Testing the properties of selection criteria: an application to copy number polymorphism measurements

A Dissertation Presented

By

Rose Edy Saint Fleur

to

The Graduate School

In Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

August 2010

Stony Brook University

The graduate school

Rose Edy Saint Fleur

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, herby recommend

acceptance of this dissertation

Stephen J. Finch – Dissertation Advisor
Professor of Statistics
Department of Applied Mathematics and Statistics

Nancy R. Mendell – Chairman of Defense
Professor of Statistics
Department of Applied Mathematics and Statistics

Wei Zhu
Professor of Statistics
Department of Applied Mathematics and Statistics

Derek Gordon
Department of Genetics
Rutgers University

The dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Testing the properties of selection criteria: an application to copy number polymorphism measurements

By

Rose Edy Saint Fleur

Doctor of Philosophy

In

Applied Mathematics and Statistics

Stony Brook University

2010

Variation in the human genome is present in many forms, including single-nucleotide polymorphisms (SNPs) and copy number polymorphisms (CNPs). CNPs have many categories such as small insertion-deletion polymorphisms, variable number of repetitive sequences, and genomic structural alterations. A major question that researchers in the field of statistical genetics need to answer is the number of CNP categories in a given dataset. In this study, I compare five information criteria (BIC, AIC, NEC, CLC, and ICL-BIC) to find if there is a "best" measure among them in finding the correct number of components (correct number of CNP categories). I consider six design factors: equal/unequal within-component variances, high/low separations, sample size, mixture proportion, multiple random starting values, and transformation using two known number of components (3 and 6). The result indicates that under "ideal" conditions (that is, small number of components, large separation between components, constant within component variance, and no subsequent transformation of mixture data), each criterion performs well. When the data is a monotonic transformation of data from a mixture, the BIC criterion, which is the most commonly used criterion in CNP research, has a low component number accuracy rate. I then considered the application of the Box-Cox transformation whether or not it was needed. The application of the Box-Cox transformation did not reduce the component number accuracy rate of the CLC, ICL-BIC, and BIC when it was not needed. The component number accuracy rates for the BIC criterion with Box-Cox transformation applied were improved when the mixture data was transformed. The Box-Cox transformation should be used routinely with CLC, ICL-BIC, or BIC criterion to estimate the number of components in a CNP mixture analysis.

In loving memory of my maternal grandmother, Mrs. Altimise Louis-Bien Aimé!

Table of Contents

List of Tables

List of Figures

Acknowledgment

I would like to thank everyone who has supported me throughout my tenure as a graduate student in the applied mathematics and statistics program at Stony Brook University. I would like to start by thinking my advisor, Dr. Stephen J. Finch for being so patient with me throughout the years. I have learned so much from you and I will forever be grateful. My thanks go to Dr. Nancy R. Mendell, Dr. Wei Zhu, and Dr. Derek Gordon for their feedback on my project.

My special thanks go to my parents Mr. Edy and Mrs. Denise Saint Fleur. Thank you for loving me unconditionally! I would like to thank my three wonderful sisters Shella, Emmanuela, and Juliette who have shared their word of wisdom with me throughout the years. Special thanks to Dr. Shella Saint Fleur-Lominy for being my role model and my inspiration!

I would like to thank all my friends who have in one way or another helped me through the years. A special shout out to Joseph and Lory Ortiz, Marianne, Wendy, and Yoldie for being there for me! Thank you to everyone at Huntington Church of God who has given me special words of encouragement.

Most importantly, I would like to thank my loving husband, Stanley J. R. Calixte, for encouraging me to always reach for the stars. You are my rock and I love you. Last, but not least, I want to thank my God for given me the strength to battle it out for six long years. I dedicate my dissertation to my son, Stanley J.E. Calixte.

Chapter 1 Introduction

1.1 Copy number polymorphism and methods of detection

Variation in the human genome is present in many forms, including single-nucleotide polymorphisms (SNPs), small insertion-deletion polymorphisms, variable number of repetitive sequences, and genomic structural alterations (Iafrate et al., 2004; Sebat et al., 2004). Copy number polymorphisms (CNPs) can be simple deletions or replications leading to copy number changes at several locations in the human genome (Iafrate et al., 2004). Recent studies have suggested that CNPs may be an underlying factor in genetic diseases (Iafrate et al., 2004; Sebat et al., 2004). For instance, Fanciulli et al., (2007) studied CNP in the Fcgr3 gene and linked CNP variation to glomerulonephritis. Glomerulonephritis is a type of kidney disease in which the kidney cannot filter toxins from the blood (Patel, 2009). Fanciulli et al. found that in humans with systemic lupus a low count in the Fcgr3 gene was associated with glomerulonephritis.

There are different methods of measuring CNPs. One of the widely used methods is array Comparative Genomic Hybridization (aCGH) (Pinkel et al., 1998). In aCGH data, each measurement is the log of the ratio of two measurements. Different issues arrive when analyzing CNP data. First, researchers have to deal with quality control issues. In the preprocessing steps, one has to normalize the data. The purpose of normalizing the intensity data is to adjust the signal intensities so that measurements from different arrays are on the same scale. There are different normalization techniques such as global normalization or loess normalization that can be used (Wineinger et al., 2008). Then, one has to apply smoothing techniques to the data. Smoothing techniques are used to remove variability within an array. Then, a calling algorithm is used to assign copy number state to each locus. A state can be either normal, gain or loss of copy number (Wineinger et al., 2008). There are two types of calling algorithm: change points methods and Hidden Markov Models.

There are two well known change-point methods: the Circular Binary Segmentation (CBS) developed by (Olshen et al., 2004) and the CGH segmentation (CGHseg) developed by (Picard et al., 2005). The CGHseg algorithm assigned a copy number state to the intensity data based on a selection criterion. It is assumed that the number of components (copy number segment) is known in advance. Picard et al. (2005) compared the Bayesian information criterion (Schwarz,1978) with CGHseg in a simulation study. In this study, the Bayesian Information Criterion had a tendency to overestimate the number of segments. CGHseg is a selection model that considers homogeneous signal variability. Lai et al. (2005) did a comparative simulation study of 11 different models used in the analysis of array CGH data including CGHseg. They found that CGHseg and CBS performed better than other model. But, CGHseg is sensitive to outliers (Ben-Yaacov and Eldar, 2008).

1.2 Objective

Finding the correct CNP category number is a fundamental task. In modern genetic research, many researchers rely on the Bayesian Information Criterion (BIC) as the method to estimate how many different copy number categories exist in a data set (Kim et al., 2008). My simulation study is designed to be an extension of a simulation study in McLachlan and Peel (2000) that compared different selection criteria to see which one is the best. Picard et al. (2005) found that the method they called CGHseg worked best for intensity data with equal within-component variances. They compared two models of intensities. One model has equal within component variance, and the other model had unequal within component variances. They found that in the model with equal within-component variance there was an overestimation of the number of segments (components).

This research is of particular interest because it considers more experimental factors. Picard et al. (2005) studied only two models (equal and unequal within-component variances). In Lai et al. (2005), the comparative study was based on uniform distances between the components. Also, the number of components was set at K=5. My goal is to find the best selection method when dealing with data like that observed in CNPs studies under different experimental condition.

My first objective is to estimate the probability that the BIC criterion finds the correct number of CNP categories. My second objective is to estimate the probability that available criteria find the correct number of components and perform a comparative study of the selected criteria. Akaike (1974) studied the Akaike Information Criterion (AIC). Celeux and Soromenho (1996) considered the Normalized Entropy Criterion (NEC). Biernacki and Govaert (1997) considered the Classification Likelihood Criterion (CLC). McLachlan and Peel (2000) estimated the Integrated Classification Likelihood (ICL) (Biernacki, et al., 1998) using the BIC version of ICL (ICL-BIC).

1.3 Research Questions

The following are my research questions.

1. Is there a measure that is always has the highest probability of correctly specifying the number of categories?

   a. If yes, which one?

   b. If no, then

      i. What are the settings for which a criterion has high probability of correct specification of the number of categories?

      ii. What are the settings for which a criterion has low probability of correct specification of the number of categories?

2. How important is separation, where separation is the number of standard deviations between adjacent component means?

   a. What is the minimum sample size required for high separation?

   b. What is the minimum sample size required for low separation?

   c. How large should the separation be to have at least 50% correct probability of correct specification of the number of categories?

3. What is the effect of heterogeneity of component variance on the probability of correct specification of the number of categories?

4. Does adding more random starting value improve the probability of correct specification of the number of categories?

5. What is the effect of departure from normality on the probability of correct specification of the number of categories?

   a. Does Box-Cox transformation improve the probability of correct specification of the number of categories?

   b. What is the effect of using Box-Cox when data do not require it?

6. What should researchers do when dealing with CNP data?

Chapter 2 Simulation study methodology

2.1 Settings of design factors

In order to assess the statistical properties of the measures to identify the number of components in a sample, I used a factorial design with seven factors (A through G):

A. Number of components. The number of components is set to 3 (the number of genotypes a Single Nucleotide Polymorphism study) or 6. My three-component model is based on a gene with two alleles (1 and 2), with mean intensity 1 for allele 1 and the mean intensity 2 for allele 2. My model is that the three possible genotypes (11, 12 and 22) have average intensities 2, 3 and 4 respectively. For the six-component model, I assume that I have a gene with 3 alleles (1, 2, and 3) with average intensities 1, 2 and 3 respectively. The possible genotypes (11, 12, 13, 22, 23, 33) then have average intensities 2, 3, 4, 4, 5, and 6 respectively. Based on the average intensities, there are five detectable components since the average intensity of the 22 genotype is the same as the 13 genotype. I studied the presence of a copy number gain (six-component model) which is the case where 12 and 33 have different average intensities. As a result, I set the mean intensities of the 2/2 and the 1/3 genotype to different values. Specifically, I set the means at 2, 3, 4, 5, 6 and 7 respectively so that all genotypes are differentiated.

B. Constant Variance (yes(+), no(-)). In the constant variance setting, I generate my components with different means μ_i and equal within component variances, σ^2. In the non-constant variance setting, I generate my component values with a variance that is proportional to the mean μ_i of each component;

that is $\sigma_i^2 = k\mu_i$. For example, this relationship of variance to mean occurs in the chi-squared distribution.

C. High Separation at Middle Component (yes(+), no(-)). In the high separation setting, the components adjacent to the middle component have 4-standard deviation separation; that is $\frac{\mu_{i+1} - \mu_i}{(\sigma_{i+1} + \sigma_i)/2} = 4$. In the low separation setting, the components adjacent to the middle component have 2-standard deviation separation.

D. Sample size. The four sample size settings are: n = 250, 500, 1000, and 2000.

E. Mixture Proportion. There are three mixture probability vectors. I use equal probability for each component for the first setting. I use a skewed pattern with more subjects in the first group than in the other groups for the second setting. In the three-component model, the proportion in each component is (0.4615, 0.3077, 0.2308). In the six-component model, I use the vector of proportions (0.3139, 0.2093, 0.1569, 0.1256, 0.1046, 0.0897). For the third setting, I use Hardy-Weinberg Equilibrium (HWE), and I set the proportion of each allele in the dataset as follows:

a. In the three-component model, the 1-allele occurs with a 50% probability and the 2-allele occurs with a 50% probability. The HWE proportion vector (0.25, 0.50, 0.25) is used for intensities 2, 3 and 4 respectively.

b. In the six-component case, I assume that allele 1 is the most frequent allele with 50% frequency. Allele 2 has a 25% frequency, and allele 3 also has a 25% frequency. Then, the

HWE proportion vector (0.25, 0.25, 0.25, 0.0625, 0.125, 0.0625) is used with intensities 2, 3, 4, 5, 6, and 7 respectively.

F. Multiple Random Starting Values (yes (+), no (-)). I use the MCLUST package in my study (Fraley and Raftery, 2002; Fraley and Raftery, September 2006, revised December 2009). The multiple random starting values (RSVs) factor has two settings. At the no setting, the default setting of the R package MCLUST is used. At the yes setting, the number of RSVs is based on the number of components of the simulated data set. When the simulated data set is from a three-component model, the number of RSVs is set to 10. When the data set is simulated from a six-component model, the number of RSVs is set to 50.

G. Transformation (yes (-), no (+)). The transformation factor has two settings. The (+) setting is that the observed intensity is a mixture of normally distributed components. The (-) setting is that the observed intensity is the square of a mixture of normal components. In other word, when using the yes level for transformation, $\sqrt{X}$ is a mixture of normal components.

2.2 Setting the number of random starting values (RSVs)

Adding RSVs is computationally expensive (about 2.5 hours using 50 RSVs for 100 replicates each with 2000 observation using a 2.53 GHz dual core personal computer). For 50 RSVs applied to data on 2000 observations using the three-component model, I estimate that the computing time to complete one out of 24 settings would be about 22 hours. In order to control computational cost, I conducted a pilot study using 100 replicates each with 250 observations from a three-component mixture and 100 replicates each with 250 observations from a six-component mixture to test how much the observed maximum likelihood is increased as the number of RSVs increased.

Let *ML(R)* be the maximum observed log likelihood value with *R* random starting values for a given replicate. For the three-component model, $ML(20) - ML(10) \le 0.0000054$, $ML(100) - ML(20) \le 0.0000335$, and $ML(100) - ML(10) \le 0.0000388$. Since using more than 10 RSVs did not appreciably increase the maximum log likelihood value in the three component model, I use 10 RSVs for the three-component analysis.

For the six-component model, $ML(60) - ML(50) \le -0.000057$, $ML(100) - ML(60) \le 0.000075$, and $ML(100) - ML(50) \le 0.000018$. This suggests that using more than 50 RSVs in the six-component analysis would not appreciably increase the maximum log likelihood value. Therefore, I use 50 RSVs.

2.3 Model selection criteria

I compare five model selection criteria to determine *g*, the number of components in the selected model.

1. The BIC criterion, as proposed by Schwarz (1978), is defined to be $BIC = 2ln[L(g)]-kln(n).$ In the formula, *L(g)* is the maximized value of the likelihood function for the model with *g* components, *k* is the number of parameters in the model to be estimated, and *n* is the sample size. The model *g* with largest BIC is the BIC selection.
2. The AIC criterion, as proposed by Akaike (1974), is a measure of goodness of fit of an estimated statistical model, and is defined to be $AIC = 2ln[\,L(\,g)]\, \text{-}\, 2k$. The AIC criterion differs from the BIC criterion when

n is large. In this model, *L(g)* is the maximized value of the likelihood function with *g* components and *k* is the number parameters in the model. The model *g* with largest AIC is the AIC selection.

3. The Normalized Entropy Criterion (NEC) has been proposed by Celeux and Soromenho (1996) to find the number of clusters in a mixture model. The criterion is defined to be $NEC(g) = \frac{E(g)}{ln[L(g)] - ln[L(1)]}$, where $2 \leq g \leq g_{sup}$. Here, g_{sup} is a user specified maximum number of components (here 9), $L(g)$ is the maximum value of the likelihood function of a *g*-component mixture, $L(1)$ is the maximized likelihood function using a one-component normal mixture, and $E(g)$ is the corresponding entropy of the *g*-component mixture model defined by $E(g) = -\sum_{i=1}^{g}\sum_{j=1}^{n}\hat{\tau}_{ij}\, ln\, \hat{\tau}_{ij},$ where $\hat{\tau}_{ij}$ are the posterior probabilities of component membership for subject *j* belonging to group *i*. We choose g^* if $NEC(g^*) < 1,$ otherwise, *g* = 1 (Biernacki et al., 1999).

4. The Classification Likelihood Criterion (CLC) was proposed by Biernacki and Govaert (1997). They used the relationship linking the likelihood for the mixture data with the complete data likelihood. The model is defined to be $CLC = 2\, ln[L(g)] - 2E(g)$ The model with largest CLC is the CLC selection

5. The Integrated Classification (Completed) Likelihood (ICL) was proposed by Biernacki *et al.* (1998). It is defined to be $ICL = 2\, ln[L(g)] - 2E(g) - k\, ln(n) - (g-1)\, ln(n),$ where k is defined to be

the number of parameters in the model and g is the number of clusters. The model with largest ICL is the ICL selection. They also introduced a new way to approximate the ICL criterion by using the BIC criterion with the entropy of the model. McLachlan and Peel (2000) call this method the ICL-BIC and defined it to be $ICL-BIC = 2\ln[L(g)] - k\ln(n) - 2E(g) = BIC - 2E(g)$. The model g with largest ICL-BIC is the ICL-BIC selection. This criterion should differ very little from the ICL version.

