraley and Raftery (1999).
Model Selection

In conventional cluster analysis, the data analyst must deal with the issue of selection of the clustering method and that of determining the number of clusters. In model-based clustering, the two issues reduce to a single concern of model selection. In Fraley and Raftery’s (2002) approach, Bayesian model selection via Bayes factors and posterior probabilities is preferred. This in practiced is evaluated by way of the Bayesian Information Criterion (BIC), which is implemented in the MCLUST software (Fraley and Raftery 1999, 2002).

Density Estimation

Estimating the number of clusters and individuals’ memberships in the clusters is probably the major purpose of model-based clustering methods. However, it is also possible to obtain density estimation, in which the value of mixture likelihood at individual points is of interest. Roeder and Wasserman (1997) used normal mixtures for univariate density estimation and BIC to decide the number of components. Fraley and Raftery’s (1999, 2002) method can be viewed as a multivariate extension because the parameter estimates for the best model describes a multivariate mixture density for the data.

Uncertainty of Classification

Fraley and Raftery (1999) described the software MCLUST which implements the model-based clustering, using the EM algorithm. The software also computes a quantity known as uncertainty, which is defined by subtracting the probability of the most likely group or cluster for each observation from 1. A descriptive analysis of uncertainty can indicate how well the observations are classified. Uncertainty plots can be produced for single or multidimensional data.
Analyzing Income Inequality with Model-Based Clustering

When our concern is only with the shape of inequality, such analysis is relatively simple because there are only two possible models to consider—equal variance (E) or varying variance (V). For an empirical example, we apply Fraley and Raftery’s (1999, 2002) model-based clustering to household income data of the Ilocos region in the Philippines. The data come from the 1997 Family and Income and Expenditure Survey and the 1998 Annual Poverty Indicators Survey (APIS), conducted by the National Statistics Office of the Philippines. There are four provinces covered in the data: Ilocos Norte, Ilocos Sur, Pangasinan, and La Union, and the data set is available as megadata accompanying the R ineq package for inequality analysis which produces statistics such as the Gini coefficient and some generalized entropy measures.

As recent research has shown, income inequality may vary spatially because of the development policy there (Balisacan and Fuwa 2003). This necessitates analyzing patterns of income inequality separately for the regions or provinces. Using the same 1997 survey data, Balisacan (2001) showed that Ilocos Norte (poverty line=7,084 pesos per capita or ppc) and Ilocos Sur (poverty line=7,906 ppc) were among the 10 provinces with the lowest poverty incidence while the other two provinces, La Union (poverty line=7,669 ppc) and Pangasinan (poverty line=7,542 ppc) were neither among the highest nor the lowest 10 provinces, all with lower poverty lines lower than Metro Manila (poverty line=10,577 ppc). The cost of living index in these four provinces of the Ilocos region ranged from 67.0 to 74.7 (Metro Manila=100).
First, let us analyze the issue of income inequality by using the Gini index and the Lorenz curve. Figure 1 presents the Lorenz Curves for the four provinces of Pangasinan, La Union, Ilocos Norte, and Ilocos Sur, with corresponding Gini indices listed below the panels. It is rather difficult to see much difference among the four Lorenz curves, except that the one for La Union may indicate a slightly greater amount of inequality than the others. The Gini indices for the four sets of data can be lined up from low to high inequality: Pangasinan, Ilocos Norte, Ilocos Sur, and La Union. The greatest pairwise comparison is between La Union and Pangasinan, with a mere difference about 0.06. For the sake of comparison, the empirical distributions of income in the four provinces are presented in Figure 2.

---Figures 1 & 2 about here---

However, the issue of stratification is entirely ignored by such analysis. Next, we move on to analyzing the same data with model-based clustering by using the MCLUST software written for R. For the province of Pangasinan, the province with the lowest inequality among the four, the model assuming unequal variances fits better than the one assuming equal variances at most number of possible clusters, and suggests the model with four latent classes as the best fitting (Figure 3). The uncertainty plot clearly indicates three uncertain regions although the density plot gives only three obvious peaks and a long tail with a slightly elevated end region.