2.4 Simulation study material

I use the R2.8.0 (2008) statistical package to generate 1000 samples at each of the 576 settings of the seven (7) factors. Then, I use the MCLUST package to perform the mixture calculations. For each sample, I fit 1, 2, …, 9 components. Then, I calculate each criterion. For each criterion, I find the correct component selection rate over the 1,000 replicates. I use Minitab® 15 to compute an ANOVA table to find the significant factors and interactions. In the ANOVA table, sample size is defined as a categorical variable with four levels. I define an F test as significant if the *p*-value is less than 0.01. Additionally, I compute the correlation coefficient of the correct component selection rate with sample size.

2.5. Effect of Box-Cox transformation on classification accuracy rate

Since researchers often use a Box-Cox transformation in Copy Number Variation analysis (Kim et al., 2008), I perform a simulation study to document the effects of Box-Cox transformation on the component number accuracy rate. I use a factorial design with 6 factors selected from the 7 above (A through E and G). Multiple RSVs were used. The settings are the same for factors A (number of components), B (constant variance), and G (transformation). I use the three highest sample sizes for D (sample size); that is, $n = 500, 1000, 2000$. I use the equal and skewed mixture proportions for E (mixture proportion), but not the Hardy-Weinberg proportions. I use three settings for the separation of components: two, three, and four standard deviation separation between middle component means.

The Box-Cox transformation (Box and Cox, 1964) is implemented on the data using the R functions "*box.cox.powers*" and "*box.cox*". The first function finds the power ($\hat{\lambda}$) that will transform the original data set into a normally distributed sample. Then, I use $\hat{\lambda}$ in the second function to transform the original data regardless of whether or not the data came from a transformed normal mixture. When data are normally distributed, there is a chance of having negative numbers. To deal with that situation, I find the minimum number of the data set and if that number is negative, I add that number so that all sample values are non-negative. I perform a simulation study with 1000 replicates at each setting of the six factors that I use. Using the MCLUST software, I perform a mixture analysis after transforming the data using the Box-Cox procedure of R. For each sample, I fit 1, 2,…,9 components. Then I calculate the value for each criterion. Using Minitab, I find the significant factors and interaction terms using a 0.01 significance level.

Chapter 3 Three-component results

The scenarios for the three-component model are based on mixtures resulting from a di-allelic gene with alleles 1 and 2 and the assumption that allele 1 has average intensity of 1 and allele two has average intensity of 2. The three possible genotypes for the gene (11, 12, 22) would then have average intensities 2, 3, and 4 respectively.

3.1 ANOVA tables of result

I report the sum of squares of the main effects and selected two-way and three-way interactions using the component number accuracy rate as the dependent variable in the ANOVA tables (Table3.1.1). The error sum of squares is the sum of the unreported interaction terms. An F-statistic greater than 6.8 corresponds to a p-value ≤ 0.01.

The variation of the BIC component number accuracy rate is principally explained by three factors in four terms: *C* (separation) explaining 31.6% of variation, *B* (constant variance) explaining 16.9%, the $C \times G$ interaction of transformation and separation explaining 21.0%, and the $B \times G$ interaction explaining 15.4% (Table 3.1.1). These four terms explained 87% of the variation. Multiple RSVs (*F*) and the interactions $B \times C$, $F \times G$ and $B \times C \times G$ are also significant based on their F-test values.

The factors explaining the variation in the AIC component number accuracy rate are transformation (*G*), multiple RSVs (*F*), constant variance (*B*), and separation (*C*) and are somewhat different from the factors explaining BIC variation (Table 3.1.1). The transformation factor *G* main effect accounts for 30.8% of AIC variation, and its interaction with constant variance B accounts for 15.9%. The main effect of multiple random starting points (F) accounts for 14.7%, and its interaction with constant variance accounts for another 10.2%. Finally, the main effect of separation (*C*) accounts for 6.5% of variation. These five sources explain roughly 67% of the TSS.

Sparation (*C*) and multiple RSVs (*F*) are the most important factors explaining the variation in the NEC component number classification accuracy rate (Table 3.1.1). Additionally, there is a significant F test for constant variance (*B*), transformation(*G*), the two-way interactions $B \times C$, $B \times E$, $B \times F$, $B \times G$, $C \times F$, $C \times G$, and $F \times G$, and the three-way interactions $B \times C \times E$, $B \times C \times F$, $B \times F \times G$, and $C \times F \times G$.

For CLC, separation (*C*) is the most important factor, with its main effect explaining about 90% of its variation (Table 3.1.1). Additionally, based on F test values, constant variance (*B*), sample size (*D*), multiple RSVs (*F*), and transformation (*G*) are significant as well as the two-way interactions $B \times C$, $B \times G$, $C \times E$, $C \times F$, and $C \times G$. The three- way interactions $B \times C \times E$, $B \times C \times F$, $B \times C \times G$, and $C \times F \times G$ are also significant.

Under the BIC approximation of ICL, the variation of the component number accuracy rates is mostly explained by separation (*C*), with that main effect explaining over 90% of the variation (Table 3.1.1). Using the F test, the main effects of *B*, *D*, *F*, and *G*, the two-way interactions $B \times C$, $B \times G$, $C \times E$, $C \times F$, and $C \times G$ and the three-way interactions $B \times C \times E$, $B \times C \times F$, and $B \times C \times G$ are significant.

Table 3.1.1: ANOVA tables for BIC, AIC, NEC, CLC, and ICL-BIC

		BIC		AIC		NEC		CLC		ICL-BIC	
Source	DF	% TSS	F	%TSS	F	%TSS	F	%TSS	F	%TSS	F
B: Constant Variance	1	16.9%	269.4	1.4%	23.9	0.5%	12.3	0.3%	16.0	0.3%	17.4
C: Separation	1	31.6%	502.9	6.5%	108.8	68.5%	1554.7	89.7%	4708.6	90.2%	4859.8
D: Sample Size	3	0.03%	0.2	0.7%	3.6	0.2%	1.4	0.2%	3.2	0.4%	7.7
E: Mixture Proportion	2	0.2%	1.4	0.4%	3.0	0.0%	0.5	0.0%	0.4	0.1%	3.6
F: Multiple RSVs	1	0.7%	10.4	14.7%	246.6	5.8%	132.7	0.1%	3.6	0.0%	0.2
G: Transformation	1	0.6%	10.2	30.8%	518.2	1.6%	36.2	0.4%	18.6	0.3%	15.9
$B\times C$	1	0.4%	6.0	0.0%	0.2	0.4%	8.9	1.1%	59.8	1.0%	55.6
$B\times E$	2	0.0%	0.2	0.1%	1.1	0.6%	7.2	0.1%	1.4	0.0%	0.1
$B\times F$	1	0.0%	0.3	10.2%	171.9	0.2%	4.1	0.0%	1.2	0.0%	0.3
$B\times G$	1	15.4%	244.7	15.9%	267.1	1.3%	29.4	0.3%	16.9	0.3%	16.7
$C\times E$	2	0.2%	1.8	0.1%	0.6	0.0%	0.1	0.3%	8.1	1.0%	25.9
$C\times F$	1	0.0%	0.5	1.9%	32.5	6.4%	144.6	0.6%	33.1	0.3%	13.8
$C\times G$	1	21.0%	334.6	1.6%	27.2	1.9%	42.4	1.3%	68.0	1.0%	53.8
$F\times G$	1	0.3%	4.6	0.3%	5.6	0.9%	21.1	0.0%	2.1	0.0%	0.2
$B\times C\times E$	2	0.1%	1.0	0.04%	0.4	0.9%	9.9	0.2%	5.7	0.4%	11.2
$B\times C\times F$	1	0.0%	0.01	2.2%	37.2	0.1%	2.3	0.5%	23.7	0.3%	14.8
$B\times C\times G$	1	2.2%	34.3	3.1%	52.5	1.5%	35.0	1.2%	61.6	1.0%	54.4
$B\times F\times G$	1	0.0%	0.2	0.2%	3.8	0.7%	15.2	0.0%	1.5	0.0%	0.2
$C\times F\times G$	1	0.0%	0.02	0.0%	0.5	1.1%	25.7	0.5%	27.9	0.3%	14.6
Remainder	166	10.4%		9.9%		7.3%		3.2%		3.1%	

3.2 Interaction of significant factors

In this section, I report the means of component number accuracy rates using sources that explain more than 5 % of variation. For criteria where only one source is significant, I use the factors important for the BIC criterion (i.e., *B*, *C*, and *G*). In tables 3.2.1-3.2.6, I report the mean component number accuracy rates ($\pm SD$) averaged over the factors that are not reported.

Table 3.2.1 contains the mean component number classification accuracy rates for the BIC criterion for its three most important factors--constant variance (*B*), separation (*C*), and transformation (*G*). The separation factor (*C*) explained the greatest fraction of variation as shown by the higher classification accuracy averages for 4-standard deviation separation between adjacent component means. There was a significant interaction between separation (*C*) and transformation (factor *G*) as shown in the table of averages. The constant variance factor (*B*) was also highly significant as shown by the higher average classification rate for unequal within-component variances. This factor also had a significant interaction with the transformation factor (*G*). The average BIC accuracy classification rate was well over 90% for all 4-standard deviation separation between adjacent component means settings except for the constant variance with transformation setting where the average was 11.4%. The BIC mean classification rate was also relatively high for the 2-standard deviation separation between adjacent component means with transformation and unequal within-component variances (accuracy rate 68.8%).

Table 3.2.1: Table of average component number accuracy rates for separation, constant variance, and transformation setting for the BIC criterion using 1000 replicates.

		C: Separation			
		4σ		2σ	
		G: Transformation		*G*: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
B: Constant Variance	Yes	11.4±20.6	99.7±2.6	20.0±21.1	9.1±9.1
	No	94.3±6.7	94.3±6.5	68.8±20.0	17.7±15.9

Average accuracy rate ± SD, average over 24 settings

Table 3.2.2 contains the mean component number classification accuracy rates for the AIC criterion for the constant variance factor (*B*), multiple RSV factor (*F*), and the transformation factor (*G*), the most significant variables for AIC. The average AIC classification accuracy rate was much lower than the average BIC component number classification accuracy rate, with all average rates below 70%. The most important factor is the transformation factor: the accuracy rate was higher when there was no transformation. Using multiple RSVs was associated with a lowered component number classification accuracy rate. There were significant interactions between the transformation and constant variance factors ($B \times G$) and between the RSV factor and the constant variance ($B \times F$).

Table 3.2.2: Table of average component number accuracy rates for constant variance, multiple RSVs, and transformation setting for the AIC using 1000 replicates

		F: Multiple RSVs			
		Yes		No	
		G: Transformation		*G*: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
B: Constant Variance	Yes	0.6±2.0	65.0±23.5	4.2±8.5	69.9±25.5
	No	1.1±1.4	4.6±3. 3	41.8±27.7	59.7±22.7

Average accuracy rate ± SD, average over 24 settings

Table 3.2.3 displays the mean component number accuracy rate for the NEC criterion for separation (*C*) and multiple RSVs (*F*). The NEC criterion has lower component number accuracy rates than the BIC criterion. The separation factor was the most significant factor with component number accuracy rates under 1% on average for a 2 standard deviation separation between component means. Use of multiple RSVs reduced the component number accuracy rate for a 4-standard deviation separation between component means.

Table 3.2.3: Table of average component number accuracy rates for separation and multiple RSVs setting for the NEC criterion using 1000 replicates

	F: Multiple RSVs		
		Yes	No
C: Separation	4σ	49.3±33.1	90.9±16.9
	2σ	0.9±3.9	0.0±0.1

Average accuracy rate ± SD, average over 48 settings

In table 3.2.4, I report the average component number accuracy rate for the CLC criterion for constant variance (*B*), separation (*C*), and transformation (*G*) to facilitate comparison of its rates with the BIC rates. As seen in the 4-standard deviation separation between component means, the CLC criterion performs slightly better than the BIC criterion. From the ANOVA table for the CLC component number accuracy rate, only the separation main effect explains more than 5% of the variation. For 2-standard deviation separation between adjacent component means, the average classification rate is under 10%. Regardless of B setting (constant variance) or transformation setting, the average accuracy classification rate is above 66% for 4-standard deviation separation between adjacent component means. Comparing 4-standard deviation separation between adjacent component means to 2-standard deviation separation between adjacent component means, the average classification accuracy rate is 2.5% for a 2-standard deviation separation between component means and 90.0% for a 4-standard deviation separation between component means. The unequal within-component variances average component number accuracy rate when there is no transformation and a 4-standard deviation separation between component means is much higher for CLC comparing to the BIC average.

Table 3.2.4: Table of average component number accuracy rates for separation, constant variance, and transformation setting for the CLC criterion using 1000 replicates

		C: Separation			
B: Constant Variance		4σ		2σ	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
	Yes	67.0±23. 1	98.2±2.7	9.8±21.3	0.0±0.0
	No	97.1±3.6	97.9±3.1	0.2±0.8	0.0±0.0

Average accuracy rate ± SD, average over 24 settings

In table 3.2.5, I report the average component number classification accuracy rate for the BIC approximation to the ICL criterion (ICL-BIC) for constant variance (*B*), separation (*C*), and transformation (*G*) using the same format as table 3.2.1 for direct comparison with BIC. As seen in the CLC criterion, the ICL-BIC criterion performs slightly better than the BIC criterion when data are from unequal within-component variances, no transformation and a 4-standard deviation separation between component means.

The separation factor (*C*) is highly significant as shown in the table with average component number accuracy rate for a 2-standard deviation separation between component means below 10%. For a 4-standard deviation separation between component means, the average accuracy classification rate is above 67%. The component number accuracy rates for both the CLC and the ICL-BIC criteria are comparable to the BIC accuracy rates, especially for settings with 4-standard deviation separation between adjacent component means, equal within-component variance, and transformation.

Table 3.2.5: Table of average component number accuracy rates for separation, constant variance, and transformation setting for the ICL-BIC criterion using 1000 replicates

		C: Separation			
B: Constant Variance		4σ		2σ	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
	Yes	67.1±21.7	95.4±9.1	8.2±19.0	0.0±0.0
	No	95.8±6.2	95.6±7.6	0.1±0.3	0.0±0.0

Average accuracy rate ± SD, average over 24 settings

3.3 Correlation coefficient of component number accuracy rate of the five criteria with sample size for three equiprobable components

One expectation of a classification criterion is that the component number accuracy rate increase as the sample size increases. Consequently, I report the correlation coefficient of the accuracy rate of the six criteria with sample size for the equiprobable component mixture components. The tables also contain in parentheses the component number accuracy percentage for a sample size of 2000. Similar results hold for the other two component probability distributions. The complete results can be seen in appendix A. An entry of “N” means that the component number accuracy rate was 0 for all sample sizes.

Table 3.3.1 contains the results for equal within-component variance, equal component proportions, and 4-standard deviation separation between component means. Each correlation coefficient is positive for data that did not require transformation and estimation based on the MCLUST default starting value, with AIC having the highest correlation coefficient (see line 3). The pattern of correlations was very similar for data not requiring transformation processed with multiple RSVs (line 1). Using multiple RSVs was associated with a slight decrease in the component number accuracy rate for samples of size 2000. For data requiring transformation, using only the default starting value (line 4) led to negative correlations of component number accuracy rate and sample size for the AIC and BIC criteria. For data requiring transformation analyzed with multiple RSVs (line 2), only the CLC and ICLBIC had a positive correlation between component number accuracy rate and sample size. The accuracy rate of the ICL-BIC for 2000 observations was somewhat higher than the CLC rate (87.8% compared to 82.8%).

Table 3.3.1: Correlations of accuracy rate of measure with sample size for selected experimental conditions (equal within-component variance, equal mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICLBIC
1. Multiple RSVs, no Transformation	0.68(99.7)	0.98(90.2)	0.91(99.9)	0.79(99.7)	0.72(99.8)
2. Multiple RSVs, Transformation	-0.61(0.0)	N (0.0)	-0.83(4.2)	0.96(82.8)	0.94(87.8)
3. Default SV, no transformation	0.59(100.0)	0.99(90.6)	0.59(100.0)	0.59(100.0)	0.64(100.0)
4. Default SV, transformation	-0.65(0.0)	-0.59(0.0)	0.75(100.0)	0.93(97.5)	0.92(99.8)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000

One interesting aspect of the data is the component number accuracy rate for a sample requiring transformation with estimates computed from multiple RSVs. Figure 3.3.1 displays the component number accuracy rates. The CLC and ICLBIC criteria have higher component number accuracy rates than the other four criteria in this situation.

Figure 3.3.1: 3-D representation of probability of component number accuracy rates by sample size (250, 500, 1000, and 2000) for samples requiring transformation using multiple RSVs, equal within-component variance and 4-standard deviation separation between adjacent component means

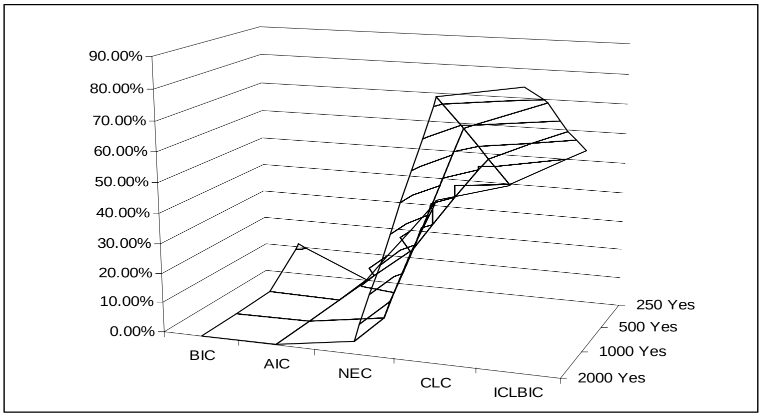


Table 3.3.2 presents the results for equiprobable components with equal within-component variance and a 2-standard deviation separation between component means. BIC and AIC had high correlations for data not requiring transformation (line 1), with the AIC having a much higher component number accuracy rate for a sample of 2000 (70.7%) compared to BIC (20.1%). There was not an effective criterion for data requiring transformation, as shown in Figure 3.3.2. Only the NEC criterion had a positive correlation, but its component number accuracy rate was only 0.3% with 2000 observations.

Table 3.3.2: Correlations of accuracy rate of measure with sample size for selected experimental conditions (equal within-component variance, equal mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICLBIC
1. Multiple RSVs, no Transformation	0.99(20.1)	0.97(70.7)	0.92(0.4)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	-0.89(0.0)	-0.60(0.0)	0.92(0.3)	-0.59(0.0)	N (0.0)
3. Default SV, no transformation	0.98(20.7)	0.99(77.7)	N (0.0)	N (0.0)	N (0.0)
4. Default SV, transformation	-0.80(2.6)	-0.72(0.0)	-0.59(0.0)	-0.59(0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000

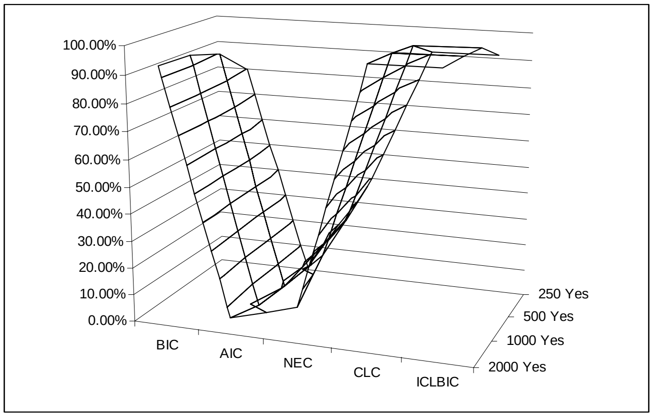
Table 3.3.3 presents the correlation coefficients of the component number accuracy rate of each criterion with sample size for equiprobable component distributions with the variance of each component proportional to the mean μ_i of the component. The average separation of the two extreme components with the middle component is 4 standard deviations. The BIC, CLC, and ICL-BIC had high component number accuracy rate for samples of size 2000 and strong correlation of accuracy and sample size. The use of multiple RSVs greatly diminished the accuracy rates of the AIC and NEC criteria. For data requiring transformation, BIC, CLC, and ICL-BIC had high component number accuracy rates for samples of size 2000 and strong correlations. The other criteria had low component number accuracy rates when multiple RSVs were used, as shown in Figure 3.3.2.

Table 3.3.3: Correlations of accuracy rate of measure with sample size for selected experimental conditions (unequal within-component variance, equal mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICLBIC
1. Multiple RSVs, no Transformation	0.93(97.4)	1.00(6.2)	-0.97(34.4)	0.87(99.9)	0.91(99.9)
2. Multiple RSVs, Transformation	0.76(93.8)	0.97(6.2)	-0.99(13.1)	0.76(99.4)	0.72(99.6)
3. Default SV, no transformation	0.95(99.6)	0.93(94.5)	0.59(100.0)	0.63(100.0)	0.67(100.0)
4. Default SV, transformation	0.69(100.0)	-0.76(60.7)	0.59(100.0)	0.68(100.0)	0.67(100.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000

Figure 3.3.2: 3-D representation of probability of component number accuracy rates by sample size (250, 500, 1000, and 2000) for samples requiring transformation using multiple RSVs, unequal within-component variance and 4-standard deviation separation between adjacent component means



In Table 3.3.4, I report the correlation coefficients of the component number accuracy rate of each criterion with sample size are reported for equiprobable component distributions with the variance of each component proportional to the mean μ_i of component$_i$. The separation is 2 standard deviations between adjacent components. The NEC, CLC, and ICL-BIC had component number accuracy rate 0 for sample size 2000. The BIC was the only criteria with minimal component number accuracy rate, as shown in Figures 3.3.3 and 3.3.4.

Table 3.3.4: Correlations of accuracy rate of measure with sample size for selected experimental conditions (unequal within-component variance, equal mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICLBIC
1. Multiple RSVs, no Transformation	0.90(26.5)	0.97(10.1)	N (0.0)	-0.59(0.0)	N (0.0)
2. Multiple RSVs, Transformation	-0.45(48.1)	-0.59(0.0)	N (0.0)	N (0.0)	N (0.0)
3. Default SV, no transformation	1.00(35.4)	0.99(46.3)	N (0.0)	N (0.0)	N (0.0)
4. Default SV, transformation	0.03(71.6)	-0.87(0.0)	N (0.0)	-0.59(0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000

Figure 3.3.3: 3-D representation of probability of component number accuracy rates by sample size (250, 500, 1000, and 2000) using multiple RSVs, unequal within-component variance, no transformation, and 2-standard deviation separation between adjacent component means

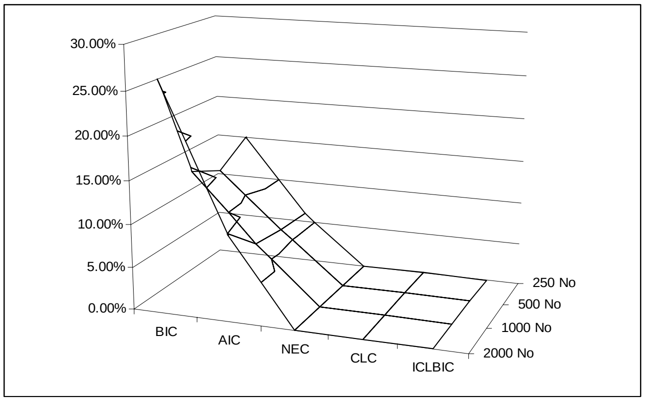
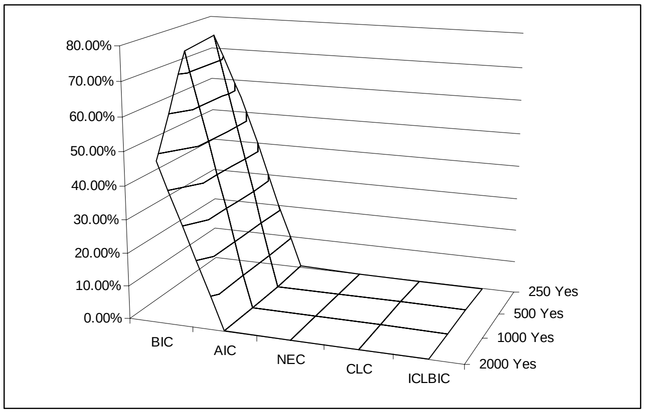


Figure 3.3.4: 3-D representation of probability of component number accuracy rates by sample size (250, 500, 1000, and 2000) using multiple RSVs, unequal within-component variance, transformation, and 2-standard deviation separation between adjacent component means



In summary, when there was 4-standard deviation separation between adjacent component means and the intensity data did not require transformation and had equal within-component variances, each of the six criteria tested had high component number accuracy rates. When data were transformed, CLC and ICL-BIC had the highest component number accuracy rates. When intensity data were from a mixture distribution with unequal within-component variances, BIC, CLC, and ICL-BIC were the best measures whether or not the data were transformed.

When there was 2-standard deviation separation between adjacent components means, none of the criterion had component number accuracy rates greater than 50% for intensity data from a mixture distribution with equal within-component variances. When data were transformed, and the within-component variances were unequal, the BIC criterion had the largest average component number accuracy rate (69%).

Chapter 4 Six-component results

The scenario for the six-component model is derived from the model for a gene with three alleles (1, 2, and 3) with average intensities 1, 2 and 3 respectively. The six possible genotypes (11, 12, 13, 22, 23, and 33) would then have average intensities 2, 3, 4, 4, 5, and 6 respectively. Assuming that the 1/3 and the 2/2 genotype have different mean intensities, I set the means at 2, 3, 4, 5, 6 and 7 respectively. In the next three sections, I report the ANOVA tables for each criterion, the average component number accuracy rates for selected factors explaining 5% or more of the total variation, the average component number accuracy rates for sources used in Chapter Three and the correlation coefficient of component number accuracy rate with sample size.

4.1 ANOVA tables of result

Table 4.1.1 contains the sum of squares of all main effects and selected two-way and three-way interactions using the component number accuracy rate as the dependent variable. The two-way and three-way interactions have to either represent at 5% of the total variation, or are essential for a hierarchical model. The remainder source is the sum of the unreported interaction terms and is used as the error sum of squares. An F-statistic greater than 6.8 corresponds to a p-value ≤ 0.01.

For BIC, among the main effects, those for separation (*C*), distribution of components (*E*), and multiple RSVs (*F*) explain at least 5% of the variation (Table 4.1.1). Together, they explain 37%. Each of the interactions $C \times E$, $C \times G$, $E \times F$, and $B \times C \times G$ explains 5% or more of the variation. Together, these 7 sources explain 65% of the variation. The F test is also highly significant for $G, B \times C$, $B \times G$, $C \times F$, and $C \times E \times F$.

Under AIC, mixture pattern (*E*) and transformation (*G*) are the main effects with more than 5% contribution to the total variation (Table 4.1.1). Additionally, the interactions $B \times F$, $B \times G$, $C \times E$, $C \times G$, $E \times F$, $B \times C \times F$, $B \times C \times G$, and $C \times E \times F$ contribute more than 5% each to the total variation. Together, these ten sources explain roughly 68% of total variation. The F test is also significant for *B*, *C*, *F*, $B \times C$, and $B \times E \times G$.

For the NEC criterion, the main effects of separation (*C*), mixture pattern (*E*) and transformation (*G*) each explain at least 5% to the total variation (Table 4.1.1). Each of the interactions $C \times E$, $C \times G$, $E \times F$, $E \times G$, $C \times E \times F$, and $C \times E \times G$ also explains at least 5 % of the variation. Together, these eight sources explain roughly 67% of the variation. The F test is also highly significant for *B*, $B \times C$, $B \times E$, $B \times F$, $B \times G$, $B \times C \times E$, $B \times C \times F$, and $B \times C \times G$.

For the CLC criterion, each of the main effects of separation (*C*), multiple RSVs (*F*), and transformation (*G*) explained more than 5% of total variation (Table 4.1.1). Each of the interactions $C \times F$ and $C \times G$ also explained more than 5% of total variation. These five sources explained approximately 64% of total variation. The F test is also significant for *B*, *E*, $B \times C$, $B \times F$, $B \times G$, $C \times E$, $E \times F$, $E \times G$, $B \times C \times F$, $B \times C \times G$, $C \times E \times F$ and $C \times E \times G$.

For the BIC version of ICL (ICL-BIC), the main effects of separation (*C*), multiple RSVs (*F*) and transformation (*G*) individually explained more than 5% of variation (Table 4.1.1). The interactions $C \times F$ and $C \times G$ also explained more than 5% of variation. These four sources explain about 58% of variation. The F test was also significant for B, E, $B\times C$, $B\times F$, $B\times G$, $C\times E$, $E\times F$, $E\times G$, $B\times C\times F$, $B\times C\times G$, $C\times E\times F$ and $C\times E\times G$.

Table 4.1.1: ANOVA tables for BIC and AIC

		BIC		AIC		NEC		CLC		ICL-BIC	
Source	DF	%TSS	F	%TSS	F	%TSS	F	%TSS	F	%TSS	F
B: Constant Variance	1	0.4%	3.9	1.1%	9.0	3.2%	34.4	2.0%	19.6	0.7%	4.9
C: Separation	1	24.4%	229.6	3.6%	28.2	11.5%	122.2	18.5%	183.4	19.7%	140.2
D: Sample Size	3	0.9%	2.9	0.2%	0.5	0.2%	0.7	0.1%	0.4	0.8%	2.0
E: Mixture Proportion	2	5.6%	26.6	7.4%	29.0	8.0%	42.8	2.3%	11.4	3.1%	10.9
F: Multiple RSVs	1	7.1%	67.3	0.8%	6.3	0.3%	3.7	5.6%	55.4	5.5%	39.4
G: Transformation	1	1.5%	14.0	8.2%	64.7	9.1%	97.3	16.1%	159.3	13.5%	96.0
$B \times C$	1	1.7%	16.5	2.0%	15.7	3.2%	34.4	1.4%	13.9	0.6%	4.2
$B \times E$	2	0.6%	3.0	0.5%	1.9	1.3%	7.0	0.1%	0.3	0.1%	0.4
$B \times F$	1	0.2%	2.0	6.6%	52.1	1.0%	10.7	1.0%	10.0	0.2%	1.5
$B \times G$	1	3.0%	27.9	8.1%	63.7	2.0%	21.7	0.8%	7.8	1.4%	10.3
$C \times E$	2	7.5%	35.5	6.1%	23.9	8.0%	42.8	2.7%	13.4	3.1%	11.2
$C \times F$	1	4.2%	39.5	0.0%	0.4	0.3%	3.7	4.6%	45.3	5.3%	37.5
$C \times G$	1	9.3%	87.3	6.1%	47.8	9.1%	97.1	17.9%	177.2	13.9%	98.9
$E \times F$	2	5.1%	24.1	7.3%	28.8	5.4%	28.8	2.4%	11.9	2.0%	7.0
$E \times G$	2	0.7%	3.3	0.4%	1.5	5.9%	31.3	2.4%	11.9	2.0%	7.1
$B \times C \times E$	2	0.6%	3.0	0.1%	0.5	1.3%	7.0	0.2%	1.0	0.2%	0.6
$B \times C \times F$	1	0.0%	0.1	6.9%	54.1	1.0%	10.7	0.6%	6.1	0.2%	1.1
$B \times C \times G$	1	6.0%	56.2	5.5%	43.0	2.0%	21.8	1.2%	12.3	1.6%	11.3
$B \times E \times G$	2	0.4%	2.0	1.2%	4.9	0.6%	3.4	0.1%	0.7	0.1%	0.3
$C \times E \times F$	2	3.6%	16.9	6.9%	27.1	5.4%	28.8	2.0%	9.7	1.8%	6.4
$C \times E \times G$	2	0.2%	1.1	0.9%	3.4	5.9%	31.3	2.0%	10.1	1.9%	6.7
Remainder	159	16.9%		20.2%		14.9%		16.0%		22.4%	

4.2 Interaction of significant factors

In general, the component number accuracy rates are much lower for the six component model than the three component model. Table 4.2.1 contains the average component number accuracy rates by separation (*C*), transformation (*G*), and constant variance (*B*) factors for the BIC criterion. For example, when separation is 4-standard deviations and no transformation of intensities was made, the accuracy rates were 66.9% for equal within component variance (compared to 99.7% for three components) and 50.8% (compared to 94.3% with three components). Otherwise, the patterns were similar. That is, the component number accuracy rates were uniformly small when the component separation was 2-standard deviations. Additionally, the component number accuracy rate for data with equal within-component variance requiring transformation was the lowest among the 4-standard deviation difference settings.