---Figure 3 about here---

For the province of La Union, however, the two-cluster unequal-variance model fits the best (Figure 4). Judging by the uncertainty and the density plots, there appear to be two classes, even though the second cluster is rather spread out.
For the province of Ilocos Norte, the two-cluster unequal-variance model also fits the best (Figure 5). The uncertainty and the density plots indicate a similar pattern of clustering to that existing in the La Union data, with perhaps less elevated second cluster in the Ilocos Norte data.

For the province of Ilocos Sur, it is the four-cluster equal-variance model that fits the data the best (Figure 6). The three clusters or latent classes for the province are clearly shown by either the density plot or the uncertainty plot, indicated by the three peak uncertainty regions between the four clusters, with the last cluster being an elevated right tail.

Finally, Figure 7 summarizes the overall income distribution in the entire Ilocos region. For the region as a whole, the unequal variance model with four latent classes fits the best, supported by any of the panels in the graph. When data from subregions are combined, it is likely that the maximum number of clusters from a subregion is maintained, unless the data, when combined, happen to smooth out the troughs.

Interestingly, the model-based clustering method allows us to see much greater detail in how the provinces differ in their patterns of stratification. Such detail is ignored by the Gini index or the Lorenz curve because these measures of inequality focus on only one kind of inequality.
Two Types of Inequality

It must have become clear by now that there exist two types of inequality, which we may term “individual inequality” and “class inequality”. The first kind is measured by methods including pairwise differences between individuals. The second type, “class inequality”, can be broadly conceived as whether there exist classes or clusters of individuals in the sample (or population), and of the absolute distances between individuals of the classes only partially reflect the distinction of classes. We may simple use the number of classes as an indicator of such inequality. However, a more informative measure is desirable. In general, the decomposition of Gini indices into the between-class and within-class components when the classes are ordered can be used as a means to partition income inequality for the purpose of measuring stratification. Here we follow the decomposition methods presented in Dagum (1997a) and Mussard, Alperin, Terraza, and Seyte (2005).

The overall Gini is computed as (3), and the within and between components of inequality are computed respectively as:

\[ G_w = \sum_{k=1}^{K} \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \left| y_i - y_j \right| \frac{1}{2n_i^2 \mu} \]  

(7)

where \( n_j \) is the size of each \( j \) class or cluster, and \( K \), the number of estimated classes or clusters, and

\[ G_b = \sum_{k=2}^{K} \sum_{j=1}^{n_i} \sum_{b=1}^{n_j} \frac{\left| y_b - y_j \right|}{n_i n_j \mu} \]  

(8)

The model based latent class method is first used to obtain the ordered classes in a population. For such classes, no consideration of transvariation or overlapping decomposition is necessary. Once the total amount of inequality is allocated into the
individual (or within) and class (or between) components, an index of relative stratification can be simply computed as:

\[ S = \frac{G_b}{G} = \frac{G - G_w}{G} \]  

(9)

where \( G_b \) is the net or total between-class component because all classes are ordered and there is no transvariation. This measure ranges from 0, where all inequality is individual-based and there is no stratification in the population to 1, where all inequality is contributed by stratification and there is no variation within classes. The measure is relative because it is the amount of class inequality expressed as a proportion of total inequality in terms of the Gini.

To illustrate, I present in Table 1 the three inequality measures—the Gini, the within- and between-class components of the Gini, and the index of relative stratification, \( S \)—for the four provinces of the Ilocos region (Table 1). The model-suggested classification schemes were used to obtain the ordering of the classes in the income distributions before computing decompositions. According to the Gini, Pangasinan has the lowest individual inequality while La Union the highest. However, Pangasinan actually has much a higher level of relative stratification according to the relative stratification index when compared with La Union (0.98 versus 0.87). Their between-class components are quiet similar; it’s the difference in their within-class components that determines the difference in the index of relative stratification. It is also revealing to contrast Ilocos Norte and Ilocos Sur: The two provinces have about the same amount of total inequality as measured by Gini, 0.45, but Ilocos Norte has a higher amount of individual inequality while Ilocos Sur has a higher mount of class inequality. The former province has a stratification
index of 0.90 while the latter has a relative stratification index of 0.97, indicating different levels of stratification.

---Table 1 about here---

To further study the issue, we consider below three simulated samples, each of which has 600 cases, generated as a random normal variate:

Sample 1 \((n=600)\): \(y_i = N(250,150)\)

Sample 2 \((n_1=300; n_2=300)\): \(y_{2i} = N(200,150); y_{2i} = N(400,150)\)

Sample 3 \((n_1=200; n_2=200; n_3=200)\): \(y_{1i} = N(100,50); y_{2i} = N(300,50); y_{3i} = N(500,50)\)

By design, the first sample has no stratification; the sample has two classes; the third sample contains three classes. The BIC, classification, uncertainty, and density plots for the three samples are presented in Figures 7 to 9.

---Figures 7, 8, and 9 about here---

For the first simulated sample, there can be only one cluster identified, as it should because of the condition under which the data are generated. Since the model is estimated with the minimum number of classes (i.e., 1), no uncertainty or density plots are produced. For the second sample, a two-class model provides the best fit, as it should, according to the equal-variance model. The density and uncertainty plots further confirm the choice. Similarly, a three-class model is the one with the best fit for the third sample, according to either the equal- or varying-variance assumption. Again, that choice is confirmed by the uncertainty and density plots. All the results should come as no surprise because the data are simulated with one, two, and three clusters for the three samples respectively.

Next, we compare the Gini and the class inequality measures for the three samples (Table 2). For the sake of comparison, we also include the decomposition results of Gini
components for the first three hypothetical examples presented at the outset of the paper. The comparison should shed further light on the within- and between-class measures and the relative stratification index because of the simulated, controlled nature of the three datasets.

---Table 2 about here---

Obviously, the three initial hypothetical distributions have no individual within-class inequality, thus producing an index of stratification of 1, with all variation explained by the between-class component. For the three simulated samples, the comparison is informative. According to the Gini, there is no discernable difference between the three samples, all yielding a coefficient of around 33%. According to the model-based latent class analysis, Sample 1 has only one stratum, and thus has no class inequality, with a stratification index of 0. Sample 2 and Sample 3 both have a relatively high degree of class inequality because of the obvious stratification in the data. The difference in stratification between the two samples is 14% when using the relative stratification index as the gauge, with Sample 3 having a higher amount between-class inequality. Note the stratification index measures the amount of inequality between classes relative to the total amount of inequality anywhere in the sample. A high value may be found even though the total amount inequality is low. In the simulated Samples of 2 and 3, the amount of inequality is moderate although the majority of which, especially in the case of Sample 3, is found in stratified inequality. For an absolute measure of class inequality, the between-class component of the Gini index can be consulted. The absolute and relative measures can be used together to assess inequality in a population.
Concluding Remarks

In this paper I identified a limitation of the Gini index for studying inequality—that the measure is sensitive to individual-based differences and may not be responsive to class-based stratification. I proposed a model-based latent class method to estimate order of classes within a distribution, and further utilize this information to decompose the Gini index into the within- and between-class components of inequality. The components are further used to form an index of relative stratification.

Both the empirical income data from the Philippines and the simulated data illustrate that the Gini may be misleading in understanding stratification when only the overall amount of inequality is considered. Using the method proposed, we can fine-tune the measurement of inequality by separating the individual-based from the class-based inequality. The proposed index of relative stratification also gives us a simple and useful assessment of the amount stratification existing in the sample, thereby releasing the true power of the Gini index. The Gini index—when extended and informed by the model-based latent cluster analysis—can not only shed measure the amount of the individual based but also class-based inequality. The Gini, as the unbound Prometheus, has potentially more power, and possess much potency for the study of social stratification.

References


Figure 1: Gini Indices and Lorenz Curves for the four Filipino provinces

The Province of Pangasinan; Gini=0.4016569

The Province of Ilocos Norte; Gini=0.4496271

The Province of La Union; Gini=0.4626395

The Province of Ilocos Sur; Gini=0.4595246
Figure 2: Empirical Income Distributions of the Four Province

Annual Household Income in Pangasinan

Annual Household Income in La Union

Annual Household Income in Ilocos Norte

Annual Household Income in Ilocos Sur
Figure 3: Income Clustering in Pangasinan
Figure 4: Income Clustering in La Union

The graph on the left shows the Bayesian Information Criterion (BIC) values for different numbers of clusters. The x-axis represents the number of clusters, and the y-axis represents the BIC values. The dashed line indicates the optimal number of clusters based on the BIC criterion.