Table 4.2.2 shows that the average component number accuracy rates are below 20% for 2-standard deviation separation between adjacent components means for each distribution of component probabilities. Additionally, with HWE and skewed component mixture proportions, the average component number accuracy rates are above 50% when multiple RSVs are used with 4-standard deviation separation between adjacent component means. With equal component mixture proportion, the average component number accuracy rates are essentially the same under the multiple RSVs factor.

Table 4.2.1: Table of average component number accuracy rates for separation, transformation, and constant variance setting for the BIC criterion using 1000 replicates

	C: Separation				
B: Constant Variance		4σ		2σ	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
	Yes	5.2±6.7	66.9±47.1	18.9±13.9	0.3±0.9
	No	50.1±40.5	50.8±40.2	8.6±17.8	0.5±0.8

Average accuracy rate ± SD, average over 24 settings

Table 4.2.2: Table of average component number accuracy rates for separation, mixture proportion, and multiple RSVs setting for the BIC criterion using 1000 replicates

	C: Separation				
E: Mixture Proportion		4σ		2σ	
		F: Multiple RSVs		F: Multiple RSVs	
		Yes	No	Yes	No
	Equal	65.6±37.1	73.6±38.0	5.5±9.7	4.5±10.6
	HWE	55.5±41.7	1.8±3.2	6.4±11.1	5.3±12.0
	Skew	60.5±40.8	2.5±3.9	16.2±21.9	4.6±9.9

Average accuracy rate ± SD, average over 16 settings

The component number accuracy rates for the AIC criterion are typically lower than the rates for the BIC. As shown in Table 4.2.3, the average component accuracy rates are below 50%. The use of multiple RSVs typically reduces the component number accuracy rate, except for no transformation with equal within-component variance. As shown in table 4.3.4, the average component number accuracy rates are slightly higher for multiple RSVs with skewed and HWE component mixture proportions.

Table 4.2.3: Table of average component number accuracy rates for constant variance, transformation, and multiple RSVs setting for the AIC criterion using 1000 replicates

	F: Multiple RSVs				
B: Constant Variance		Yes		No	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
	Yes	4.0±7.3	45.5±35.9	7.5±9.2	24.7±32.4
	No	4.4±4.3	7.7±6.9	25.3±27.6	22.3±25.7

Average accuracy rate ± SD, average over 24 settings

Table 4.2.4: Table of average component number accuracy rates for separation, mixture proportion, and multiple RSVs setting for the AIC criterion using 1000 replicates

		C: Separation			
		4σ		2σ	
		F: Multiple RSVs		F: Multiple RSVs	
		Yes	No	Yes	No
E: Mixture Proportion	Equal	20.3±34.8	62.3±37.0	10.6±6.0	17.0±10.7
	HWE	21.4±36.0	6.5±5.4	11.0±6.2	13.5±10.8
	Skew	20.8±35.8	4.0±8.6	8.4±6.7	16.6±11.1

Average accuracy rate ± SD, average over 16 settings

With multiple RSVs, the average component number accuracy rates using the NEC criterion were below 30% (Table 4.2.5). The NEC criterion for six components typically had lower average component number accuracy rate when compared to the three component model. Using the default setting of the MCLUST package, the average component number accuracy rate was 84% when the separation was at 4-standard deviation between adjacent component means with equi-probable components and the transformation factor at the no setting. For every other setting, the average component number accuracy rate was below 10%

Table 4.2.5: Table of average component number accuracy rates for multiple RSVs, separation, transformation, and mixture proportion setting for the NEC criterion using 1000 replicates

a.-F: Multiple RSVs (yes)

		C: Separation			
		4σ		2σ	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
E: Mixture Proportion	Equal	0.0±0.0	29.5±31.6	0.0±0.0	0.1±0.4
	HWE	0.6±1.6	10.2±12.9	0.0±0.0	0.3±0.5
	Skew	0.4±1.1	28.4±30.4	0.0±0.0	0.1±0.4

Average accuracy rate ± SD, average over 8 settings

b.-F: Multiple RSVs (No)

		C: Separation			
		4σ		2σ	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
E: Mixture Proportion	Equal	7.9±13.0	83.9±18.3	0.0±0.0	0.0±0.0
	HWE	0.1±0.2	5.3±12.4	0.0±0.0	0.0±0.0
	skew	0.1±0.3	0.9±2.2	0.0±0.0	0.0±0.0

Average accuracy rate ± SD, average over 8 settings

For the CLC criterion, component number accuracy rates for the six component model with 2-standard deviation separation were less than 5%. With 4-standard deviation separation and data not requiring transformation, using multiple RSV's had an average component number accuracy rate of 75.9% (Table 4.2.6). The average component number accuracy rate for the yes level of the transformation setting was less than 5%.Table 4.2.7 contains the average component number accuracy rates for separation (*C*), constant variance (*B*), and transformation (*G*) and corresponds to table 3.2.5. The largest average component number accuracy rate was 65% for equal within-component variance with no transformation and a 4-standard deviation separation between adjacent component means.

Table 4.2.6: Table of average component number accuracy rates for separation, multiple RSVs, and transformation setting for the CLC criterion using 1000 replicates

		C: Separation			
		4σ		2σ	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
F: Multiple RSVs	Yes	3.5±4.6	75.9±25.8	2.7±5.4	0.0±0.1
	No	0.0±0.1	25.8±38.1	0.0±0.0	0.0±0.0

Average accuracy rate ± SD, C = 24 cases

Table 4.2.7: Table of average component number accuracy rates for separation, constant variance, and transformation setting for the CLC criterion using 1000 replicates

		C: Separation			
B: Constant Variance		4σ		2σ	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
	Yes	3.5±4.7	64.5±46.8	2.7±5.6	0.0±0.0
	No	0.0±0.0	37.2±29.7	0.0±0.0	0.0±0.1

Average accuracy rate ± SD, C = 24 cases

The component number accuracy rates for ICL-BIC criterion were very similar to those of the CLC criterion, as shown in Table 4.2.8. The largest average component number accuracy rate was 76% using multiple RSVs on data not transformed with 4-standard deviation separation between adjacent component means. The other settings are all below 50%.

Table 4.2.9 contains the average component number accuracy rates for separation (*C*), constant variance (*B*), and transformation (*G*) and corresponds to table 3.2.6. With one exception, the average component number accuracy rates were below 40%. The largest average component number accuracy rate is 64.5% for data with equal within-component variance without transformation and 4-standard deviation separation between adjacent component means.

Table 4.2.8: Table of average component number accuracy rates for separation, multiple RSVs, and transformation setting for the ICL-BIC criterion using 1000 replicates

		C: Separation			
F: Multiple RSVs		4σ		2σ	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
	Yes	9.7±25.3	76.2±29.3	0.7±1.6	0.0±0.0
	No	0.0±0.1	27.1±40.0	0.0±0.0	0.0±0.0

Average accuracy rate ± SD, C = 24 cases

Table 4.2.9: Table of average component number accuracy rates for separation, constant variance, and transformation setting for the ICL-BIC criterion using 1000 replicates

		C: Separation			
B: Constant Variance		4σ		2σ	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
	Yes	2.1±2.8	64.5±46.8	0.7±1.6	0.0±0.0
	No	7.6±25.8	38.8±34.5	0.0±0.0	0.0±0.0

Average accuracy rate ± SD, C = 24 cases

4.3 Correlation coefficient of component number accuracy rate for the six criteria with sample size for six equiprobable components

As in chapter three, I report the correlation coefficients of the component number accuracy rate with sample size for the six criteria and report the component number accuracy rate for sample size 2000 in parenthesis after the correlation coefficient when the six components are equiprobable. The corresponding results for HWE and skewed mixture proportions are reported in appendix B.

Table 4.3.1 contains the results for intensity data with 4-standard deviation separation between adjacent component means and equal within-component variance. For intensity data not requiring transformation, the component number accuracy rate increased as sample size increased (lines 1 and 3). For intensity data requiring transformation, the component number accuracy rate decreased with an increase in sample size in all criteria except for the NEC criterion.

Table 4.3.1: Correlations of accuracy rate of measure with sample size for selected experimental conditions (equal within-component variance, equal mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.63 (99.9)	0.96 (85.2)	0.62 (63.0)	0.71 (99.9)	0.65 (100.0)
2. Multiple RSVs, Transformation	-0.88 (0.8)	-0.69 (0.0)	N (0.0)	-0.97 (2.3)	-0.99 (0.9)
3. Default SV, no transformation	0.69 (100.0)	0.95 (94.0)	0.73 (100.0)	0.81 (99.7)	0.65 (100.0)
4. Default SV, transformation	-0.94 (1.3)	-0.80 (0.2)	-0.92 (0.0)	-0.76 (0.0)	-0.74 (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000

Table 4.3.2 contains the correlation coefficients of component number accuracy rate with sample size for intensity data with 2-standard deviation separation between adjacent component means and equal within-component variance. The performance of each criterion was poor for these settings. For intensity data not requiring transformation (lines 1 and 3), although the component number accuracy rate increased as sample size increased for the BIC and the AIC criteria, the accuracy rate was below 30% for $n = 2000$. For the other criteria, the component number accuracy rates were 0 for each sample size studied. For intensity data requiring transformation (lines 2 and 4), the component number accuracy rates decreased with increased sample size for BIC, AIC, CLC and ICL-BIC (line 2), decreased with an increase in sample size for AIC (lines 4), increased with sample size for BIC line (4) and 0 for all other cases.

Table 4.3.2: Correlations of accuracy rate of measure with sample size for selected experimental conditions (equal within-component variance, equal mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.84 (0.1)	0.97 (22.4)	N (0.0)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	-0.12 (5.8)	-0.76 (0.0)	N (0.0)	-0.59 (0.0)	-0.59(0.0)
3. Default SV, no transformation	0.92 (0.1)	0.98 (25.9)	N (0.0)	N (0.0)	N (0.0)
4. Default SV, transformation	0.32 (15.7)	-0.95 (1.8)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000

Table 4.3.3 contains the correlation coefficients for intensity data from a population with 4-standard deviation separation between adjacent component means and unequal within-component variance. The BIC criterion performed well in this case for both transformed and un-transformed data. Specifically, the component number accuracy rate increased as sample size increased, and the component number accuracy rate was almost 94% for $n = 2000$. For the AIC criterion, the use of multiple RSVs was associated with a decrease in component number accuracy rate as sample size increased. When intensity data did not require transformation (lines 1 and 3), the component number accuracy rate increased as sample size increased for the CLC criterion. The use of multiple RSVs was associated with an increase in the component number accuracy rate as sample size increase for the ICL-BIC criterion (lines 1 and 3). The NEC criterion had component number accuracy rates 0 for almost all cases.

Table 4.3.3: Correlations of accuracy rate of measure with sample size for selected experimental conditions (unequal within-component variance, equal mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.84 (93.8)	-0.46 (0.2)	N (0.0)	0.72 (61.6)	0.42 (61.3)
2. Multiple RSVs, Transformation	0.84 (93.9)	-0.45 (0.3)	N (0.0)	N (0.0)	0.05 (0.0)
3. Default SV, no transformation	0.74 (100.0)	0.78 (92.8)	-0.61 (62.7)	0.47 (62.7)	-0.59 (62.7)
4. Default SV, transformation	0.70 (100.0)	0.87 (92.9)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000

Table 4.3.4 contains the correlation coefficients for intensity data with 2-standard deviation separation between adjacent component means and unequal within-component variance. None of the criteria performed well, with maximum component number accuracy rate for $n = 2000$ equal to 42.8%. The component number accuracy rates were zero for each sample size studied for CLC, and ICL-BIC. For un-transformed data (lines 1 and 3), the component number accuracy rate using the AIC criterion increased as sample size increased.

Table 4.3.4: Correlations of accuracy rate of measure with sample size for selected experimental conditions (unequal within-component variance, equal mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	-0.84 (0.4)	0.38 (13.2)	0.92 (0.1)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	-0.72 (3.1)	-0.10 (9.4)	N (0.0)	N (0.0)	N (0.0)
3. Default SV, no transformation	-0.59 (0.0)	0.97 (28.8)	N (0.0)	N (0.0)	N (0.0)
4. Default SV, transformation	0.98 (0.2)	1.00 (42.8)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000

In summary, when there was 4-standard deviation separation between adjacent component means and the intensity data were not transformed and had equal within-component variances, each of the six criteria tested had high component number accuracy rates for equiprobable components . In the skewed and HWE cases, the use of multiple RSVs increased the component number accuracy rate for BIC, AIC, CLC, and ICL-BIC when intensity data are from a mixture distribution with equal within-component variances (see Appendix B). When data were transformed, none of the six criteria had component number accuracy rates above 43%. When intensity data were from a mixture distribution with unequal within-component variances, BIC was the best criterion whether or not the data were transformed. In the skewed and HWE cases, multiple RSVs were required once again for the BIC to have high component number accuracy rate at sample size 1000 or more. The component number accuracy rates were 90.4 for skewed proportions and 85.6 for HWE proportions with 4-standard deviation separation between adjacent component means.

When there was 2-standard deviation separation between adjacent components means, none of the criterion had component number accuracy rates greater than 50% for intensity data from a mixture distribution with equal within-component variances regardless of the transformation setting of data.

Chapter 5 Box-Cox transformation results

It is common in analyzing CNP intensity data to apply the Box-Cox transformation to the data (Kim et al., 2008). Since the BIC criterion had low component number classification accuracy rate when applied to data requiring transformation, I studied the extent to which the use of the Box-Cox transformation to normality would change the component number accuracy rate. In this work, I use the Box-Cox transformation (as programmed in R) to make the sample as close to a sample from a single component normal distribution as possible. I want to address the research questions: How much does this Box-Cox transformation change the component number accuracy rate of the criteria when it is not needed? Does this Box-Cox transformation improve the component number accuracy rate when it is needed?

For data not requiring transformation, the estimated power transform ($\hat{\lambda}$) is expected to be 1. The expected value of $\hat{\lambda}$ equals ½ for data requiring transformation since the transformed data is a normal mixture squared. One thousand samples have been simulated at each design setting. The average value of $\hat{\lambda}$ is ($\bar{\hat{\lambda}} = \frac{\sum_{i=1}^{1000} \hat{\lambda}_i}{1000}$) for each setting should be close to 1 for a normal mixture and close to ½ for a normal mixture squared. For each mixture model, I report the mean of the average $\hat{\lambda}$ with its standard deviation in table 5.1.

Table 5.1: Mean of Average power ($\bar{\hat{\lambda}}$) ±*SD* used in the Box Cox transformation for each n-component model (c = 36, the number of cases in each group of data)

	Normal Mixture	Normal Mixture Squared
Three-component model	0.39±0.36	0.19±0.19
Six-component model	0.67±0.23	0.67±0.22

As in chapters 3 and 4, I report the analysis of variance table for the six classification criteria after application of a Box-Cox transformation in tables 5.2.1, 5.2.2, 5.5.1 and 5.5.2. Also, I report tables of averages corresponding to the factors explaining at least 5% of the total sums of squares of each criterion. Finally, I report the correlation coefficient of the component number accuracy rates with sample size for each criterion.