The graph on the right displays the uncertainty and density plots. The x-axis represents the number of clusters, and the y-axis represents the uncertainty and density values. The uncertainty plot shows a peak at a certain number of clusters, indicating the optimal point. The density plot shows the distribution of data points across different clusters.
Figure 5: Income Clustering in Ilocos Norte
Figure 6: Income Clustering in Ilocos Sur

---

**Graph 1:**
- X-axis: Number of clusters
- Y-axis: BIC

**Graph 2:**
- X-axis: Number of clusters
- Y-axis: Uncertainty

**Graph 3:**
- X-axis: Number of clusters
- Y-axis: Density
Figure 7: Income Clustering in the Ilocos Region

[Graph showing BIC values against the number of clusters, indicating a potential number of clusters.]

[Graph showing uncertainty against a range of values, with peaks indicating higher uncertainty.

[Graph showing density against a range of values, with a peak indicating a dense region.]
Figure 8: Income Clustering of Simulated Data of $N(250, 150)$

Gini: 0.332649
Figure 9: Income Clustering of Simulated Data of $N(200,150)$ & $N(400,150)$

Gini: 0.331760
Figure 10: Income Clustering of Simulated Data of $N(100,50) \& N(300,50) \& N(500,50)$

Gini: 0.330702
Table 1: Comparing the Gini, Its Components, and Relative Stratification, the Filipino Income Data

<table>
<thead>
<tr>
<th>Province</th>
<th>Gini</th>
<th>Gini\textsubscript{within}</th>
<th>Gini\textsubscript{between}</th>
<th>Stratification Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pangasinan</td>
<td>0.401657</td>
<td>0.042992</td>
<td>0.358665</td>
<td>0.979048</td>
</tr>
<tr>
<td>La Union</td>
<td>0.462640</td>
<td>0.102090</td>
<td>0.360550</td>
<td>0.874710</td>
</tr>
<tr>
<td>Ilocos Norte</td>
<td>0.449627</td>
<td>0.142308</td>
<td>0.307319</td>
<td>0.904987</td>
</tr>
<tr>
<td>Ilocos Sur</td>
<td>0.459525</td>
<td>0.058951</td>
<td>0.400574</td>
<td>0.979035</td>
</tr>
<tr>
<td>The Ilocos Region</td>
<td>0.426951</td>
<td>0.046137</td>
<td>0.380813</td>
<td>0.977517</td>
</tr>
</tbody>
</table>
Table 2: Comparing Gini Decomposition and Relative Stratification, Simulated Samples

<table>
<thead>
<tr>
<th>Sample</th>
<th>Gini</th>
<th>Gini(_{\text{within}})</th>
<th>Gini(_{\text{between}})</th>
<th>Stratification Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: [1, 1, 4, 4, 4]</td>
<td>0.300000</td>
<td>0.000000</td>
<td>0.300000</td>
<td>1.000000</td>
</tr>
<tr>
<td>2: [1, 1, 5, 5, 5]</td>
<td>0.333333</td>
<td>0.000000</td>
<td>0.333333</td>
<td>1.000000</td>
</tr>
<tr>
<td>3: [1, 2.5, 2.5, 4, 4]</td>
<td>0.266667</td>
<td>0.000000</td>
<td>0.266667</td>
<td>1.000000</td>
</tr>
<tr>
<td>4: (N(250,150))</td>
<td>0.332649</td>
<td>0.332649</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>5: (N(200,150),\ N(400,150))</td>
<td>0.331760</td>
<td>0.092507</td>
<td>0.239253</td>
<td>0.840307</td>
</tr>
<tr>
<td>6: (N(100,50),\ N(300,50),\ N(500,50))</td>
<td>0.330702</td>
<td>0.028861</td>
<td>0.301841</td>
<td>0.976629</td>
</tr>
</tbody>
</table>