5.1 Results of Box-Cox transformation for the three-component analysis

Table 5.1.1., as usual, is the ANOVA table for the component number accuracy rate for the BIC, AIC, NEC, CLC, and ICL-BIC criteria. An F-value > 7.5 corresponds to a p-value ≤ 0.01.

For the BIC criterion, constant variance (*B*), separation (*C*), and mixture proportion (*E*) are most significant main effects (Table 5.1.1). A model that that includes these main effects, the three two-factor interactions ($B \times C$, $B \times E$, and $C \times E$) and the three-factor interaction ($B \times C \times E$) explains roughly 96% of variation of component number accuracy rate. The F-test was significant for the interactions $B \times D$, $C \times D$, $D \times E$, $B \times C \times D$, $B \times D \times E$, $C \times D \times E$, and $B \times C \times D \times E$. The main effect of the transformation factor *G* accounts for 0.01of the variation and is not involved in any significant interactions, thus documenting the effectiveness of the Box-Cox transformation.

Using the AIC criterion, the main effects of constant variance (*B*), separation (*C*), and mixture proportion (*E*) explained each more than 5% of the variation (Table 5.1.1). Also, the interactions $B \times C$, $B \times E$, $C \times D \times E$, and $B \times C \times D \times E$ explained 5% or more of that total variation of AIC component number accuracy rate. Together, these seven sources accounted for approximately 91% of TSS. The F test was significant for the interactions $B \times D$, $C \times D$, $C \times E$, $B \times C \times D$, $B \times C \times E$, and $B \times D \times E$. Again, the main effect of the transformation factor *G* explains a very low percentage of the variation (i.e., 0.1%), and none its interactions are significant.

For the NEC criterion, separation (*C*), and mixture proportion (*E*) each explained 55.1% and 10.1 % respectively of the variation (Table 5.1.1). The interactions $C \times E$ and $B \times C \times E$ explained an additional 29.5% of the variation. The F-test was significant for constant variance (*B*) and the interactions $B \times C$, $B \times D$, $B \times E$, $D \times E$, $B \times C \times D$, $B \times D \times E$, and $C \times D \times E$.

For the CLC criterion, separation (*C*) explained 68.4% of the variation (Table 5.1.1). Additionally, the interactions $B \times C$, $C \times E$, and $B \times C \times E$ explained together 21% of variation. The F-test was significant for constant variance (*B*), mixture proportion (*E*) and the interactions $B \times D$, $B \times E$, $C \times D$, $B \times C \times D$, $B \times D \times E$, $C \times D \times E$, and $B \times C \times D \times E$.

Under the BIC approximation of ICL, separation (*C*) explained 67.8% of the total ICL-BIC variation (Table 5.1.1). The interaction $B \times C$, $C \times E$, and $B \times C \times E$ explained an additional 21.4% of the variation. The F-test was significant for constant variance (*B*), sample size (D), mixture proportion (*E*), and the interactions $B \times D$, $B \times E$, $C \times D$, $B \times C \times D$, $B \times D \times E$, $C \times D \times E$, and $B \times C \times D \times E$.

The transformation factor G and its interactions were not significant for any criteria documenting the effectiveness of this Box-Cox transformation. It is also notable that sample size is included in some of the significant interactions.

Table 5.1.1: ANOVA tables for BIC, AIC, NEC, CLC, and ICL-BIC

		BIC		AIC		NEC		CLC		ICL-BIC	
Source	DF	%TSS	F	%TSS	F	%TSS	F	%TSS	F	%TSS	F
B: Constant Variance	1	17.5%	17585.9	28.9%	3912.9	0.0%	10.9	3.5%	12667.0	3.5%	63044.3
C: Separation	2	34.8%	17513.3	15.5%	1047.6	55.1%	33750.0	68.4%	122989.0	67.8%	615469.0
D: Sample Size	2	0.0%	4.2	1.2%	81.2	0.0%	0.6	0.0%	0.0	0.0%	17.3
E: Mixture Proportion	1	14.9%	14967.6	17.5%	2377.4	10.1%	12308.0	3.7%	13230.2	3.8%	68408.4
G: Transformation	1	0.0%	5.9	0.1%	14.0	0.0%	0.0	0.0%	0.5	0.0%	0.6
$B\times C$	2	8.2%	4109.6	6.3%	429.8	0.0%	10.9	6.8%	12240.8	6.9%	62581.8
$B\times D$	2	0.1%	46.0	1.9%	127.5	0.1%	64.2	0.0%	7.2	0.0%	71.9
$B\times E$	1	5.3%	5378.8	12.5%	1693.9	4.7%	5745.4	3.4%	12185.3	3.5%	63088.5
$C\times D$	4	2.0%	508.5	1.2%	40.8	0.0%	0.6	0.0%	5.2	0.0%	25.0
$C\times E$	2	9.9%	4973.5	1.1%	76.3	20.1%	12308.0	7.4%	13209.9	7.6%	68546.5
$D\times E$	2	0.0%	15.1	0.1%	3.3	0.1%	60.8	0.0%	3.3	0.0%	1.3
$B\times C\times D$	4	0.4%	108.7	1.4%	48.1	0.2%	64.2	0.0%	20.2	0.0%	86.7
$B\times C\times E$	2	5.2%	2622.9	1.3%	87.5	9.4%	5745.4	6.8%	12126.9	6.9%	62956.1
$B\times D\times E$	2	1.3%	638.0	0.3%	22.0	0.0%	0.5	0.0%	15.8	0.0%	57.6
$C\times D\times E$	4	0.1%	11.4	5.2%	175.9	0.2%	60.8	0.0%	3.9	0.0%	1.0
$B\times C\times D\times E$	4	0.4%	88.5	5.3%	179.3	0.0%	0.5	0.0%	17.3	0.0%	60.2
Remainder	35	0.0%		0.3%		0.0%		0.0%		0.0%	

5.2 Interaction of significant factors

Table 5.2.1 presents the classification accuracy rates for the BIC after using the Box-Cox transformation analogous to Table 3.2.1. Only multiple RSVs are used. As expected, the performance of the BIC criterion is not affected by transformation. That is, on average, the component number accuracy rate is the same for transformation (yes and no). BIC has good component number accuracy rate for unequal within-component variances at 3-standard deviation and 4-standard deviation separation between adjacent means. BIC has relatively low component number accuracy rates for equal within-component variance even with 4 standard deviation separation.

Table 5.2.2 documents the component number accuracy rates for sources explaining 5% or more of BIC variation (*i.e.*, *B* (constant variance), *C* (separation), and *E* (component mixing proportions). The component number accuracy rates are greater than 60% when there is at least a 3-standard deviation separation between adjacent component means and the components are equiprobable. The accuracy rates are highest when the component variances are non-constant. At 2-standard deviation separation between adjacent means, the average component number accuracy rates are below 20%. For any line of table 5.2.2., the accuracy rate is higher for equal component proportions than skewed proportions. The component number accuracy rate is very low for skewed mixing proportions even when the separation between components is large.

Table 5.2.1: Table of average component number accuracy rates for separation, constant variance, and transformation setting for the BIC criterion (1000 replicates after Box-Cox transformation)

		B: Constant Variance			
		Yes		No	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
C: Separation	2σ	13.9±12.8	14.1±12.0	15.6±5.7	15.5±4.8
	3σ	45.1±48.0	44.5±47.1	87.3±8.1	86.3±8.5
	4σ	32.3±37.7	31.4±37.1	84.0±14.1	83.1±14.0

Average accuracy rate ± SD, average over 6 settings

Table 5.2.2: Table of average component number accuracy rates for separation, constant variance, and mixture proportion setting for the BIC criterion (1000 replicates after Box-Cox transformation)

		B: Constant Variance			
		Yes		No	
		E: Mixture Proportion		E: Mixture Proportion	
		Equal	Skewed	Equal	Skewed
C: Separation	2σ	9.2±6.4	18.7±14.5	10.7±6. 7	15.1±3.2
	3σ	87.9±7.2	1.7±2.4	93.0±5.0	80.3±4.1
	4σ	63. 7±19.1	0.0±0. 1	94.3±3.9	72.8±10.1

Average accuracy rate ± SD, average over 6 settings

Table 5.2.3 contains the AIC component number accuracy rates for settings of the constant variance (*B*) and transformation (*G*) factors averaged over all other factors and is analogous to the multiple RSVs section of table 3.2.2. Since I only use multiple RSVs in this chapter, this table is only half of the size of Table 3.2.2. The component number accuracy rates are lower than those of the BIC. Box-Cox transformation has made the transformation factor *G* not significant for the AIC as shown by the very similar accuracy rates. On average, equal within component variance is associated with a higher component number accuracy rate. Without Box-Cox transformation, the AIC component number accuracy rate was $65.0\% \pm 23.5\%$, compared to the $20.1\% \pm 21.0\%$ reported in Table 5.2.3.

Table 5.2.4 contains the means component number accuracy percentages of factors explaining 5% or more of the AIC criterion variation. The component number accuracy rate is higher for equiprobable components than for a skewed distribution. The component number accuracy rate is higher for equal within-component variance. The component number accuracy rates decrease as the separation of components increases. That is, the AIC is not a good criterion. As seen in table 5.2.4, all average component number accuracy rates are below 50%.

Table 5.2.3: Table of average component number accuracy rates for constant variance and transformation setting for the AIC criterion (1000 Replicates after Box-Cox Transformation)

		B: Constant Variance	
G: Transformation		Yes	No
	Yes	19.6±19.9	3.6±3.3
	No	20.1±21.0	2.5±2.5

Average accuracy rate ± SD, average over 18 settings

Table 5.2.4: Table of average component number accuracy rates for separation, constant variance, and mixture proportion setting for the AIC criterion (1000 Replicates after Box-Cox Transformation)

		B: Constant Variance			
		Yes		No	
		E: Mixture Proportion		E: Mixture Proportion	
		Equal	Skewed	Equal	Skewed
C: Separation	2σ	45.4±11.5	23.7±9. 6	8.0±1.7	5.0±1.0
	3σ	35.3±14.7	0.2±0.2	2.1±1.6	0.5±0.4
	4σ	18.8±14.0	0.0±0.0	2.2±1.7	0.4±0.3

Average accuracy rate ± SD, average over 6 settings

Table 5.2.5 contains the average component number accuracy rates for the NEC criterion and corresponds to table 3.3.3. The table contains the average component number accuracy rates using multiple RSVs for separation 2, 3, and 4 standard deviations. At 2- and 3-standard deviation separations between adjacent means, the NEC criterion did not have any correct component number classification. At 4-standard deviation separation, the average component number accuracy rate is 45%.

Table 5.2.6 contains the average component number accuracy rates for the sources explaining 5% or more of the NEC criterion total variation. The best component number accuracy is 91.6% at 4-standard deviation separation between adjacent means, equal within-component variance and equiprobable mixture proportion. With unequal within-component variances, equiprobable mixture proportion and a 4-standard deviation separation between adjacent component means, the means component number accuracy percentage is about 53%. The average component number accuracy rate is below 50% for every other combination of the three sources.

Table 5.2.5: Table of average component number accuracy rates for the separation setting for the NEC criterion (1000 Replicates after Box-Cox Transformation)

C: Separation		
2σ	3σ	4σ
0.0±0.0	0.0±0.0	45.0±33.9

Average accuracy rate ± SD, average over 24 settings

Table 5.2.6: Table of average component number accuracy rates for separation, constant variance, and mixture proportion setting for the NEC criterion (1000 Replicates after Box-Cox Transformation)

		B: Constant Variance			
		Yes		No	
		E: Mixture Proportion		E: Mixture Proportion	
		Equal	Skewed	Equal	Skewed
C: Separation	2σ	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	3σ	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	4σ	91.6±5.6	0.1±0.1	52.8±1.2	35.6±6.5

Average accuracy rate ± SD, average over 6 settings

Table 5.2.7 contains the average component number accuracy rate for the CLC criterion and corresponds to table 3.2.4. As expected, the average component number accuracy rate is no longer affected by transformation. The average component number accuracy rate is below 1% for 2-standard deviation and 3-standard deviation separation. It is at least 50% for cases where there is a 4-standard deviation separation between adjacent component means. The accuracy rate is highest (98%) for unequal within-component variances and 4-standard deviation separation between adjacent component means.

Table 5.2.8 contains the average component number accuracy rates for sources explaining 5% or more of the total variation of the CLC criterion. The CLC has an extremely low accuracy rate (3.2%) for skewed component probability distribution even with equal component variance and 4-standard deviation separation.

Table 5.2.7: Table of average component number accuracy rates for separation, constant variance, and transformation setting for the CLC criterion (1000 replicates after Box-Cox transformation)

		B: Constant Variance			
C: Separation		Yes		No	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
	2σ	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	3σ	0.1±0.1	0.1±0.1	0.6±0.9	0.6±0.8
	4σ	51.5±51.9	50.4±52.8	97.9±2.6	98.4±1.6

Average accuracy rate ± SD, average over 6 settings

Table 5.2.8: Table of average component number accuracy rates for separation, constant variance, and mixture proportion setting for the CLC criterion (1000 Replicates after Box-Cox Transformation)

		B: Constant Variance			
C: Separation		Yes		No	
		E: Mixture Proportion		E: Mixture Proportion	
		Equal	Skewed	Equal	Skewed
	2σ	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	3σ	0.1±0.2	0.0±0.0	0.6±0.8	0.6±0.9
	4σ	98.7±1.2	3.2±3.1	99. 2±0.9	97.2±2.5

Average accuracy rate ± SD, average over 6 settings

Table 5.2.9 contains the average component number accuracy rates for the ICL-BIC criterion using the same factors as used in table 3.2.5. Table 5.2.10 contains the averages of the component number accuracy rates for the sources explaining 5% or more of the variation of the ICL-BIC criterion. The pattern of results for the ICL-BIC criterion is similar to the pattern for the CLC criterion. That is, after Box-Cox transformation, the component number accuracy rate is not affected by transformation status. The average component number accuracy rate is below 1% for 2-standard deviation separation and 3-standard deviation separation. At 4-standard deviation separation between adjacent means, the average component number accuracy rate is at least 50%. The ICL-BIC criterion has a very low component number accuracy rate (2.1% in table 5.2.10) for a skewed mixing proportion distribution even when the separation is 4-standard deviations and the component variance is constant.

Table 5.2.9: Table of average component number accuracy rates for separation, constant variance, and transformation setting for the ICL-BIC criterion (1000 replicates after Box-Cox transformation)

		B: Constant Variance			
C: Separation		Yes		No	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
	2σ	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	3σ	0.0±0.0	0.0±0.0	0.1±0.17	0.13±0.23
	4σ	50.8±53.0	50.5±53.5	98.2±2.6	98.1±2.4

Average accuracy rate ± SD, average over 6 settings

Table 5.2.10: Table of average component number accuracy rates for separation, constant variance, and mixture proportion setting for the ICL-BIC criterion (1000 Replicates after Box-Cox Transformation)

	B: Constant Variance				
C: Separation		Yes		No	
		E: Mixture Proportion		E: Mixture Proportion	
		Equal	Skewed	Equal	Skewed
	2σ	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	3σ	0.0±0.0	0.0±0.0	0.1±0.2	0.2±0.2
	4σ	99.2±0.7	2.1±2.0	99.1±1.0	97.1±0. 7

Average accuracy rate ± SD, average over 6 settings

5.3 Correlation coefficient of component number accuracy rate of the six criteria with sample size for three equiprobable components

In this section, I report the correlation coefficient of component number accuracy rates with sample size for equiprobable components for the six criteria after using Box-Cox transformation. I also report in parentheses the component number accuracy rate for sample size 2000. The corresponding tables for the skewed distribution are given in appendix C.

Table 5.3.1 contains the correlation coefficients for intensity data with 4-standard deviation separation between component means and equal within-component variances. Regardless of the transformation status of the data, the component number accuracy rates increased as sample size increased for NEC, CLC and ICL-BIC (lines 1 and 2). For $n = 2000$, component number accuracy rates were above 95% for these criteria. For the BIC and AIC criteria, the component number accuracy rates decreased as sample size increased. The BIC component number accuracy rate was below 50% for $n = 2000$.

Table 5.3.1: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, equal mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	-1.00 (38.9)	-0.97 (3.6)	0.96 (97.3)	0.79 (99.6)	0.81 (99.7)
2. Multiple RSVs, Transformation	-1.00 (42.6)	-0.96 (4.3)	0.98 (97.9)	0.76 (99.4)	0.63 (99.5)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table 5.3.2 contains the correlation coefficient of the component number accuracy rate with sample size using intensity data with 3-standard deviation separation between adjacent component means and equal within-component variances. With three-standard deviation separation between adjacent component means, the component number accuracy rates decreased as sample size increased except for NEC and ICL-BIC where the component number accuracy rates were all 0. The BIC component number accuracy rate was above 77.3% for $n = 2000$, even though the correlation coefficients were strongly negative

Table 5.3.2: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, equal mixture proportion, and 3-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	-0.98 (77.3)	-0.99 (17.9)	N (0.0)	-0.76 (0.0)	N (0.0)
2. Multiple RSVs, Transformation	-1.00 (80.4)	-0.99 (18.8)	N (0.0)	-0.76 (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table 5.3.3 contains the correlation coefficients for intensity data with 2-standard deviation separation between adjacent component means and equal within-component variance. Regardless of transformation status, component number accuracy rates increased with an increase in sample size for the BIC and the AIC criteria. For $n = 2000$, the BIC component number accuracy rate was below 17.7%, and the AIC accuracy rate was below 58.7%. The component number accuracy rates were all 0 for the entropy based criteria.

Table 5.3.3: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, equal mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	1.00 (17.7)	0.97 (58.7)	N (0.0)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	0.98 (17.0)	0.96 (56.5)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table 5.3.4 contains the correlation coefficients for intensity data with 4-standard deviation separation between adjacent component means and unequal within-component variances. The BIC, CLC, and ICL-BIC had the best measures of performance, with component number accuracy rates above 95% for $n = 2000$. Regardless of transformation status, the component number accuracy rates increased as sample size increased when using the BIC, AIC, CLC and ICL-BIC. When using the NEC criterion, the component number accuracy rate increased as sample size increased for only for intensity data not requiring transformation.

Table 5.3.4: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, equal mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.92 (97.6)	0.99 (2.7)	0.05 (53.5)	0.76 (99.8)	0.76 (99.8)
2. Multiple RSVs, Transformation	0.88 (98.0)	0.99 (5.0)	-0.59 (52.2)	0.79 (99.8)	0.76 (99.8)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table 5.3.5 contains the correlation coefficients for intensity data unequal within-component variances and 3-standard deviation separation between adjacent component means. The BIC criterion was the only one with good performance statistics for this case, with component number accuracy rate above 96.8% for $n = 2000$. Regardless of transformation status, the component number accuracy rate increased as sample size increased when using the BIC and the AIC criteria. When the CLC and ICL-BIC criteria were used, the component number accuracy rates decreased as sample size increased. When using the NEC criterion, the component number accuracy rates were 0 for all sample sizes studied.

Table 5.3.5: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, equal mixture proportion, and 3-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.87 (96.8)	0.96 (2.9)	N (0.0)	-0.79 (0.0)	-0.76 (0.0)
2. Multiple RSVs, Transformation	0.82 (97.2)	0.98 (4.7)	N (0.0)	-0.76 (0.0)	-0.76 (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table 5.3.6 contains the correlation coefficients for intensity data with unequal within-component variances and a 2-standard deviation separation between adjacent component means. Component number accuracy rates were low for all criteria with for $n = 2000$. The BIC criterion had the highest component number accuracy rate (23.3%) for $n = 2000$. Regardless of transformation status, the component number accuracy rate increased as sample size increased when using the BIC and the AIC criteria. The component number accuracy rates were 0 for all sample sizes for all other criteria.

Table 5.3.6: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, equal mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.99 (23.3)	0.86 (7.4)	N (0.0)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	0.98 (25.7)	0.91 (11.1)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

In summary, when there was 4-standard deviation separation between adjacent component means and the intensity data had equal within-component variances, NEC, CLC, and ICL-BIC had the highest component number accuracy rates for $n = 2000$ regardless of transformation status. When intensity data were from a mixture distribution with unequal within-component variances, BIC, CLC, and ICL-BIC were the best measures whether or not the data were transformed.

When there was 3-standard deviation separation between adjacent components means, regardless of transformation status, BIC is the only criterion that had strong performance measures for intensity data from a mixture distribution with either equal or unequal within-component variances.

When there was 2-standard deviation separation between adjacent components means, regardless of transformation, none of the criterion had component number accuracy rates greater than 50% for intensity data from a mixture distribution with equal and unequal within-component variances. The BIC had low component number accuracy rate and positive correlation coefficient.

5.4 Result of Box-Cox transformation for six-components

Table 5.4.1., as usual, is the ANOVA table for the component number accuracy rate for the BIC, AIC, NEC, CLC, and ICL-BIC criteria. An F-value > 7.5 corresponds to a p-value ≤ 0.01.

For the BIC criterion, the main effects of constant variance (*B*) and separation (*C*) were the two largest sources of variation, explaining 12.2% and 44.3% of variation respectively (Table 5.4.1). Each of the interactions $B \times C$, $B \times D$, $C \times D$, and $C \times E$ explained more than 5% of the variation. Together, these six sources explained roughly 83.8% of variation. The F-test was also significant for sample size (D), mixture proportion (E) and the interaction $B \times E$, $B \times C \times E$, and $B \times C \times D$. The main effect of transformation (*G*) explained 0.04% of total variation.

Using the AIC criterion, the constant variance factor (*B*) and separation factor (*C*) explained 18.4% and 35.5% respectively of the variation (Table 5.4.1). The interactions $C \times D$, $B \times C \times D$, $B \times C \times E$, and $C \times D \times E$ each explained more than 5% of the variation. Together, these six sources explained approximately 78% of variation. The F-test was also significant for sample size (D), mixture proportion (E) and the interaction $B \times C$, $B \times D$, $B \times E$, $C \times E$, and $B \times C \times D \times E$. The transformation (*G*) main effect and all of its interactions were non-significant.

For the NEC criterion, constant variance (*B*), separation (*C*), and mixture proportion (*E*) explained 6.8%, 13.7% and 6.8% respectively of the variation (Table 5.4.1). The interactions $B \times C$, $B \times E$, $C \times E$ and $B \times C \times E$ explained each 5% or more of the variation. These seven sources explained roughly 75% of variation. The F-test was significant sample size (D), and the interaction $B \times D$, $C \times D$, $D \times E$, $B \times C \times D$, $B \times D \times E$, $C \times D \times E$ and $B \times C \times D \times E$. The transformation (*G*) effect was removed.

Using the CLC criterion, separation (*C*) and mixture proportion (*E*) and the interaction $C \times E$ explained 43%, 18.1%, and 37.8% respectively of the variation (Table 5.4.1). The F-test was significant for constant variance (*B*), separation (*C*), sample size (*D*), and mixture proportion (*E*). It was also significant for $B \times C$, $B \times D$, $B \times E$, $C \times D$, $D \times E$, $B \times C \times D$, $B \times C \times E$, $B \times D \times E$, $C \times D \times E$, and $B \times C \times D \times E$. The transformation (*G*) effect was removed.

Under the BIC approximation of ICL, separation (*C*) and mixture proportion (*E*) explained 41.9% and 18.2% respectively of the variation (Table 5.4.1). The interaction $C \times E$ explained more than 5% of the variation. These three sources accounted for approximately 99% of variation. The F-test was significant for constant variance (*B*), separation (*C*), sample size (*D*), and mixture proportion (*E*). It was also significant for $B \times C$, $B \times D$, $B \times E$, $C \times D$, $D \times E$, $B \times C \times D$, $B \times C \times E$, $C \times D \times E$, and $B \times C \times D \times E$. The transformation (*G*) effect was removed.

Table 5.4.1: ANOVA tables for BIC, AIC, NEC, CLC, and ICL-BIC

		BIC		AIC		NEC		CLC		ICL-BIC	
Source	DF	%TSS	F	%TSS	F	%TSS	F	%TSS	F	%TSS	F
B: Constant Variance	1	12.2%	173.7	18.4%	300.9	6.8%	688.2	0.1%	375.5	0.0%	34.5
C: Separation	2	44.3%	315.1	35.5%	289.3	13.7%	688.2	43.0%	56501.8	41.9%	43150.5
D: Sample Size	2	1.3%	9.1	4.8%	39.0	2.0%	102.1	0.1%	108.2	0.2%	220.7
E: Mixture Proportion	1	4.6%	66.1	3.1%	50.4	6.8%	688.2	18.1%	47592.7	18.2%	37398.0
G: Transformation	1	0.0%	0.5	0.0%	0.4	0.0%	3.1	0.0%	5.7	0.0%	2.2
$B\times C$	2	8.5%	60.6	3.7%	30.0	13.7%	688.2	0.1%	151.2	0.0%	27.5
$B\times D$	2	5.6%	39.7	2.1%	17.3	2.0%	102.1	0.0%	8.5	0.0%	21.4
$B\times E$	1	0.8%	12.1	1.0%	17.0	6.8%	688.2	0.0%	74.3	0.1%	238.9
$C\times D$	4	7.6%	26.9	6.2%	25.5	4.1%	102.1	0.1%	76.8	0.3%	144.9
$C\times E$	2	5.6%	40.2	0.7%	5.6	13.7%	688.2	37.8%	49615.0	38.6%	39758.2
$D\times E$	2	0.3%	1.9	0.3%	2.1	2.0%	102.1	0.1%	78.4	0.1%	115.2
$B\times C\times D$	4	4.3%	15.3	5.7%	23.3	4.1%	102.1	0.0%	4.1	0.0%	7.5
$B\times C\times E$	2	1.3%	9.6	6.6%	53.6	13.7%	688.2	0.0%	21.6	0.1%	101.8
$B\times D\times E$	2	0.4%	3.2	0.3%	2.8	2.0%	102.1	0.0%	40.2	0.0%	0.1
$C\times D\times E$	4	0.5%	1.6	5.1%	20.6	4.1%	102.1	0.5%	307.8	0.4%	214.9
$B\times C\times D\times E$	4	0.1%	0.5	4.4%	17.9	4.1%	102.1	0.0%	25.6	0.0%	19.8
Remainder	35	2.5%		2.1%		0.3%		0.0%		0.0%	

5.5 Interaction of significant factors

Table 5.5.1 contains the average component number accuracy rates for separation (*C*), constant variance (*B*) and transformation (*G*) for the BIC criterion and corresponds to table 3.2.1. As expected, the Box-Cox transformation removed the effect of transformation. As seen in table 5.5.1, the average component number accuracy rates were higher for unequal within-component variance. Each average component number accuracy rate was below 50% except for data with unequal within-component variances and a 4-standard deviation separation between adjacent component means.

Table 5.5.2 contains the BIC component number accuracy counts (not rates) for constant variance (*B*), separation (*C*), sample size (*D*), and component probability distribution (*E*). These numbers are the counts (out of 1000 replicates) for the component number accuracy. The first number is the count when transformation in not needed. And the second number is the count when transformation is needed. As shown in the table, the component number accuracy counts for transformation status (yes or no) were almost identical showing that the transformation status factor was not significant. The highest accuracy rate (94%) was for unequal within-component variance, equal component mixture distribution, and 4-standard deviation separation between adjacent component means.

Table 5.5.1: Table of average component number accuracy rates for separation, constant variance, and transformation setting for the BIC criterion (1000 replicates after Box-Cox transformation)

		B: Equal Component Variance			
C: Separation		Yes		No	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
	2σ	3.2±7.1	3.8±8.0	7.8±10.5	1.2±0.7
	3σ	26.2±23.3	25.0±22.0	43.5±27.3	44.5±25.7
	4σ	31.1±30.0	30.2±24.9	67.6±21.4	81.4±9.6

Average accuracy rate ± SD, average over 6 settings

Table 5.5.2: Table of component number accuracy count for separation, constant variance, mixture distribution, and sample size setting for the BIC criterion (1000 replicates after Box-Cox transformation)

				B: Constant Variance					
E: Mixture Distribution				Yes			No		
				D: Sample Size			D: Sample Size		
				$n = 500$	$n = 1000$	$n = 2000$	$n = 500$	$n = 1000$	$n = 2000$
	Equal	C: Separation	2σ	0, 0	1, 0	5, 1	5, 11	8, 6	7, 3
			3σ	305, 324	588, 612	399, 426	179, 174	431, 411	809, 828
			4σ	692, 693	422, 473	155, 159	771, 766	870, 878	944, 946
	skew		2σ	2, 1	19, 16	201, 176	21, 65	11, 107	20, 274?
			3σ	15, 22	88,78	106, 107	188, 164	382, 341	682, 692
			4σ	392, 363	151, 173	2, 4	660, 455	830, 495	808, 513

Table 5.5.3 contains the average component number accuracy rates for the AIC criterion for each setting of factor B (equal within-component variance or not) and factor G, transformation status. It corresponds to half of table 3.2.2. As expected, the Box-Cox transformation removed the transformation effect. However, the average component number accuracy rates were lower than those reported in table 4.2.3. All the component number accuracy rates were below 20%.

Table 5.5.4 contains the component number accuracy counts (out of 1000 replicates) for constant variance (*B*), separation (*C*), sample size (*D*), and mixture distribution (*E*) factors. The accuracy counts for transformation status (yes and no) were almost identical, confirming that the transformation status factor was not significant after application of the Box-Cox transformation. None of the accuracy counts were above 300 (i.e., 30%).

Table 5.5.3: Table of average component number accuracy rates for constant variance and transformation setting for the AIC criterion (1000 replicates after Box-Cox transformation)

		B: Constant Variance	
G: Transformation		Yes	No
	Yes: $\sqrt{X}$ is a normal mixture	11.8±10.3	4.0±4.1
	No	11.1±10.2	4.2±5.0

Average accuracy rate ± SD, average over 18 settings

Table 5.5.4: Table of component number accuracy count for separation, constant variance, mixture distribution, and sample size setting for the AIC criterion (1000 replicates after Box-Cox transformation)

				B: Constant Variance					
E: Mixture Distribution				Yes			No		
		C: Separation		D: Sample Size			D: Sample Size		
				$n = 500$	$n = 1000$	$n = 2000$	$n = 500$	$n = 1000$	$n = 2000$
	Equal		2σ	106, 119	171, 172	286, 270	128, 131	118, 108	124, 122
			3σ	263, 286	115, 149	22, 32	37, 27	18, 18	9, 12
			4σ	169, 198	55, 66	5, 7	1, 0	1, 0	2, 1
	Skew		2σ	216, 236	272, 248	150, 186	111, 43	85, 40	82, 22
			3σ	127, 127	30, 26	0, 1	31, 35	11, 15	1, 6
			4σ	9, 8	1, 0	0, 0	2, 56	0, 45	1, 43

Table 5.5.5 contains the average component number accuracy rates for the main effects explaining 5% or more of the total NEC variation. The average component number accuracy rates were below 15%. Table 5.5.6 contains the average component number accuracy rates for separation and corresponds to table 5.2.5. The average component number accuracy rates were all below 5%.

Table 5.6.5: Table of average component number accuracy rates for separation, constant variance, and mixture proportion setting for the NEC criterion (1000 replicates after Box-Cox transformation)

		B: Constant Variance			
		Yes		No	
		E: Mixture Proportion		E: Mixture Proportion	
		Equal	Skewed	Equal	Skewed
C: Separation	2σ	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	3σ	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	4σ	12.8±7.7	0.0±0.0	0.0±0.0	0.0±0.0

Average accuracy rate ± SD, average over 6 settings

Table 5.5.6: Table of average component number accuracy rates for the separation for the NEC criterion (1000 replicates after Box-Cox transformation)

C: Separation		
2σ	3σ	4σ
0.0±0.0	0.0±0.0	3.2±6.7

Average accuracy rate ± SD, average over 24 settings

Table 5.5.7 contains the average component number accuracy rates for sources (separation (*C*) and mixture distribution (*E*)) explaining 5% or more of the total CLC criterion variation. With the exception of one setting, all average component number accuracy rates were below 5%. At 4-standard deviation separation between component means and equal component mixture distribution, the average component number accuracy rate was 92.3%.

Table 5.5.8 contains the average component number accuracy rate for the CLC criterion using separation (*C*), constant variance (*B*) and transformation (*G*) and corresponds to table 3.2.4. As expected, the Box-Cox transformation removed the effect of transformation. The component number accuracy rates for transformed and non-transformed data were essentially the same. At 4-standard deviation separation between adjacent component means, the average component number accuracy rates are above

45%. At the other separation settings, the average component number accuracy rates are below 6%.

Table 5.5.7: Table of average component number accuracy rates for separation and mixture proportion setting for the CLC criterion (1000 replicates after Box-Cox transformation)

	E: Mixture Proportion		
C: Separation		Equal	Skewed
	2σ	0.0±0.0	0.0±0.0
	3σ	2.2±2.8	3.4±4.6
	4σ	92.3±6.8	4.6±3.0

Average accuracy rate ± SD, average over 12 settings

Table 5.5.8: Table of average component number accuracy rates for separation, constant variance, and transformation setting for the CLC criterion (1000 replicates after Box-Cox transformation)

	B: Constant Variance				
C: Separation		Yes		No	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
	2σ	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	3σ	0.1±0.2	0.0±0.0	5.5±3.5	5.8±4.0
	4σ	47.4±49.6	47.5±49.4	48.8±47.5	50.3±46.9

Average accuracy rate ± SD, average over 6 settings

Table 5.5.9 contains the average component number accuracy rates for separation (*C*) and mixture distribution (*E*), both of which explaining 5% or more of the total ICL-BIC criterion variation. With the exception of one setting, all average component number accuracy rates were below 5%. At 4-standard deviation separation between component means and equal component mixture distribution, the average component number accuracy rate was 90.8%.

Table 5.5.10 contains the average component number accuracy rate for the ICL-BIC criterion using separation (*C*), constant variance (*B*) and transformation (*G*) and corresponds to table 3.2.5. As expected, the Box-Cox transformation removed the effect of transformation. The component number accuracy rates for transformed and non-transformed data are essentially the same. At 4-standard deviation separation between adjacent component means, the average component number accuracy rates are above 45%. At the other separation settings, the average component number accuracy rates are below 5%.

Table 5.5.9: Table of average component number accuracy rates for separation and mixture proportion setting for the ICL-BIC criterion (1000 replicates after Box-Cox transformation)

		E: Mixture Proportion	
C: Separation		Equal	Skewed
	2σ	0.0±0.0	0.0±0.0
	3σ	0.3±0.4	2.1±3.2
	4σ	90.8±9.1	2.9±3.0

Average accuracy rate ± SD, average over 12 settings

Table 5.5.10: Table of average component number accuracy rates for separation, constant variance, and transformation setting for the ICL-BIC criterion (1000 replicates after Box-Cox transformation)

		B: Constant Variance			
C: Separation		Yes		No	
		G: Transformation		G: Transformation	
		Yes: $\sqrt{X}$ is a normal mixture	No	Yes: $\sqrt{X}$ is a normal mixture	No
	2σ	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	3σ	0.0±0.0	0.0±0.0	2.5±2.8	2.4±3.4
	4σ	46.9±51.3	46.6±51.1	46.1±46.0	47.8±46.0

Average accuracy rate ± SD, average over 6 settings

5.6 Correlation coefficient of component number accuracy rate of the six criteria with sample size for six equiprobable components

In this section, I report the correlation coefficients of the component number accuracy rates with sample size for intensity data from mixtures with six equiprobable components with Box-Cox transformation automatically applied. As before, I also report the component number accuracy rate for sample size 2000 for all six criteria. Results for mixtures with skewed component distributions are given in Appendix D.

Table 5.6.1 contains the correlation coefficient for the component number accuracy rate with sample size for intensity data with equal within-component variances and 4-standard deviation separation between adjacent component means. The CLC and ICL-BIC criteria performed well with component number accuracy rates for $n = 2000$ equal to at least 98.4% and 99.4% respectively. Regardless of transformation status, the component number accuracy rate increased as sample size increased when using CLC and ICL-BIC. On the other hand, the component number accuracy rate decreased as sample size increased when using BIC, AIC and NEC.

Table 5.6.1: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, equal mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	-0.98 (15.5)	-0.92 (0.5)	-0.97 (3.8)	0.92 (99.5)	0.89 (99.9)
2. Multiple RSVs, Transformation	-1.00 (15.9)	-0.92 (0.7)	-0.97 (5.1)	0.89 (98.4)	0.87 (99.4)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table 5.6.2 contains the correlation coefficients for intensity data with equal within-component variances and 3-standard deviation separation between adjacent component means. Only the BIC criteria had positive correlation and appreciable component number accuracy rate for $n = 2000$ (specifically, at least 39.9%). When using AIC and CLC, the component number accuracy rate decreased as sample size increased. The component number accuracy rates were 0 for all sample sizes when using NEC and ICL-BIC.

Table 5.6.2: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, equal mixture proportion, and 3-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.14 (39.9)	-0.95 (2.2)	N (0.0)	-0.76 (0.0)	N (0.0)
2. Multiple RSVs, Transformation	0.17 (42.6)	-0.97 (3.2)	N (0.0)	-0.76 (0.00	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table 5.6.3 contains the correlation coefficients for intensity data with equal within-component variances and 2-standard deviation separation between adjacent component means. None of the criteria had component number accuracy rate above 30% for $n = 2000$. The AIC had better performance than the other criteria. Only the BIC and the AIC had positive correlation coefficients. The component number accuracy rates were 0 for all other criteria.

Table 5.6.3: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, equal mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.99 (0.5)	1.00 (28.6)	N (0.0)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	0.94 (0.1)	1.00 (27.0)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table 5.6.4 contains the correlation coefficients for intensity data with unequal within-component variances and a 4-standard deviation separation between adjacent component means. The BIC, CLC, and ICL-BIC criteria had positive correlation coefficient and component number accuracy rates above 94.6% for $n = 2000$. The component number accuracy rate increased as sample size increased when using BIC, AIC, CLC, and ICL-BIC. The component number accuracy rates were 0 for the NEC criteria.

Table 5.6.4: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, equal mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.96 (94.4)	0.94 (0.2)	N (0.0)	0.87 (98.8)	0.89 (98.3)
2. Multiple RSVs, Transformation	0.95 (94.6)	0.94 (0.1)	N (0.0)	0.89 (98.4)	0.91 (97.6)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table 5.6.5 contains the correlation coefficients for intensity data with unequal within-component variances and a 3-standard deviation separation between adjacent component means. The BIC criterion was the only one with positive correlation coefficients and component number accuracy rate above 80.9% for $n = 2000$. The component number accuracy rate decreased in sample size increased when using AIC, CLC, and ICL-BIC. The component number accuracy rates were 0 for NEC.

Table 5.6.5: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, equal mixture proportion, and 3-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	1.00 (80.9)	-0.92 (0.9)	N (0.0)	-0.99 (1.4)	-0.94 (0.2)
2. Multiple RSVs, Transformation	1.00 (82.8)	-0.95 (1.2)	N (0.0)	-1.00 (1.4)	-0.98 (0.3)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table 5.6.6 contains the correlation coefficients for intensity data with unequal within-component variances and a 2-standard deviation separation between adjacent component means. None of the criteria worked well. The component number accuracy rate decreased as sample size increased for the AIC, CLC, and ICL-BIC criteria. The

BIC component number accuracy rate was below 0.7% for $n = 2000$. The component number accuracy rates were 0 for NEC, CLC, and ICL-BIC criteria.

Table 5.6.6: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, equal mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.50 (0.7)	-0.22 (12.4)	N (0.0)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	-0.94 (0.3)	-0.21 (12.2)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

In summary, when there was 4-standard deviation separation between adjacent component means and the intensity data had equal within-component variances, CLC and ICL-BIC had the highest component number accuracy rates regardless of transformation status. When intensity data were from a mixture distribution with unequal within-component variances, BIC, CLC, and ICL-BIC were the best measures regardless of transformation status.

When there was 3-standard deviation separation between adjacent components means, regardless of transformation status, none of the six criteria had component number accuracy rates greater than 50% for intensity data from a mixture distribution with equal within-component variance. With unequal within-component variances, the BIC criterion had component number accuracy rates greater than 80.9% for $n = 2000$ regardless of transformation status.

When there was 2-standard deviation separation between adjacent components means, regardless of transformation status, none of the six criteria had component number accuracy rates greater than 50% for intensity data from a mixture distribution with either equal or unequal within-component variances.

Chapter 6 Discussions and conclusions

6.1. Procedure for assessing a criterion.

Since standard research practice is to base model selection on a maximized likelihood function using a large number of RSVs, I will assess the six criteria based on their results with multiple RSVs. It is notable, however, that increasing the number of RSVs sometimes reduced the component number accuracy rate. My goal was to identify those criteria whose component number accuracy rate increases with increasing sample size and that has a high component number accuracy rate for a large sample when the global maximum of each competing model is used. Consequently, I focus on the tables of correlation coefficients of component number accuracy rates with sample size and component number accuracy rates at large samples as assessed with multiple RSVs.

Each criterion worked well when separation was large, within-component variances were equal, the components were equi-probable and the sample size was large. There were settings in which the NEC, and AIC performed extremely poorly. These were low separation, unequal within-component variances, skewed and HWE mixture distributions, and data that were the square of a normal mixture. Consequently, I have excluded these criteria from further discussion.

6.2 Data from a three-component mixture

My results documented that a Box-Cox transformation should always be applied. Tables 3.3.1-3.3.4 present the critical information for data from a three-component model without Box-Cox transformation. When Box-Cox is not used and transformation status is on (that is $\sqrt{X}$ is a mixture of normal components) then CLC and ICL-BIC were the only effective criteria for equal within-component variance. When transformation status was off (that is X is a mixture of normal components) each criterion was effective for equal within-component variance (table 3.3.1 and table 3.3.2). When Box-Cox is not used and $\sqrt{X}$ is a mixture of normal components, then BIC, CLC and ICLBIC were effective criteria for unequal within-component variance.

Tables 5.4.1-5.4.6 present the critical information for data with Box-Cox transformation applied. When Box-Cox transformation was used, regardless of the data used, CLC, and ICL-BIC were effective criteria for equal within-component variance (table 5.4.1). For unequal within-component variance, BIC, CLC, and ICL-BIC were effective criteria (table 5.4.4).

6.3 Data from a six-component mixture

As in the three-component model, the results show that one should automatically use the Box-Cox transformation. Tables 4.3.1-4.3.4 present the results from a six-component model without Box-Cox transformation. When Box-Cox transformation is not used and $\sqrt{X}$ is a mixture of normal components, none of the six criteria used had component number accuracy rates greater than 34.7% for equal within-component variance. When X is a normal mixture, each criterion had high component number accuracy rates for equal within-component variance (Table 4.3.1). When Box-Cox transformation is not used, regardless of transformation status, BIC was the only criterion with high component number accuracy rates for unequal within-component variances (Table 4.3.2).

Tables 5.7.1-5.7.6 present results from a six-component model with Box-Cox transformation. When Box-Cox transformation is used, regardless of data used, BIC, CLC, and ICL-BIC were the criteria with high component number accuracy rates for unequal within-component variances (Table 5.7.4). Two criteria, CLC and ICL-BIC, had high component number accuracy rates for equal within component variance (Table 5.7.1).

6.4 Is there a criterion which always has high component number accuracy rates?

Based on the overall component number accuracy rates for the each model with or without Box-Cox transformation, there was no criterion that always had high component number accuracy rate in all cases tested. For large separation (4-standard deviation separation between adjacent component means), BIC, CLC, and ICL-BIC had high component number accuracy rates. For intermediation separation (3-standard deviation separation between adjacent component means), the BIC criterion was the only effective criterion for the three-component model. For small separation (2-standard deviation separation between adjacent component means), no criterion had high component number accuracy rate for $n = 2000$. BIC performed best of these, albeit with very low accuracy rate.

For the BIC criterion, the settings required to have high component number accuracy rates were 4-standard deviation separation between adjacent component means and no transformation. The settings for which it had low component number accuracy rates were 2-standard deviation separation between adjacent component means, equal within-component variance, and transformation. Also, when using a Box-Cox transformation, the BIC criterion had high component number accuracy rates for unequal within-component variances.

For the CLC and the ICL-BIC criteria, the settings that yielded high component number accuracy rates were 4-standard deviation separation between adjacent component means with and without Box-Cox transformation, equal within-component variances and equi-probable distribution. For example, component number accuracy rate was at least 91% for samples of size 500. The settings for low component number accuracy rates were 2-standard deviation separation between adjacent component means. For instance, the component number accuracy rate was less than 9% for $n = 2000$ even though the correlation was positive.

Unequal within-component variances did not affect BIC, CLC and ICL-BIC when there was 4-standard deviation separation between adjacent component means.

6.5 What should one do?

1. Always use Box-Cox. When $\sqrt{X}$ is a mixture of normal components, Box-Cox effectively addressed the problem in that component number accuracy rates were similar to those with transformation status off. When *X* is a mixture of normal components, the component number accuracy rates after transformation were similar to those without the unnecessary Box- Cox transformation. In table 5.1, one should note that the average power transform $\hat{\lambda}$ for the normal mixture squared is about half the average power transform $\hat{\lambda}$ for the normal mixture in the three-component model. The average power transform $\hat{\lambda}$ for the normal mixture squared is almost to the average power transform $\hat{\lambda}$ for the normal mixture in the six-component model. This is probably an indication as to why the three-component model after Box-Cox transformation has better result.

2. Use BIC, ICL-BIC, or CLC as criteria to select the number of components. Next one should fit the mixture model and examine the estimated component probabilities, component means, and component variances. Separation of component means of 4 or more standard deviations are indicative of situations with high component number accuracy rates. In this event, each of the three measures works well. For 3 standard deviation settings, the BIC worked better than the other two criteria. Assessing the number of components based on results with an estimated 2-standard deviation separation between components is a task with high component number error rate. All criteria had low component number accuracy rate, even with sample sizes of 2000. In such a case, while the BIC criterion is better than the other two, its probability of correct component number classification was below 35% for equal within-component variances. Looking at cases with component number accuracy rates greater than 50%, BIC was the best about 19% of the time after Box-Cox transformation compared to 22% of the time before Box-Cox transformation. CLC and ICL-BIC were the best about 10% and 11% of the time respectively before Box-Cox transformation compared to about 10% and 14.6% respectively after Box-Cox transformation.
3. Always use multiple RSVs. There were cases where multiple RSVs reduced the component number accuracy rates for all criteria. However, there were numerous cases where using multiple RSVs actually increased the chance of having high component number accuracy rates by finding the global maximum. Such cases includes skewed and HWE mixture distributions. Knowing the mixture distributions a priori is unlikely. Therefore, one should use multiple RSVs.

6.6 Future research

There are other approaches that one can use in addressing this problem. One is to use a commingling type analysis to transform the data (Barrett,et al., 1996). To date, one can use the SAGE program to do the analysis on a three-component model. Software needs to be implemented to tackle other numbers of components. With a commingling type analysis, one can answer the question does using a commingling type approach to transform the data increase the component number accuracy rate?

References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6): 716-723.

Barrett, J. H., Foy, C. A., Grant, P.J. (1996). Commingling analysis of the distribution of a phenotype conditioned on two marker genotypes: application to plasma angiotensin-converting enzyme levels. Genetic Epidemiology 13(6): 615-625.

Ben-Yaacov, E. and Eldar, Y. C. (2008). A fast and flexible method for the segmentation of aCGH data. Bioinformatics 24(16): i139-145.

Biernacki, C., Celeux, G., Govaert, G. (1998). Assessing a Mixture Model for Clustering with theIntegrated Classification Likelihood, Institut National de Recherche en informatique et en automatique.

Biernacki, C., Celeux, G., Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. Pattern Recognition Letters 20(3): 267-272.

Biernacki, C. and Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. Computing Science and Statistics 29(2): 451-457.

Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations Journal of the Royal Statistical Society Series B (Methodological) 26(2): 211-252.

Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. Journal of Classification 13(2): 195-212.

Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J. M., Gough, S. C. L., de Smith, A., Blakemore, A. I. F., Froguel, P., Owen, C. J., Pearce, S. H. S., Teixeira, L., Guillevin, L., Graham, D. S. C., Pusey, C. D., Cook, H. T., Vyse, T. J., Aitman, T. J. (2007). FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39(6): 721-723.

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

Fraley, C. and Raftery, A. E. (September 2006, (revised December 2009)). MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering, Department of Statistics, University of Washington.

Iafrate, A. J.[1,2], Feuk , L.[3], Rivera, M. N.[1,2], Listewnik, M. L.[1], Donahoe, P. K.[2,4], Qi, Y.[3], Scherer, S. W.[3,5], Lee, C.[1,2,5]. (2004). Detection of large-scale variation in the human genome. Nature Genetics 36(9): 949-951.

Kim, W.[1], Gordon, D.[2*], Sebat, J.[3], Ye, K.[4], Finch, S. J.[5] (2008). Computing Power and Sample Size for Case-Control Association Studies with Copy Number Polymorphism: Application of Mixture-Based Likelihood Ratio Test. PlosOne 3(10): e3475.

Lai, W. R., Johnson,M. D., Kucherlapati, R., Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics 21(19): 3763-3770.

Lo, Y., Mendell, N. R., Rubin, D. B. (2001). Testing the number of components in a normal mixture. Biometrika 88(3): 767-778.

McLachlan, G. and Peel, D. (2000). Finite Mixture Models, John Wiley and Sons.

Olshen, A. B., Venkatraman, E. S., Lucito, R. Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostat 5(4): 557-572.

Patel, P. (2009, 8/12/2009). Glomerulonephritis.   Retrieved 06/02/2010, from http://www.nlm.nih.gov/medlineplus/ency/article/000484.htm.

Picard, F., Robin,S., Lebarbier, E., Daudin, J.-J. (2005). A Segmentation/Clustering Model for the Analysis of Array CGH Data. Biometrics 63(3): 758-766.

Picard, F., S. Robin, et al. (2007). A Segmentation/Clustering Model for the Analysis of Array CGH Data. Biometrics 63(3): 758-766.

Pinkel, D.[1,2], Segraves, R.[1], Sudar, D.[2], Clark, S.[1], Poole, I.[3], Kowbel, D.[2], Collins, C.[2], Kuo, W.-L.[1], Chen, C.[1], Zhai, Y.[1], Dairkee, S. H.[4], Ljung, B.-M.[5], Gray, J. W.[1,2], Albertson, D. G.[1,2,6] (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature Genetics 20: 207-211.

R Development Core Team (2008). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Schwarz, G. (Mar. ,1978). Estimating the Dimension of a Model. The Annals of Statistics 6(2): 461-464.

Sebat, J., Lakshmi,B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., Wigler, M. (2004). Large-Scale Copy Number Polymorphism in the Human Genome. Science 305(5683): 525-528.

Wineinger, N. E. [A1], Kennedy, R. E. [A2], Erickson, S. W.[A3], Wojczynski, M. K.[A4], Bruder, C.[A5], Tiwari, H.[A6] (2008). Statistical issues in the analysis of DNA Copy Number Variations. International Journal of Computational Biology and Drug Design 1(4): 368-395.

Appendix A. coefficient for the component number accuracy rates with sample size for HWE and skewed mixture proportions (three-component model)

Table A.1: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, HWE mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.47 (99.6)	0.92 (88.4)	0.89 (99.0)	0.84 (98.7)	0.82 (99.2)
2. Multiple RSVs, Transformation	-0.66 (0.0)	-0.59 (0.0)	-0.85 (1.1)	0.99 (69.7)	0.99 (62.4)
3. Default SV, no transformation	0.38 (100.0)	0.96 (96.0)	0.85 (100.0)	0.84 (100.0)	0.83 (99.9)
4. Default SV, transformation	-0.78 (0.0)	-0.59 (0.0)	0.98 (76.7)	0.97 (76.7)	0.98 (65.9)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table A.2: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, skewed mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICLBIC
1. Multiple RSVs, no Transformation	0.65 (99.8)	1.00 (90.7)	0.84 (99.6)	0.82 (99.7)	0.73 (99.9)
2. Multiple RSVs, Transformation	-0.60 (0.0)	N (0.0)	-0.83 (2.0)	0.92 (42.2)	0.88 (53.5)
3. Default SV, no transformation	0.90 (100.0)	0.86 (94.5)	0.59 (100.0)	0.68 (100.0)	0.63 (100.0)
4. Default SV, transformation	-0.71 (0.0)	-0.59 (0.0)	0.48 (99.7)	0.63 (100.0)	0.65 (100.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table A.3: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, HWE mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.98 (27.6)	0.98 (80.8)	0.92 (0.2)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	-0.66 (0.0)	-0.59 (0.0)	-0.87 (0.4)	0.99 (72.3)	0.98 (65.2)
3. Default SV, no transformation	0.97 (29.0)	0.97 (0.1)	N (0.0)	N (0.0)	N (0.0)
4. Default SV, transformation	0.72 (34.4)	-0.92 (0.1)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table A.4: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, skewed mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.99 (25.0)	0.99 (74.6)	0.92 (0.5)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	-0.78 (0.0)	-0.59 (0.0)	0.84	-0.79 (0.0)	-0.79 (0.0)
3. Default SV, no transformation	0.97 (24.2)	0.98 (73.3)	N (0.0)	N (0.0)	N (0.0)
4. Default SV, transformation	-0.80 (0.0)	-0.60 (0.0)	-0.59 (0.0)	-0.59 (0.0)	-0.59 (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table A.5: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, HWE mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.82 (97.6)	0.99 (3.8)	-0.59 (52.0)	0.74 (99.5)	0.76 (99.7)
2. Multiple RSVs, Transformation	0.78 (96.0)	0.93 (3.0)	0.99 (71.7)	0.83 (98.6)	0.75 (98.7)
3. Default SV, no transformation	0.29 (99.3)	0.96 (90.5)	0.65 (100.0)	0.79 (100.0)	0.83 (99.9)
4. Default SV, transformation	0.59 (100.0)	0.73 (79.3)	0.64 (99.9)	0.75 (99.9)	0.80 (99.9)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table A.6: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, skewed mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.85 (98.2)	1.00 (4.7)	-0.97 (16.4)	0.72 (99.9)	0.71 (99.9)
2. Multiple RSVs, Transformation	0.09 (86.0)	0.84 (2.1)	-0.97 (9.0)	0.73 (99.0)	0.74 (99.3)
3. Default SV, no transformation	0.59 (100.0)	0.87 (86.4)	0.59 (100.0)	0.64 (100.0)	0.63 (100.0)
4. Default SV, transformation	0.39 (99.4)	-0.96 (23.4)	0.59 (99.7)	0.68 (100.0)	0.70 (100.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table A.7: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, HWE mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.99 (46.6)	0.89 (9.0)	N (0.0)	N (0.0)	0.05 (0.0)
2. Multiple RSVs, Transformation	-0.45 (53.3)	-0.79 (0.0)	N (0.0)	-0.59 (0.0)	N (0.0)
3. Default SV, no transformation	0.99 (71.8)	0.91 (79.1)	N (0.0)	N (0.0)	N (0.0)
4. Default SV, transformation	0.67 (91.9)	-0.96 (0.8)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table A.8: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, skewed mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.53 (18.2)	0.99 (11.9)	N (0.0)	-0.59 (0.0)	N (0.0)
2. Multiple RSVs, Transformation	-0.85 (24.9)	-0.92 (0.0)	N (0.0)	-0.59 (0.0)	-0.63 (0.0)
3. Default SV, no transformation	0.93 (26.6)	0.86 (44.6)	N (0.0)	N (0.0)	N (0.0)
4. Default SV, transformation	0.35 (69.3)	-0.88 (0.0)	N (0.0)	-0.59 (0.0)	-0.59 (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Appendix B. Correlation coefficient for the component number accuracy rates with sample size for HWE and skewed mixture proportions (six-component model)

Table B.1: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, HWE mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.65 (99.7)	0.93 (88.3)	-0.98 (7.5)	0.73 (99.4)	0.61 (99.6)
2. Multiple RSVs, Transformation	-0.87 (0.0)	-0.74 (0.0)	0.91 (4.6)	-0.94 (3.4)	-0.40 (2.3)
3. Default SV, no transformation	-0.86 (0.0)	-0.83 (0.0)	-0.84 (0.2)	-0.73 (0.0)	-0.92 (0.1)
4. Default SV, transformation	-0.74 (0.0)	-0.68 (0.0)	-0.59 (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table B.2: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, skewed mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.71 (99.8)	0.99 (88.0)	0.92 (60.3)	0.82 (99.9)	0.63 (99.9)
2. Multiple RSVs, Transformation	-0.72 (0.0)	-0.71 (0.0)	0.92 (3.2)	-0.96 (1.8)	-0.94 (1.3)
3. Default SV, no transformation	-0.59 (0.0)	-0.59 (0.0)	-0.59 (0.0)	N (0.0)	N (0.0)
4. Default SV, transformation	-0.62 (0.0)	-0.65 (0.0)	-0.59 (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table B.3: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, HWE mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.92 (0.1)	0.99 (16.7)	N (0.0)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	-0.29 (8.8)	-0.83 (0.2)	N (0.0)	-0.59 (0.0)	-0.59 (0.0)
3. Default SV, no transformation	0.92 (0.6)	0.99 (40.4)	N (0.0)	N (0.0)	N (0.0)
4. Default SV, transformation	-0.25 (7.1)	-0.96 (1.9)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table B.4: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, skewed mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.92 (0.2)	0.99 (17.7)	N (0.0)	-0.59 (0.0)	N (0.0)
2. Multiple RSVs, Transformation	-0.86 (0.4)	-0.67 (0.0)	N (0.0)	-0.98 (8.8)	-0.15 (4.0)
3. Default SV, no transformation	0.94 (4.4)	0.99 (41.8)	N (0.0)	N (0.0)	N (0.0)
4. Default SV, transformation	-0.87 (0.1)	-0.79 (0.0)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table B.5: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, HWE mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.89 (93.9)	-0.27 (1.7)	N (0.0)	0.88 (69.3)	0.94 (64.6)
2. Multiple RSVs, Transformation	0.90 (92.0)	-0.64 (1.3)	N (0.0)	N (0.0)	N (0.0)
3. Default SV, no transformation	1.00 (1.0)	-0.87 (8.9)	-0.59 (0.0)	-0.98 (0.0)	-0.37 (0.0)
4. Default SV, transformation	1.00 (0.7)	-0.59 (6.1)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table B.6: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, skewed mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.85 (94.7)	-0.47 (0.9)	N (0.0)	0.81 (76.9)	0.88 (96.6)
2. Multiple RSVs, Transformation	0.83 (91.5)	-0.79 (0.3)	N (0.0)	N (0.0)	0.05 (0.0)
3. Default SV, no transformation	-0.41 (1.8)	-0.84 (0.0)	-0.46 (0.0)	-0.74 (0.0)	-0.84 (0.0)
4. Default SV, transformation	-0.48 (1.1)	-0.73 (0.0)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table B.7: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, HWE mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	-0.77 (0.1)	0.97 (16.1)	0.28 (0.1)	-0.59 (0.0)	N (0.0)
2. Multiple RSVs, Transformation	-0.96 (0.3)	0.61 (10.6)	N (0.0)	N (0.0)	N (0.0)
3. Default SV, no transformation	N (0.0)	-0.94 (2.8)	N (0.0)	N (0.0)	N (0.0)
4. Default SV, transformation	N (0.0)	0.83 (15.7)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table B.8: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, skewed mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	-0.93 (0.3)	0.69 (17.5)	0.92 (0.1)	-0.78 (0.0)	-0.59 (0.0)
2. Multiple RSVs, Transformation	0.85 (61.3)	-0.87 (0.8)	N (0.0)	N (0.0)	N (0.0)
3. Default SV, no transformation	N (0.0)	0.95 (30.6)	N (0.0)	-0.59 (0.0)	N (0.0)
4. Default SV, transformation	0.97 (2.7)	0.07 (19.6)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Appendix C. coefficient for the component number accuracy rates with sample size for skewed mixture proportion (three-component Box-Cox model)

Table C.1: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, skewed mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	-0.76 (0.0)	N (0.0)	-0.76 (0.0)	-0.76 (0.3)	-0.76 (0.4)
2. Multiple RSVs, Transformation	-0.76 (0.0)	N (0.0)	-0.76 (0.0)	-0.98 (0.9)	-0.98 (0.2)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table C.2: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, skewed mixture proportion, and 3-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	-0.80 (0.0)	-0.76 (0.0)	N (0.0)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	-0.79 (0.0)	-0.76 (0.0)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table C.3: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, skewed mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	1.00 (35.9)	-1.00 (10.5)	N (0.0)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	1.00 (37.6)	-0.97 (13.3)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table C.4: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, skewed mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	-0.98 (59.8)	-0.93 (0.3)	-0.97 (28.3)	0.86 (99.0)	0.85 (99.3)
2. Multiple RSVs, Transformation	-1.00 (60.4)	-0.98 (0.0)	-0.99 (28.4)	0.92 (99.6)	0.87 (99.9)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table C.5: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, skewed mixture proportion, and 3-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	-0.90 (74.0)	-0.94 (0.1)	N (0.0)	-0.85 (0.0)	-0.85 (0.0)
2. Multiple RSVs, Transformation	-0.46 (77.7)	-0.34 (0.5)	N (0.0)	-0.76 (0.0)	-0.76 (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table C.6: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, skewed mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	1.00 (19.0)	-0.91 (3.6)	N (0.0)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	0.97 (19.1)	-0.97 (5.1)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Appendix D. Correlation coefficient for the component number accuracy rates with sample size for skewed mixture proportion (six-component Box-Cox model)

Table D.1: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, skewed mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	-0.95 (0.2)	-0.82 (0.0)	N (0.0)	-0.98 (0.1)	-0.95 (0.0)
2. Multiple RSVs, Transformation	-0.98 (0.4)	-0.76 (0.0)	N (0.0)	-0.92 (0.1)	-0.84 (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table D.2: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, skewed mixture proportion, and 3-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.86 (10.6)	-0.88 (0.0)	N (0.0)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	0.93 (10.7)	-0.87 (0.1)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table D.3: Correlations of accuracy rate of measure with sample size for selected experimental conditions (constant variance, skewed mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.97 (20.1)	-0.6915.0)	N (0.0)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	0.97 (17.6)	-0.87 (18.6)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table D.4: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, skewed mixture proportion, and 4-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.67 (80.8)	-0.33 (0.1)	N (0.0)	-0.71 (6.1)	-0.19 (5.7)
2. Multiple RSVs, Transformation	0.92 (51.3)	-0.88 (4.3)	N (0.0)	-0.87 (3.3)	-0.57 (3.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table D.5: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, skewed mixture proportion, and 3-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	1.00 (68.2)	-0.93 (0.1)	N (0.0)	1.00 (12.2)	1.00 (8.9)
2. Multiple RSVs, Transformation	1.00 (69.2)	-0.92 (0.6)	N (0.0)	0.99 (11.4)	1.00 (7.6)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.

Table D.6: Correlations of accuracy rate of measure with sample size for selected experimental conditions (non-constant variance, skewed mixture proportion, and 2-standard deviation separation between component means)

	BIC	AIC	NEC	CLC	ICL-BIC
1. Multiple RSVs, no Transformation	0.10 (2.0)	-0.81 (8.2)	N (0.0)	N (0.0)	N (0.0)
2. Multiple RSVs, Transformation	0.99 (27.4)	-0.98 (2.2)	N (0.0)	N (0.0)	N (0.0)

Note: Value in parenthesis is the component number accuracy rate for sample size 2000.